Research and Methodology Directorate

Legacy Techniques and Current Research in Disclosure Avoidance at the U.S. Census Bureau

By Laura McKenna and Matthew Haubach

Issued April 2019

United States™ Census Bureau

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

CONTENTS

INTRODUCTION[1]

The U.S. Census Bureau's mission is to serve as the nation's leading provider of quality data about its people and economy. The Census Bureau honors privacy, protects confidentiality, shares expertise globally, and conducts work openly. The Census Bureau is guided on this mission by scientific objectivity, a strong and capable workforce, a devotion to research-based innovation, and an abiding commitment to customers. The Census Bureau operates 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.

The Census Bureau's Disclosure Review Board (DRB) supports the Data Stewardship Executive Policy Committee in its efforts to protect Title 13 respondent confidentiality by proposing protection policies and methodologies, and reviewing external products, such as microdata and tabulation releases, for potential disclosure. The DRB coordinates activities that inform decisions made to protect confidentiality through data collection, linking, and dissemination.

PURPOSE STATEMENT

This report is an introduction into the DA procedures previously and currently practiced by the Census Bureau and the ongoing research into new DA practices (a modernization). The purpose is to explain the processes related to the application and development of DA procedures, as practiced by the DRB. The report exists to promote greater understanding of the importance of data stewardship and to encourage formal and transparent DA procedures at all federal statistical agencies.

The increasing risk of reidentification of respondents and their data means it is no longer adequate to employ ad hoc DA techniques. Protecting data confidentiality now necessitates proactive collaboration across agencies to move towards establishing shared, standardized DA methods with proven effectiveness. This evaluation of current and future DA techniques is a move towards greater transparency by the Census Bureau of the processes used in the publication of public data.

In an age that is increasingly data-driven, the Census Bureau is committed to maintaining the highest standards of confidentiality protection. Disclosure concerns are especially paramount on the scale at which the Census Bureau operates. The Census Bureau published at least 5.4 billion independent statistics on 308 million people after the 2010 Census of Population and Housing.

None of the information in this paper is confidential.

DESIRED OUTCOME OF DISCLOSURE AVOIDANCE TECHNIQUES AND RESEARCH: PUBLISHING QUALITY INFORMATION WHILE MAINTAINING CONFIDENTIALITY

The Census Bureau's mission is to serve as the nation's leading provider of quality data about its people and economy. The Census Bureau honors privacy, protects confidentiality, shares expertise globally, and conducts work openly. The Census Bureau is guided on this mission by scientific objectivity, a strong and capable workforce, a devotion to research-based innovation, and an abiding commitment to customers. All data, once collected or otherwise acquired by the Census Bureau, are protected by Title 13 and Title 26 of the U.S. Code.

To compile a more comprehensive and complete data resource, the Census Bureau uses data obtained from other sources to supplement data collected in censuses and surveys. Through an agreement with the Internal Revenue Service, the Census Bureau obtains Federal Tax Information, data whose disclosure is prohibited by Title 26 of the U.S. Code. Administrative records provided by other federal and state agencies are also used in conjunction with commercial data purchased from various data brokers to further support the Census Bureau's mission.

Under these regulations, no one other than the sworn officers and employees of the Census Bureau with a business need-to-know may examine the individual response reports. The Census Bureau must ensure that the privacy and confidentiality of responding individuals and companies (firms) are not compromised in any way.

[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 authors and not necessarily those of the U.S. Census Bureau.

This is supplemented by the Confidential Information Protection and Statistical Efficiency Act of 2002 that requires all federal agencies protect against the disclosure of information collected or acquired for exclusively statistical purposes under a pledge of confidentiality. Altogether, these laws mandate that the Census Bureau both publish and protect data, but they do not indicate how to balance these two requirements. Inherently, any release of data poses a risk to confidentiality. Balancing the trade-off between data utility and confidentiality is the imperative of policymakers.

These mandates have led to the development of DA techniques to prevent the reidentification of respondents after a public release of the data. These measures are necessary to uphold the public's trust and expectation that information will not be divulged in a manner inconsistent with the mission of the Census Bureau. For more information on reidentification, see the Section on Reidentification Studies.

CENSUS BUREAU STAFF WORKING ON DISCLOSURE AVOIDANCE PROTECTION TECHNIQUES, REVIEW, AND RESEARCH

The Center for Disclosure Avoidance Research

The Center for Disclosure Avoidance (CDAR) seeks to be the internationally recognized leader in DA research and technology, and to assist in the Census Bureau's mission in implementing methods that protect confidentiality in its data products. Located within the Census Bureau's Directorate for Research and Methodology, CDAR is charged with assisting the Census Bureau's programs in performing DA for their data products.

The CDAR develops and improves DA methods to ensure that the Census Bureau effectively disseminates the maximum amount of high-quality data about the nation's people, housing, and economy, while fully meeting the Census Bureau's legal and ethical obligation to protect the confidentiality of respondents and the information they provide. CDAR also runs the Census Bureau's DRB, which oversees data releases made by the Census Bureau to assure compliance with Title 13 and Title 26 requirements.

Disclosure Review Board

The mission of the DRB is to support the Data Stewardship Executive Policy Committee (DSEP) in its efforts to ensure that the Census Bureau protects Title 13 and Title 26 respondent confidentiality. This will include proposing policies and setting methodologies underlying confidentiality protection; reviewing external products for potential disclosure; identifying policy and research issues; and coordinating the confidentiality-related activities needed to inform decision-making on data collection, data linking, and data dissemination.

The DRB consists of representatives from each of the Census Bureau's program areas that publish data. Its chair can be contacted at <cdar.drb.chair@census.gov>. It has six voting members representing the Census Bureau's demographic, decennial, and economic directorates and at least three additional members representing the Census Bureau's research and policy areas.

The DRB serves as the focal point for issue identification, research coordination, and policy development on issues related to DA regarding the public release of all data products. It provides a mechanism for a comprehensive and consistent approach to DA in tabulations, microdata, and statistical products to ensure respondent confidentiality. The DRB reviews and clears all Census Bureau microdata, tabular data, and other data releases under its purview for confidentiality. It provides formal written responses with clearance for such requests. It reports to the DSEP and coordinates its efforts with division chiefs within affected directorates.[2]

The Federal Statistical Research Data Centers

The Census Bureau's Center for Economic Studies operates 25 Federal Statistical Research Data Centers (FSRDCs) across the United States.[3] FSRDCs are secure environments that allow approved outside researchers to access and explore confidential microdata not available to the public. Researchers seeking to use the FSRDCs must submit a proposal to the Census Bureau stating the research they wish to conduct, which restricted data sets they will need, and what type of results are to be published. Qualified researchers at these FSRDCs, like Census Bureau employees, must hold Special Sworn Status, and are therefore sworn for life to protect the confidentiality of the data they access.

[2] Disclosure Review Board Charter; available upon request from <CDAR.DRB.Chair@census.gov>.

[3] Federal Statistical Research Data Centers, U.S. Census Bureau, <www.census.gov/about/adrm/fsrdc/locations.html>.

Absolutely no data or research results may leave a FSRDC without approval from a CDAR Disclosure Avoidance Officer (DAO) or a designated FSRDC Administrator. Certain requests to release data from an FSRDC must be brought before the DRB, while others may be approved by the CDAR DAO. Requests containing household survey statistical data that contain any cells with counts smaller than 10 unweighted individuals, or contain output at a geographic level lower than the state level must petition the DRB for approval.

Large, complex requests for tabular data may be brought before the DRB for approval depending on the complexity of the request. This is done at the discretion of the CDAR DAO. The data released from FSRDCs are held to the same DA standards as the data released publicly by the Census Bureau.

