Journal of Official Statistics, Vol. 23, No. 2, 2007, pp. 253–265

Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update

Laura Zayatz[1]

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code 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, our goal is to release as much high quality data as possible without violating the pledge of confidentiality. We apply disclosure avoidance techniques prior to publicly releasing our data products to protect the confidentiality of our respondents and their data (Willenborg and de Waal 2001). This article discusses the various types of data we release, our disclosure review process, restricted access procedures, disclosure avoidance techniques currently being used, and current disclosure avoidance research.

Key words: Confidentiality; microdata; frequency counts; magnitude data; data protection.

1. Introduction

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code. This prevents the U.S. 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. In addition, the agency has the responsibility of releasing data 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 without violating the pledge of confidentiality (Duncan, Keller-McNulty, and Stokes 2003; Kaufman, Seastrom, and Roey 2005). We apply disclosure avoidance techniques prior to publicly releasing our data products to protect the confidentiality of our respondents and their data. This article discusses the various types of data we release, our disclosure review process, restricted access procedures, disclosure avoidance techniques currently being used, and current disclosure avoidance research. It is an update to Zayatz, Massell, and Steel (1999).

[1] U.S. Census Bureau, Commerce/Census/SRD/3209-4, Washington, DC 20233, U.S.A. Email: laura.zayatz@census.gov
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. Publicly Released U.S. Census Bureau Data

If the U.S. Census Bureau releases a data set to an outside data user, those data are considered publicly available. We cannot release a data set to some outside users and deny the same data set to others. Unlike some statistical agencies, we cannot use data licensing (Massell and Zayatz 2000). The most common forms of data release are microdata, frequency count data, and magnitude data.

2.1. Microdata

The U.S. Census Bureau releases microdata files from our demographic surveys and the long form of the decennial census. Note that beginning with Census 2010, there will be no long form as these data will be collected in the American Community Survey, a survey that collects the same data that were on the long form but will be collected every year from a smaller sample of people. We do not release microdata files from economic surveys and censuses, because economic data are highly skewed, and establishments can often be easily identified by just a few characteristics. 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 (Federal Committee on Statistical Methodology 1994). Typical variables for a demographic microdata file are age, race, sex, income, and occupation of a respondent. Occasionally, files will focus on specific issues and might include variables describing crime victimization and alcohol abuse.

2.2. Frequency Count Data

The U.S. Census Bureau publishes frequency count data mainly from the decennial census and the American Community Survey (ACS). Other (much smaller) demographic surveys do not support (in terms of data quality) tables at low levels of geography, and tables at higher levels of geography can simply be obtained from the public use microdata files through the use of tabulation software. Economic data are typically published in tables of magnitude data (see below). Tables of frequency count data present the number of units in each table cell. For example, a table may have columns representing the marital status of respondents and rows representing their 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.

2.3. Magnitude Data

The U.S. Census Bureau publishes magnitude data from its economic censuses and surveys. Tables of magnitude data often contain the frequency counts of establishments in each cell, but they also contain the aggregate of some quantity of interest over all units of analysis (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 within state. 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. The magnitude values, however, are considered sensitive and must be protected. Magnitude data are generally

nonnegative quantities. A given company may have establishments that are in more than one table cell. Protection is given at the company level (rather than the establishment level). Disclosure avoidance techniques are used to ensure published data cannot be used to estimate an individual company's data too closely.

3. The Disclosure Review Board

The U.S. Census Bureau has a Disclosure Review Board (DRB) to ensure consistency in the disclosure review of all publicly released Census Bureau data products. The Board establishes and reviews the U.S. Census Bureau's disclosure avoidance policy and procedures for all data products. The Board consists of six members representing the U.S. Census Bureau's demographic, decennial, and economic directorates, and its Research Data Centers (RDCs). These members serve six-year terms. An additional three members representing the research and policy areas are permanent members.

Almost all publicly released data products must be reviewed by the DRB. The exceptions are outlined in Zayatz (2004). They include data products produced at the Research Data Centers and those generated by the Advanced Query System described in Sections 4.1 and 4.2, respectively. U.S. Census Bureau staff 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 (if microdata), table outlines (if tabular data), and perhaps some cross-tabulations of the variables of interest. The DRB checklist asks basic questions about the content of the data file to be released. It has sections for microdata, frequency count data, and magnitude data. It helps to ensure consistency in the DRB's decision making process. The Confidentiality and Data Access Committee (CDAC) under the U.S. Office of Management and Budget has generalized the U.S. Census Bureau's Checklist for Disclosure Potential of Data for use by other federal statistical agencies. See www.fcsm.gov/committees/cdac/cdac.html.

