Research and Methodology Directorate

A History of the U.S. Census Bureau's Disclosure Review Board

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 its censuses and surveys 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.

Additionally, the products must be approved before dissemination by the Disclosure Review Board (DRB). The Board ensures that standard DA techniques have been applied, but its members also discuss each product to determine if it presents any additional disclosure risks. The Board may suggest additional procedures to address these risks, or it may refer the requestor to the Center for Disclosure Avoidance Research (CDAR), where staff can also help identify procedures to address the risks. The DRB and CDAR work closely together.

THE MICRODATA REVIEW PANEL (MRP)

Almost no documentation can be found on the MRP. Thus, much of this section is based on the author's and others' memories.

The MRP was established in 1981, <https://nces.ed.gov/FCSM/pdf/CDAC_DRB_Panel.pdf>. As the name states, it only reviewed microdata files prior to their release. At that time, microdata was only available from the demographic and decennial areas of the Census Bureau. That meant that no tables or any other type of product were officially reviewed for potential disclosure problems. DA techniques were being applied to the other products, but there was no group of people reviewing the tabular products before they were released to the public to ensure they were adequately protected before the DRB was established.

The MRP consisted of approximately five members from the demographic and decennial areas, one from the policy office, and one from Data User Services Division. Members would rotate as the chair of the panel every 2 years. The members were part of the MRP until they changed jobs, retired, or died, often jokingly requesting black robes and gavels. The author recalls the two last chairs as Chet Bowie and Jerry Gates, and based on their memories, Brian Greenberg and Paul Zeissett also served as chairs. The MRP met once a month to discuss microdata requests. The requests could be discussed at perhaps three meetings before a final decision was made. One member (Brian Greenberg) was in charge of the confidentiality research staff, then in the Statistical Research Division (SRD). His staff worked to help the MRP assess the potential risks of microdata files and to establish thresholds on geographic area population sizes and topcode[2] thresholds.

HOW THE MRP BECAME THE DRB

As stated previously, the MRP only reviewed microdata files before their release. Two problems began surfacing in the early 1990s that concerned MRP members and the confidentiality staff in SRD.

First, an MRP member discovered that the decennial area was about to publish a set of tables with six dimensions for small geographic areas. The more dimensions there are in such a table, the closer that is to essentially publishing microdata because there can be many cells with a count of one, thus linking six variables together for a person or household. The MRP members felt that they should review this as microdata. The decennial area that wanted to publish the data felt that this was out of the MRP's jurisdiction. The issue went to the executive staff, and the number of dimensions was reduced, but it was obvious that this issue would come up again.

Secondly, at about the same time, the confidentiality staff in SRD was developing new cell suppression methodology and software to be used for products from the 1992 Economic Census. The new software could be easily modified and used for products from economic surveys. But in speaking to different divisions in the economic area, three issues became clear. First, most staff members were doing cell suppression by hand (which can easily lead to a disclosure problem because of the large number of tables and the fact the relationships between tables can be very complicated). Second, staff members were using different methods and rules for applying suppression. Third, no staff members wanted to change anything.

[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.

[2] The Census Bureau uses topcoding and bottom-coding to eliminate outliers in a file for continuous variables such as wages and salary.

Members of the MRP and the confidentiality staff proposed to the executive staff that the MRP be changed to a DRB that would review all Census Bureau data products before public release. A charter was developed, and the DRB was formed in 1995, <https://nces.ed.gov/FCSM/pdf/CDAC_DRB_Panel.pdf>. Since its formation, the four DRB chairs have been Easley Hoy (1995–2001), Laura Zayatz (now McKenna) (2002–2016), Simson Garfinkel (2016–2018), and Rob Sienkiewicz (2018–present). The DRB works very closely with the CDAR staff members who work to identify potential disclosure risks in data products and develop new DA techniques that can be used to protect various types of data products.

Five of the most recent and important DRB issues are discussed in the next five sections. These issues lead to the creation of the Data Stewardship Executive Policy Committee (DSEP) in 2001. The DSEP ensures the Census Bureau maintains its commitment to protect the confidentiality of respondent's information by fulfilling the legal, ethical, and reporting obligations levied by Title 13 of the U.S. Code. It is the focal point for decision-making and communication on policy issues related to privacy, security, confidentiality, and administrative records. It oversees several staff committees, such as the DRB, that focus on these important issues. It acts on behalf of the full executive staff in setting policy and making decisions on policy-related matters within the scope of the committee.

ISSUE 1: PROMISING CONFIDENTIALITY

Two related problems affected the Census Bureau's promise of confidentiality to a subset of respondents, and both became apparent from the Survey of Income and Program Participation (SIPP).