THE RISK OF DISCLOSURE

Each national statistical institute and agency faces the social welfare problem of determining the balance between increased data utility and decreased disclosure risk. The Census Bureau's philosophy has always been to publicly release as much high-quality data as possible for the purpose of statistical analysis, while continuing to maintain its pledge of confidentiality. DA techniques are applied before data sets leave the Census Bureau for public release, cross-agency research, or otherwise. These techniques minimize instances of unauthorized disclosure, which can take three forms:

Identity disclosure—Reveals the identity of a data subject (individual or household).

Attribute disclosure—Reveals sensitive information about a data subject (individual, household, establishment, or company).

Inferential disclosure—The value of some characteristic of a respondent can be more accurately determined than otherwise possible (basic statistics).

The Census Bureau focuses on identity disclosure for individuals and households and attribute disclosure for individuals, households, establishments, and companies (firms). The proliferation of publicly available and/or proprietary data and advances in computing power have increased the risk of unauthorized disclosure. Two high-profile reidentification cases in recent years have emphasized these increased risks. In one instance, over 400,000 Netflix users were placed at unacceptably high risks of disclosure when publicly available Netflix data were linked with data from Internet Movie Database accounts and other public sources. In another instance, previously assumed anonymous health insurance data, in conjunction with publicly available voter rolls, demonstrated disclosure protection failure in the minimization of the risk of reidentification (Sweeney, 2001). By applying mathematical models, researchers were able to identify the private medical records of then-Massachusetts Governor William Weld. In subsequent research, health records could be successfully linked to news coverage in the state of Washington and analyzed in a similar fashion to reidentify 35 records out of a total 81 records.[4]

In addition to minimizing the risk of unauthorized disclosure, the Census Bureau must also minimize the perception of disclosure, even if none occurs. If respondents perceive that their confidential data will be disclosed or is at a high risk of disclosure, they will be less likely to participate in future Census Bureau censuses and surveys. Maintaining the public's trust is an integral step in ensuring that the Census Bureau continues to collect high-quality data.

TYPES OF DATA PRODUCTS

The Census Bureau makes most of the data collected available to the public for the benefit of both the private and public sector. The data that eventually gets published helps the Census Bureau to fulfill its mandated mission to help appropriate seats for the U.S. Congress with the additional benefit of also being beneficial to community planning and informative to the analysis of the U.S. economy. The structure of the data that are collected and published takes three primary forms, all requiring their own specific DA procedures. Microdata, frequency count data, and magnitude data make up the bulk of the published data that the Census Bureau makes available to the public.

Microdata

Statisticians use the term microdata to refer to any record-level data. At the Census Bureau, the term microdata has a narrower definition: it refers to collected data that have been cleaned, edited, and sometimes imputed so that they can be used to produce statistical tabulations and analyses. These data are still presented at the record level. A microdata file consists of data at the respondent level, as opposed to aggregate counts or magnitudes. Each record represents one respondent, such as a person or household, and consists of values of characteristic

[4] <https://dataprivacylab.org/projects/disclosurecontrol/paper1.pdf>.

variables for this respondent. Typical variables for a person-level demographic microdata file are age, race, sex, and income, and a household-level file might include mortgage amount or rent, year house built, and type of electricity. Microdata files may include hundreds of such variables for each respondent.

The Census Bureau releases microdata files from the decennial census and from many of its demographic surveys and a few economic surveys. Typical demographic surveys include the American Community Survey (ACS), the Current Population Survey, the Survey of Income and Program Participation, and the American Housing Survey.[5] Specific to either the decennial census or the ACS, the publicly available microdata files are called Public Use Microdata Samples (PUMS). The publicly available microdata files from all other demographic surveys are called Public Use Files (PUFs).[6, 7] The difference is because the PUMS from the decennial census and the ACS do not contain records from each respondent. They contain records from a sample of their respondents that can be released with an underlying layer of uncertainty. The added uncertainty exists from the inability to discern whether an individual respondent is captured in the PUMS files. This creates a scenario where a record with a unique combination of certain variables in the PUMS may not necessarily represent a unique person or household in the population (decennial census) or full sample (ACS). PUFS from other demographic surveys contain records for all respondents.

All microdata sets have the distinct purpose of providing the data user with the ability to analyze the data contained within the survey, but not always immediately available through the summary tables. The exceptions are the decennial census and the ACS (see the section on Frequency Count Data below).[8] The other surveys and their data sets, when compared to the ACS, are much smaller. Tabulating their data for small geographic areas would lead to low-quality tabulations. Thus, they are tabulated only for very large geographic areas, such as states, but have been perturbed or curtailed in detail through data DA measures specific to their microdata nature and the mathematical preferences of the survey sponsors and data users.

[5] <www.census.gov/ces/dataproducts/economicdata.html>.

[6] <www.census.gov/content/dam/Census/programs-surveys/ahs/tech-documentation/2015/Getting%20Started%20with%20the%20AHS%20PUF.pdf>.

[7] <www.census.gov/programs-surveys/acs/technical-documentation/pums.html>.

[8] <www.census.gov/programs-surveys/acs/technical-documentation/pums.html>.

Since microdata are used as the underlying data with which tables are built and analyses performed, protecting microdata file against unauthorized disclosure also has the effect of protecting any tables or results from statistical analyses constructed from those files. Current practices at the Census Bureau take this approach, which is especially helpful for online query systems that generate tables in real time to a user's specifications. By protecting the underlying microdata, this guarantees that any generated tables will pose no higher disclosure risk than the public-use microdata itself.

First steps in minimizing the risk of unauthorized disclosure of microdata include removing direct identifiers such as names, addresses, and Social Security numbers. High-risk records (e.g., individuals with very large incomes or unusual jobs) are identified to ensure their visibility within the file is decreased. Other characteristics are weighed for their uniqueness and their contribution to any increase in reidentification risk. Reidentification studies may be conducted to audit the disclosure risk associated with different combinations of released variables, taking into account publicly available or commercial data. Microdata associated with longitudinal or panel surveys that track the same group of respondents over time pose a greater risk as data can be linked over time.

Typically, the Census Bureau does not release microdata from economic surveys and censuses because the skewness of economic data usually makes it easy to identify establishments by only a few characteristics. Because of this, most published microdata is sourced from demographic surveys with few exceptions. Research has been undertaken to improve the utility of economic microdata while minimizing disclosure risk, but there is still no formal process for DA as of yet. Still, the Census Bureau has released microdata files from a few economic surveys: The Survey of Business Owners, the Commodity Flow Survey, the Survey of Construction, and the Manufactured Housing Survey.

Frequency Count Data

Frequency count data present the number of units of analysis in a cell. This can be the number in a population of individuals or establishments that have certain characteristics. This is the most basic data published by the Census Bureau and is used mainly to publish results from the decennial census, the ACS, and other demographic surveys. For example, a table may have columns representing marital status

and rows representing age ranges. Frequency count data are found in each cell and reflect the number of people that have a given combination of marital status and age range. Most tables have one, two, or three dimensions. A few have more.

Frequency count data are generally published at high levels of geography with the two exceptions being the decennial census and the ACS. All other demographic surveys are conducted on a much smaller scale than the decennial census and the ACS. They do not have adequate sample sizes at low levels of geography (e.g., at the block or county level) to support data quality for small geographic areas, thus frequency count data are only published at higher geographic levels (e.g., at the state or national level). All demographic surveys and the mandated decennial census and ACS must undergo DA procedures to defend against the risk of reidentification.

Magnitude Data

Like frequency count data, magnitude data are also presented in tables. However, magnitude data aggregates quantities of interest from individuals, households, or establishments within a cell instead of just a general count of respondents in a cell. For example, frequency count data may display how many establishments are operating in a given state broken down by industry, while magnitude data displays the count and the total gross revenues of all of those establishments within a cell. In tables of magnitude data, individual establishments and the companies they are affiliated with will vary, but the sum of their efforts remains the focus.

Magnitude data are most common for economic surveys and censuses, and those that ask about income, sales, revenues, shipments, and related values. These data are vulnerable to reidentification attacks because there often exists publicly available information that can be linked to Census Bureau data releases. For example, it may be publicly known which establishments are in a given cell.

