The working paper *Report on Statistical Disclosure Limitation Methodology (Revised 2005), 1994, (WP22)*, which follows, is available for historical reference. This FCSM working paper has been revised as the Data Protection Toolkit, available at https://www.fcsm.gov/resources/safe-guard-data/

STATISTICAL POLICY
WORKING PAPER 22 (Second version, 2005)

Report on Statistical Disclosure Limitation Methodology

Federal Committee on Statistical Methodology

Originally Prepared by Subcommittee on Disclosure Limitation Methodology 1994

Revised by Confidentiality and Data Access Committee 2005

Statistical and Science Policy
Office of Information and Regulatory Affairs
Office of Management and Budget

December 2005

The Federal Committee on Statistical Methodology
(December 2005)

Members

Brian A. Harris-Kojetin, Chair, Office of Management and Budget

Wendy L. Alvey, Secretary, U.S. Census Bureau

Lynda Carlson, National Science Foundation

Steven B. Cohen, Agency for Healthcare Research and Quality

Steve H. Cohen, Bureau of Labor Statistics

Lawrence H. Cox, National Center for Health Statistics

Robert E. Fay, U.S. Census Bureau

Ronald Fecso, National Science Foundation

Dennis Fixler, Bureau of Economic Analysis

Gerald Gates, U.S. Census Bureau

Barry Graubard, National Cancer Institute

William Iwig, National Agricultural Statistics Service

Arthur Kennickell, Federal Reserve Board

Nancy J. Kirkendall, Energy Information Administration

Susan Schechter, Office of Management and Budget

Rolf R. Schmitt, Federal Highway Administration

Marilyn Seastrom, National Center for Education Statistics

Monroe G. Sirken, National Center for Health Statistics

Nancy L. Spruill, Department of Defense

Clyde Tucker, Bureau of Labor Statistics

Alan R. Tupek, U.S. Census Bureau

G. David Williamson, Centers for Disease Control and Prevention

Expert Consultant

Robert Groves, University of Michigan and Joint Program in Survey Methodology

Preface

The Federal Committee on Statistical Methodology (FCSM) was organized by the Office of Management and Budget (OMB) in 1975 to investigate issues of data quality affecting Federal statistics. Members of the committee, selected by OMB on the basis of their individual expertise and interest in statistical methods, serve in a personal capacity rather than as agency representatives. The committee conducts its work through subcommittees that are organized to study particular issues. Statistical Policy Working Papers are prepared by the subcommittee members and are reviewed and approved by FCSM members.

The Confidentiality and Data Access Committee (CDAC) is a special interest subcommittee of the FCSM that was formed in 1995 as a result of recommendations contained in the original Statistical Policy Working Paper 22. The committee consists primarily of statisticians working in federal agencies who are involved with issues relating to protecting data confidentiality, and providing selective and controlled access to confidential data. CDAC provides a unique forum for discussing these issues and sharing information and research ideas among the federal agencies. CDAC's website may be accessed at http://www.fcsm.gov/committees/cdac. The 2005 revision to Statistical Policy Working Paper 22 is the second version of the 1994 work by the Subcommittee on Disclosure Limitation and Methodology. The Subcommittee on Disclosure Limitation Methodology was formed in 1992 to describe and evaluate existing disclosure limitation methods for tabular and microdata files and to update previous work presented in Statistical Policy Working Paper 2, "Report on Statistical Disclosure and Disclosure-Avoidance Techniques" published in 1978. See Cover and Introductory Material in the 1994 version of Statistical Policy Working Paper 22 for a discussion of the Subcommittee on Disclosure Limitation Methodology.

The Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper 22, discusses both tables and microdata and describes current practices of the principal Federal statistical agencies. The original report includes a tutorial, guidelines, and recommendations for good practice; recommendations for further research; and an annotated bibliography. In 2004, the Confidentiality and Data Access Committee (CDAC) revised Statistical Policy Working Paper 22 to include research and new methodologies that were developed over the past ten years, and to reflect current agency practices. The annotated bibliography was partially updated. The CDAC members who worked on the revision:

Jacob Bournazian, Energy Information Administration
Nancy Kirkendall, Energy Information Administration
Steve Cohen, Bureau of Labor Statistics
Philip Steel, Bureau of Census
Alvan O. Zarate, National Center for Health Statistics
Arnold Reznek, Bureau of Census
Paul Massell, Bureau of Census

Acknowledgements

We thank the agency representatives of CDAC for their contributions to this working paper and updating the descriptions of agency practices in Chapter 3.

Table of Contents

CHAPTER I - Introduction

A. Subject and Purposes of This Report

Federal agencies and their contractors who release statistical tables or microdata files are often required by law or established policies to protect the confidentiality of individual information. This confidentiality requirement applies to releases of data to the general public; it can also apply to releases to other agencies or even to other units within the same agency. The required protection is achieved by the application of statistical disclosure limitation procedures whose purpose is to ensure that the risk of disclosing confidential information about identifiable persons, businesses or other units will be very small.

During 2004, the Confidentiality and Data Access Committee (CDAC), a special interest committee on data confidentiality and access issues for the Federal Committee on Statistical Methodology (FCSM), revised Statistical Policy Working Paper 22 to incorporate new developments in statistical disclosure limitation methodologies, and to update agency data confidentiality practices and procedures. A description of CDAC and its activities is contained in Appendix D. Statistical Policy Working Paper 22 was written in 1994 by the Subcommittee on Disclosure Limitation Methodology. The 1994 subcommittee's purpose was to review and evaluate statistical disclosure limitation methods used by federal statistical agencies and to develop recommendations for their improvement. A description of that subcommittee is contained in the Cover and Introduction Material of the original 1994 Statistical Policy Working Paper 22.

Legislation passed by Congress after the original release of Statistical Policy Working Paper 22 in 1994 added to the federal agencies' need to protect the confidentiality of the data they collect. The Health Insurance Portability and Protection Act (HIPPA) originally enacted in 1996 had a strong impact on setting requirements for protecting health data. The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002 created a new mechanism for agencies to protect data confidentiality while at the same time limited the data sharing activity to statistical purposes only. Over this same time period, the interest in federal statistical data within the research and data user community continued to grow. The need for greater data access led to the development of new disclosure avoidance methods so that more data could be released to the public while agencies maintain the protection of respondent information. This revision of Statistical Policy Working Paper 22 updates the discussion of these issues by incorporating current research and new developments in the field.

The goals in revising this report were to:

- describe and evaluate existing disclosure limitation methods for tables and microdata files;
- provide recommendations and guidelines for the selection and use of effective disclosure limitation techniques;
- promote the development, sharing and use of software for the applications of disclosure limitation methods; and
- encourage research to develop improved statistical disclosure limitation methods, for

both tabular as well as public-use microdata files.

Every agency or unit within an agency that releases statistical data should be capable of selecting and applying suitable disclosure limitation procedures to all the data it releases. Each agency should have one or more employees with a clear understanding of the methods and the theory that underlies them. This report is directed primarily at employees of federal agencies and their contractors who are engaged in the collection and dissemination of statistical data, especially those who are directly responsible for the selection and use of disclosure limitation procedures. This report is also useful to employees with similar responsibilities in other organizations that release statistical data, and to data users so that they may better understand and use disclosure protected data products.

B. Some Definitions

In order to clarify the scope of this report, we define and discuss here some key terms that will be used throughout the report.

B.1. Confidentiality and Disclosure

A definition of **confidentiality** was given by the President's Commission on Federal Statistics (1971:222):

> [Confidential should mean that the dissemination] of data in a manner that would allow public identification of the respondent or would in any way be harmful to him is prohibited and that the data are immune from legal process. Duncan et. al., 1993, *Private Lives and Public Policies*, p. 24.

Confidentiality differs from privacy because it applies to business as well as individuals. Privacy is an individual right whereas confidentiality often applies to data on organizations and firms. The second element of this definition, immunity from mandatory disclosure through legal process, is a legal question and is outside the scope of this report.

A second definition is also provided to assist users in understanding this concept.

