Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research

Amy Lauger[1], Billy Wisniewski[1], Laura McKenna[1]
[1]U.S. Census Bureau, 4600 Silver Hill Rd, Washington, DC 20233

Abstract

The U.S. Census Bureau collects its survey and census data under the U.S. Code's Title 13, which promises confidentiality to its respondents. The agency also has the responsibility of releasing data for the purpose of statistical analysis. In common with most national statistical institutes, the Census Bureau's goal is to release as much high quality data as possible while maintaining the pledge of confidentiality. We apply disclosure avoidance techniques prior to releasing our data products publicly to protect the confidentiality of our respondents and their data. This paper discusses the various types of data we release, the disclosure review process, restricted access procedures, disclosure avoidance techniques currently being used, and current disclosure avoidance research.

Key Words: Confidentiality, Disclosure Avoidance, Microdata, Synthetic Data, Noise Infusion, Data Swapping

1. Introduction

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code. This title prevents the Census Bureau from releasing any data "...whereby the data furnished by any particular establishment or individual under this title can be identified." In addition to Title 13, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) requires the protection of information collected or acquired for exclusively statistical purposes under a pledge of confidentiality. However, the agency certainly also has the responsibility and aim of releasing high quality data to the public for the purpose of statistical analysis. In common with most national statistical institutes, our goal is to release as much high quality data as possible while maintaining the pledge of confidentiality. We apply disclosure avoidance techniques prior to releasing our data products publicly to protect the confidentiality of our respondents and their data. This paper discusses the various types of data we release, our disclosure review process, restricted access procedures, disclosure avoidance techniques currently used, and recent and current disclosure avoidance research. It is an update to Zayatz (2007).

2. Publicly Released Census Bureau Data

Unlike some statistical agencies, the Census Bureau does not use data licensing (Massell and Zayatz, 2000) to provide data to some users but not to others. Therefore, all data released to any external party is considered publicly available. The Census Bureau uses different disclosure avoidance methods for each type of data before release to the public. The most common forms of data release are microdata, frequency count data, and magnitude data. The following sections will discuss the types of data we typically publish, the current methods we use to protect them, and recent and current research to improve our methods.

3. Microdata

3.1 Description

The Census Bureau releases microdata files from the decennial census, many demographic surveys, and some economic surveys. A microdata file consists of data at the respondent level. Each record represents one respondent and consists of values of characteristic variables for that respondent. Typical variables for a demographic microdata file are age, race, sex, income, and home ownership / tenure. Sometimes, files will focus on specific issues and might include variables about topics such as crime victimization and alcohol abuse.

Typically, the Census Bureau does not release microdata from economic surveys and censuses because the skewness of economic data makes it often easy to identify establishments by only a few characteristics. However, in recent years, the Census Bureau has produced a public use microdata file for the 2007 Survey of Business Owners and synthetic economic microdata files, such as the Survey of Income and Program Participation Synthetic Beta (SSB) and the synthetic Longitudinal Business Database (synLBD).

3.2 Current Disclosure Avoidance Methods

The Census Bureau currently uses several disclosure avoidance techniques for our microdata files including geographic thresholds, rounding, noise infusion, categorical thresholds, topcoding, and data swapping. This paper primarily describes the procedures used for the Census 2010 and American Community Survey Public Use Microdata Samples (PUMS) files but many of these techniques are also used for other microdata files. Of course, all direct identifiers (name, address, etc.) are removed before public release.

3.2.1 Geographic Thresholds

All geographic areas identified on public-use microdata files must have a population of at least 100,000 (Hawala, 2001). Several data sets have an even higher geographic threshold, which may, for example, only allow for the identification of the four Census Regions or the nine Census Divisions. Applicable thresholds are determined depending on the level of detail of the variables on the file, whether the survey is longitudinal, and the public availability of other similar data.

3.2.2 Rounding

The Census Bureau uses a traditional rounding scheme. For example, dollar amounts are rounded in this way:

$0 remains $0
$1-7 rounded to $4
$8-$999 rounded to nearest $10
$1,000-$49,999 rounded to nearest $100
$50,000+ rounded to nearest $1,000

Census 2000 data were used to develop this rounding scheme and the resulting rounded categories were deemed to have enough values in them. Rounding is done prior to all summaries and ratio calculations. Because the variable Property Taxes is readily and publicly available, it has larger categories than those resulting from the rounding described above. The variable Departure Time for Work is also rounded.