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

4. Restricted Access Procedures

Some data sets cannot be publicly released because of confidentiality concerns. One example is economic microdata. Because we still want users to have access to the data, we have developed some restricted data procedures.

4.1. Research Data Centers (RDCs)

At times, data users need more detailed data than can be publicly released due to confidentiality protection requirements. This is particularly true for those who would like to use economic microdata. In these instances, users can access the data at Census Bureau RDCs. To do this, a researcher must submit a proposal to the U.S. Census Bureau stating what research he or she wishes to conduct, what data sets he or she will need, and what type of results he or she wants to publish. The research must, in some way, be beneficial to

the U.S. Census Bureau, such as 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 researcher is then bound by law to maintain confidentiality, just like any other Census Bureau employee. Results of research performed at the RDCs are reviewed for disclosure problems before they are publicly released. Currently, there are nine RDCs spread across the country. See www.ces.census.gov/ces.php/rdc.

4.2. The Advanced Query System

The American FactFinder (Rowland and Zayatz 2001) was developed to allow for broader and easier access to the standard Summary Files (frequency count data) from Census 2000 and to allow data users to generate their own tabular data products from Census 2000. See http://factfinder.census.gov/home/saff/main.html.

One part of American FactFinder is the Advanced Query System (AQS). The goal of the AQS is to allow users to submit requests for user-defined tabular data electronically. A request passes through a firewall to an internal Census Bureau server, which holds a previously swapped, recoded, and top coded microdata file. These disclosure avoidance techniques are described in Section 5.1. The table is created and electronically reviewed for disclosure problems (see Section 5.4). If it is judged to have none, it is sent back electronically to the user.

The AQS is currently only available to the Census Bureau's State Data Centers and Census Information Centers as well as a group of beta testers. Data users can contact these Centers to request free tabulations.

5. Current Disclosure Avoidance Practices

5.1. Microdata

There are several disclosure avoidance techniques that we are currently using for our microdata files including geographic thresholds, rounding, noise addition, categorical thresholds, top coding, and data swapping. This article describes the procedures used for the Census 2000 Public Use Microdata Samples (PUMS), but many of these techniques are also used for our other microdata files from demographic surveys. Obviously, all direct identifiers (name, address, etc.) are removed.

5.1.1. Geographic Thresholds

All geographic areas identified on our microdata files must have a population of at least 100,000 (Hawala 2000). This is the minimum. Several files are published at higher levels of geography, such as Census Division (there are nine) and Census Region (there are four), depending on the level of detail in the variables on file, whether or not the data are longitudinal, and whether or not we have identified other data files available to the public with some of the same variables on our file.

5.1.2. Rounding

We use traditional rounding. For example, dollar amounts are rounded according to the following scheme:

$1–7 = $4
$8–$999 rounded to nearest $10
$1,000–$49,000 rounded to nearest $100
$50,000 + rounded to nearest $1,000

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

5.1.3. Noise Addition

Noise is added to the age variable for persons in households with 10 or more people (Fuller 1993). Ages are required to stay within certain groupings so program statistics are not affected. 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.

5.1.4. Categorical Thresholds

All categorical variables must have at least 10,000 people nationwide in each published category. Otherwise categories must be recoded.

5.1.5. Top Coding

Top coding is used to reduce the risk of identification by means of outliers in continuous variables (for example someone with an income of five million dollars). All continuous variables (age, income amounts, travel time to work, etc.) are top coded using the half-percent/three-percent rule. Top codes for variables that apply to the total universe (for example age) should include at least 1/2 of 1 percent of all cases. For variables that apply to subpopulations (for example farm income), top codes should include either 3 percent of the nonzero cases or 1/2 of 1 percent of all cases, whichever is the higher top code. Distributions of data from the 1990 Decennial Census were used to develop this rule. Some variables, such as year born, are likewise bottom coded.

5.1.6. Data Swapping

We examine the records, looking for what are often called "special uniques" (Elliott, Skinner, and Dale 1998). These are household records which remain unique based on certain demographic variables at very high levels of geography and, therefore, have a disclosure risk. Any such household we find is swapped with some other household in a different geographic area. This typically does not affect many records, but those that it does need this added protection. See more on data swapping in the next section.