"Confidentiality pertains to the treatment of information that an individual has disclosed in a relationship of trust and with the expectation that it will not be divulged to others in ways that are inconsistent with the understanding of the original disclosure without permission." IRB Guidebook, Part III.D, Department of Health and Human Services, Office of Human Research Protections.

An agency's need to protect the confidentiality of data it collects is based upon various legislative requirements. Statistical disclosure occurs when released statistical data (either tabular or individual records) reveal confidential information about an individual respondent. This paper is concerned with minimizing the risk of **disclosure** (public identification) of the identity of individual reporting units and information about them.

Section 512 of Title V of the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) requires all federal agencies to protect data or information acquired by the agency under a pledge of confidentiality for exclusively statistical purposes from being disclosed in identifiable form. Section 502 of CIPSEA defines "**identifiable form**" as any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means.

The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule was implemented on April 14, 2003. This rule obligates most "covered entities," such as Medicare providers, to protect the confidentiality of health care information that they possess. The Privacy Rule subjects the providers of health care information to certain requirements to protect the confidentiality of the data being released. Regardless of the basis used to protect confidentiality, federal statistical agencies as well as some private information organizations involved with health care information, must balance two objectives: to provide useful statistical information to data users, and to assure that the responses of individuals are protected.

The Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99) was enacted to protect the privacy of student education records. The law applies to all schools that receive funds under an applicable program of the U.S. Department of Education. FERPA gives parents and eligible students (i.e. students over the age of 18 or who attend a school beyond the high school level) certain rights with respect to their education records. Generally, schools must have written permission from the parent or eligible student in order to release any information from a student's education record. However, FERPA allows schools to disclose those records, without consent, to certain designated parties or when specific conditions are present. Schools may also disclose, without consent, "directory" information such as a student's name, address, telephone number, date and place of birth, honors and awards, and dates of attendance. However, schools must tell parents and eligible students about directory information and allow parents and eligible students a reasonable amount of time to request that the school not disclose directory information about them.

The release of statistical data inevitably reveals some information about individual data subjects. Disclosure occurs when confidential information is revealed. Sometimes disclosure can occur based on the released data alone; other times disclosure may result from combining the released data with publicly available information; and sometimes disclosure is possible only through combining the released data with detailed external data sources that may or may not be available to the general public. The accessing and/or linking by the public to electronic data bases creates some degree of risk that disclosure of confidential information may occur even though personal identifiers are removed from a file. At a minimum, each statistical agency must assure that the risk of disclosure from the released data when combined with other relevant publicly available data is very low.

Several different definitions of disclosure and of different types of disclosure risk have been proposed. Duncan et al. (1993: 23-24) provides a definition that distinguishes three types of disclosure:

Disclosure relates to inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file (**identity disclosure**), sensitive information about a data subject is revealed through the released file (**attribute disclosure**), or the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (**inferential disclosure**).

Note that each type of disclosure can occur in connection with the release of either tables or microdata. The definitions and implications of these three kinds of disclosure are examined in more detail in the next chapter.

B.2. Tables, Microdata, and On-Line Query Systems

The choice of statistical disclosure limitation methods depends on the nature of the data products whose confidentiality must be protected. Most statistical data are released in the form of tables, microdata files, or through on-line query systems. Tables can be further divided into two categories: tables of frequency (count) data and tables of magnitude data. For either category, data can be presented in the form of numbers, proportions or percentages.

A microdata file consists of individual records, each containing values of variables for a single person, business establishment or other unit. Some microdata files include direct identifiers, such as name, address or Social Security number. Removing any of these identifiers is an obvious first step in preparing for the release of a file for which the confidentiality of individual information must be protected.

Historically, disclosure limitation methods for tables were applied directly to the tables. Methods include redesign of tables, suppression, controlled and random rounding. More recent methods have focused on protecting the microdata underlying the tables using some of the microdata protection techniques. In this way all tables produced from the protected microdata are also protected. This may be done whether there is an intention to release the microdata or not. It is a particularly useful way to protect tables produced from on-line query systems.

B.3. Restricted Data and Restricted Access

The confidentiality of individual information can be protected by restricting the amount of information provided or by adjusting the data in released tables and microdata files (**restricted data**) or by imposing conditions on access to the data products (**restricted access**), or by some combination of these. The number of federal agencies that have implemented restricted access programs have increased during the past ten years and this report provides some references. However, the main trhust of this report is to discuss the disclosure limitation methods that provide confidentiality protection by restricting the data. The fact that this report deals primarily with disclosure limitation procedures that restrict or adjust data content should not be interpreted to mean that restricted access procedures are of less importance. Readers interested in the latter can find detailed information in Duncan et. al., 1993, *Private Lives and Public Policies*,. p. 157 and "Restricted Access Procedures" by the Confidentiality and Data Access Committee (April 2002) at http://www.fcsm.gov/committees/cdac/cdacra9.doc.

As a brief summary, there are four main methods that agencies use to provide restricted access to confidential data: Research Data Centers (RDCs), Remote Access, Research Fellowships and Post Doctoral Programs, and Licensing Agreements. **RDCs** permit use of confidential files in a physically secure environment with specialized equipment. Users agree to terms and conditions governing the access and use of the confidential data. Research products are reviewed by the agency to assure no confidential information is revealed. **Remote access** over secure electronic lines to dedicated computers is a second method. Users can apply statistical techniques to confidential data. The statistical products are reviewed by the agency to assure no confidential data are revealed. **Fellowships and post-doctoral programs** are a third method, and researchers sign agreements to allow them to be treated as agency employees, subject to the same restrictions as employees. Similar to RDC access, researchers may be given limited access and products are reviewed by the agency to make sure no confidential data are released. Fourth, **licensing agreements** permit a researcher to use confidential data offsite, but under highly restricted conditions as spelled out in a legally binding agreement. Arrangements that place restrictions on who has access, at what locations, and for what purposes access is allowed normally require written agreements between agency and users. These agreements usually subject the user to fines, being denied access in the future and/or other penalties for improper disclosure of individual information and other violations of the agreed conditions of use. Users may be subject to external audits conducted by the agency to assure terms of the agreement are being followed. Users in violation may be required to pay fines or be subject to other legal penalties.

Most **public-use** data products are released by statistical agencies to anyone usually without restrictions on use or other conditions, except for payment of fees to purchase publications or data files in electronic form. Both NCHS and NCES require users of public use data files to signify that they will not use the data being made available to them to try to identify an individual respondent. Agencies require that the disclosure risks for public-use data products be very low. In meeting this requirement the application of the disclosure limitation methods described in this document may substantially restrict data content, to the point where the data may no longer be of value for some purposes. The National Center for Education Statistics provides public-use data products that involve access to confidential data. Though these are "Public Use", users must sign agreements assuring that they will maintain the confidentiality of the data. Users may be audited to make sure they are following proper procedures.

C. Organization of the Report

Chapter II, "Statistical Disclosure Limitation Methods: A Primer," provides a simple description and examples of disclosure limitation techniques that may be used to limit the risk of disclosure in releasing tables and microdata.

Chapter III, "Current Federal Statistical Agency Practices," describes disclosure limitation methods used by fourteen (14) major federal statistical agencies and programs. Among the factors that explain variations in agencies' practices are differences in types of data and respondents, different legal requirements and policies for confidentiality protection, different technical personnel and different historical approaches to confidentiality issues.

Chapter IV, "Methods for Tabular Data," provides a systematic and detailed description and evaluation of statistical disclosure limitation methods for tables of frequency and magnitude data. Chapter V, "Methods for Public-Use Microdata Files," describes various statistical disclosure limitation methods used to protect confidentiality in the public release of microdata files. These chapters will be of greatest interest to readers who have direct responsibility for the application of disclosure limitation methods or are doing research to evaluate and improve existing methods or develop new ones.

