RESEARCH REPORT SERIES
(Statistics #2007-7)

Imputation for Disclosure

Laura Zayatz

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: September 4, 2007

Imputation for Disclosure Avoidance

Laura Zayatz[1]
U.S. Department of Commerce, U.S. Census Bureau, Statistical Research Division
4600 Silver Hill Road, Washington, DC 20233-9100
laura.zayatz@census.gov

Key Words: Confidentiality, Microdata, Tabular Data, Data Protection

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. The 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. This paper discusses the various types of data we releases, the disclosure avoidance techniques currently being used, and how they may be seen as a form of imputation.

1. Introduction to Confidentiality, Census Bureau Data Products, and a Broad Definition of Imputation

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. Thus, the 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. The most common forms of data release are microdata, frequency count data, and magnitude data.

The Census Bureau releases microdata files from our demographic 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 (Federal Committee on Statistical Methodology, 1994). Typical variables for a demographic microdata file are age, race, sex, income, and occupation of a respondent.

---

[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 Census Bureau.

The Census Bureau publishes frequency count data from the decennial census and the American Community Survey (ACS). 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.

The 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. Disclosure avoidance techniques are used to ensure published data cannot be used to estimate an individual company's data too closely.

For this paper, we will view imputation as the substitution of one value for another (missing or present).

2. Using Noise to Protect Establishment Magnitude Data

A noise addition technique is currently being used for our Quarterly Workforce Indicator data and will soon be adopted for several economic surveys. This technique results in all tables cells (and underlying microdata values) being replaced by another value (a form of imputation). 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 (a method commonly used for this type of data). 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. We are currently using the noise method to protect Quarterly Workforce Indicators, Non-Employer data, and Survey of

Business Owners data, and intend to use the method more extensively in the future (Massell and Funk, 2007a and 2007b).

3. Disclosure Techniques for Microdata

Noise Addition

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 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.

Topcoding

Topcoding 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 topcoded using the half-percent/three-percent rule. Topcodes 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), topcodes should include either 3 percent of the non-zero cases or 1/2 of 1 percent of all cases, whichever is the higher topcode. Some variables, such as year born, are likewise bottomcoded. Topcoded values are typically replaced with the mean or median of all topcoded values.

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 (or replaced) with some other household in a different geographic area. This typically does not effect many records, but those that it does need this added protection. See more on data swapping in the next section.

4. Using Data Swapping to Protect Frequency Count Data

The main procedure used for protecting Census 2000 tabulations was data swapping. 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. 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.

5. 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. Using all of the original data, we develop a regression model for a given variable. 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, 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 which looks exactly like the original file, but with synthetic data; or we can generate several implicates that could be released together. Multiple synthetic replicates can be analyzed using multiple imputation analysis techniques. We are currently using synthetic data to protect our "On The Map" data product, Group Quarters data from the American Community Survey, and a file which links some of our data with data from the Social Security Administration, and we anticipate more extensive use of it in the future (Hawala and Funk, 2007).

6. Conclusion

Using our broad definition of imputation (the substitution of one for another), we see that disclosure avoidance procedures very often involve imputation.

7. 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.

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 1998.

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.

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.

Hawala, S. and Funk, J. (2007), "Model Based Disclosure Avoidance for Data on Veterans," Proceedings of the 2007 Federal Committee on Statistical Methodology Conference.

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).

Massell, P. and Funk, J. (2007b), "Recent Developments in the Use of Noise for Protecting Magnitude Data Tables," Proceedings of the 2007 Federal Committee on Statistical Methodology Conference.

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