DA protection is given at the company (firm) level. Values are commonly highly skewed at both the individual establishment and company level. Totals of establishment values within a given company in a cell figure into the calculation of cells that are at risk of disclosure and their required protection. See the section on Current Techniques for Magnitude Data to follow.

Other Types of Data

Special circumstances exist for the Census Bureau's Federal Statistical Research Data Centers as they are responsible for helping the Census Bureau publish a wide range of statistical models, research papers, and other data products that are not normally considered under standard or mandated data processing. These data products, like the data upon which they are built, must undergo the same disclosure review process to ensure confidentiality protections are in place before they are released. While the underlying data might exist in either microdata, frequency count, or magnitude form, it is presented in a way that is unique to other Census Bureau products and requires DA practices specific to each product.

Special circumstances also exist for comingled data. Comingled data result when the Census Bureau links its data to data from other agencies. In this case, Census Bureau DA rules and regulations must be applied to the data product(s), and the DA rules and regulations from the agency providing the data that are linked to the Census Bureau data must also be applied.

Finally, special circumstances include published data that are not presented in the typical forms of microdata, frequency count data, and magnitude data. Information in the form of means, medians, quantiles, aggregates, percentages, negative values, and graphs can follow the procedures for the three main types of data. This will be described in the three sections on Current Techniques.

CURRENT TECHNIQUES FOR MICRODATA

For any given microdata product, the Census Bureau may use a combination of the techniques described below. Note that almost all Census Bureau microdata products are from the decennial census and demographic surveys. Economic data from establishments are very difficult to protect because many values of interest are publicly released and could be used to link Census Bureau data with data from outside sources. Also, economic data are skewed and thus difficult to hide.

Removing Information for Microdata

Remove Direct Identifiers

Beginning with the obvious, the Census Bureau removes direct identifiers such as name, address, telephone number, social security number, establishment identification numbers, etc.

Topcoding and Bottom-Coding

Topcoding and bottom-coding are used to eliminate outliers in a file. They are used for continuous variables such as age and dollar amounts. When topcoding, the top 0.5 percent of all values or the top 3 percent of all nonzero values are cut off, whichever is the higher topcode cutoff. They can be replaced with the topcode cutoff value, or the mean or median of all topcoded values. At least three values must be included in the topcode or it will be lowered to meet this criterion. Bottom-codes are the same except on the other end of the distribution. An example of a bottom-code might be the year that a building was built or gross income. For variables that are part of a sum, the individual parts are topcoded before anything is summed.

Beginning in 2001, the Current Population Survey (CPS) Annual Social and Economic supplement (ASEC) topcoded values were replaced with values generated from a technique called "Rank Proximity Swapping."[9] The technique preserves the distribution of values while maintaining adequate disclosure protection.[10] People/households with values above the topcode are sorted and ranked by those values from lowest to highest, and those values are swapped between the people/households within a given interval of rank. The bounded interval is large enough to include many people/households in order to protect the data and small enough to ensure that the swapped values are within "proximity" of each other. The parametric details of this are confidential. All values must be swapped with another value, and all of the values are also rounded to two significant digits. A value cannot remain the same through randomness, unless due to the rounding to two significant digits.[11, 12]

Recoding and Rounding

Recoding is done for categorical and continuous variables. Each category of a variable must contain nationwide at least 10,000 weighted people or households (depending on the universe for that variable). This is imposing a categorical threshold. Otherwise, the category must be combined with another until the rule is met. For continuous data values that the Census Bureau knows are public information and some dollar amounts, recoding is also applied. One value that is publicly available is property taxes. This follows the recoding scheme found in Appendix A.

Other dollar amounts may follow one of two rounding/recoding schemes.

Round to the nearest 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.

Geographic Population Thresholds

All geographic areas identified on Public Use Microdata Samples (PUMS)/Public Use Files (PUFS) must have a weighted population of 100,000 or more. The thresholds are larger for some surveys. There are some surveys, such as the Survey of Income and Program Participation (SIPP), that have an extraordinary amount of variables and detail of those variables. The geographic areas identified in SIPP must have at least a 250,000 weighted population or more.

For the 2000 Census, a PUMS file was released containing information on 5 percent of the population with a population threshold of 100,000. A second PUMS file was released containing information on 1 percent of the population with much more detail in variables than the 5 percent file and a population threshold of 400,000.

There are some surveys that are potentially linkable to outside files, for example, the National Survey of College Graduates and the National Crime and Victimization Survey. The geographic areas can be even larger for those (e.g., census division or region).

When figuring out the population of an identified area, all geography-related variables on the file must be crossed to obtain the final population count. For example, other geographic variables may be urban/rural, Metropolitan Statistical Area Status, and other geographic areas named such as Congressional District. All geographic pieces identified after crossing

[9] <https://cps.ipums.org/cps/inctaxcodes.shtml>.

[10] <https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar17.pdf>.

[11] <https://www2.census.gov/programs-surveys/demo/datasets/income-poverty/time-series/data-extracts/pu-swaptopcodes-readme.docx>.

[12] The CPS ASEC rank swapping of topcoded values was reviewed in stages (six different meetings) by the DRB from April 26, 2010, to March 7, 2011, when it was finalized and approved.

all geographic variables must meet the required threshold for that PUMS/PUFS.

Altering Information for Microdata

Data Swapping and Synthetic Data

Data swapping and the generation of partially synthetic data are current methods for the protection of frequency count data from the decennial census and ACS. This will change with the introduction of formal privacy (Nissim et al., 2018). See the section on current research. While the two methods are used mainly to protect tables for very small geographic areas, both methods are performed on the underlying microdata before tabulation. The PUMS files are sampled from the swapped and partially synthesized data. Details on the two procedures are found in the section on Altering Information for Frequency Count Data.

Noise Infusion

At this time, noise infusion is not widely used for the protection of microdata. It is used to hide very unusual characteristics of a person or household at a given point in time that is not caught by the 10,000-threshold rule for individual categories described above. For example, consider a person who gave birth to seven children at one time or a person who is a practicing physician at the age of 15 (both very unusual circumstances that would probably be in the news). Also, very large households may present a disclosure problem. Editing procedures capture and alter many but not all of these unusual occurrences.

Noise is also used in longitudinal files to hide a change in a personal or household circumstance that could be found in publicly available records, for example, a birth, death, marriage, or divorce that would be reflected in the longitudinal microdata file. The Census Bureau does not publicly describe precisely how noise is added to protect this type of data.

CURRENT TECHNIQUES FOR FREQUENCY COUNT DATA

Removing Information for Frequency Count Data

Rules for Ratios, Graphs, Geographic Levels, Number of Table Dimensions, Means, Aggregates, and Medians

For ratios, see the section on Rounding below. For graphs, usually the counts are being released as well, so standard rules are applied to the counts, and then the graphs are created from that data set. If the data are not going to be published with the graph, then rasterized images are preferred unless the underlying data pass all rules.

The other rules are found in memos on tabular data from the ACS standard tabular data products (those from a list compiled by the Census Bureau) and special (customized) tabulations (requested and paid for by a specific data user but then made public) from the decennial census and ACS. These are summarized below.

Rules for ACS standard tabular data products:

1. There must be at least 50 unweighted cases in the geographic area for the data to be released for that area. When data are suppressed by this rule, complimentary suppression must be performed on other areas so that the suppressed areas' data cannot be derived via subtraction. To streamline the process, the Census Bureau has decided to suppress whole tables rather than to perform complementary suppression.
2. All medians in the ACS program are calculated using interpolation. A distribution of the variable in question is created, and the median is interpolated based on that distribution. Any medians that match a particular respondent's reported value occur due to coincidence only.
3. For Rule 2, medians are calculated using a linear method. The bins for the distributions used for medians vary by topic, and are not necessarily uniform. The bins for the income medians are of an interval of at least 2,500, which allows the ACS program to not have to use the Pareto method of interpolation.
4. Estimates in the form of means or aggregates, defined here as a sum of the values for each of the elements in the universe (e.g., the sum of the income of all households in a given geographic area), must be based on either zero unweighted cases or three or more unweighted cases in a geographic area to show the mean or aggregate for that area.
5. When there is a mean or aggregate in a table for a given geographic area that is suppressed by Rule 4, complimentary suppression must be performed on other means or aggregates so that the suppressed mean or aggregate cannot be derived via addition and subtraction due to totals. In practice, the Census Bureau does not

do complementary suppression, but instead, suppresses the whole table.