Due in part to the stimulus provided by previous subcommittee's reports (including Statistical Policy Working Papers 2 and 22), improved methods of disclosure limitation have been developed and used by some agencies over the past 25 years. Based on a review of these methods, guidelines are provided in Chapter VI as recommended practice for all agencies. The development and production of public use microdata files continues to grow and has increased the need to review the possibility of data linkage to external files and the role of identifiers on files.

Three appendices are also included. Appendix A contains technical notes on practices the statistical agencies have found useful in extending primary suppression rules to other common situations. Appendix B is a list of websites and government references on statistical disclosure. Appendix C is a reference list. Appendix D contains a description of CDAC and its accomplishments.

D. Underlying Themes of the Report

Five principal themes underlie the guidelines in Chapter VI:

There are differences between the disclosure limitation requirements that apply to federal agencies. Federal agencies that have specific legislation covering their data collection activities are bound to maintain the confidentiality of all survey responses. Other agencies that do not have specific legislation covering their data collection activities can determine which data may need protection. Nevertheless, agencies that need to protect data should move as far as possible toward the use of a small number of standardized disclosure limitation methods whose effectiveness has been demonstrated.

Statistical disclosure limitation methods have been developed and implemented by individual agencies over the past 40 years. Information and research in this field needs to be shared across all federal agencies. The documentation and the corresponding software used by a statistical agency should then be shared among federal agencies.

Disclosure-limited products should be auditable to determine whether or not they meet the intended data protection objectives of the procedure that was applied. For example, linear programming software can be used to perform disclosure audits for some kinds of tabular data. At the same time, the data utility of the disclosure-limited products should be assessed as part of the evaluation of the applied procedure.

Several agencies have formed disclosure review boards, statistical or review panels, and designated agency confidentiality officers to ensure that appropriate disclosure limitation policies and practices are in place and being properly used. Each agency should centralize its oversight and review of the application of disclosure limitation methods through the development of a standardized list of questions or areas of inquiry. The "Checklist on Disclosure Potential of Proposed Data Releases" by CDAC and located at http://www.fcsm.gov/committees/cdac/resources.html is a useful guide for agencies to structure their review.

CHAPTER II - Statistical Disclosure Limitation Methods: A Primer

This chapter provides a basic introduction to the disclosure limitation techniques that are commonly used to limit the possibility of disclosing identifying information about respondents in tables and microdata files. The techniques are illustrated with examples. The tables or microdata files produced using these methods are usually made available to the public with no further restrictions. Section B presents some of the basic definitions used in theses sections and subsequent chapters. It includes a discussion of the distinction between tables of frequency data and tables of magnitude data, a definition of table dimensionality, and hierarchical variables, and a summary of different types of disclosure. Section C discusses the disclosure limitation methods applied to tables of counts or frequencies. Section D addresses tables of magnitude data, Section E discusses microdata, and Section F summarizes the chapter. Readers who are already familiar with the methodology of statistical disclosure limitation may prefer to skip directly to Chapter III, which describes agency practices, Chapter IV which provides a more mathematical discussion of disclosure limitation techniques used to protect tables, or Chapter V which provides a more detailed discussion of disclosure limitation techniques applied to microdata.

A. Background

One of the functions of a federal statistical agency is to collect individually identifiable data, process it and provide statistical summaries, and/or public use microdata files to the public. Some of the data collected are considered proprietary by respondents.

On the other hand, not all data collected and published by the government are subject to disclosure limitation techniques. Some data on businesses that is collected for regulatory purposes are considered public. In addition, some data are not considered sensitive and are not collected under a pledge of confidentiality. The statistical disclosure limitation techniques described in this paper are applied whenever confidentiality is required and data or estimates are made publicly available. All disclosure limitation methods result in some loss of information, and sometimes the publicly available data may not be adequate for certain statistical studies. However, the intention is to provide as much data as possible, without revealing individually identifiable data. (See Chapter I for a brief discussion of the use of restricted access as opposed to restricted data.)

The most common method of providing data to the public is through statistical tables. With the development of powerful computers with large memory capability and high processing speeds, agencies have started providing an on-line query system with access to a statistical data base. Data users create their own tabulations by customized queries. In most of these systems only data that have already had disclosure limitation applied are available to users. If the unprotected microdata are used as the basis for a query system, disclosure limitation rules must be applied automatically to the requested tables. The concern with the later approach is that users may be able to discern confidential data if they use a sequence of queries in which disclosure limitation is applied independently.

M**icrodata files** are another way agencies attempt to provide user-friendly products. These products have become indispensable to the research community as the release of microdata files for public use has grown. In a microdata file, each record contains a set of variables that pertain to a single respondent and are related to that respondent's reported values. However, names, addresses and other **direct identifiers** are removed from the file and the data may be disguised in some way to make sure that individual data items cannot be uniquely associated with a particular respondent.

B. Definitions

Each entry in a statistical table represents the aggregate value of a quantity over all units of analysis belonging to a unique statistical cell. For example, a table that presents counts of individuals by 5-year age categories and the total annual income in increments of $10,000 is comprised of statistical cells such as the cell (35-39 years of age, $40,000 to $49,999 annual income). The number in the cell is the count or frequency of the number of people in the population with the cell characteristic. A table that displays value of construction work done during a particular period in the state of Maryland by county and by 4-digit North American Industry Classification System (NAICS) groups is comprised of cells such as the cell {NAICS 4231, Prince George's County}. In this case the number in the cell would be the average value (or aggregate value) of the construction work for companies in the population with the cell characteristics.

B.1. Tables of Magnitude Data Versus Tables of Frequency Data

The selection of a statistical disclosure limitation technique for data presented in tables (**tabular data**) depends on whether the data represent frequencies or magnitudes. Tables of **frequency count data** present the number of units of analysis in a cell. Equivalently the data may be presented as a percent by dividing the count by the total number presented in the table (or the total in a row or column) and multiplying by 100. Tables of **magnitude data** present the aggregate of a "quantity of interest" that applies to units of analysis in the cell. Equivalently the data may be presented as an average by dividing the aggregate by the number of units in the cell.

To distinguish formally between **frequency count data** and **magnitude data**, the "quantity of interest" must measure something other than membership in the cell. Thus, tables of the number of establishments within the manufacturing sector by SIC group and by county-within-state are frequency count tables, whereas tables presenting total value of shipments for the same cells are tables of magnitude data.

B.2. Table Dimensionality

If the values presented in the cells of a statistical table are aggregates over two variables, the table is a **two-dimensional** table. Both examples of detail cells presented above, (35-39 years of age, $40,000-$49,999 annual income) and (NAICS 4231, Prince George's County) are from two-dimensional tables. Typically, categories of one variable are given in columns and categories of the other variable are given in rows.

If the values presented in the cells of a statistical table are aggregates over three variables, the table is a **three-dimensional** table. If the data in the first example above were also presented by county in the state of Maryland, the result might be a detail cell such as (35-39 years of age, $40,000-$49,999 annual income, Montgomery County). For the second example if the data were also presented by year, the result might be a detail cell such as (NAICS 42, Prince George's County, 2002). The first two-dimensions are said to be presented in rows and columns, the third variable in "layers" or “pages,” with the layers being a separate table for each category of the third variable.

B.3. Hierarchical Structure of Variables

Most tables are cross tabulations of two or three classification variables such as geography. Classification variables may have a hierarchical structure. A hierarchical coding structure produces subtotals with the variable’s coding structure. For example, the North American Industry Classification System (NAICS) classification variables are variables with a hierarchical structure. Four digits industry codes can be collapsed into three digit codes for major industries and two digits for industry groups. An interior table cell might relate to a specific 4 digit NAICS code, with subtotals given by 3-digit NAICS codes, and the marginal total given by the appropriate 2-digit code. Identifying any hierarchical structure within the classification variables on a file is necessary for applying disclosure limitation techniques, and for assessing protection.

Geography is commonly referred to as a variable with a hierarchical structure. However, this may not always be technically correct depending upon the classification structure. If geography is broken down into states, regions, and national level, then geography would be a hierarchical variable because each state is classified within specific regions. However, if the geographic classification provides locality, metropolitan area, county, state, and region, then the classification may not necessarily be hierarchical because the counties, localities, and metropolitan areas may not be component parts of each other.