3.2.3 Noise Infusion

Sometimes, noise is added to demographic survey variables when other, more traditional protection methods are not suitable. For example, noise is added to the age variable for persons in households with 10 or more people. Ages are required to stay within certain groupings so certain statistics are not affected. The original ages are blanked and new ages are chosen from a given distribution of ages within their particular grouping. Noise is also added to a few other variables to protect small but well-defined populations but we do not disclose those procedures.

3.2.4 Categorical Thresholds

All categorical variables must have at least 10,000 people nationwide in each published category. Any categories not meeting this threshold must be recoded into broader intervals.

3.2.5 Topcoding

Topcoding is used to reduce the risk of identification by masking outliers in continuous variables. For example, someone with an income of five million dollars would appear to have a much lower income in the public data set. All continuous variables (age, income, travel time to work, etc.) are topcoded using the half-percent/three-percent rule. Topcodes for variables that apply to the total universe (e.g. age) must include at least 1/2 of 1 percent of all cases. For variables that apply to subpopulations (e.g. farm income), topcodes must include either 3 percent of the non-zero cases or 1/2 of 1 percent of all cases, whichever is the higher value. Distributions of data from the 1990 Census were used to develop this rule. Some variables, such as year born, are likewise bottomcoded.

3.2.6 Data Swapping

In data swapping, a small number of households are swapped with other households in a different geographic area. Any household has at least a small chance of being swapped but targeted households are "special uniques" (Elliott, et al. 1998), which are household records unique based on certain demographic variables at high levels of geography and thus have a substantial disclosure risk. Swapping occurs at the microdata stage for the decennial census and for the American Community Survey but is performed primarily to protect aggregate data. See more about swapping in section 4.2.

3.3 Recent and Current Research

3.3.1 Re-identification Studies

The Census Bureau regularly conducts re-identification studies to assess the disclosure risk for our publicly available microdata. In light of the ever-changing amount, characteristics, and quality of other publicly available data, it is imperative for the Census Bureau to be situationally aware regarding the risk of our microdata products.

Most recently, the Census Bureau conducted a re-identification study using public use, anonymized, microdata for the 2008 American Community Survey (ACS) other public information freely available on the Internet, and a demographic data set for three counties available for purchase. The researchers used record linkage techniques to attempt to match individuals between the ACS dataset and other data sets. A person on the ACS dataset was considered re-identified if identifying information could be matched to their record. The researchers were only able to re-identify 87 people out of a couple of million. While this study shows that re-identification is fairly straightforward and possible, large-scale re-identification is not. Additionally, if an outsider intruder finds a possible match, it usually isn't a true match. Often survey records are unique within the sample but not in the population (Ramachandran, 2012). The Census Bureau will use the results of this research to continue to evaluate and adapt our disclosure avoidance procedures.

Census Bureau researchers are currently conducting a new re-identification study using the American Housing Survey. Studies on other surveys will follow.

3.3.2 Synthetic Data

Creating synthetic data is one method to protect confidentiality by replacing original microdata values by data that have been simulated. This method was introduced by Rubin (1993) and Little (1993). Rubin's technique is considered "fully synthetic," in that all values of the dataset are replaced. Rubin's approach used a Bayesian bootstrap method while Feinberg expanded on this method by using a posterior predictive distribution (Feinberg, 1994). Rubin's method preserved statistical properties by reusing actual survey responses while Feinberg's method used modelling to predict values that were not necessarily actual survey responses. In contrast to the methods of both Rubin and Feinberg, Little's technique is known "partially synthetic" approach, in that only values deemed particularly sensitive or risky are replaced with a modeled value.

Generally, demographic data are modeled and synthesized more easily than economic data. Geographic information is often difficult to synthesize. Data can be synthesized with a goal of releasing the synthetic microdata or some other product generated from the synthetic microdata. Finally, one synthetic data set or implicate, which looks exactly like the original file, can be synthesized, or, alternatively, several different implicates can be released together. Multiple synthetic implicates can be analyzed using multiple imputation analysis techniques.

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 but albeit often less precise than those that would be made with the original microdata (Weinberg, et al., 2007).

