An Update on Statistical Disclosure Avoidance Methodologies for Tabular and Microdata Files

Jacob D. Bournazian
Energy Information Administration

Abstract

In 1994, the Federal Committee on Statistical Methodology (FCSM) released Statistical Policy Working Paper No. 22, "Report on Statistical Disclosure Limitation Methodology." This working paper has been a valuable resource and educational tool for statisticians working in the field of data confidentiality. There has been a great deal of research since 1994 on developing new methodologies for protecting the confidentiality of survey responses, for assessing the risk of disclosure from proposed data releases, and for improving the analytical properties of "protected" tabular data and microdata files. This paper provides an overview of the work by the Confidentiality and Data Access Committee, a sub-committee of the FCSM, and the revisions made by the committee to Statistical Working Paper No. 22. There is a brief description of some of the new developments over the past 10 years in methodologies that may be applied to either tabular and microdata files and a description of the use of various disclosure avoidance methodologies by federal agencies to protect confidential data.

Keywords: Confidentiality, Disclosure Avoidance, Disclosure Risk

1. Introduction

The Federal Committee on Statistical Methodology (FCSM) released Statistical Policy Working Paper 22 (SPWP 22), "Report on Statistical Disclosure Limitation Methodology" in 1994. The FCSM, **http://www.fcsm.gov/**, is an interagency committee that recommends standards for statistical methodology as guidance for the federal agencies and investigates problems that affect the quality of Federal statistical data. The original version of SPWP 22 was an important contribution to the field of data confidentiality. It served as an update to SPWP 2, "Report on Statistical Disclosure and Disclosure Avoidance Techniques" that was released in 1978. More importantly, SPWP 22 became a cornerstone document that was routinely relied upon for summaries and explanations of statistical methodologies that protect confidential data from being released in tabular and micro-level data products. It contained a "primer" or summary chapter on statistical disclosure limitation methods for statisticians unfamiliar with data confidentiality issues, a summary of federal agency data confidentiality practices, recommendations for applying disclosure avoidance methodologies, and a research agenda that categorized areas for future work.

Significant developments occurred since the first release of SPWP 22 in many areas of data confidentiality. First, new disclosure avoidance methodologies were developed that provided statisticians with additional tools for protecting data confidentiality in response to a growing demand from users for more data. Second, tremendous growth occurred in the electronic availability of external files that may be used for matching against data products released by the federal agencies. Third, major legislative changes occurred that affect how Federal statistical agencies collect and release data collected under a pledge of confidentiality.

The most significant recent legislation were The Health Insurance Portability and Accountability Act (HIPAA) in 1996 and the Confidential Information Protection and Statistical Efficient Act of 2002 (CIPSEA). HIPAA led to the first set of national standards to protect the privacy of Americans' personal health records. The standards applied to medical records created by health care providers, hospitals, health plans and health care clearinghouses that are either transmitted or maintained electronically, and the paper printouts created from these records. After several years of public comment, the HIPAA Privacy Rule was implemented on April 14, 2003. The Privacy Rule requires most "covered entities" who provide data, to take reasonable steps to protect the confidentiality of health care information that they possess. SPWP 22 is cited in the preamble of the Privacy Rule for guidance in what constitutes "reasonable steps" to protect the confidentiality of the data. Title V of the Confidential Information Protection and Statistical Efficient Act of 2002 (CIPSEA) applies to all Federal agencies. It requires an agency to protect information it collects pursuant to Title V from disclosing that information in identifiable form to anyone not authorized to receive it or from having that information used for any purpose other than a statistical purpose. CIPSEA also created a process for certain agencies to share information that may be used only for statistical purposes.

With the rapid developments occurring in the methodologies being applied as well as other broad changes occurring in the field of data confidentiality, the task of updating SPWP 22 was undertaken by the Confidentiality and Data Access Committee (CDAC), a sub-committee of the FCSM, beginning in 2004. CDAC, **http://www.fcsm.gov/committees/cdac/**, was formed in 1996 as a result of Recommendations 3 and 4 in Chapter 7 of the original SPWP 22. It has grown over the years as a forum for staff members who work on confidentiality and data access topics and includes representatives from approximately 25 Federal agencies.

The revisions to SPWP 22 include a discussion of new disclosure avoidance methodologies as well as other data confidentiality and disclosure risk issues. This paper discusses the main revisions to SPWP 22 and highlights the new methodologies that were added. Methodologies that were discussed in the original version of SPWP 22 are referenced in this paper but not discussed.