B.4. What is Disclosure?

Although the definition of disclosure given in Chapter I is broad, this report documents the methodology used to limit disclosure and is concerned only with the disclosure of confidential information through the public release of data products. In Chapter I, the three types of disclosure presented in Duncan, et. al (1993) were briefly introduced. These are identity disclosure, attribute disclosure and inferential disclosure.

Identity disclosure occurs if a third party can identify a subject or respondent from the released data. Revealing that an individual is a respondent or subject of a data collection may or may not violate confidentiality requirements. For tabulations, revealing identity is generally not disclosure, unless the identification leads to divulging confidential information (attribute disclosure) about those who are identified. For microdata, identification is generally regarded as disclosure, because microdata records are usually so detailed that identification will automatically reveal additional attribute information that was not used in identifying the record. Hence disclosure limitation methods applied to microdata files limit or modify information that might be used to identify specific respondents or data subjects.

Attribute disclosure occurs when confidential information about a data subject is revealed and can be attributed to the subject. Attribute disclosure occurs when confidential information about a person or firm's business operations is revealed or may be closely estimated. Thus, attribute disclosure comprises identification of the subject and divulging confidential information pertaining to the subject.

Attribute disclosure is the primary concern of most statistical agencies in deciding whether to release tabular data. Disclosure limitation methods applied to tables assure that respondent data are published only as part of an aggregate with a sufficient number of other respondents to disguise the attributes of a single respondent.

The third type of disclosure, **inferential disclosure**, occurs when individual information can be inferred with high confidence from statistical properties of the released data. For example, the data may show a high correlation between income and purchase price of home. As purchase price of home is typically public information, a third party might use this information to infer the income of a data subject. There are two main reasons that some statistical agencies are not concerned with inferential disclosure in tabular or micro data. First a major purpose of statistical data is to enable users to infer and understand relationships between variables. If statistical agencies equated disclosure with inference, very little data would be released. Second, inferences are designed to predict aggregate behavior, not individual attributes, and thus are often poor predictors of individual data values. Inferential disclosure is still a concern where cases of exceptionally close statistical associations exist and regression models can be used to generate predictions. Inference disclosure is a consideration for reviewing analytical products produced from either a research data center or research project with an agency's restricted access data program. The risk of disclosure may exist in regression models that contain only fully-interactive sets of dummy variables as independent variables. In these cases, agencies need to further examine the potential disclosure risks from the use of certain regression models.

C. On-Line Query Systems

The dissemination of data through the availability of on-line query systems requires special application of disclosure limitation methods. On-line query systems may have multiple capabilities. The simplest form is where the system accesses summary files containing aggregated data that have already been tested for sensitivity and disclosure limitation methods applied. Another capability is the dissemination of tabulations from online queries of microdata files that have already been protected. Applications that access unprotected microdata can introduce a risk of identity disclosure when restricting the query to a small geographic area or category. This is of particular concern for sequences of independent queries about small geographic areas or categories. Specialized tabulations generated from queries to unprotected microdata files must pass through a series of filters where the disclosure limitation rules are applied.

Four agencies have developed on-line query systems with various capabilities for users to generate special tabulations. The Centers for Disease Control and Prevention developed "CDC Wonder" ((Wide-ranging OnLine Data for Epidemiologic Research (WONDER)) at http://www.cdc.gov/nchs/index.htm. 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 and the requested data are readily summarized. The data are previously tested for sensitivity with disclosure limitation methods applied prior to being added to the database. Users of the CDC wonder system are subject to the agency's data use restrictions that prohibits linking the data with other data sets or information for the purpose of identifying an individual. The Bureau of Labor Statistics also has an online query system available at http://www.bls.gov/data/sa.htm which allows users to access first level summary data (disclosure limitation applied) to generate customized tables.

The Economic Research Service in conjunction with the National Agricultural Statistics Service developed a system available at http://www.ers.usda.gov/Data/ARMS/ for users to generate customized data tables by accessing data from the Agricultural Resource Management Survey (ARMS) program. In the ARMS system, disclosure limitation has already been applied to the microdata. The Census Bureau developed the "American Fact Finder" available at **http://www.census.gov** that provides users with access to both summary tabular data as well as microdata files. The Advanced Query System of American Fact Finder has the sensitivity rules and disclosure methods built into the system so that queries submitted by users must pass disclosure review before the user can view the results. At the National Center for Education Statistics (NCES) all postsecondary sample survey data are available through the use of data analysis tools that produce tables up to three-dimensions and give correlation matrices. In addition, elementary and secondary level data from the National Assessment of Educational Progress (NAEP) are also available in an on-line data tool. A more detailed description of on-line query systems is contained in Chapter 4 Section C.

D. Tables of Counts or Frequencies

The data collected from most surveys about people are published in tables that show counts (number of people by category) or frequencies (fraction or percent of people by category). A portion of a table published from a sample survey of households that collects information on energy consumption is shown in Table 1 below as an example.

D.1. Sampling as a Statistical Disclosure Limitation Method

One method of protecting the confidentiality of data is to conduct a sample survey rather than a census. Disclosure limitation techniques are not applied in Table 1 even though respondents are given a pledge of confidentiality because it is a large-scale **sample** survey. Estimates are calculated by multiplying a respondent's data by a sampling weight and then aggregating all the weighted responses. When data are used to make estimates concerning the population from which a sample is drawn, they are generally adjusted by sample weights that take into account the peculiarities of the sampling procedure. Weighted totals take the place of actual frequencies in published tables. The use of sample weights makes an individual respondent's data less identifiable from published totals when the values of the weights themselves are not disclosed. In particular, if the weighting of the survey responses is complex, the published estimate may hide the fact that there are only one or two contributors to a cell. Because the weighted numbers represent all households in the United States, the counts in Table 1 are given in units of millions

of households. They were derived from a sample survey of less than 7000 households. This illustrates the protection provided to individual respondents by sampling and estimation.

Table 1: Example Without Disclosure

Number of Households by Heated Floor Space and Family Income (Million U.S. Households)

1997 Family income

Heated Floor Space sq ft	Total	Less than $10000	$10000 to $24999	$25000 to $49999	$50000 or more	Below Poverty Line	Eligible for Federal Assistance
Fewer than 600	7.9	2.9	3.1	1.6	0.3	2.7	4.9
600 to 999	21.5	4.3	8.6	6.0	2.6	4.6	10.2
1000 to 1599	30.4	2.8	9.7	10.8	7.0	3.7	9.9
1600 to 1999	15.3	.6	3.2	5.4	6.1	0.9	2.8
2000 to 2399	7.9	.2	1.2	2.5	4.0	0.3	1.1
2400 to 2999	5.3	Q	0.3	1.4	3.4	0.2	0.5
3000 or more	4.1	Q	0.3	.9	2.8	Q	0.4

NOTE: Q -- Data withheld because relative standard error exceeds 50% or fewer than 10 households were sampled.
SOURCE: "Housing Characteristics 1997", Residential Energy Consumption Survey, Energy Information Administration, DOE/EIA-0632(97), page 58.

When it is known with certainty that an individual is a study respondent, the task of identifying the person and his/her attributes is much simpler than when there is a high probability that the person is not represented in the table or microdata at all. Should the complete count data reveal that respondent to be unique using information that an individual was a respondent, his or her identity would be confirmed and their attributes revealed. Data collection based upon a sample of persons is protective because the presence of a given person's records is not certain and a respondent who appears to be unique may not be the person he/she is thought to be.

Additionally, many agencies require that estimates must achieve a specified accuracy before they can to be published. In Table 1 cells with a "Q" are withheld because the relative standard error is greater than 50 percent. Sample survey accuracy requirements such as this one result in more cells being withheld from publication than would a disclosure limitation rule. In Table 1 the values in the cells labeled Q can be derived by subtracting the other cells in the row from the marginal total. The purpose of the Q is not necessarily to withhold the value of the cell from the public, but rather to indicate that any number so derived does not meet the accuracy requirements of the agency.