First, in 1996, President Clinton signed the Personal Responsibility and Work Opportunity Reconciliation Act (also known as the Welfare Reform Act). One section of the Act charged the Census Bureau with continuing the collection of data from the 1992 and 1993 SIPP panels to evaluate the impact of the law with a focus on welfare and children. The Census Bureau then developed the Survey of Program Dynamics to carry out this mandate, <www.census.gov/history/www/programs/demographic/survey_of_program_dynamics.html>. Part of the survey collected data about whole households and part of it, the Self-Administered Questionnaire, collected data about adolescents 12–17 years of age in those households <www.census.gov/srd/papers/pdf/sm98-08.pdf>. When staff presented both microdata files (one from each part) to the DRB, the Board was very concerned. The adolescents had been told that their answers to the survey questions (several about sex, alcohol consumption, and drug use) would be confidential. However, due to overlapping variables in the adolescent survey and the household survey, the two microdata files could be easily linked, meaning that parents in the households could identify their children's responses. The DRB denied the request, and staff that had worked on the adolescent part of the survey were very concerned. The DRB took the issue to the executive staff. It was decided that the adolescent Survey of Program Dynamics microdata file could not be publicly released, but would be available only at the Research Data Centers only, a great disappointment for many people.

Second, in 1997, staff became concerned about a situation that occurred during a SIPP interview. A female member of the household had given previous interviews, answering questions for the whole household. When she was unavailable for a subsequent interview, her husband was interviewed, and Census Bureau staff reminded him of his wife's responses in order to reduce respondent burden by shortening the length of the interview and making the questions easier to answer for the new respondent. This was common practice. Unfortunately, in previous interviews, the wife had revealed that she had been previously married. She had never revealed this information to her husband about the previous marriage and he was very upset. This lead to the implementation of the Respondent Identification Policy, <www.researchgate.net/publication/237521296>.

Because of these two incidents, the Census Bureau rewrote the documentation on confidentiality protection that accompanies all household surveys and censuses, explaining that confidentiality between members of the same household cannot be protected. In addition, the Census Bureau explained that a respondent can find himself in a data product.

ISSUE 2: REIDENTIFICATION STUDIES

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 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.

A few small reidentification attempts were made with microdata files by summer interns in the early 1990s, but they yielded nothing of substance. Reidentification studies that yielded useful results were subsequently conducted on microdata files from the SIPP, the American Community Survey (ACS), and the American Housing Survey.

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.

ISSUE 3: NOISE AS AN ALTERNATIVE TO CELL SUPPRESSION FOR ECONOMIC TABULAR DATA PRODUCTS

In 1996, the confidentiality staff in SRD introduced an alternative to cell suppression for economic (establishment) tables. This technique, commonly referred to as EZS noise, is applied to the underlying microdata prior to tabulation (Evans et al., 1998). Each responding company's data are perturbed by a small amount, e.g., say approximately 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 (sensitive cells), 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. Because noise is added at the microdata level, additivity of the table is maintained.

To perturb an establishment's data by about say 10 percent, the Census Bureau would multiply its microdata values (prior to tabulation) by a random number that is close to either 1.1 or 0.9 for this example. Any of several types of distributions may be used 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 (with multipliers all near either 1.1 or 0.9, for this example). 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. One could incorporate this information into published coefficients of variation. For more information about suggested improvements to the original EZS noise technique, see Massell and Funk (2007).

In 1998, John Fowler, who was a member of the DRB from the economic area, approached the confidentiality staff asking them to test the noise methodology and software on tables from the Commodity Flow Survey. Test results were very good, and for the first time, the DRB and the Internal Revenue Service (IRS) (Title 13, Section 26 permits the Census Bureau to use IRS data for sampling or imputation) approved the use of multiplicative noise as an alternative to cell suppression for economic tabular data.

Building on SRD'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.

ISSUE 4: SYNTHETIC DATA

In 2003, John Abowd (Cornell University, and now at the Census Bureau as well) introduced to the DRB the idea of using synthetic data to protect respondent confidentiality and still produce and release very valuable data products (Abowd and Lane, 2004). It was a bit of a learning curve for DRB members, but after a few presentations, the Board understood how this could be a great asset to the Census Bureau. Creating synthetic data is one method of protecting confidentiality by replacing original microdata values by data that have been simulated. Synthetic datasets are required to serve two purposes. First, they must provide adequate protection from disclosure. Secondly, they must allow for statistically valid inferences, consistent with, albeit often less precise than, those that would be made with the original microdata (Lauger et al., 2015). Since then, and thanks to Dr. Abowd, the Census Bureau has used synthetic data for several data products.

Through the LEHD partnership with states known as Local Employment Dynamics partners, the Census Bureau has released a data product called OnTheMap, which is an online mapping and reporting tool that provides a user with data on where people are employed and where they reside. OnTheMap is protected by strict confidentiality protection requirements. For example, residential address information for each workplace address is based on synthetic data, while workplace information is protected by noise infusion.