2. Methodologies for Tabular Data

Chapter IV discusses the methodological issues concerning confidentiality protection for tabular data. Tabular data are classified into two categories for purposes of disclosure risk analysis: tables of frequency (or count) data and tables of magnitude data. This chapter was re-organized to discuss methodologies that protect sensitive cells after tabulation and those methodologies that protect sensitive cells prior to tabulation. New methodologies are discussed for protecting tabular data within these categories. The methodologies that protect sensitive cells after tabulation include cell suppression and **controlled tabular adjustment.**

Controlled tabular adjustment (CTA) is a useful methodology for protecting tables of magnitude data and is similar to the controlled rounding approach that has been successfully applied to tables of frequency data. CTA replaces each sensitive original value in a table with an imputed safe value that is a sufficient distance from the true sensitive value. Some of the remaining non-sensitive cell values are adjusted from their true values by as small an amount as possible to restore additivity to the published totals. CTA can be applied to produce solutions where marginal sums are minimally changed. However, allowing minor adjustments to the marginal values reduces the need for larger adjustments to the internal non-sensitive cells in a table.

There are two different approaches that apply CTA methodology. The original CTA method uses a linear programming method to restore additivity to the table. Initially, the **LP-based Controlled Tabular Adjustment** procedure used the reciprocal of the cell values as a cost function to minimize the overall deviation of non-sensitive cells from the true cell value. For example, another appropriate optimization function may be to minimize the sum of the absolute values of the data adjustments. The common function uses the reciprocal of the cell value because it allows for larger changes to large cells and causes smaller changes to small cells when compared with other cost functions. Most LP based procedures review the solution quality and feasibility using the underlying table structure. The algorithm systematically changes sensitive and non-sensitive cells by first seeking to obtain a feasible solution, and then once feasibility is reached, then it moves on to optimize the quality of the adjustment using a pre-specified cost function. Software that use some type of adaptive memory process for reviewing the optimal adjustments provide better results in terms of minimal adjustments to cell values than those methods that apply a "rigid memory" design such as a branch and bound technique.

During the first phase of applying CTA methodology, the sensitive cells are ordered from largest to the smallest. By using an alternating sequence, the ordered sensitive cell values are then changed to either lower or upper protection bounds. After completing the changes to all the sensitive cells in the table, non-sensitive table cells are considered to restore the additive table structure.

A second approach, called **simplified Controlled Tabular Adjustment**, was developed as a cost effective alternative to the original LP-based CTA method. The simplified CTA minimizes the percentage deviation from the true cell value for non-sensitive cells as its optimization function. The minimum percent deviation criteria used in simplified controlled tabular adjustment produces similar results as the reciprocal of the cell value-based cost function used in the LP-based approach. Simplified CTA is easier to implement and more computationally efficient than the LP-based CTA procedure, although further research is needed on different table structures to further evaluate these two approaches. LP based CTA and simplified CTA use different approaches to restore additivity to the table structure. The original CTA method uses a linear programming method to restore table additivity. The simplified CTA method, on the other hand, re-computes all the marginal table cell values to restore additive table structure.

Another category of methodologies takes a different approach to protect sensitive cells by modifying the micro-level data prior to tabulation. Those methodologies include **data swapping** and **noise addition**. Data swapping is a useful methodology for protecting tables of frequency data. The data are swapped at the microdata level prior to calculating the aggregates shown in the tables. The adjusted files themselves are not released. They are only used to prepare tables. Targeted data swapping involves selecting a set of records, finding a match in the data base on a set of predetermined variables and swapping the values for all other variables. For example, records identified from different regions or counties that match on race, sex, and income, could be candidates for swapping. Data swapping was used to protect the confidentiality of the Census 2000 tabulations. This approach is particularly useful if there are many tabulations being created from the same data shown in frequency tables.

Noise addition is a similar approach that has been successfully applied to magnitude data by the U.S. Census Bureau and National Agricultural Statistical Service. In this approach, noise is added to the underlying micro-level data reported by companies prior to tabulation. This approach is different from noise procedures used to protect and release public use microdata files. In the noise addition procedure used to protect tabular data, each reported value is perturbed by a small amount, (the adjustment parameter should remain confidential with the statistical agency). Each company in the sample or census is assigned a multiplier, or noise factor. All companies have their values multiplied by their corresponding noise factors before the data are tabulated. Since the same multiplier is used with a company where ever that company is tabulated, values will be consistent from one table to another. Adding noise to the underlying company-level data also protects the reported values of companies that dominate cells. If a cell contains only one company, or if a single company dominates the cell, the value in the cell will not be a close approximation to the dominant establishment's value because each value has been perturbed.