Sampling may lower the disclosure risks from published data depending on the sampling rate, the number and detail of variables tabulated, and whether or not there exists a public listing of the complete population from which the sample is drawn. The sample should also be free of any outlier values such as individuals or establishments with unusual characteristics. The use of sampling methodology does not ensure that the published data are free from disclosure risks and any published tables from a sample should still be reviewed.

D.2. Defining Sensitive Cells

In the discussion below we identify two classes of disclosure limitation rules for tables of counts or frequencies. The first class consists of special rules designed for specific tables to protect against the potential harm to an agency or respondent from disclosing confidential information. Such rules differ from agency to agency and from table to table. These special rules are generally designed to provide protection to data considered particularly sensitive by the agency. The second class is more general where the number in a cell is considered to represent an unacceptable disclosure risk such as: a cell is defined as sensitive if the number of respondents is less than some specified threshold (the threshold rule).

D.2.a Special Rules

Special rules impose restrictions on the level of detail that may be provided in a table. For example, Social Security Administration (SSA) rules prohibit tabulations in which a cell value inside a row or column of a table is equal to a marginal total or which would allow users to determine an individual's age within a five-year interval, earnings within a $1000 interval or benefits within a $50 interval. Tables 2 and 3 illustrate these rules. They also illustrate the method of restructuring tables and combining categories to limit disclosure in tables.

Table 2 is a two-dimensional table showing the number of beneficiaries by county and size of benefit. This table could not be released to the public because the data shown for counties B and D violate Social Security's disclosure rules. For county D, there is only one cell with a positive value, and a beneficiary in this county is known to be receiving benefits between $40 and $59 per month. This violates two rules. First the detailed cell is equal to the row total; and second, this reveals that all beneficiaries in the county receive between $40 and $59 per month in benefits. This interval is less than the required $50 interval. For county B, there are 2 cells with positive values, but the range of possible benefits is from $40 to $79 per month, an interval of less than the required $50.

Table 2: Example -- With Disclosure

Number of Beneficiaries by Monthly Benefit Amount and County

	Monthly Benefit Amount						
County	$0-19	$20-39	$40-59	$60-79	$80-99	$100+	Total
A	2	4	18	20	7	1	52
B	--	-	7	9	-	-	16
C	--	6	30	15	4	-	55
D	-	-	2	--	-	-	2

SOURCE: FCSM Statistical Policy Working Paper 2.

To protect confidentiality, Table 2 could be restructured and rows or columns combined (sometimes referred to as "rolling-up categories" or "collapsing"). Combining the row for county B with the row for county D would still reveal that the range of benefits is $40 to $79. Combining A with B and C with D does offer the required protection, as illustrated in Table 3.

Table 3: Example -- Without Disclosure

Number of Beneficiaries by Monthly Benefit Amount and County

	Monthly Benefit Amount						
County	$0-19	$20-39	$40-59	$60-79	$80-99	$100+	Total
A and B	2	4	25	29	7	1	68
C and D	--	6	32	15	4	-	57

SOURCE: FCSM Statistical Policy Working Paper 2.

D.2.b. The Threshold Rule

With the threshold rule, a cell in a table of frequencies is defined as **sensitive** if the number of respondents is less than some specified number. Some agencies require at least 5 respondents in a cell, while others require 3. Under certain circumstances the number may be much larger. The choice of the minimum number is generally made in consideration of: (a) the sensitivity of the information that the agency is considering to publish, (b) the amount of protection the agency determines to be necessary given the degree of precision required to achieve disclosure.

D.3. Protecting Sensitive Cells After Tabulation

In tables of frequency data, if cells have been identified as being sensitive, the agency must take steps to protect the sensitive data. There are generally two approaches for doing this. One consists of making changes to the table itself. This is done as part of, or after tabulation. These methods include restructuring tables and combining categories (as illustrated above), cell suppression, random rounding, controlled rounding, or controlled tabular adjustment. The second approach that has evolved more recently is the application of microdata methods to the data file prior to tabulation. These methods are particularly efficient for use with on-line query systems or where multiple tables will be created from a single data file. This approach is illustrated in section D.4 of this chapter.

Table 4 is a fictitious example of a table with disclosures. The fictitious data set consists of information concerning delinquent children. Cells in Table 4 with fewer than 5 respondents are defined as sensitive and are identified with an asterisk. This table is used to illustrate cell suppression, random rounding, controlled rounding, and controlled tabular adjustment in the sections below.

Table 4: Example -- With Disclosure

Number of Delinquent Children by County and Education Level of Household Head

	Education Level of Household Head				
County	Low	Medium	High	Very High	Total
Alpha	15	**1***	**3***	**1***	20
Beta	20	10	10	15	55
Gamma	**3***	10	10	**2***	25
Delta	12	14	7	**2***	35
Total	50	35	30	20	135

SOURCE: Numbers taken from Cox, McDonald, and Nelson (1986). Titles, row and column headings are fictitious.

D.3.a. Suppression

One of the most common methods of protecting sensitive cells is by **suppression**. In a row or column with a suppressed sensitive cell, at least one additional cell must be suppressed, or the value in the sensitive cell could be calculated exactly by subtraction from the marginal total. For this reason, certain other non-sensitive cells must also be suppressed. These are referred to as

complementary suppressions. While it is possible to select cells for complementary suppression manually, in all but the simplest of cases, it is difficult to guarantee that the result provides adequate protection.

Table 5 shows an example of a system of suppressed cells for Table 4 that has at least two suppressed cells in each row and column. This table appears to offer protection to the sensitive cells, however, a closer review shows disclosure of sensitive data still occurs

Table 5: Example -- With Disclosure, Not Protected by Suppression

Number of Delinquent Children by County and Education Level of Household Head

Education Level of Household Head

County	Low	Medium	High	Very High	Total
Alpha	15	D1	D2	D3	20
Beta	20	D4	D5	15	55
Gamma	D6	10	10	D7	25
Delta	D8	14	7	D9	35
Total	50	35	30	20	135

NOTE: D indicates data withheld to limit disclosure.
SOURCE: Numbers taken from Cox, McDonald, and Nelson (1986). Titles, row and column headings are fictitious.

Consider the following linear combination of row and column entries: Row 1 (county Alpha) + Row 2 (county Beta) - Column 2 (medium education) - Column 3 (high education), can be written as

(15 + D1 + D2 + D3) + (20 + D4 + D5 + 15) - (D1 + D4 + 10 + 14) - (D2 + D5 + 10 + 7) = 20 + 55 - 35 - 30.

This reduces to D3 = 1.

This example shows that selection of cells for complementary suppression is a complicated process. Mathematical methods of linear programming are used to automatically select cells for complementary suppression and also to **audit** a proposed suppression pattern (e.g. Table 5) to see if it provides the required protection. Chapter IV provides more detail on the mathematical issues of selecting complementary cells and auditing suppression patterns.

Table 6 shows our table with a system of suppressed cells that does provide adequate protection for the sensitive cells. However, Table 6 illustrates one of the problems with suppression. Out of a total of 16 interior cells, only 7 cells are published, while 9 are suppressed.

Table 6: Example -- Without Disclosure, Protected by Suppression

Number of Delinquent Children by County and Education Level of Household Head

Education Level of Household Head

County	Low	Medium	High	Very High	Total
Alpha	15	D	D	D	20
Beta	20	10	10	15	55
Gamma	D	D	10	D	25
Delta	D	14	D	D	35
Total	50	35	30	20	135

NOTE: D indicates data withheld to limit disclosure.
SOURCE: : Numbers taken from Cox, McDonald, and Nelson (1986).. Titles, row and column headings are fictitious.

D.3.b. Random Rounding

In order to reduce the amount of data loss that occurs from suppressing sensitive cells in a table alternative data perturbation methods such as random rounding and controlled rounding are available to protect sensitive cells in tables showing frequency data. In **random rounding** cell values are rounded, but instead of using standard rounding conventions a random decision is made as to whether they will be rounded up or down. (A more theoretical discussion of this method is contained in "Elements of Statistical Disclosure Control" by Leon Willenborg and Ton de Waal, 2001).