5.2. Frequency Count Data

The main procedure used for protecting Census 2000 tabulations is data swapping (Dalenius and Reiss 1982). It was applied to both the short form (100%) data and the long form (sample) data independently. It is also currently being used to protect American Community Survey tabulations. In each case, a small percent of household records is swapped. Pairs of households that are in different geographic regions are swapped across those geographic regions. The selection process for deciding which households should be swapped is highly targeted to affect the records with the most disclosure risk. For example, these include households in very small geographic areas and those that are racially isolated. Pairs of households that are swapped match on a minimal set of demographic variables. All data products (tables and microdata) are created from the swapped data files. After performing the data swapping, we did an extensive evaluation of the procedure and the resulting tables in terms of preserving data quality. The results of this evaluation are confidential, but the effects of the data swapping were minimal compared to nonresponse and response errors.

In addition to the swapping, thresholds are used for disclosure avoidance in our standard Summary Files 2 and 4. Summary File 2 iterates a set of tables from the short form (100%) data by universe groups, such as race, ancestry, and ethnicity. There must be at least 100 people of a given race (or ancestry or ethnicity) in a given geographic area for those tables to be released. Summary File 4 also iterates a set of tables from the long form data by groups such as race, ancestry, and ethnicity. There must be at least 50 unweighted sampled people of a given race (or ancestry or ethnicity) in a given geographic area for those tables to be released.

The U.S. Census Bureau publishes billions of tables from the short form and the long form data and (in the near future) a large amount from the American Community Survey, as well. Still, users may not find the tables they want in the standard Summary Files. When this happens, they can request and pay for a special tabulation. All special tabulations are generated from the swapped data files. All cell values are rounded according to the following scheme:

0 rounds to 0
1–7 rounds to 4
8 or larger rounds to the nearest multiple of 5

Totals are constructed before rounding; thus, universes remain the same from table to table, but the tables may no longer be additive. For Census 2000, Group Quarters data are rounded to the nearest multiple of ten, and only the categories Institutional and Non-Institutional are available.

Quantiles (percentiles) 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. This is the technique used in the standard Summary Files. If they are point quantiles generated using SAS and Proc Univariate, they are rounded to two significant digits, and there must be five nonoverlapping cases on either side of each quantile point. Means and aggregates must be based on at least three values. Thresholds on universes are often applied to avoid showing data for small geographic areas or small population groups.

We often require 100 cases for 100% data and 50 unweighted cases for sample data. Occasionally we require three unweighted cases for sample data for very small tables, say ten cells. Percents and rates are calculated after rounding. We allow some exceptions when the numerator and/or denominator is not shown. Usually tables have no more than three or four dimensions, and the DRB does consider mean cell size (at least three and sometimes more). For demographic profiles from user defined areas, all areas must have a population of 300 and boundaries must not overlap with standard Census Bureau geographic areas, creating geographic "slivers" with small populations.

5.3. Magnitude Data

5.3.1. Cell Suppression

The U.S. Census Bureau uses cell suppression for disclosure avoidance for almost all of its tables of magnitude data. 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 values are called primary suppressions or sensitive cells. They are identified using the *P%* rule (Federal Committee on Statistical Methodology 1994). This rule is designed to ensure that a user cannot estimate a respondent's value to within *P%* of that value.

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. Software based on network flow theory is used to find complementary suppressions for two-dimensional tables. Software based on linear programming theory is used to find complementary suppressions for small three-dimensional tables. For large three-dimensional tables, the linear programming software runs too slowly. In this case, the network flow software is used, followed by an auditing program to find any primary suppression that did not receive adequate protection because network flow theory only guarantees 100% coverage for two-dimensional tables. If the auditing program finds any primary suppression that did not receive full protection, linear programming is used to add suppressions where necessary.

5.3.2. Noise Addition

A different technique is being used for our Quarterly Workforce Indicator data and may be used in the near future for other magnitude data products. Noise is added to the underlying microdata prior to tabulation (Evans, Zayatz, and Slanta 1998). Each responding company's data are perturbed by a small amount, say 10% (the actual percent is confidential), in either direction. 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. 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 guaranteed.