6. Estimates in aggregate tables involving figures in dollar amounts are rounded to the nearest 100. Estimates in aggregate tables involving travel time to work are rounded to the nearest 5 minutes. Estimates of aggregate travel time to work are rounded to 5 minutes. Estimates of the aggregate number of vehicles used in commuting by workers 16 years and over by sex is rounded to the nearest 5. Estimates of aggregate hours worked, aggregate number of rooms, and aggregate number of vehicles available by tenure are not rounded. For estimates rounded to the nearest 100, estimates of –100 to 100, not inclusive, are rounded to zero. Estimates from –7.5 to zero, not inclusive, are rounded to –4 and estimates of zero to 7.5, not inclusive, are rounded to 4.
7. Tables on the number of people in a household and average household and family size, although technically aggregates and means, are not subject to the rules for means and aggregates.
8. If an aggregate income table is suppressed, then the per capita income table associated with that aggregate table is also suppressed.
9. Tables involving a geographic area other than current place of residence (such as workplace, place of birth, residence 1 year ago) crossed with characteristics other than current place of residence must have at least 50 unweighted cases in the universe of the table.
10. Tables with more than 100 categories for a non-geographic characteristic variable (excluding totals and subtotals) cannot be released for block groups and tribal block groups. If a table is iterated by a variable, such as race/ethnicity or gender, the set of iterated tables should be considered as a single table. The iterated variable should be considered a dimension when counting the number of lines in the table to determine if the set of iterated tables can be released for block groups and tribal block groups.
11. Certain other tables will not be published for block groups and tribal block groups. They include:
    - Tables where the universe is restricted to the foreign-born or a subset of the foreign born.
    - Tables containing estimates of or characteristics of noncitizens.
    - Tables containing characteristics of unmarried partners. Tables containing estimates of, or characteristics of, people that were married, widowed, divorced, or became mothers within the last 12 months. Tables containing characteristics of people living in group quarters (GQs).
    - Tables containing detailed type of GQs (categories that can be shown at the block-group level are institutional and noninstitutional).
    - Tables containing detailed language categories (categories that can be shown at the block-group level are English, Spanish, Other Indo-European, Asian/Pacific Islander, and All Else).
    - Tables containing specific type of disability, disabled by race, or categories of number of disabilities other than “0,” “1,” or “2 or more.”
12. Tables of unweighted counts of people and housing units may only be shown for areas where there are no occupied housing units or three or more unweighted occupied housing units.

Rules for special tabulations from the decennial census:

1. Special (customized) tabulations are sets of tables created under reimbursable agreements, those done for working papers, tables, professional papers, etc. All decennial census special tabulations must be reviewed by the DRB.
2. All cells in any 2000 Census or 2010 Census special tabulations must be rounded. The rounding schematic is:
    - Zero remains zero.
    - 1 to 7 rounds to 4.
    - Eight or greater rounds to the nearest multiple of 5 (i.e., 864 rounds to 865, 982 rounds to 980).
    - Any number that already ends in “5” or “0” stays as is.
        - This rounding applies to all special tabulations that pertain to the population in households or the population in group quarters.
        - Any totals or subtotals needed should be constructed before rounding. This assures

that universes remain the same from table to table, and it is recognized that cells in a table will no longer be additive after rounding.

3. Medians or other quantiles may be calculated as:
   - An interpolation from a frequency distribution of unrounded data (these are not subject to additional rounding).
   - As a point quantile. These must be rounded to two significant digits: 12,345 would round to 12,000; 167,452 would round to 170,000. There must be at least five cases on either side of the quantile point. It is recognized that the interpolated quantile may coincidentally be an individual's response.
4. Thresholds on universes will normally be applied to avoid showing data for very small geographic areas or for very small population groups (often 50 unweighted cases for sample data). Tables may normally not have more than three or four dimensions, and mean cell size lower limits may also be required (mean cell size of each table is at least three for 100 percent data, or 20 weighted for sample data).
5. Percentages, rates, etc., should be calculated after rounding, but the DRB has granted exceptions to this rule when the numerator and/or denominator of the percent or rate is not shown.
6. Means and aggregates must be based on at least three values.
7. The finest level of detail shown for group quarters data will be institutional/noninstitutional.
8. For Demographic Profiles from user-defined geographic areas (neighborhoods), all areas must have at least 300 people in them. Using a computer program, the user-defined areas will be compared with standard Census Bureau areas to make sure users cannot obtain data from very small geographic areas by subtraction. If such small areas are found, the boundaries of the user-defined areas must be changed.

Rules for special tabulations from the ACS:

1. All ACS special tabulations must be reviewed by the DRB. After the tabulation has been created, if the program area identifies any potential disclosure problems, they will refer them back to the DRB.
2. All cells in any ACS special tabulation must be rounded. The rounding schematic for all tables is:
   - Zero remains zero.
   - 1 to 7 rounds to 4.
   - 8 or greater rounds to nearest multiple of 8 (i.e., 864 rounds to 865, 982 rounds to 980).
   - Any number that already ends in "5" or "0" stays as is.

   Any totals or subtotals needed should be constructed before rounding. This assures that universes remain the same from table to table, and it is recognized that cells in a table will no longer be additive after rounding.
3. Medians or other quantiles may be calculated as:
   - An interpolation from a frequency distribution of unrounded data (these are not subject to additional rounding).
   - As a point quantile. These must be rounded to two significant digits: 12,345 would round to 12,000; 167,452 would round to 170,000. There must be at least five cases on either side of the quantile point.

   It is recognized that a quantile may coincidentally be an individual's response.
4. Thresholds on universes will normally be applied to avoid showing data for very small geographic areas or for very small population groups (often three or 50 unweighted cases). Tables may normally not have more than three or four dimensions, and mean cell size lower limits may also be required (mean cell size of each table is at least three unweighted cases).
5. Percentages, rates, etc., should be calculated after rounding, but the DRB has granted exceptions to this rule when the numerator and/or denominator of the percent or rate is not shown.
6. Means and aggregates must be based on at least three values.
7. Universes allowed for GQs data are as follows.
   - Noninstitutional:
     - College dormitory facilities.
     - Military facilities.
     - Other facilities.

- Institutional:
    - Nursing facilities and skilled-nursing facilities.
    - Adult correctional facilities.
    - Juvenile correctional facilities.
    - Other facilities.

    For a given geographic area and a given data product (1-, 3-, or 5-year), there must be at least 50 unweighted cases in any given type of facility (as well as 50 in an Other category) and those 50 cases must come from at least three different facilities. Categories may be combined to reach these thresholds. Previously released requests will be considered to ensure that there are no complementary disclosure problems.

8. For demographic profiles from user-defined geographic areas (neighborhoods), all areas must have at least 300 (weighted) people in them. Using a computer program, the user-defined areas will be compared with standard Census Bureau areas to make sure users cannot obtain data from very small geographic areas by subtraction. If such small areas are found, the boundaries of the user-defined areas must be changed.

Cell Size Thresholds

The Census Bureau often requires a minimum unweighted count for each cell. Counts of zero are not considered disclosures, but very small counts are. For example, the minimum unweighted cell size must be at least three. If the count was one, that particular unit, say a person, can easily find himself in that table. If the count was two, one of the units can find herself in that table, remove herself from the cell and perhaps identify the second person in that cell. She may know the person and know that they were in that same table cell for a given census or survey.