Through a partnership with Local Employment Dynamics (LED) partner states, the Census Bureau has released a data product called OnTheMap, which is an online mapping and reporting tool that provides a user with data both on where people are employed and where they reside. Currently, data are generally available for all 50 states and U. S. territories for 2002-2013, down to the Census block level. The underlying data come from a variety of sources, such as the LEHD Origin-Destination Employment Statistics (LODES), the Office of Personnel Management, and private workforce data from the Bureau of Labor Statistics.

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 some noise infusion. The Census Bureau is confident that the output does not disclose any confidential information.

Research led by John Abowd of Cornell University recently led to the update of an existing public-use microdata file called the Survey of Income and Program Participation Synthetic Beta (SSB). This product links individual-level microdata from the Census Bureau's Survey of Income and Program Participation, administrative tax data from the Internal Revenue Service, 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. This new version cannot be linked to original SIPP public use files or to earlier versions of the SSB versions (Benedetto, et al, 2013). The Census Bureau approved the release of the current version, SSB 6.0, in June 2014.

The Synthetic Longitudinal Business Database (SynLBD) was the first business establishment-level public-use microdata file ever released by a U.S. statistical agency and was developed between researchers at Cornell University, Duke University, the National Institute of Statistical Standards (NISS), and the Census Bureau's Center for Economic Studies (Jarmin, et al, 2014). This data set is fully synthetic, with all establishments and their characteristics modeled after the values in the confidential LBD. It contains information on 21 million establishment records across all sectors from 1976-2000. The current version does not include any geographic or firm-level variables.

3.3.3 Microdata Analysis System

For a brief time following Census 2000, a tool named the Advanced Query System (AQS) was accessible through American FactFinder. The AQS allowed users to submit requests for user-defined tabular data. A request passed through a firewall to an internal Census Bureau server, which held a previously swapped, recoded, and topcoded Census 2000 microdata file. The system created the requested table and reviewed it for disclosure risk. If the table passed disclosure review, it was released to the user in almost real-time.

The Census Bureau is developing a successor to the AQS named the Microdata Analysis System (MAS). Like the AQS, the MAS will allow users to make queries on microdata residing on an internal server and to receive the results if the query passes disclosure review. Unlike the AQS, the goal is to make demographic and economic surveys available through the MAS, and options other than tables will be presented. The on-demand disclosure avoidance application will allow users to have much greater access to and flexibility with Census Bureau data than previously allowed while allowing the Census Bureau to continue to uphold confidentiality standards.

4. Frequency Count Data

4.1 Description

The Census Bureau publishes frequency count data mainly from the decennial census and demographic surveys. Tables of frequency count data present the number of units in each table cell. For example, a table may have columns representing marital status and rows representing age groups. The cell values reflect the number of people in a given geographic area having the various combinations of marital status and age group. The decennial census and the American Community Survey have a multitude of published tables. However, other demographic surveys do not have a large enough sample to support tables at low levels of geography with sufficient data quality so only a limited number of tables at higher levels of geography are published.

4.2 Current Disclosure Avoidance Methods

Data swapping is the main procedure used to protect decennial census and American Community Survey tabulations. A small amount of household records is 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 Census 2010, 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 the sampling, measurement, coverage, and non-response error already present.

The Census Bureau continually conducts research to adapt and improve the swapping procedures. Over the past few years, we have 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.

Synthetic data are used to protect some of the data from the decennial census and the American Community Survey. Both programs collect data for both residential households and group quarters. Swapping is infeasible for group quarters so we now use partially synthesized group quarters data for these programs (Hawala, 2008). The Census Transportation Planning Products (CTPP) special tabulations also use synthetic data (Li, et al., 2011).

Tables are often required to meet certain thresholds in order to be released. For example, Summary File 2 for the decennial census iterates a set of tables by universe groups such as race, ancestry, and ethnicity. For these tables, each universe must contain at least 100 people in a given geographic area to be released. The American Community Survey has several types of rules, including population thresholds and geographical restrictions, some for data quality for its 1- and 3-year data products and some for disclosure avoidance (U.S. Census Bureau, 2013).

Often the standard products for the decennial census and the American Community Survey do not include the data particular users need. These users can request and pay for

a special tabulation. All special tabulations are generated from the swapped data files and must meet certain criteria before release.