For this example, it is assumed that each cell will be rounded to a multiple of 5. Each cell count, X, can be written in the form

$$X = 5q + r,$$

where q is a nonnegative integer, and r is the remainder (which may take one of 5 values: 0, 1, 2, 3, 4). This count would be rounded up to 5*(q+1) with probability r/5; and would be rounded down to 5*q with probability (1-r/5). A possible result is illustrated in Table 7.

Table 7: Example -- Without Disclosure, Protected by Random Rounding

Number of Delinquent Children by County and Education Level of Household Head

Education Level of Household Head

County	Low	Medium	High	Very High	Total
Alpha	15	0	0	0	20
Beta	20	10	10	15	55
Gamma	5	10	10	0	25
Delta	15	15	10	0	35
Total	50	35	30	20	135

SOURCE: Numbers taken from Cox, McDonald, and Nelson (1986). Titles, row and column headings are fictitious.

Because rounding is done separately for each cell in a table, the rows and columns do not necessarily add to the published row and column totals. In Table 7 the total for the first row is 20, but the sum of the values for the interior cells in the first row is 15. A table prepared using random rounding could lead the public to lose confidence in the numbers: at a minimum it looks as if the agency cannot add.

D.3.c. Controlled Rounding

To solve the additivity problem, a procedure called **controlled rounding** was developed. It is a form of random rounding, but it is constrained to have the sum of the published entries in each row and column equal the appropriate published marginal totals (see Cox and Ernst, 1982). Linear programming methods are used to identify a controlled rounding for a table. Controlled rounding is used by the Social Security Administration in statistical tables showing frequency counts. Table 8 illustrates controlled rounding where the sum of the cell values in each row and column are constrained to equal the sum of the published totals.

D.3.d. Controlled Tabular Adjustment

Controlled tabular adjustment is a relatively new approach, similar to controlled rounding, but it is most valuable when applied to tables of magnitude data. This method was initially referred to as "synthetic tabular data." It was described as controlled tabular adjustment in subsequent work (Cox and Dandekar, 2004). For magnitude data, a linear sensitivity rule is used to determine which cells are sensitive. With controlled tabular adjustment each original sensitive value of a table is replaced with a safe value that is a "sufficient distance" away from the true value; and non-sensitive cell values are minimally adjusted to ensure that the published marginal totals are additive. A "sufficient distance" from the true value would be the value needed to be added to

the cell total that would make the cell not sensitive according to the linear sensitivity rule being applied. For frequency data, most linear sensitivity rules are equivalent to a threshold rule of 3 respondents and a "sufficient distance" from the true value would involve changing the value by either 1 or 2. That is, the value of a sensitive cell would be changed to either 0 or 3. This is identical to rounding to the base 3.

Table 8: Example -- Without Disclosure, Protected by Controlled Rounding

Number of Delinquent Children by County and Education Level of Household Head

Education Level of Household Head

County	Low	Medium	High	Very High	Total
Alpha	15	0	5	0	20
Beta	20	10	10	15	55
Gamma	5	10	10	0	25
Delta	10	15	5	5	35
Total	50	35	30	20	135

SOURCE: Numbers taken from Cox, McDonald, and Nelson (1986). Titles, row and column headings are fictitious.

Table 9 illustrates a simplified way to implement controlled tabular adjustment, as described in Dandekar (2004). The internal sensitive cells are first listed in descending order from most sensitive to least sensitive (2, 2, 1, 1). Adjustments are applied sequentially beginning with the first cell. The first cell is changed at random to 0 or 3 (by either subtracting 2, or by adding 1.) Subsequent adjustments will be implemented with alternate signs. So if the first cell is altered by adding 1, the second cell is altered by subtracting 2, the third is altered by adding 2, the last is altered by subtracting 1. Once the internal sensitive cells have been altered, no additional changes are needed in the interior non-sensitive cells (as is typically done for controlled rounding). Marginal table totals are re-computed to account for the changes made to the internal sensitive cells. These changes are needed so that the tables add. In Table 9 the marginal totals are adjusted to minimize the percent by which cells are changed. In this example, no changes are needed to the grand total.

Table 9: Example – Without Disclosure -- Protected by Controlled Tabular Adjustment

Number of Delinquent Children by County and Education Level of Household Head

	Education Level of Household Head				
County	Low	Medium	High	Very High	Total
Alpha	15	**1* - 1 = 0**	3	**1* + 2 = 3**	**20 + 1 = 21**
Beta	20	10	10	15	55
Gamma	3	10	10	**2* - 2 = 0**	**25 - 2 = 23**
Delta	12	14	7	**2* + 1 = 3**	**35 + 1 = 36**
Total	50	**35 - 1 = 34**	30	**20 + 1 = 21**	135

Controlled tabular adjustments to individual cell values are shown in **Bold** font**.**

D.4. Protecting Sensitive Cells Before Tabulation

Tabular data can be protected by applying disclosure protection methods to the underlying microdata files to assure that any tables that are generated from the microdata files are fully protected. This approach is particularly efficient if there are many tabulations being created from the same data.

The Census bureau has been the leader in applying microdata methods to protect files based on the Decennial Census. Data swapping is illustrated in section II.F.2.c, and is also described in Domingo-Ferrer, (2002). The decennial Census collects basic data from all households in the U.S. It collects more extensive data via the long-form from a sample of U.S. households. Both sets of data are subjected to a data swapping procedure. This technique was used for short form data in the 1990 census, but was revised and extended to the long form data in 2000. The procedure now takes a targeted approach to swapping which increases the effectiveness of the procedure with some cost in terms of bias of variance. All Decennial tabulations come from the swapped files, this guarantees the consistency of the tables and avoids problems associated with protecting interrelated tables.

In 1990, a different procedure was used in the confidentiality edit for the sample data, called "blank and impute", see section II.F.2.d. In this technique, selected records have particular values blanked and treated as missing. Since there are usually pre-existing procedures for imputation of missing data, "blank and impute" has some advantage in economy. However, the procedure reduces effective sample size and the compensation in the calculation of variance is sometimes difficult to accomplish. In some sense, "blank and impute" is a precursor of the synthetic data techniques currently being researched at the Census Bureau and elsewhere (Raghunthan, et. al. 2003). The advantage of data swapping is that it maximizes the information that can be provided in tables. Additionally, all tables are protected in a consistent way.

E. Tables of Magnitude Data

Tables showing magnitude data have a unique set of disclosure problems. Magnitude data are generally nonnegative quantities reported in surveys or censuses of business establishments, farms or institutions. The distribution of these reported values is likely to be skewed, with a few entities having very large values. Disclosure limitation in this case concentrates on making sure that the published data cannot be used to estimate within too close of a range the values reported by the largest, most highly visible respondent. By protecting the largest reported values, we, in effect, are able to protect all values.

Linear sensitivity rules are used to identify cells that are "sensitive" and need to be protected. Recent research has focused on applying protections to the microdata file prior to tabulation. This provides a great advantage, especially if tabulations will be provided through a query system. Historically cell suppression was used to protect sensitive cells in tables. Cell suppression is done as part of the construction of a table.

E.1. Defining Sensitive Cells – Linear Sensitivity Rules

For magnitude data it is less likely that sampling alone will provide disclosure protection because most sample designs for economic surveys include a stratum of the larger volume entities that are selected with certainty. Thus, the units that are most visible because of their size do not receive any protection from sampling. For tables of magnitude data, rules called **primary suppression rules** or **linear sensitivity measures**, have been developed to determine whether a given table cell could reveal individual respondent information. Cells that do not pass the linear sensitivity test are defined as **sensitive** cells, and are withheld from publication.