3. Methodologies for Public-use, Micro-level Data Files

Chapter V of SPWP 22 discusses the methodological issues concerning the release of public-use micro-level data files. There are two main sources of disclosure risk associated with releasing microdata files. The first risk is the possibility that the file contains high visibility records with unique characteristics that may lead to identification of a respondent. Removing the variables that directly identify respondents may not be enough to reduce this type of risk. Many times indirect variables may be used to identify respondents. The second source of risk is the possibility of matching the microdata file with external files.

The section on measures of disclosure risk was expanded to include new approaches. One method was the **R-U Confidentiality Map**. This method attempts to measure the simultaneous impact on disclosure risk and data utility that results from applying a specific disclosure limitation methodology. The R-U Map can also serve as a tool by a data provider for choosing the appropriate parameter value for the primary disclosure rule(s). R is a numerical measure of the statistical disclosure risk in a proposed release of a data file. This could be measured by the percentage of records that can be correctly re-identified using record linkage software. U is a numerical measure of the data utility of the released file. This could be measured by comparing the mean values or the variance-covariance matrix of the original data and the perturbed data. By mapping the values of R and U on the Y and X axis, a confidentiality map is generated which shows the trade-offs between reducing disclosure risk by changing the parameters of the disclosure limitation procedure and the loss in the usefulness of the data by changes in the analytical properties of the file. R-U Confidentiality Maps may be constructed for different disclosure limitation techniques and can provide guidance and insights for applying a specific disclosure limitation methodology.

Another measure of risk was developed using the **Micro Agglomeration, Substitution, Subsampling, and Calibration (MASSC)** disclosure limitation method (discussed in more detail later in this paper). This approach creates sets of identifying variables, called strata, to find records that may be at risk of disclosure. A unique record in a stratum is a record whose profile is unique for a given set of identifying variables. The record is at risk of disclosing personal information if the record is unique among the set of identifying variables and if any values of the sensitive variables are sensitive. After categorizing the database into a series of strata represented by different sets of identifying variables, a disclosure risk measure is calculated for each stratum.

Unique records falling in a stratum are then assigned a disclosure risk associated with that stratum., A disclosure risk measure can be calculated for a strata or even an individual record. A disclosure risk measure can also be calculated for an entire database by collapsing over the strata.

A section on disclosure risks associated with regression models was also added to Chapter V. Coefficients of models that contain only full-interactive sets of dummy variables on the right-hand side of the equation can be used to obtain entries in cross-tabulations of the dependent variable broken down by the categories defined by the dummy variables. These types of models can present disclosure risks if the tabular data also contain disclosure risks.

The methodologies used for protecting public-use microdata files were expanded to include three categories: Methods of reducing risk by reducing the amount of information released; methods of reducing risk by disturbing the microdata; and methods of reducing risk by using simulated microdata. The section on methods of reducing risk by disturbing the microdata was expanded to include a discussion of the following methodologies: **Data Shuffling** and **Micro Agglomeration, Substitution, Subsampling, and Calibration (MASSC)**. The section on methods of reducing risk by using simulated microdata was added and includes a discussion of the following methodologies: **Latin Hypercube Sampling**; **Inference-Valid Synthetic data**; and the **FRITZ algorithm**.

Data Shuffling is a data masking procedure that has been applied to numerical data. The procedure involves two steps: first the values of the confidential variables are modified and second, a data shuffling procedure is applied to the confidential variables on the file. The shuffling of data records occurs after the values for the confidential variable have been perturbed using some imputation method and then ranked. Data shuffling is useful for preserving the rank order correlation between the confidential and non-confidential attributes of a file.

Micro Agglomeration, Substitution, Subsampling, and Calibration (MASSC) is a disclosure limitation methodology that consists of the following four major steps. The first step, Micro Agglomeration, partitions the records into risk strata. Some recoding of variables may be done during this phase if needed. Individuals in each risk stratum are grouped so that the variance is small with respect to a given key set of identifying variables. In the second step, Substitution, values of sensitive variables are swapped with values from records that are the closest to them in terms of a certain distance measure. In the third step, Subsampling, records are randomly selected for subsampling within each strata. In the fourth step, Calibration, weights are assigned to the selected records using certain key variables to preserve the domain counts from the original dataset. The calibration step is used to reduce bias due to the substitution and to reduce variance due to the subsampling step.

Latin Hypercube Sampling (LHS) involves creating a file containing replacement values for the sensitive variables in the microdata file. The LHS method generates a synthetic data set that has similar univariate statistical characteristics as the original data such as mean, standard deviation and coefficient of skewedness. LHS can be used to generate a synthetic data set for a group of uncorrelated variables. If the variables are correlated, a restricted pairing algorithm is first applied to reproduce the rank correlation structure of the real data. Variables are first shuffled on the file and a cumulative distribution function is created for selected variables and used to generate the synthetic values.