To perturb an establishment's data by about 10%, we multiply its data by a random number that is close to either 1.1 or 0.9. We could use any of several types of distributions 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 company level, all establishments within a given company 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.

5.4. The Advanced Query System

The AQS does not provide an open-ended or unconstrained opportunity to construct any or all possible tabulations from the full microdata files. A query for a table through the AQS would pass through a firewall to an internal Census Bureau server with a previously swapped, recoded, and top coded microdata file. All tables generated from the sample data are weighted. The incoming query and the outgoing resulting table must each pass through a filter.

5.4.1. The Query Filter

If a user requests a tabulation for more than one geographic area or for a combination of areas, each area must individually pass the query filter.

The external user is advised in the user interface that the block group is the lowest level of geography permitted for 100% data and the tract is the lowest level of geography permitted for sample data for an external user. Requests for split block groups or split tracts are not permitted. A minimum population requirement (confidential parameter) is also imposed for each area. The user interface permits no more than three dimensions (page, column, and row) and one universe, not including geography. For example, a user could ask for a table of race (page) by income (column) by occupation (row) for all females (universe) living in Delaware (geography).

The query filter also delimits the use of variables such as race, Hispanic origin, group quarters, cost of electricity, gas, water, fuel, property taxes, property insurance cost, mortgage payments, condo fees/mobile home costs, gross rent, selected monthly owner cost, household/family income and individual income types. External users may obtain only predefined categories or recoded values of these variables. Most variables have several sets of recodes that the user can choose from. They are the same sets of recodes found in our standard Summary File tables. So if the user is requesting a table from a large geographic area, he or she can choose a very detailed list of recodes. If a user is requesting a table from a small geographic area, he or she can choose a short list of recodes, to try to ensure that the table will pass the results filter.

If the query passes the query filter rules, the query is sent from the external server outside the firewall to the internal server inside the firewall to the full microdata files. The full microdata files contain all of the predefined categories for race, Hispanic origin, group quarters, etc.

5.4.2. The Results Filter

Each resulting tabulation selected from the full microdata files obtained through the Advance Query System must meet certain criteria or the AQS will not provide the user with the tabulation. If a user requests a tabulation for more than one geographic area or for a combination of areas, each area must individually pass the results filter. The criteria are designed to prevent the release of sparse tabulations which can lead to disclosure. If a tabulation does not meet the criteria, the user will receive a message stating that the tabulation cannot be released for confidentiality reasons.

The system computes the total mean and median population cell sizes of the tabulation. For both mean and median calculations, only the internal cell counts are used (not the marginal totals). For both the mean and median calculations, cells with zero are included. If either the mean or median is less than some confidential number, the system does not permit the tabulation.

As stated previously, our disclosure avoidance rules are designed to prevent the release of sparse tables. They do not guarantee that there will be no cell values of size one. In fact, many of our standard Summary File tables contain cell values of size one, and for those we rely on the data swapping procedure to protect the data. The Advanced Query System uses the swapped file in generating tables. The third rule in the results filter limits the proportion of cells with values of one. The ratio of the number of unweighted cell counts of one to the number of nonzero cells must be less than some given confidential parameter.

In our testing, we found that the mean rule is unnecessary. Whenever it failed, either the median or the ratio of ones rule also failed. It was taken out of the system.

6. Current Disclosure Avoidance Research

6.1. Microdata

6.1.1. Data Integration

Data integration is putting together data originating from different sources. Data may have been gathered by different collection mechanisms and may be located online or in other data repositories. The data integration research project involves identifying data sets available to the public at no cost, or at a minimal cost, and linking them together. We then compare the integrated data to our public use microdata files to identify any data on our files at risk of disclosure. This work helps to develop new disclosure avoidance procedures for entities (individuals or households) that may currently be at risk of reidentification by outside intruders. Staff members working on this project have been involved in simulating and automating the steps an intruder could take to reidentify records.

We are currently locating publicly available data on the internet through web searches, as well as automating the search and download of data records if those are available through online queries. We are writing scripts to download and transform the data into a usable format for linking and reidentification. We are attempting to find records with a risk of disclosure and trying to attach names to the records to see if additional protection is needed. As more and more data have become available on the internet, we have been modifying our disclosure techniques using many of the methods already described.