She may also be able to connect overlapping, related tables, that can help connect a given person to that cell, and discover additional information about that person. Another problem with small cells that appear in many different overlapping tables is that a data user could potentially link the other information in those tables to form a microdata record for a small geographic area or even a partial microdata record that could be linked to a microdata file released from that same census or survey. In the past, the Census Bureau has used a cell size threshold of three, but the Census Bureau is now considering increasing that to five due to results of an empirical study by a member of the DRB.

For Title 26 counts and estimates from Internal Revenue Service (IRS) data and comingled data (from the Census Bureau and the IRS), IRS requires these thresholds:

For establishment data:

- 5 companies (firms) for national estimates.
- 10 companies for state-level estimates.
- 20 companies for substate-level estimates, except for zip codes.
- 100 companies for ZIP code-level estimates.

For housing unit data:

- 3 housing units for national estimates.
- 10 housing units for state-level estimates.
- 20 housing units for substate-level estimates, except for zip codes.
- 100 housing units for ZIP code-level estimates.

Cell Suppression

Cell suppression is the most common, and oldest, applied tabular DA technique for protecting economic magnitude data at the company or establishment level (see Cell Suppression in the section on Magnitude Data). Cell suppression may also be used for frequency count data. Utilizing Census Bureau software, sensitive cells (those with small unweighted counts) are recognized and then suppressed (the estimate is replaced with the letter “D”) from the published data. The Census Bureau has excellent methods and software for cell suppression for magnitude data that can be easily modified to use on frequency count data. Previous iterations of these methods relied on a network flow-based system for two-dimensional tables, that same system with a heuristic for large tables of three or more dimensions, a linear programming system for small tables of three of more dimensions, and an auditing system based on linear programming to identify any cases of undersuppression for all tables up to four dimensions. They are now conducted within a much-improved linear programming-based system for all tables. Linear programming systems satisfy disclosure requirements for higher-dimensional tables, where network flow systems are most appropriate for two-dimensional

tables. Additionally, linear programming systems for cell suppression can handle linked tables without a significant decrease in data protection and quality. Issues of scale do arise, and the economic census tables are the largest scale with which the system can be applied. The majority of cell suppression cases arise in economic magnitude data (Massell, 2011; Steel, 2013). As mentioned above, it is occasionally used for frequency count data, and again, the current cell suppression software is quite easy to modify for this purpose.

The cell size threshold rule is integral in determining the sensitivity of a primary cell targeted for suppression. Cells of size zero are never suppressed. There is nothing to protect, and they cannot offer protection to other cells. The primary basis of the threshold rule is that a cell must be given enough protection so that a data user cannot estimate that the cell has an unweighted count of one to the threshold minus one. For example, if the threshold is three, a user should not be able to estimate that a cell value can be estimated to (1, 3-1), that is (1, 2). Additional cells almost always must be selected and suppressed to ensure that primary suppression values cannot be derived via addition and subtraction of published values. These are called complementary suppressions and are necessary due to a table's published marginal totals.

The shift from network flow to linear programming aggregates protection through more efficient processes than the ones previously under use. The new software for frequency count data protection through linear programming greatly reduces undersuppression and oversuppression. Thus, more data can be published while still fulfilling protection requirements (Steel, 2013). Note that unless tables are extraordinarily small, unrelated and with extremely few primary suppressions, software is necessary to identify complementary suppressions. It is practically impossible to do this correctly by hand.

Collapsing or Recoding of Rows, Columns, and Other Dimensions

A second way of eliminating small unweighted cell counts is collapsing or recode rows, columns, and other dimensions until all small cells are gone. If the release involves many related (overlapping) tables, then a collapsing/recoding scheme that works for one table may not work for all tables. If the schemes do not match for the related tables, then possibly the tables could be examined together and small counts revealed. When dealing with a single table and very few small cells, this may be the best way to go.

The FSRDCs' DA Guidelines on Rounding

Most counts and estimates in released outputs must be rounded. Rounding helps minimize disclosure risk within and between FSRDC projects, usually with minimal loss to the usefulness of the data.

FSRDC administrators have code in both Stata and SAS that they can share with researchers to aid in this matter. It is important to understand that merely changing the format of a number so that it appears rounded is not sufficient, as the unformatted value is often retained. The underlying value itself must be rounded.

Number of observations and related integers

All reported Numbers of Observations (Ns), whether weighted or unweighted, must be rounded. This is true even for large Ns.

An exception is made if the count is at a level substantially different from microdata and does not reveal microdata counts. For example, sometimes researchers use tract, county, or industry as the unit of analysis.

Other integers related to observation counts must also be rounded. For instance, degrees of freedom are often closely related to sample size.

The rounding scheme is as follows:

- If N is less than 15, report N < 15.
- If N is between 15 and 99, round to the nearest 10.
- If N is between 100-999, round to the nearest 50.
- If N is between 1,000-9,999, round to the nearest 100.
- If N is between 10,000-99,999, round to the nearest 500.
- If N is between 100,000-999,999, round to the nearest 1,000.
- If N is 1,000,000 or more, round to four significant digits.

Summary statistics/model-based output

Summary statistics and model-based estimates must be rounded to four significant digits. These include, but are not limited to:

- Means.
- Standard deviations/standard errors.
- Correlations.
- Test statistics.
- Degrees of freedom.
- Model coefficients.

In most cases, this rounding scheme will not have a noticeable effect on the researchers' ability to make inferences with their data. However, it may be a hindrance in special circumstances, for instance in simulation studies. If a researcher has a demonstrated analytical need for an exception, the request can be sent to the DRB for review.

Ratio estimates use the FSRDCs rounding rules but can be brought before the DRB if the sponsor wishes to ask for more detail. Rounding of both the numerators and the denominators is applied before the ratios are calculated. Ratios can be shown to four significant digits if the denominator is not reported or otherwise follows the FSRDC rounding rules.

Tables of Percentiles and Quantiles

Percentiles and other quantiles may be calculated in one of two ways. If they are calculated as an interpolation from a frequency distribution of unrounded data, no additional rounding is required. Otherwise, point quantiles must be rounded to two significant digits and at least five nonoverlapping observations must be on either side of each quantile point.

Altering Information for Frequency Count Data

The data swapping and partially synthetic data techniques described below are mainly used to protect frequency count data for very small areas from the decennial census and the ACS. The swapping is used to protect household data, and the synthetization is used to protect group quarters data. As mentioned above, the PUMS for the census and ACS are taken from the swapped (household) and synthesized group quarters data.

Data Swapping

The purpose of any swapping methodology is to introduce uncertainty into the tables so that the data user does not know whether real data values correspond to certain respondents. Household records with a high risk of disclosure are typically identified through software and called uniques because they have a unique combination of certain variables. Those records are targeted for the swapping procedures. In the swapping procedure, a small percentage of records are matched with other records in the same file on a set of predetermined variables used as swapping attributes. A set of one (or more) other variables are then swapped between the two records without disturbing the responses for nonsensitive and nonidentifying fields. The variables may be continuous or categorical. A household record is typically swapped with another household within a large area but in a different smaller area within the larger one, for example across tracts but within the same county.

Thus, data swapping can impact data quality at the smallest geographic areas, but it has the advantage that tabulations at larger geographic areas are unaffected. As another example, if a pair of households are swapped between two blocks in different towns but within the same county, the town tabulations will be changed, but the county tabulation will be unchanged.

The swapping procedure is simple and requires only the microdata file and a random number generating routine to implement with straightforward programming. However, depending on the number of records and variables, it may take a significant amount of time and computer resources to swap and store the original file and the swapped version. Due to size, this would be more likely to happen for the decennial census than the ACS.

The greater the percentage of swapped records, the greater the losses in data utility of the tables, and while swapping does not change most marginal distributions of any variable in a file, it does distort joint distributions involving both swapped and unswapped variables. If used, arbitrary swaps may produce a large number of records with unusual combinations, for example swapping a clerk's income with a brain surgeon's income.