Inference-Valid Synthetic data is another method that uses modified data for releasing public-use data files by drawing samples from the posterior predictive distribution of the adjusted confidential data. In this approach, the actual values of the confidential variable(s) in the microdata file, Y, are replaced using some controlled data adjustment constraint algorithm. The initial step generates a predicted value for Y and a residual for each Y variable 10 times, called 'implicates." Statistical models using the generated predicted data average the results from the ten implicates to generate standard error estimates. Depending on the variables which need protection and the variables that the researcher is interested in, the values for the confidential variable can be replaced by a posterior predictive distribution for that confidential variable based on a given set or combinations of variable keys. By customizing the distribution of the predicted Y values plus the residuals for the relevant confidential variable, i.e., the posterior predictive distribution, various micro-level datasets can be created. The statistical inferences from the synthetic data are consistent with the inferences generated by the actual reported values. Multiple public-use files can be created from the same underlying data using this method with each public use file customized for different groups of users.

The **Federal Reserve Imputation Technique Zeta**

(FRITZ) system is a unique approach that has been successfully used for both missing value imputation and disclosure limitation in the Survey of Consumer Finances (SCF). A survey sponsored by the Board of Governors of the Federal Reserve System in cooperation with the Statistics of Income Division of the IRS (SOI). The FRITZ model reviews the data along a sequential pre-determined path and imputes values one (sometimes two) at a time. The model is also iterative in that it imputes for the missing values in the data file, and then uses that information as a basis for imputing values in the second step, and continues the process until all values for the missing or sensitive estimates are stabilized and final. The file is reviewed for variable keys that cause excessive disclosure risks and those records are selected for protection.

4. A Primer on Statistical Disclosure Limitation Methods

Chapter II was re-organized to include examples that illustrate some of the above referenced methodologies discussed in Chapter IV and V. A new section was added to Chapter II that discusses on-line query systems. Disseminating data through on-line query systems requires special application of disclosure limitation methods because on-line query systems have multiple capabilities. The simplest form is where the system accesses summary files containing aggregated data that have already been tested for sensitivity. Another capability is the dissemination of tabulations from on-line queries of microdata files that have already been protected. An example of this type of on-line query system is the system developed by the Economic Research Service in conjunction with the National Agricultural Statistics Service that allows users to generate customized data tables by accessing microdata from the Agricultural Resource Management Survey (ARMS) program. The confidentiality of the reported values in the ARMS database is protected by applying a data perturbation method to the reported values prior to generating the tabulation. Another example is the "CDC Wonder" ((Wide-ranging OnLine Data for Epidemiologic Research (WONDER)) system developed by the Centers for Disease Control and Prevention. The CDC Wonder system allows users to submit queries to public-use data sets about mortality (deaths), cancer incidence, HIV and AIDS, behavioral risk factors, diabetes, natality (births), and census data on CDC's mainframe. The data are previously tested for sensitivity with disclosure limitation methods applied prior to the public-use file being added to the database. Applications that access unprotected microdata can introduce a risk of identity disclosure when restricting the query to a small geographic area or category. Specialized tabulations generated from queries to unprotected microdata files must apply appropriate disclosure limitation rules. The Advanced Query System of American Fact Finder developed by the Census Bureau has the sensitivity rules and disclosure methods built into the system so that queries submitted by users must pass through a series of filters where disclosure limitation rules are applied before the user can view the results.

books, websites, reports of conferences and

5. Recommendations

Chapter VI contains 13 recommendations relating to disclosure limitation practices. The recommended practices for federal agencies were revised to address several issues. For example, agencies should consult data users on issues relating to: balancing the risk of disclosure against the loss in data utility; increasing the availability of public-use microdata files; the need for restricted data access procedures so that researchers may access microdata in a controlled and safe environment, and the development of on-line public use data base query systems through the Internet. Other issues that affect data utility include whether users would prefer disclosure limitation methods that modify, replace, or adjust the data in some manner rather than methods that suppress data.

Interagency cooperation is also encouraged and should be expanded where possible. For example, the release of identical or similar data by different agencies or groups within agencies (either from identical or similar data sets) and the availability to match to external files are factors that contribute to the need for interagency cooperation. Agencies need to share information on what external files are available to a user for matching to agency data products. Information on external files should be updated and widely circulated among the statistical agencies so that disclosure review boards, confidential officers, and other ad-hoc disclosure review boards can properly assess the disclosure risk from a proposed data release. Agencies should consider procedures for expanding the shared use of research data centers as a method for increasing access to confidential data by researchers.

References

The reference section was also revised. Recent papers that were cited in the chapters were added to the bibliography list and some articles were deleted. A list of references was added that includes important workshops, special issues of journals, and manuals.