Staff also recently designed and developed a prototype user-interface system for information visualization. The system facilitates the identification of risky records by matching and linking microdata files through visualization techniques. The system allows users to easily see and understand the data via graphs, and identify outliers that may be at risk of disclosure. Such records are then masked through swapping or noise addition.

6.1.2. Synthetic Data

Given a data set, one can develop posterior predictive models to generate synthetic data that have many of the same statistical properties as the original data (Abowd and Woodcock 2001). Generating the synthetic data is often done by sequential regression imputation, one variable in one record at a time (Rubin 1993). Using all of the original data, we develop a regression model for a given variable (Raghunathan, Reiter, and Rubin 2003). Then, for each record, we blank the value of that variable and use the model to impute for it. Then, we go to the next variable and repeat the process (Reiter 2003 and Reiter 2004).

Synthesizing data can be done in different ways and for different types of data products. One can synthesize all variables for all records (full synthesis) or a subset of variables for a subset of records (partial synthesis). If doing partial synthesization, we target records that have a potential disclosure risk and those variables that are causing this risk. We can synthesize demographic data and establishment data, though demographic data are easier to model and synthesize. We can synthesize data with a goal of releasing the synthetic microdata or some tabulation or other type of product (such as a map) generated from the synthetic microdata. And finally, we can generate one implicate (one synthetic data set) which looks exactly like the original file, but with synthetic data; or we can generate several implicates (several different synthetic data sets) that could be released together. Multiple synthetic implicates can be analyzed using multiple imputation analysis techniques.

John Abowd (Cornell University) is leading a group which is trying to develop a public use microdata file containing linked Social Security Administration earnings data and the Census Bureau's Survey of Income and Program Participation (SIPP) data with the goal of releasing multiple synthetic implicates. If we want to begin releasing public use files that link our data with data from other agencies, synthetic data are probably our only choice. Other disclosure avoidance techniques are not sufficient to protect the confidentiality of such files. The vast majority of the variables on the file will be synthesized. The two agencies are responsible for judging the quality of the final data product. The Census Bureau's Disclosure Avoidance Research Group will be using record linkage software to ensure the resulting data cannot be linked to any of our SIPP public use microdata files.

The DRB recently approved the release of the U.S. Census Bureau's first data product based on partially synthetic data. John Abowd developed the product, which is a set of maps of transportation data. The maps are based on partially synthetic data. The DRB looked at the data underlying the maps and decided that the synthetic data were sufficiently different from the original data, especially in small geographic areas. John compared the resulting maps and decided they looked almost identical, so everyone was pleased with the product. In developing this product, it helped knowing its intended use, and one should also note that only a handful of variables needed to be synthesized.

6.2. Frequency Count Data

Previously, we have used a data swapping technique as our main disclosure avoidance procedure for tabulations from the decennial census and the ACS. We are currently researching the possibility of changing from swapping to partially synthesizing the ACS data. It would be nice to have another option for protecting this type of data. We could even use a mixture of both techniques. Once we have developed the best models for the ACS data, we will compare the two techniques and decide which technique is best in terms of both protecting the data and maintaining data quality and utility.

6.3. Magnitude Data

Recall that we use the *P%* rule to identify sensitive cells (primary suppressions). This rule is designed to ensure that a user cannot estimate a respondent's value to within *P%* of that value. Recently, staff analyzed sliding interval protection for cell suppression (Massell 2005). Currently, we are using fixed interval protection. Under fixed interval protection, the lower bound of the interval of uncertainty around any respondent's value must be at most Value$(1 - P/100)$ and the upper bound must be at least Value$(1 + P/100)$. This ensures that both bounds are a given distance from the true value. Under sliding protection, the interval of uncertainty must be at least as wide as $2*\text{Value}*P/100$, but the true value may be anywhere within that interval, even very close to one of the bounds.

We showed that using sliding protection in our current cell suppression production programs will not work if we continue under our current assumption that data users can estimate a responding company's value to within 100%. If this assumption is relaxed in the future, sliding protection would have certain advantages (e.g., fewer suppressions), and so it should be seriously considered.

Currently, we are developing a tabular statistical disclosure control method that combines some of the best features of cell suppression, noise addition, and rounding. The resulting table would have numerical entries for each cell (i.e., no suppressions), but each value would have an uncertainty associated with it. This uncertainty would be expressed in the way that statistical errors are often expressed, viz., value $+/-$ error, and would be published along with the cell value.

