WP.15
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (ii): Tabular data protection

NEW IMPLEMENTATIONS OF NOISE FOR TABULAR MAGNITUDE DATA, SYNTHETIC TABULAR FREQUENCY AND MICRODATA, AND A REMOTE MICRODATA ANALYSIS SYSTEM

Invited Paper

Prepared by Laura Zayatz, U.S. Census Bureau, United States of America[1]

[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.

New Implementations of Noise for Tabular Magnitude Data, Synthetic Tabular Frequency and Microdata, and a Remote Microdata Analysis System

Laura Zayatz

U.S. Census Bureau[1], Commerce/Census/SRD/5K011, 4600 Silver Hill Road, Washington, DC 20233, Laura.zayatz@census.gov

Abstract: 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 paper discusses three areas of current disclosure avoidance research: noise for tabular magnitude data, synthetic tabular frequency and microdata, and a remote access system. It also discusses how the methods developed needed to be altered when we applied them to real data, and how they are currently being used on real data products.

Key Words: Disclosure Avoidance, Confidentiality, Public Use Data Products

1 Introduction

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code. This 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. 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

[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.

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 paper discusses three areas of current disclosure avoidance research: noise for tabular magnitude data, synthetic tabular frequency and microdata, and a remote access system. For each technique, we give an introduction to the method. We describe what happened when we applied the method to real data. We discuss how we needed to alter the method for use on real data products. We then list the public use data products that currently use the method.

2 Noise for Tabular Magnitude Data

2.1 Introduction to the Method

This technique is an alternative to cell suppression which we have used for decades for tabular magnitude data. Noise is added to the underlying microdata prior to tabulation (Evans, Zayatz, and Slanta, 1998; Massell, Zayatz, and Funk, 2006). 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.

2.2 Applications to Real Data

A problem we found with the noise technique is that it can add excessive amounts of distortion to cells that would be shown much more precisely under deterministic methods such as cell suppression and controlled tabular adjustment. A natural question to ask then is whether or not there is a way to modify the method such that it adds less noise to the non-sensitive cells, while retaining the amount of protection provided to the sensitive cells. One of the primary benefits of the noise technique over cell suppression is that it allows for the release of more usable data. Suppression has the advantage that all published cell values are the exact estimates collected by an agency, and so if noisy cell values are highly distorted then the benefit of noise over suppression is significantly reduced. As a result, we investigated possible methods to reduce the overall amount of noise add to the data without compromising the level of protection (Massell and Funk, 2007a).

The Census Bureau’s magnitude data is almost always published in some rounded form, often in integer form representing thousands or millions of dollars. This type of rounding can be done at the record level prior to any tabulations, or applied directly to the unrounded table values. Noise is designed to protect individual respondents by changing their response values by small percentages. Rounding can therefore systematically remove the effect of noise on small response values. In some cases, this may not be an issue of concern, but under certain circumstances, this occurrence could result in serious damage to the level of protection provided by noise. We investigated a few types of rounding methods that could work to sustain the adequate protection provided by noise (Massell and Funk, 2007b).

2.3 Modifying the Method

In order to decrease the amount of distortion to non-sensitive cells, we developed a balanced noise technique, the details of which are described in (Massell and Funk, 2007a). The technique involves choosing one or more tables for the balancing application. This choice is determined after experimenting with several options and can be different for different surveys and censuses, but it is typically at a lower level in the hierarchy of related tables and has a trickle up effect. In step 1, random noise multipliers are applied to establishments in cells with only 1 or 2 contributors and to establishments from companies represented in more than 1 cell. In the second step, multipliers are applied to all other establishments (single unit establishments not in cells with only 1 or 2 contributors) taking into account the multipliers that have been previously assigned in such a way that the distortion in non-sensitive cells is minimized. If multipliers assigned in step one would lead to an increase (or decrease) in a non-sensitive cell's value, then multipliers assigned in step two would be created to make the final cell value have the least amount of distortion possible.