Research lead by John Abowd recently lead to the update of an existing public-use microdata file called the Survey of Income and Program Participation Synthetic Beta. This product links individual-level microdata from the Census Bureau's SIPP, administrative tax data from the IRS, and retirement and disability benefit data from the Social Security Administration. Almost all variables on the file are synthesized, except for sex and the first marital link observed in the SIPP, yet reliable analytic results may be generated with known measures of error.

The Synthetic Longitudinal Business Database was the first major business establishment-level public-use microdata file ever released by a U.S. statistical agency and was developed by researchers at Cornell University, Duke University, the National Institute of Statistical Standards, and the Census Bureau's Center for Economic Studies. This data set is fully synthetic, with all establishments and their characteristics modeled after the values in the confidential Longitudinal Business Database.

Partially synthetic data was also used to protect Group Quarters[3] tabular data and microdata products from the 2010 decennial census.

ISSUE 5: DISCLOSURE AVOIDANCE OFFICERS (DAOS)

In 2010, an error was made in carrying out a decision of the DRB for a given data product. The decision was that a small amount of random noise was to be added to a variable for a particular subset of respondents. At that time, once the DRB made a ruling, it was up to the division requesting the publication of the product to carry out the ruling before releasing the data. This job often went to computer programmers, rather than statisticians. Unfortunately, the programmer for this job (probably the best at the Census Bureau at the time), added systematic noise rather than random noise. This biased the estimates from the data product, and a few users let the Census Bureau know that something was wrong. The Census Bureau quickly pulled the data from its Web site, fixed the problem, and released the corrected data. The Census Bureau realized that someone was needed for each data product to make sure the DRB rulings were carried out correctly and to examine the data for any potential problems before they were released. In 2011, with approval from the DSEP, the DRB began recruiting Disclosure Avoidance Officers (DAOs).

Divisions that produce data releases or publications based on confidential data must designate one or more DAOs who are charged with overseeing data product DA activities, record keeping, and the preparation of data product review submissions to the DRB.

Divisions may specify any number of DAOs, but each DAO must comply with the position's training and record keeping requirements. A DAO may serve as a DRB alternate, but must cease being a DAO if they become a DRB member.

DAOs act as intermediaries between the Disclosure Avoidance Coordinators or DACs (those requesting the release of a given data product) and the DRB in the DA review process. The critical nature of

[3] Group Quarters data include information about people living in nursing homes, prisons, college dormitories, military barracks, etc. (somewhere other than a household).

this position means that all DAOs are entrusted in maintaining the confidentiality of the data products assigned to them. To do this, DAOs need to proactively work with DACs to implement approved DA techniques and processes.

The DAO serves as the point person for all DRB review requests under their mentorship. Prior to data product submissions, the DAO must thoroughly review all referenced statistical tables and microdata for disclosure risk. A fully authorized DAO has two primary channels for processing confidential data products for public release:

- The DRB review request process.
- The DAO Data Product Bypass process, but only if the DAO is certified to review data products for DRB bypass and characteristics of the data and statistics to be released meet DRB bypass eligibility rules.

DAOs currently do not have the authority to approve the release of microdata. All proposed releases of microdata must be approved by the DRB.

CONCLUSION

The jobs of the DRB members, the assistant to the DRB, and the DAOs are quite time-consuming, but being involved in the disclosure review process has many benefits. The most prevalent are the chance to meet and work with people in all areas of the Census Bureau (all program areas, the policy area, the DSEP, sometimes even the field office) and the chance to learn about the many data products that the Census Bureau publicly releases. Members also have contact with others working in the field of DA in other U.S. agencies, academia, and around the world, because many groups of people are working on DA issues, face the same problems, and have the same goals.

Since 2016, the DRB has begun a complete overhaul of its procedures and documents that aid in the review process. New documents include a new charter, checklists, and a cover sheet to be filled out when DACs (requestors) wish to send a data product to the DRB for review, instructions and sign off forms for DAOs, and a document describing certain data products that DAOs can approve, and thus do not need to be sent to the full DRB (they may be bypassed). Most of these documents are not confidential and can be obtained through a request to <CED.DRB.Coordinator@census.gov>. There are also new and improved methods for maintaining information from previous requests.

The new charter discusses the make-up of the DRB. Some might think that all members are statisticians, but that is incorrect. While the Board needs statisticians, it also needs researchers and representatives from the policy area, the Federal Statistical Research Data Centers (FSRDCs), and all of the program areas (demographic, decennial, and economic) that release data products. The DRB must consider new DA techniques and data quality, must know and adhere to all policies concerning confidentiality, and must keep abreast of projects and expected data releases from the FSRDCs. The DRB could not function without representatives from the program areas with extensive knowledge of their area's particular data (such as sampling, weighting, editing, imputation, processing, proposed data products, and users).

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 Census Bureau's current 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.

REFERENCES

J. M. Abowd and J. Lane, "New Approaches to Confidentiality Protection: Synthetic data, remote access, and research data centers," Lecture Notes in Computer Science, 3050, 2004, pp. 282–289.

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.

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.

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.

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>.