The primary suppression rules most commonly used to identify sensitive cells by government agencies are the **(n) threshold rule**, **(n, k) rule**, and the **p-percent** or **pq** rules. See Cox, (1981). All are based on the desire to make it difficult for one respondent to estimate the value reported by another respondent too closely. The largest reported value is the most likely to be estimated accurately. Primary suppression rules can be applied to frequency data. However, since all respondents contribute the same value to a frequency count, the rules default to a threshold rule and the cell is sensitive if it has too few respondents. The p% and pq rules default to a threshold rule of 3 when applied to count data. Primary suppression rules are discussed in more detail in Section VI.B.1.

E.2 Protecting Sensitive Cells After Tabulation

Tables for publication are populated from the microdata files. During aggregation, a linear sensitivity rule is used to identify any sensitive cells. Once sensitive cells have been identified, there are 3 options: restructure the table and collapse cells until no sensitive cells remain, use cell suppression, or apply controlled tabular adjustment. With cell suppression, once the sensitive cells have been identified they are withheld from publication. These are called **primary suppressions**. Other cells, called **complementary suppressions** are selected and suppressed so that the sensitive cells cannot be derived by addition or subtraction from published marginal totals. Problems associated with cell suppression for tables of count data were

illustrated in Section C.3.a of this chapter. The same problems exist for tables of magnitude data.

Controlled tabular adjustment was illustrated for frequency data in Section C.3.d. of this chapter. For magnitude data, the "sufficient distance" is the amount that would need to be added to the cell total so that the linear sensitivity rule would classify the cell as not sensitive.

An administrative way to avoid cell suppression is used by a number of agencies. They obtain written permission, or "**informed consent**" to publish a sensitive cell from the respondents that contribute to the cell. The written permission is called a "waiver" of the promise to protect sensitive cells and specific authorization or consent to the agency for publicly releasing the confidential information. In this case, respondents are requested by an agency to voluntarily give their consent after being informed of the need to release the confidential information, and the proposed statistical or non-statistical use of the information. This method is most useful with small surveys or sets of tables involving only a few small cells, where only a few waivers are needed. Of course, respondents must be informed of the proposed use of the data prior to giving their consent.

E.3. Protecting Sensitive Cells Before Tabulation

There are few microdate products for establishment surveys because of the skewed nature of the population. However, applying microdata methods to protect files of establishment data prior to tabulation has simplified the protection of tabular data and provided new data products.

The Census Bureau was the first to apply microdata methods to protect establishment level data files prior to tabulation. The technique of noise addition, section II.F.2.b, has been the primary method used, in conjunction with other methods. In particular, noise addition has been used to protect quarterly workforce indicators released from the Longitudinal Employer Household Dynamics project. Magnitude data for establishments tends to be skewed and dominated by large companies. This can lead to a situation where applying linear sensitivity rules flags many cells for protection against disclosure. Noise addition adds noise to each responding establishment's data by a small percentage. The amount of the perturbation of the reported value depends on the magnitude of the reported value, and the value of the linear sensitivity rule for the cells containing that respondent's data. If a cell contains only one establishment, or if a single establishment dominates a cell, the published value in a cell will not be a close approximate to the dominant establishment's value because that value has had noise added to it. The dominant establishment's true reported value is protected by the noise addition. It is important to note that all establishments have their values multiplied by a corresponding noise factor, or adjusted weight, before the data are tabulated. The noise multipliers can be randomly assigned to control the effects of the noise on different types of cells within a table.

Noise addition was also used by the U.S. Department of Agriculture's Economic Research Service to protect the reported values in their annual Agricultural Resource Management Survey (ARMS) that is available through an on-line query system. The values are adjusted alternating between adding and subtracting noise following the order of observations in the data set so that the cell totals are approximately the same after the noise addition is applied.

The method has several advantages over cell suppression in that in provides some information in more cells of the table, and it eliminates the need to coordinate cell suppression patterns. This methodology provides consistency in the tables generated from the microdata, but it is important that the initial microdata have been sufficiently perturbed so that the tables produced are safe for release. One limitation of this methodology is that marginal values can show large changes as a result of adjusting the underlying weights. The relationship between the actual unadjusted cell values and adjusted cell values using noise addition should be reviewed prior to releasing the data.

F. Microdata

Information collected about establishments is primarily magnitude data. These data are likely to be highly skewed, and there are likely to be high risk respondents that could easily be identified via other publicly available information. As a result, special care must be taken when considering the release of microdata files containing establishment data. Examples of the public release of microdata files from establishment surveys include data from the Commercial Building Energy Consumption Survey, which is provided by the Energy Information Administration, and files from the 1997 Census of Agriculture provided by the Census Bureau. Disclosure protection is provided using the techniques described below in addition to removing variables that serve as direct identifiers of respondents to the survey.

It has long been recognized that it is difficult to protect a microdata set from disclosure because of the possibility of matching to outside data sources (Bethlehem, Keller and Panekoek, 1990). Additionally, there are no accepted measures of disclosure risk for a microdata file, so there is no "standard" which can be applied to assure that protection is adequate. A “Checklist on Disclosure Potential of Proposed Data Releases” was developed by the Confidentiality and Data Access Committee to assist agencies in reviewing the disclosure potential of proposed public use microdata files and is available for download at http://www.fcsm.gov/committees/cdac/. The Bureau of Labor Statistics, Bureau of Transportation Statistics, National Center for Health Statistics, Census Bureau, and Social Security Administration use the CDAC checklist or some modified format of the checklist for reviewing proposed data releases for any disclosure potential. The National Science Foundation also uses the CDAC checklist as guidelines for their contractors to follow when reviewing a proposed file for public release. The methods for protection of microdata files described below are used by all agencies which provide public use data files. To reduce the potential for disclosure, most public-use microdata files:

1. Include data from only a sample of the population,
2. Do not include obvious identifiers,
3. Limit geographic detail, and
4. Limit the number and detailed breakdown of categories within variables on the file.

Additional methods used to disguise high risk variables include:

1. Truncation of extreme codes for certain variables (Top or bottom-coding),
2. Recoding into intervals or rounding,
3. Adding or multiplying by random numbers (noise),
4. Swapping or rank swapping (also called switching),
5. Selecting records at random, blanking out selected variables and imputing for them (also called blank and impute),
6. Aggregating across small groups of respondents and replacing one individual's reported value with the average (also called blurring).

These will be illustrated with the fictitious example we used in the previous section.

F.1. Sampling, Removing Identifiers and Limiting Geographic Detail

First: include only the data from a sample of the population. For this example we used a 10 percent sample of the population of delinquent children. Second: remove identifiers that directly identify respondents such as name, address, and identification numbers. In this case the identifier is the first name of the child. Third: consider the geographic detail. We decide that we cannot show individual county data for a county with less than 30 delinquent children in the population. Therefore, the data from Table 4 shows that we cannot provide geographic detail for counties Alpha or Gamma. As a result counties Alpha and Gamma are combined and shown as AlpGam in Table 9. These manipulations result in the fictitious microdata file shown in Table 10.

In this example we discussed only 5 variables for each child. One might imagine that these 5 were selected from a more complete data set including names of parents, names and numbers of siblings, age of child, ages of siblings, address, school and so on. As more variables are included in a microdata file for each child, unique combinations of variables make it more likely that a specific child may be identified by a knowledgeable person. Limiting the number of variables to 5 makes such identification less likely.

F.2. High Risk Variables

It may be that information available to others in the population could be used with the income data shown in Table 10 to uniquely identify the family of a delinquent child. For example, the employer of the head of household generally knows his or her exact salary. Variables such as income, race, and age are **high risk** variables and require additional protection.

Table 10: Fictitious Microdata -- Sampled, Identifiers Removed

Geographic Detail Limited - Delinquent Children

Number	County	HH Education	HH Income	Race
1	AlpGam	High	61	W
2	AlpGam	Low	48	W
3	AlpGam	Medium	30	B
4	AlpGam	Medium	52	W
5	AlpGam	Very High	117	W
6	Beta	Very High	138	B
7	Beta	Very High	103	W
8	Beta	Low	45	W
9	Beta	Medium	62	W
10	Beta	High	85	W
11	Delta	Low	33	B
12	Delta	Medium	59	B
13	Delta	Medium	59	W
14	Delta	High	72	B

NOTE: HH means head of household. Income reported in thousands of dollars. County AlpGam means either Alpha or Gamma.