All cell values are rounded according to the following scheme:

0 remains 0
1-7 rounds to 4
8 or greater rounds to the nearest multiple of 5

Totals are constructed before rounding, so the universes remain the same from table to table but the tables may no longer be additive. Percentages and rates are calculated after rounding. We allow some exceptions when the numerator, denominator, or both are not shown.

Tables usually must have no more than three or four dimensions and a mean cell size of at least three and sometimes higher than that. 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.

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, they must be rounded to two significant digits and at least five observations must be on either side of each quantile point.

4.3 Recent and Current Research

The Census Bureau continues to research ways to improve protection of frequency count data. Recent research explored two methods to improve data swapping. The research involved two new aspects. The first method is the use of "n-cycles" for swapping instead of swapping pairs of households with each other. In the current method, one could say the Census Bureau uses a swap cycle of size two, with two households, say A and B. Household A's characteristics are swapped with the characteristics of household B. In the n-cycle approach, the cycle may involve more than two households. For example, if n=3, A's characteristics are assigned to B, B's characteristics are assigned to C, and C's characteristics are assigned to A. Unlike the current method, in the case of an odd number of households for a given set, the new method will allow all households with at least one suitable match to be swapped. The second explored method for swapping involved the creation of a method to rank swaps in terms of data utility versus disclosure risk (DePersio, et al, 2012). The results were favorable but are not yet implemented into Census Bureau data products.

Additionally, researchers are currently studying the use of post-randomization (PRAM) methods as an alternative to data swapping. When using PRAM to protect categorical data, each record is stochastically transformed using pre-selected probabilities. The method protects the ability to make statistical inferences with the data while providing confidentiality protection to each record (Nayak and Adeshiyan, 2015).

5. Magnitude Data

5.1 Description

The Census Bureau publishes magnitude data from many of its surveys and the economic census. Most magnitude data comes from economic data products. However, some demographic variables such as household income is in the form of magnitude data. For economic data, 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 (e.g., 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 (NAICS) 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 non-negative quantities. A given firm may have establishments that are in more than one table cell. Protection is applied to the firm level rather than the establishment level. Disclosure avoidance techniques are used to ensure published data cannot be used to estimate an individual firm's data too closely.

5.2 Current Disclosure Avoidance Methods

5.2.1 Cell Suppression

The Census Bureau uses cell suppression for disclosure avoidance for most of its tables of magnitude data in economic data products. Any table cell value that could allow users to estimate a responding company's value too closely is not shown. The value is suppressed and replaced with a "D" for disclosure. These sensitive cells are called primary suppressions. They are identified using the p% rule, which is designed to ensure that a user cannot estimate a respondent's value to within p% of that value (Federal Committee on Statistical Methodology, 2005).

Because marginal totals are shown in the tables, other cells called complementary suppressions must be selected and suppressed, so that primary suppression values cannot be derived or estimated too closely via addition and subtraction of published values. For the past few years, researchers have worked on developing new cell suppression software. The modernized software is based on linear programming and replaces the older system that relied on network flow theory.

The new system is able to protect certain classes of tables better than the old system. Significantly, linear programming now allows for precise protection of three-dimension tables, as well as most sets of linked tables. The Census Bureau is required to protect economic data at both the firm level and the establishment level. In order to improve on this requirement, the system implements a new feature, called "protection of supercells." Here, a supercell is defined as the union of all interior primaries, along with the set of all secondaries, which exist in specified additive constraints (Massell, 2011). In addition, linear programming eliminates under-suppression and reduces over-suppression. Thus, more data can be published while still fulfilling protection requirements. The new system includes several innovative algorithmic procedures that allow the program to run quickly enough to meet production requirements (Steel, 2013).

5.2.2 Noise Infusion

A different technique is used for many of the Census Bureau's economic data products., Commonly referred to as EZS noise, this technique is applied to the underlying microdata prior to tabulation (Evans, et al, 1998). Each responding company's data are perturbed by a small amount, say up to 10% 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. Also, because noise is added at the microdata level, additivity of the table is guaranteed.

To perturb an establishment's data by about 10%, the Census Bureau multiplies 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 our multipliers and the distributions remain confidential within the agency. The overall distribution of the multipliers is symmetric about 1. The noise procedure does not introduce any bias into the cell values for census or survey data. Because we protect the data at the firm 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. One could incorporate this information into published coefficients of variation.