In (Massell and Funk, 2007b), one can find the results of testing various modifications to the standard rounding techniques. They include rounding of the underlying microdata values and rounding of tabulated cell values to ensure that standard rounding does not undo the protection offered to small cells by the noise. In general, ceiling/floor techniques seem to work best, but all of the techniques should be tested for each survey or census to identify the best technique for a given set of data.

2.4 Current Uses on Public Use Data Products

A few years ago, John Abowd (Cornell University and the Census Bureau) was developing a new data product: Quarterly Workforce Indicators. John was not a fan of cell suppression (he did not like holes in the data), so he decided to use the noise technique and the Disclosure Review Board approved it. Since then, he has been using noise and recently added a small amount of synthetic data.

Then, staff at the Census Bureau who work on the Commodity Flow Survey were considering adding many more detailed tables to their public data products but did not want them filled with suppressions, so they decided to use the noise technique. The method has since caught on. We have used it for our Non-Employer data products, and plan to use it for our Census of Island Areas, Survey of Business Owners, and Commodity Flow Survey. The Associate Director for our Economic Programs says that he sees this as the future technique for most of our tabular magnitude data products.

3 Synthetic Tabular Frequency and Microdata

3.1 Introduction to the Method

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 similar to 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.

3.2 Applications to Real Data

The biggest problems we have encountered when trying to generate synthetic data from real data are relationships between variables within a data set. For example, for many of our microdata files, we have records for households which are linked with records from all people within those households (think of a family with a mother, father, son, and daughter). Many values for many of the variables will be structurally missing (or blank) because of skip patterns in the survey instrument. For example, the number of children ever born to the father and son will be missing. The income of anyone under 15 will be missing. Other combinations of variables would not make sense. For example, we cannot have a mother who is only 7 years older than her own child after synthesizing the age variable.

3.3 Modifying the Method

We often impute some of the structurally missing values, but at the end of the synthesis procedure, restore them to missing for standard imputation and edits. For the SIPP/SSA file described in the next section, synthesizing the logical structure of the complicated longitudinal survey was not possible using the existing methods. Imposing structural zeros involving the interactions of several variables on the feasible statistical models we were using required an additional layer of programming that eventually became a nine-level collection of parent-child relationships that enforced all of the constraints.

3.4 Current Uses on Public Use Data Products

John Abowd (Cornell University) lead a group in developing 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 statistical avoidance techniques are not sufficient to protect the confidentiality of such files. The vast majority of the variables on the file are synthesized. The two agencies were responsible for judging the quality of the final data product. The Census Bureau's Disclosure Avoidance Research Group used record linkage software to ensure the resulting data cannot be linked to any of our SIPP public use microdata files.

John Abowd developed another product called "On The Map," which is a set of maps of transportation data. See. http://lehd.did.census.gov. The maps are based on partially synthetic data. The Disclosure Review Board 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.

We are using partially synthetic data to protect both frequency tabular and microdata for Group Quarters data in the American Community Survey (ACS). We are conducting research to see if we should use the method to protect more or even all ACS data products (Hawala and Funk, 2007), and to see if the synthesized data are an improvement over currently imputed values for missing data.

4 Remote Microdata Analysis System

4.1 Introduction to the Method

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 topcoded microdata file. The table is created and electronically reviewed for disclosure problems. If it is judged to have none, the table is sent back electronically to the user.

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.

4.2 Applications to Real Data

Perhaps the biggest decision that we needed to make about the Microdata Analysis System is whether we wanted a “disabled” or an “enabled” type of system. We could allow users to write and submit their SAS own code and disable some procedures such as PROC PRINT and PROC LIST (anything that would allow users to see the underlying microdata). The other option was to enable users to choose, from various menus, what they wanted to do and have the system generate the SAS code. We chose the second (enabled) option. There are advantages and disadvantages to both. A disabled system gives many more options to a data user, but it may require quite a bit of “babysitting” in that an agency should probably restrict who has access to such a system and strictly monitor the requests coming in and the results going out. An enabled system restricts the types of analyses that can be done, but could be made available to the general public without strict monitoring.