Currently, when a pair of records is swapped, the two households have the same number of people and the same number of those 18 years and over and 17 years and under. This is not a legal requirement. Statistical DA is not a prohibited technique. Utah vs. Evans upheld the use of vetted statistical methods other than sampling, even if they change the population totals used for reapportionment (e.g., response acceptances in the decennial response file and edits/imputations applied to the census unedited file).

The swapping process uses the smallest geography identified in a set of tables when identifying records at risk of disclosure. Other things are considered when thinking about the risk of a household record. For decennial data, there was a threshold value for not swapping in blocks with a high imputation rate (block groups for ACS). Records that were unique in their decennial block or ACS block group, based on a set of key demographic variables, were swapped. The probability of being swapped had an inverse relationship with block (or block group) size. In addition, records representing households containing members of a race category that appeared in no other household in that block (or block group) had an additional increase in the probability of selection. All data products were consequently created from the swapped file (Zayatz, 2003).

Partially Synthetic Data

Applying data swapping to GQ data does not work well. Imagine swapping a nursing home (or someone who lives there) with a college dorm (or someone who lives there). The resulting data would make no sense, so the Census Bureau relies on the generation of partially synthetic data to protect GQ data from the census and ACS.

The original data are modeled using a general linearized model. The process then continues with identifying unique records by cross-tabulating certain values and flagging records in the resulting cells with a count of one (in this case, representing people rather than households). Those variable values that are causing the disclosure risk problem in a given unique record are then blanked and replaced with values generated from the model. Geography and type of GQ are never altered, and the numbers of people aged less than 18 and aged 18 or more are never changed. Occasionally, a modeled (simulated) value may coincidentally be the same as the original value.

CURRENT TECHNIQUES FOR MAGNITUDE DATA

Removing Information for Magnitude Data

Cell Suppression

Cell suppression is the most common, and oldest, applied tabular DA technique for protecting economic magnitude data at the company (firm) or establishment level, or both. Utilizing this system, sensitive cells are recognized and then suppressed. The estimate is replaced with the letter "D" in the published data. Previous iterations of this method relied on a network flow-based system for two-dimensional tables with or without a hierarchical structure in one of the dimensions, a linear programming system for small tables with three or more dimensions, and an auditing system that also used linear programming methods to identify any cases of undersuppression for all tables. They are now conducted within a much-improved linear programming-based system for all tables. Linear programming systems satisfy disclosure requirements for higher-dimensional tables, where network flow systems are most appropriate for two-dimensional tables. Additionally, linear programming systems for cell suppression can handle linked tables without a significant decrease in data protection and quality. Issues of scale do arise, and the Economic Census tables are the largest scale with which the system can be applied. The majority of cell suppression cases arise in economic magnitude data (Steel, 2013). As mentioned above, it is occasionally used for frequency count data.

The p-percent (p%) rule is integral in determining the sensitivity of a primary cell targeted for suppression. The primary basis of the p% rule is that an establishment or company's value cannot be estimated more precisely than within p% of its true value (SPWP22, 2005). The value for "p" is never published and is considered confidential information. Additional cells almost always must be selected and suppressed to ensure that primary suppression values cannot be derived or estimated too closely via addition and subtraction of published values. These are called complementary suppressions and are necessary due to a table's published marginal totals. Values of zero are never suppressed because there is nothing to protect, and they offer no protection. For information on determining if a cell qualifies as a primary suppression and to determine how much protection is necessary, see Statistical Policy Working Paper 22 (SPWP22) and Appendix B. The Census Bureau publishes some economic data that have already been rounded to the nearest $1,000 or even $1,000,000. For information on identifying primaries and calculating their amount of needed protection for rounded data, see Appendix B.

The Census Bureau is required to protect economic data at the company (firm) level, as well as at the establishment level. In order to meet this requirement, DA measures consistently have to be evaluated. The shift from network flow to linear programming aggregates protection through more

efficient processes than those previously used. Aggregate company protection through linear programming greatly reduces undersuppression and oversuppression. Thus, more data can be published while still fulfilling protection requirements (Steel, 2013). Note that unless tables are extraordinarily small, unrelated, and with extremely few primary suppressions, software is necessary to identify complementary suppressions. It is practically impossible to do this correctly by hand.

Rolling up Rows, Columns, and Other Dimensions

A second way of eliminating cells that violate the p% rule is to "roll-up" (combine) rows, columns, and other dimensions until all primary suppressions are gone. If the release involves many related (overlapping) tables, then a roll-up scheme that works for one table may not work for all tables, and if the schemes do not match for the related tables, then possibly the tables could be examined together and values of primary suppressions revealed. When dealing with a single table and very few primary suppressions, this may be the best way to go. This is very similar to collapsing or recoding rows, columns, and other dimensions for frequency count data.

Altering Information for Magnitude Data

EZS-Balanced Noise Addition

A different technique is used for many of the Census Bureau's economic data products. This technique, commonly referred to as EZS noise, is applied to the underlying microdata prior to tabulation (Evans, Zayatz, and Slanta, 1998). Each responding company's data are perturbed by a small amount, say at least 10 percent in either direction. The actual percentage used by the Census Bureau is confidential. Noise is added in such a way that cell values that would normally be primary suppressions, thus needing protection, are changed by a large amount, while cell values that are not sensitive are changed by a small amount. Noise has several advantages over cell suppression. It enables data to be shown in all cells in all tables, it eliminates the need to coordinate cell suppression patterns between tables, and it is a much less complicated and less time-consuming procedure than cell suppression. Because noise is added at the microdata level, additivity of the table is maintained.

To perturb an establishment's data, for example, by about 10 percent, the Census Bureau would multiply its data by a random number that is close to either 1.1 or 0.9. Any of several types of distributions may be used from which to choose the multipliers, and the distributions remain confidential within the agency. The overall distribution of the multipliers is symmetric about one. The noise procedure does not introduce any bias into the cell values for census or survey data. Because the Census Bureau protects the data at the firm (company) level, all establishments within a given firm are perturbed in the same direction. The introduction of noise causes the variance of an estimate to increase by an amount equal to the square of the difference between the original cell value and the noise-added value. An agency could incorporate this information into published coefficients of variation.

Building on the former Statistical Research Division's work to protect magnitude data with noise, the Longitudinal Employer Household Dynamics (LEHD) program developed methods for using noise infusion to protect ratios and percentages in a systematic way that allows the effect on inferences based on the released estimates to be specified. The following surveys now use noise infusion to protect their data: Nonemployer Statistics, Integrated Longitudinal Database, the LEHD Quarterly Workforce Indicators, workplace information for a key product from the LEHD program called OnTheMap, Commodity Flow Survey, Survey of Business Owners, and County Business Patterns. Cell suppression is still the method of choice for the stateside Economic Census, but noise infusion is now used for the Economic Census of Island Areas.

In some surveys where data are protected using noise, a single table is designated to be the most important table. For these surveys, staff developed an enhanced version of the EZS methodology, called "balanced noise." Here, noise factors are not assigned randomly to each of the microdata records. Instead, select records are placed into small groups, which are defined by the unique interior cells of the table to which they contribute. The noise factors are then assigned to each of these groups by alternating the direction of the noise factors to each contributing record. This process enhances the amount of noise cancellation in most cells and results in cells closer to the true values. Balanced noise is more complicated to implement than random EZS noise, but the improved accuracy of the "most important table" is often worth the extra effort. Massell and Funk (2007) found that the effect of balanced noise on one

table does not typically hurt the accuracy on other produced tables, while guaranteeing the protection of the data.

REIDENTIFICATION STUDIES

When legacy 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 it is 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 dataset 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 (Ramachandran, 2012). A few small reidentification attempts were made with microdata files by summer interns in the early 1990s, but they yielded nothing of substance. The most recent reidentification study for microdata at the Census Bureau was done for the American Housing Survey public-use microdata file, which is funded by the Department of Housing and Urban Development. Studies were also conducted on Survey of Income and Program Participation and the ACS.