The following surveys now use noise infusion to protect their data: Nonemployer Statistics, Integrated Longitudinal Database, the LEHD Quarterly Workforce Indicators, workplace information for 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 whose data are protected using noise, a single table is considered to be the most important one. 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 table deemed most important is often worth the extra effort. Massell and Funk 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 underlying microdata (2007).

5.2.3 Synthetic Data

Many external users are interested in having the Census Bureau release more microdata from its surveys and censuses. However, releasing microdata poses many risks due to the great amount of data readily available on the Internet. Currently, the following economic data products use synthetic data to protect the underling data: OnTheMap versions 3-6,

SIPP Synthetic Beta (SSB), and the Synthetic Longitudinal Business Database (SynLBD). The SSB and the SynLBD are available through the Cornell University Virtual RDC.

5.3 Recent and Current Research

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. In most cases, 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.

However, in some recent cases, the Census Bureau has approved the use of sliding interval protection. Under sliding protection, the interval of uncertainty must be at least as wide as (2*p/100) * v, but the true value may be anywhere within that interval, even very close to one of the bounds.

Another current focus is about applying the p% rule to atypical types of data, such as percentages, rounded data, negative values, differences, net changes, and weighted averages.

6. The Disclosure Review Board

The Census Bureau has a Disclosure Review Board (DRB), which establishes disclosure avoidance policies and ensures consistency in the disclosure review of all publicly released Census Bureau data products. The board consists of at least six members representing the Census Bureau's demographic, decennial, and economic directorates, and the Research Data Centers (RDCs). These members usually serve six-year terms. At least an additional three members representing the research and policy areas are permanent members.

The Disclosure Review Board reviews almost all publicly released data products as explained in the DRB checklist (U. S. Census Bureau, 2007). These data products include those produced by Census Bureau staff and those produced at the Research Data Centers. Census Bureau staff members wishing to release data send a memo to the chair of the DRB accompanied by the DRB checklist, the questionnaire from the survey or census, a list of variables of interest, a record layout for requested microdata, table outlines for requested tabular data, and often some cross-tabulations of the variables of interest. The DRB checklist asks basic questions about the content of the data file to be released and helps to ensure consistency in the DRB's decision-making process. The Federal Committee on Statistical Methodology has created a generalized checklist (1999) for use by other federal statistical agencies.

After reviewing a request, the DRB may choose to approve it as is, approve it with modifications, or deny it. Census Bureau staff members not satisfied with a decision may appeal the decision to the Data Stewardship Executive Policy Committee (DSEP), which consists of a subset of Census Bureau Associate Directors.

7. Federal Research Data Centers

Some data sets cannot be publicly released because of confidentiality concerns. However, we have developed some restricted-use data procedures to allow researchers to use confidential data in a secure environment at what is known as Federal Statistical Research Data Centers (RDCs). In addition to Census Bureau data, the RDCs house data from other agencies, including the National Center for Health Statistics and the Agency for Healthcare Research and Quality. To use Census Bureau data within the RDCs, researchers must submit a proposal to the Census Bureau stating what research they wish to conduct, which restricted data sets they will need, and what type of results are to be published. The research must benefit the Census Bureau in some way, such as by improving data quality or improving methodology to collect, measure, or tabulate a survey, census, or estimate. If the proposal is accepted, the researcher and any associates who will work on the project at the RDC must obtain Special Sworn Status and come to one of the RDCs to work with the data they need. The researchers are then required by law to maintain confidentiality for life, just as any other Census Bureau employee is. Census Bureau staff review research results for disclosure problems before they are publicly released. The network continues to grow and, by the end of 2015, twenty-four RDCs will be open throughout the country.

8. Conclusion

Several developments have occurred in disclosure avoidance methodology at the Census Bureau since Zayatz (2007) was published. The noise infusion technique for establishment magnitude data is used in more economic data sets. Improved data swapping techniques have been performed on Census 2010 and American Community Survey data and research continues on ways to improve the technique further. Re-identification experiments on our microdata files continue. Current research focuses on synthetic data, the Microdata Analysis System, and other new disclosure avoidance alternatives for both demographic and economic data.

Acknowledgements

The authors would like to acknowledge Paul Massell, Philip Steel, and Jiashen You for their helpful contributions and feedback.