Also, in the early part of our work, we were focusing on the model statements in terms of looking for disclosure problems, but we soon realized that we needed to look at the underlying data tables that would be used in, for example, a regression. We also took a lesson already learned from the Advanced Query System that we will need to offer short, medium, and long lists of categories for certain variables so that users can obtain as much detail as possible in their analyses without running into potential disclosure problems (add AQS reference here).

4.3 Modifying the Method

We are now modifying the system to look at the tables underlying any type of analysis, and in particular to look at the marginal totals in the tables, because marginal totals of size 1 could potentially be used through several queries to put together a microdata record. We are also conducting research to find the best ways of identifying “cut points” in our short, medium, and long lists of continuous variables, again so that users can obtain as much detail as possible without disclosure problems. This includes automatic treatment of negative values, missing values, and non-monetary continuous values.

4.4 Current Uses on Public Use Data Products

The AQS is currently 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 (Weinberg, et.al. 2007). The Microdata Analysis System is still under development. It is being tested with American Community Survey and Current Population Survey data (Steel and Zayatz, 2006).

5 Conclusion

There have been several recent developments in disclosure avoidance at the Census Bureau. We are using the noise addition technique for tabular magnitude data for several data products. We have released several data products on based on partially synthetic data. The Advanced Query System was completed and is being widely used by State Data Centers and Census Information Centers, and we will continue our work on the Microdata Analysis System. The new techniques all took a few years to develop conceptually, and then took more years to adapt for use with real data. The work proved well worth it when we began using the techniques on real, public use data products.

References

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

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.

Evans, B. T., Zayatz, L., and Slanta, J. (1998), “Using Noise for Disclosure Limitation for Establishment Tabular Data,” *Journal of Official Statistics*, Vol. 14, No. 4, pp. 537-552.

Hawala, S. and Funk, J. (2007), “Model Based Disclosure Avoidance for Data on Veterans,” *Proceedings of the Research Conference of the Federal Committee on Statistical Methodology*, Arlington, Virginia, November 5-7, 2007.

Kaufman, S., Seastrom, M., and Roey, S. (2005), “Do Disclosure Controls to

Protect Confidentiality Degrade the Quality of the Data?” American Statistical Association, Proceedings of the Section on Survey Research.

Massell, P. and Funk, J. (2007a), “Protecting the Confidentiality of Tables by Adding Noise to the Underlying Microdata,” *Proceedings of the 2007 Third International Conference on Establishment Surveys (ICES-III)*, Montreal Canada, June 18-21, 2007.

Massell, P. and Funk, J. (2007b), “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 Research Conference of the Federal Committee on Statistical Methodology*, Arlington, Virginia, November 5-7, 2007.

Massell, P., Zayatz, L., and Funk, J. (2006), “Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey,” Privacy in Statistical Databases, CENEX-SDS Project International Conference, PSD 2006, Proceedings, Lecture Notes in Computer Science (LNCS) 4302, Springer 2006, ISBN 3-540-49330-1.

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

Reiter, J. P. (2003), “Inference for Partially Synthetic, Public Use Microdata Sets”, *Survey Methodology*, 29, pp. 181-188.

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

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

Rubin, D. B. (1993), “Discussion of Statistical Disclosure Limitation,” *Journal of Official Statistics*, 9, pp. 461-468.

Steel, P. and Zayatz, L. (2006), “Description of a Microdata Access System” for Presentation to the Census Advisory Committee of Professional Associations, US Census Bureau, October 27, 2006.

Weinberg, D., Abowd, J., Rowland, S., Steel, P., and Zayatz, L. (2007), "Access Methods for United States Microdata," *Proceedings of the Workshop on Data Access to Microdata, Nurembourg,* Germany, August 20-21, 2007. Also found on the Social Science Research Network http://hq.ssrn.com and US Census Bureau Center for Economic Studies Paper No. CES-WP-07-25.

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