For tabular data, reidentification studies often attempt to link tables produced from a given survey or census. The goal is to determine if there are cells appearing in several tables that could be linked together to form microdata records for people or households in small geographic areas. The most recent (completed) reidentification study for tables at the Census Bureau was done for ACS special tabulations to be produced for the Census Transportation Planning Products, funded by the American Association of State Highway and Transportation Officials.

Although results cannot be publicly released, recent studies were greatly beneficial to the DRB. They pointed to particular variables or combinations of variables on these files that could potentially be used to reidentify someone. As a result, either noise was added to the variables or the variables were recoded or dropped completely from some tables.

CURRENT SOFTWARE THAT THE CENSUS BUREAU DEVELOPED AND USED TO APPLY THE DISCLOSURE AVOIDANCE TECHNIQUES LISTED ABOVE

There is currently no standard software for microdata or other types of data. This is because these types of data currently rely on very simple protection methods such as topcoding, thresholds, and rounding.

Cell Suppression Software for Frequency Count Data

There are a few software packages that have been developed for this purpose, however some were developed in-house and some were purchased from outside companies. The DRB is waiting to see more tests of these software packages and an evaluation of results that shows proof that they work correctly.

Data Swapping Software for Frequency Count Data

This software is currently being used to protect household data from Census Bureau decennial censuses and the ACS.

Partially Synthetic Data Software for Frequency Count Data

This software is currently being used to protect Group Quarters data from Census Bureau decennial censuses and the ACS.

Advanced Cell Suppression Software for Magnitude Data

The former cell suppression software recently went through a major overhaul. The new version can process much more data at one time, and it greatly cuts down on the amount of undersuppression and oversuppression.

Software for the EZS-Balanced Noise Addition for Magnitude Data

This software is great to have as an alternative to cell suppression. It has the advantages of being simple to run, modify, and understand, and recall that with it, a value may be published for every cell in a table.

CURRENT RESEARCH IN DISCLOSURE AVOIDANCE METHODS

Most of the current Census Bureau DA research is focused on formal privacy for all types of data (Nissim et al., 2007; Dwork, 2006). It is being planned for the 2020 Census, <https://privacytools.seas.harvard.edu/forma-privacy-models-and-title-13>. Formal privacy is an expansion of differential privacy.

According to Nissim et al., 2018, "Differential privacy is a strong, mathematical definition of privacy in the context of statistical and machine learning analysis. It is used to enable the collection, analysis, and sharing of a broad range of statistical estimates, such as averages, contingency tables, and synthetic data, based on personal data while protecting the privacy of the individuals in the data.

"Differential privacy is not a single tool, but rather a criterion that many tools for analyzing sensitive personal information have been devised to satisfy. It provides a mathematical provable guarantee of privacy protection against a wide range of privacy attacks, i.e., attempts to learn private information specific to individuals from a data release. Privacy attacks include reidentification, record linkage, and differencing attacks, but may also include other attacks currently unknown or unforeseen. Those concerns are separate from security attacks, which are characterized by attempts to exploit vulnerabilities in order to gain unauthorized access to a system.

"Computer scientists have developed a robust theory for differential privacy over the last 15 years, and major commercial and government implementations have now started to emerge."

Research discussed below for Frequency Count Data and Magnitude Data will likely end soon and be replaced with research on formal privacy methods for all data products.

Microdata

See formal privacy in the section above.

Frequency Count Data

Data swapping is the main procedure used to protect decennial census and ACS tabulations. A small amount of household records are swapped with partner households in a different geographic area. The selection process to decide which households should be swapped is highly targeted to affect the records with the most disclosure risk. For example, households in very small geographic areas and those that are racially isolated are targeted. Households swapped with each other match on a minimal set of demographic variables. Public-use microdata, tables, and all other data products are created from the swapped data files. After performing the data swapping for the 2000 Census and the 2010 Census, the Census Bureau did an extensive evaluation of the procedure and the resulting tables' preservation of data quality. The results of this evaluation are confidential but the effects of the data swapping were minimal compared to sampling, measurement, coverage, and nonresponse error already present.

The Census Bureau continually conducts research to adapt and improve the swapping procedures. Over the past few years, the Census Bureau has altered the swapping routine, changed the variables used to determine which households are at risk, and slightly increased the percentage of households that are swapped. Staff also researched n-cycle swapping and rank swapping, which both increased data utility and data protection (DePersio, 2012; Lauger, 2015). These were never implemented and research stopped due to a change in direction (formal privacy) for protecting tables from the 2020 Census and future ACS tabulations.

Synthetic data are used to protect some of the data from the decennial census and the ACS. Both programs collect data for both residential households and group quarters. Swapping is infeasible for group quarters, thus the Census Bureau now uses partially synthesized group quarters data for these programs (Hawala, 2008). The ACS Census Transportation Planning Products special tabulations also use partially synthetic data, as well as other DA techniques (Li, 2011). Partially synthetic data has room for improvement.

For demographic frequency count tables, totals are constructed before rounding, so the universes remain the same from table to table but the tables may no longer be additive. Percentages, rates, and ratios are calculated after rounding. The Census Bureau allows some exceptions when the numerator, denominator, or both are not shown. Tables usually must have no more than three or four dimensions. Thresholds on universes are often applied to avoid showing data for small geographic areas or small population groups. Usually, any cells with an unweighted count of one or two are not published and for survey data, usually, only weighted estimates are published. The Census Bureau may study these requirements further to make sure the rules and procedures used are adequate.

Magnitude Data

The Census Bureau publishes magnitude tabular data from its economic surveys and the Economic Census. Most magnitude data come from economic data products. Tables of magnitude data usually contain both the frequency counts of establishments in each cell and the aggregate of some quantity of interest over all units (establishments) in each cell.

For example, a table may present the total value of shipments within the manufacturing sector by North American Industry Classification System code by county. The frequency counts in the tables are not considered sensitive because so much information about establishments, particularly classifications that would be used in frequency count tables, is publicly available. However, the magnitude values are considered sensitive and must be protected. Magnitude data are generally nonnegative quantities. A given firm may have establishments that are in more than one table cell. Protection is applied to the company level rather than the establishment level. DA techniques are used to ensure published data cannot be used to estimate an individual firm's data too closely.

Recall that in cell suppression, the Census Bureau uses the p% rule to identify sensitive cells. This rule is designed to ensure that a user cannot estimate a respondent's value to within p% of that value. Currently, the Census Bureau uses fixed interval protection, which means the lower bound of the interval of uncertainty around any respondent's value v must be at most (1-p/100) * v and the upper bound must be at least (1+p/100) * v. This rule ensures that both bounds are a given distance from the true value. An alternative for which staff has done research is sliding protection (Massell, 2005). In this case, the length of the interval of uncertainty must be at least 2p/100 * v, but it need not have a given percentage of that interval above or below v. This resulted in less complementary suppression, but there was a debate over whether it protected the primary suppressions well enough. Thus, the Census Bureau is currently not using it, but could reconsider it in the future.

Another current focus involves applying the p% rule to atypical types of data, such as sample survey data, tables with imputed data, tables with negative values reported, tables reporting differences between positive values, tables reported net changes, tables reporting weighted averages, and tables of output from models. Recommendations for these types of data are found in Statistical Policy Working Paper 22 (SPWP22) Appendix A (2005). Also covered in SPWP22 Appendix A are suggestions for ways of handling key item suppression, preliminary and final data, and time series data. These types of data and situations are being reviewed to ensure that the Census Bureau concurs with SPWP22. Again, this research may end with the introduction of formal privacy.

Other Types of Data

See formal privacy above.

SUMMARY

The Census Bureau's goal is to publish as much high-quality information as possible without violating the pledge to protect the confidentiality of its respondents and their data.

This paper discusses the past and present DA techniques used to protect the confidentiality of different types of data products and includes information on current research being conducted to greatly improve and replace those techniques for the future.