References

Benedetto, G., M. Stinson, and J. Abowd 2013. "The Creation and Use of the SIPP Synthetic Beta." Suitland, MD: U.S. Census Bureau. Available at: http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf (accessed August 2014).

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

Elliott, M., C. Skinner, and A. Dale. 1998. "Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk." *Research in Official Statistics* 1: 53-67. Luxembourg, March 1998.

Evans, B., L. Zayatz, and J. Slanta 1998. "Using Noise for Disclosure Limitation for Establishment Tabular Data." *Journal of Official* Statistics 14 no 4: 537-551. Available at http://www.jos.nu/Articles/abstract.asp?article=144537 (accessed August 2014).

Federal Committee on Statistical Methodology 1999. "Checklist on Disclosure Potential of Proposed Data Releases." Washington, DC: U. S. Office of Management and Budget. Available at http://fcsm.sites.usa.gov/files/2014/04/checklist_799.doc (accessed August 2014).

Federal Committee on Statistical Methodology 2005. *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology* Version 2. Washington, DC: U. S. Office of Management and Budget. Available at http://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf (accessed August 2014).

Fienberg, S. 1994. "A Radical Proposal for the Provision of Microdata Samples and the Preservation of Confidentiality." Carnegie Mellon University Department of Statistics. *Technical Report* No. 611.

Jarmin, R., T. Louis, and J. Miranda. 2014. "Expanding the Role of Synthetic Data at the US Census Bureau." Statistical Journal of the IAOS: Journal of the International Association for Official Statistics 30 no 2: 117-121. DOI: http://dx.doi.org/10.3233/SJI-140813.

Hawala, S. 2001. "Enhancing the "100,000 Rule: On the Variation Of The Per Cent Of Uniques in a Microdata Sample And the Geographic Area Size Identified On The File." In Proceedings of the Section on Survey Research Methods: American Statistical Association, August 2001. Available at http://www.amstat.org/sections/srms/proceedings/y2001/Proceed/00211.pdf (accessed August 2014).

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

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

Little,R. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9 no 2: 407-426. Available at http://www.jos.nu/articles/abstract.asp?article=92407 (accessed Aug 2015).

Massell, P. and L. Zayatz. 2000. “Data Licensing Agreements at U.S. Government Agencies and Research Organizations.” In Proceedings of the International Conference on Establishment Surveys II. July 2000. 1393-1410. Alexandria, VA. American Statistical Association. Available at http://www.amstat.org/meetings/ices/2000/proceedings/S29.pdf (accessed August 2014).

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

Massell, P. 2011. “Modernizing Cell Suppression Software at the U. S. Census Bureau.” Proceedings of the Section on Survey Research Methods, American Statistical Association, August 2011. 3007-3015. Alexandria, VA. American Statistical Association. Available at http://www.amstat.org/sections/srms/proceedings/y2011/Files/301855_67474.pdf (accessed August 2014).

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

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

Rubin, D. 1993. “Discussion of Statistical Disclosure Limitation.” Journal of Official Statistics 9 no 2: 461-468. Available at http://www.jos.nu/Articles/abstract.asp?article=92461 (accessed September 2014).

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

Weinberg, D., J. Abowd, S. Rowland, P. Steel, and L. Zayatz. “Access Methods for United States Microdata.” Working Papers 07-25, Center for Economic Studies, U.S. Census Bureau. ftp://ftp2.census.gov/ces/wp/2007/CES-WP-07-25.pdf (accessed August 2015).

Zayatz, L. 2007. “Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update. Journal of Official Statistics 23 no. 2: 253-265. Available at http://www.jos.nu/Articles/abstract.asp?article=232253 (accessed August 2014.)

U.S. Census Bureau 2007. “Supporting Document Checklist on Disclosure Potential of Data Disclosure Review.” Suitland, MD: U.S. Census Bureau. Available at http://www.census.gov/srd/sdc/S14-1_v1.3_Checklist.doc (accessed August 2014).

U.S. Census Bureau 2013. “American Community Survey: Data Suppression.” Suitland, MD: U. S. Census Bureau. Available at http://www2.census.gov/programs-surveys/acs/tech_docs/data_suppression/ACSO_Data_Suppression.pdf (accessed August 2015).