Another current focus is on how to apply the *P%* rule to atypical types of data, such as percentages, rounded data, negative values, differences, net changes, and weighted averages.

6.4. Microdata Analysis System

The AQS accepts queries only for tables and only from Census 2000 data. We would like to see if we can expand its capabilities to handle data from other demographic surveys and other types of statistical analysis. We are currently developing a prototype of a Microdata Analysis System (MAS) that would do just that. It is a web-based system. The user selects the data set, the geography, the universe, the type of analysis, and the variables (or transformations thereof). The web site generates the SAS code needed to arrive at the desired results. The user may see the SAS code but may not alter it. The generated code is

run against the data and the results are verified. If the output passes the results filter (we are working on this now), it is returned to the user.

7. Conclusion

Since Zayatz, Massell, and Steel (1999) was published, there have been several developments in disclosure avoidance at the U.S. Census Bureau. The Advanced Query System was completed and is being widely used by State Data Centers and Census Information Centers. We are using the noise addition technique for establishment magnitude data in our Quarterly Workforce Indicator data. We have released one data product on transportation statistics based on partially synthetic data. We successfully used the targeted swapping technique for Census 2000 data. We are performing reidentification experiments on our microdata files. Current research focuses on synthetic data, the microdata analysis system, and disclosure avoidance alternatives for magnitude data.

For information on disclosure avoidance procedures and research in other countries, see http://neon.vb.cbs.nl/casc.

8. References

Abowd, J.M. and Woodcock, S.D. (2001). Disclosure Limitation in Longitudinal Linked Data. In Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Doyle, P., Lane, J., Zayatz, L., and Theeuwes, J. (eds). The Netherlands: Elsevier Science, 215–277.

Dalenius, T. and Reiss, S.P. (1982). Data Swapping: A Technique for Disclosure Control. Journal of Statistical Planning and Inference, 6, 73–85.

Duncan, G.T., Keller-McNulty, S., and Stokes, S.L. (2003). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 2003–6, Heinz School of Public Policy and Management, Carnegie Mellon University.

Elliott, M.J., Skinner, C.J., and Dale, A. (1998). Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk. Proceedings of the 1st International Conference on Statistical Data Protection. Lisbon, March.

Evans, B.T., Zayatz, L., and Slanta, J. (1998). Using Noise for Disclosure Limitation for Establishment Tabular Data. Journal of Official Statistics, 14, 537–552.

Federal Committee on Statistical Methodology (1994). Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology. Washington, DC: U.S. Office of Management and Budget.

Fuller, W. (1993). Masking Procedures for Microdata Disclosure Limitation. Journal of Official Statistics, 9, 383–406.

Hawala, S. (2000). On the Variation of the Percent of Uniques in a Microdata Sample and the Sample Size. Statistical Research Division Internal Memo, U.S. Census Bureau.

Kaufman, S., Seastrom, M., and Roey, S. (2005). Do Disclosure Controls to Protect Confidentiality Degrade the Quality of the Data? Proceedings of the American Statistical Association, Section on Survey Research Methods.

Massell P. and Zayatz, L. (2000). Data Licensing Agreements at U.S. Government Agencies and Research Organizations. Proceedings of the International Conference on Establishment Surveys II.

Massell, P. (2005). Protecting Sensitive Cells in a Cell Suppression Program Using Sliding Protection. Statistical Research Division Report Series, SSS2005-02, U.S. Census Bureau.

Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19, 1–16.

Reiter, J.P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology, 29, 181–188.

Reiter, J.P. (2004). Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. Survey Methodology, 30, 235–242.

Rowland, S. and Zayatz, L. (2001). Automating Access with Confidentiality Protection: The American FactFinder. Proceedings of the American Statistical Association, Section on Government Statistics.

Rubin, D.B. (1993). Discussion of Statistical Disclosure Limitation. Journal of Official Statistics, 9, 461–468.

Willenborg, L. and de Waal, T. (2001). Elements of Statistical Disclosure Control. New York: Springer-Verlag.

Zayatz, L., Massell, P., and Steel, P. (1999). Disclosure Limitation Practices and Research at the U.S. Census Bureau. Netherlands Official Statistics, Volume 14, Spring. Statistics Netherlands, Voorburg/Heerlen, 26–29.

Zayatz, L. (2004). Disclosure Review. U.S. Census Bureau Standard.

Received June 2005
Revised March 2006