Since the publication of Zayatz (2007) and Lauger et al., (2015), 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 DRB is quickly learning about formal privacy and how it protects Census Bureau data products.

Several developments have occurred in DA methodology at the Census Bureau since the papers cited above were published. The noise infusion technique for establishment magnitude data is used for more economic surveys. Improved data swapping techniques have been performed on the 2010 Census and ACS data, and research continues on ways to improve the techniques further. More reidentification experiments on microdata files are being considered.

Current research focuses on synthetic data and on other new DA alternatives for demographic and economic data, microdata, frequency count data, and magnitude data with a focus on formal privacy.

REFERENCES

M. DePersio, K. Ramanayake, J. Tsay, and L. Zayatz, "n-Cycle Swapping for the American Community Survey," Privacy in Statistical Databases 2010 LNCS 7556, edited by J. Domingo-Ferrer and I. Tinnirello, Berlin, Springer Verlag, 2012, pp. 143–164.

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

B. Evans, L. Zayatz, and J. Slanta, "Using Noise for Disclosure Limitation for Establishment Tabular Data," Journal of Official Statistics, Volume 14, No. 4, 1998, pp. 537–551.

Federal Committee on Statistical Methodology, "Report on Statistical Disclosure Limitation Methodology, Second version," Statistical Policy Working Paper 22, U.S. Office of Management and Budget, Washington, DC, 2005. Available at <https://nces.ed.gov/FCSM/pdf/SPWP22_rev.pdf> (accessed February 2019).

S. Hawala, "Producing Partially Synthetic Data to Avoid Disclosure," Proceedings of the Section on Government Statistics: American Statistical Association, Alexandria, VA, 2008, pp. 1345–1350. Available at <www.amstat.org/sections/srms /Proceedings/y2008/Files/301018.pdf> (accessed August 2014).

A. Lauger, W. Wisniewski, and L. McKenna, "Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research," Proceedings of the Section on Government Statistics, American Statistical Association, Alexandria, VA, 2015, pp. 3630—3642.

J. Li, T. Krenzke, T. Brick, D. Judkins, and M. Larsen, "Variance Estimation for the Census Transportation Planning Products with Perturbed American Community Survey Data," Proceedings of the Section on Survey Research Methods: American Statistical Association, Alexandria, VA, 2011, pp. 1595–1603. Available at <www.amstat.org/sections/srms /proceedings/y2011/Files/301081_66127.pdf> (accessed September 2014).

P. Massell, "Protecting Sensitive Cells in a Cell Suppression Program Using Sliding Protection," Statistical Research Division Study Series, Statistics #2005-02, 2005.

P. Massell and J. Funk, "Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection," Proceedings of the 2007 Federal Committee on Statistical Methodology (FCSM) Research Conference, 2007, <http://fcsm .sites.usa.gov/files/2014/05/2007FCSM_Massell-IX-B .pdf>, accessed September 2014.

P. Massell, "Modernizing Cell Suppression Software at the U.S. Census Bureau," Proceedings of the Section on Survey Research Methods: American Statistical Association, Alexandria, VA, 2011, pp. 3007–3015. Available at <www.amstat.org/sections/srms>.

T. Nayak and S. Adeshiyan, "On Invariant Post-Randomization for Statistical Disclosure Control," *International Statistical Review*, Early View: pp. 1–17, 2015, DOI: 10.1111/insr.12092.

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>.

A. Ramachandran, L. Singh, E. Porter, and F. Nagle, "Exploring Re-Identification Risks in Public Domains," Tenth Annual International Conference on Privacy, Security and Trust, Institute of Electrical and Electronics Engineers, Danvers, MA, 2012, DOI:10.1109/pst.2012.6297917.

P. Steel, "The Census Bureau's New Cell Suppression System," Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference, 2013. <https://fcsm.sites.usa.gov/files /2014/05/E3_Steel_2013FCSM.pdf> (accessed August 2014).

L. Sweeney, "Information Explosion...Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies," L. Zayatz, P. Doyle, J. Theeuwes, and J. Lane, (eds.), Urban Institute, Washington, DC, 2001.

L. Zayatz, "Disclosure Limitation for Census 2000 Tabular Data," Working Paper #15, Joint ECE/Eurostat work session on statistical data confidentiality, 2003, <www.unece.org/stats/documents/2003/04 /confidentiality/wp.15.e.pdf>.

L. Zayatz, "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update," *Journal of Official Statistics*, Vol. 23, No. 2, 2007, pp. 253–265. Available at <www.jos.nu/Articles /abstract.asp?article=232253> (accessed August 2014).

Appendix A: Recodes of Property Taxes (yearly amount)

Bb	NA (Group quarters/vacant/ not owned or being bought)
01	None
02	$1–49
03	$50–99
04	$100–149
05	$150–199
06	$200–249
07	$250–299
08	$300–349
09	$350–399
10	$400–449
11	$450–499
12	$500–549
13	$550–559
14	$600–649
15	$650–699
16	$700–749
17	$750–799
18	$800–849
19	$850–899
20	$900–949
21	$950–999
22	$1,000–1,099
23	$1,100–1,199
24	$1,200–1,299
25	$1,300–1,399
26	$1,400–1,499
27	$1,500–1,599
28	$1,600–1,699
29	$1,700–1,799
30	$1,800–1,899
31	$1,900–1,999
32	$2,000–2,099
33	$2,100–2,199
34	$2,200–2,299
35	$2,300–2,399
36	$2,400–2,499
37	$2,500–2,599
38	$2,600–2,699
39	$2,700–2,799
40	$2,800–2,899
41	$2,900–2,999
42	$3,000–3,099
43	$3,100–3,199
44	$3,200–3,299
45	$3,300–3,399
46	$3,400–3,499
47	$3,500–3,599
48	$3,600–3,699
49	$3,700–3,799
50	$3,800–3,899
51	$3,900–3,999
52	$4,000–4,099
53	$4,100–4,199
54	$4,200–4,299
55	$4,300–4,399
56	$4,400–4,499
57	$4,500–4,599
58	$4,600–4,699
59	$4,700–4,799
60	$4,800–4,899
61	$4,900–4,999
62	$5,000–5,499
63	$5,500–5,599
64	$6,000–6,099
65	$7,000–7,999
66	$8,000–8,999
67	$9,000–9,999
68	$10,000 or more (topcode)

Appendix B: Locating Primary Suppressions and Calculating Their Amount of Needed Protection for Rounded Data

Unrounded Data

Using a p% rule for most data, a cell is a primary suppression if
rem < v1 * p / 100
where
Total cell value = v1 + v2 + rem,
v1 is the value for the largest company,
v2 is the value for the second largest company,
and rem is the remainder.

If including the possibility of collusion of two companies, replace v2 with (v2 + v3) and so on. This applies to this entire appendix.

If a cell is a primary suppression, the additional protection it requires is
prot = v1 * p / 100 - rem + 1.

Data Rounded to Three Digits

When data are to be rounded (the values below are not yet rounded) to three digits (i.e., to the nearest multiple of 1,000), the rounding offers additional protection. Thus, the above inequality and equation should be modified.

For this type of data, a cell is a primary suppression if:
rem + |500 - |Total - round(Total) | | < v1 * p / 100.

The additional protection it requires is:
prot = v1 * p / 100 - [rem + |500 - |Total - round(Total) | |] + 1.

Data Rounded to Six Digits

When data are to be rounded to six digits (i.e., to the nearest multiple of 1,000,000), the rounding offers additional protection. Thus, the above inequality and equation should again be modified.

For this type of data, a cell is a primary suppression if:
rem + |500,000 - |Total - round(Total) | | < v1 * p / 100.

The additional protection it requires is:
prot = v1 * p / 100 - [rem + |500,000 - |Total - round(Total) | |] + 1.

Survey Data

For survey data, v1 and v2 should be adjusted and weighted up to the company level but not weighted to represent any other companies, and rem = completely weighted adjusted unrounded total - v1 - v2.