F.2.a. Top-coding, Bottom-coding, Recoding into Intervals

In the example, large income values are **top-coded** by showing only that the income is greater than 100,000 dollars per year. Small income values are **bottom-coded** by showing only that the income is less than 40,000 dollars per year. Finally, income values are **recoded** by presenting income in 10,000 dollar intervals. The result of these manipulations yields the fictitious public use data file in Table 11. Top-coding, bottom-coding and recoding into intervals are among the most commonly used methods to protect high risk variables in microdata files.

Table 11: Fictitious Microdata -- Sampled, Identifiers Removed

Geographic Detail Limited, Income Top, Bottom and Recoded - Delinquent Children

Geographic Detail Limited Delinquent Children

Number	County	HH Education	HH Income	Race
1	AlpGam	High	60-69	W
2	AlpGam	Low	40-49	W
3	AlpGam	Medium	<40	B
4	AlpGam	Medium	50-59	W
5	AlpGam	Very High	>100	W
6	Beta	Very High	>100	B
7	Beta	Very High	>100	W
8	Beta	Low	40-49	W
9	Beta	Medium	60-69	W
10	Beta	High	80-89	W
11	Delta	Low	<40	B
12	Delta	Medium	50-59	B
13	Delta	Medium	50-59	W
14	Delta	High	70-79	B

NOTE: HH means head of household. Income reported in thousands of dollars. County AlpGam means either Alpha or Gamma.

F.2.b. Adding Random Noise

An alternative method of disguising high risk variables, such as income, is to add or multiply by random numbers. For example, in the above example, assume that we will add a normally distributed random variable with mean 0 and standard deviation 5 to income. Along with the sampling, removal of identifiers and limiting geographic detail, this might result in a microdata file such as Table 12. To produce this table, 14 random numbers were selected from the specified normal distribution, and were added to the income data in Table 10.

Table 12: Fictitious Microdata -- Sampled, Identifiers Removed

Geographic Detail Limited, Random Noise Added to Income - Delinquent Children

Number	County	HH education	HH income	Race
1	AlpGam	High	61	W
2	AlpGam	Low	42	W
3	AlpGam	Medium	32	B
4	AlpGam	Medium	52	W
5	AlpGam	Very high	123	W
6	Beta	Very high	138	B
7	Beta	Very high	94	W
8	Beta	Low	46	W
9	Beta	Medium	61	W
10	Beta	High	82	W
11	Delta	Low	31	B
12	Delta	Medium	52	B
13	Delta	Medium	55	W
14	Delta	High	61	B

NOTE: HH means head of household. Income reported in thousands of dollars. County AlpGam means either Alpha or Gamma.

F.2.c. Data Swapping and Rank Swapping

Swapping involves selecting a sample of the records, finding a match in the database on a set of predetermined variables and swapping all other variables. Swapping is illustrated in section E.2.e. In that example records were identified from different counties that matched on race, sex, and income, and the variables first name of child and household education were swapped. For purposes of providing additional protection to the income variable in a microdata file, we might choose instead to find a match in another county on household education and race and to swap the income variables.

Swapping offers the opportunity to select some statistics that will be preserved through the swapping operation. This is accomplished by forcing agreement between the swapped pairs on the variables involved in those statistics. The National Institute of Statistical Sciences (NISS) has a software package which performs and analyzes data swapping in categorical data variables that is available from their website at http://www.niss.org/software/dstk.html. The NISS technique uses random swapping; this affords one the ability to quantify the effect on statistics produced from the swapped data set. For data sets with an accurate measure of record level risk, one can employ a variation, termed targeted swapping. Those records with high risk are automatically selected for pairing in the swap process. In targeted swapping, fewer records are involved and the protection level is generally higher. However, the targeted procedure is biased and the ability to present a general statement on data quality is very limited.

Rank swapping provides a way of using continuous variables to define pairs of records for swapping. Instead of insisting that variables match (agree exactly), they are defined to be close based on their proximity to each other on a list sorted by the continuous variable. Records that are close in rank on the sorted variable are designated as pairs for swapping. Frequently in rank swapping, the variable used in the sort is the one that will be swapped.

Data Shuffling is another method for modifying micro data that has been applied to numerical data. The procedure involves two steps: first the values of the confidential variables are modified using a general perturbation technique and second, a data shuffling procedure is applied using the perturbed values of the confidential variables on the file. The perturbed values are sorted from lowest to highest value in the re-shuffled file. Then the perturbed value is replaced with the original value of the confidential variable based on the ranking of the original values from the confidential variable. Before the data are perturbed, the conditional distribution between the confidential and non-confidential variables is derived. This method preserves the rank order correlation between the confidential and non-confidential attributes, and avoids the loss in data utility that could occur from applying data swapping or rank swapping methodology. Data shuffling is discussed in more detail in Chapter V.

Data swapping was used to protect the confidentiality of the Census 2000 tabulations. The procedure was performed on the underlying microdata, and all tabulations from the 100% (short form) and from the sample (long form) data were created from the swapped files. It affected pairs of households (or partnered households) where one or both of those households had a high risk of disclosure. The set of census households that were deemed as having a disclosure risk were selected from the internal census data files. These households were unique in their geographic area (block for 100% data and block group for sample data) based on certain characteristics. The data from these households were swapped with data from partnered households that had identical characteristics on a certain set of key variables but were from different geographic locations. Which households were swapped is not public information. The swapping procedure was performed independently for the 100% data and the sample data. To maintain data quality, there was a maximum percent of records that were swapped for each state for the 100% data and another maximum percent for the sample data.

To illustrate the set of data swapping procedures that were applied to the 100 percent microdata file we use fictitious records for the 20 individuals in county Alpha who contributed to Tables 4 through 8. Table 13 shows 5 variables for these individuals. Recall that the previous tables showed counts of individuals by county and education level of head of household. The purpose of the data swapping is to provide disclosure protection to tables of frequency data. However, to achieve this, adjustments are made to the microdata file before the tables are created. The following steps are taken to apply the data swapping procedures:

1. Take a random sample of records from the microdata file (such as 10% sample). Assume that records number 4 and 17 were selected as part of our 10% sample.

Table 13: Fictitious Microdata

All Delinquent Children in County Alpha

Number	Child	County	HH education	HH income	Race	Sex
1	John	Alpha	Very high	201	B	M
2	Jacob	Alpha	High	103	W	M
3	Sue	Alpha	High	75	B	F
4	Pete	Alpha	High	61	W	M
5	Ramesh	Alpha	Medium	72	W	M
6	Dante	Alpha	Low	103	W	M
7	Larry	Alpha	Low	91	B	M
8	Marilyn	Alpha	Low	84	W	F
9	Steve	Alpha	Low	75	W	M
10	Paul	Alpha	Low	62	B	M
11	Renee	Alpha	Low	58	W	F
12	Virginia	Alpha	Low	56	B	F
13	Mary	Alpha	Low	54	B	F
14	Laura	Alpha	Low	52	W	F
15	Tom	Alpha	Low	55	B	M
16	Al	Alpha	Low	48	W	M
17	Mike	Alpha	Low	48	W	M
18	Phil	Alpha	Low	41	B	M
19	Brian	Alpha	Low	44	B	M
20	Nancy	Alpha	Low	37	W	F

NOTES: HH indicates head of household. Income shown in thousands of dollars.

2. Since we need tables by county and education level, we find a match in some other county on the other variables race, sex and income. (As a result of matching on race, sex and income, county totals for these variables will be unchanged by the swapping.) A match for record 4 (Pete) is found in County Beta. The match is with Alfonso whose head of household has a very high education. Record 17 (Mike) is matched with George in county Delta, whose head of household has a medium education. In addition, part of the randomly selected 10% sample from other counties match records in county Alpha. One record from county Delta (June with high education) matches with Virginia, record number 12. One record from county Gamma (Heather with low education) matched with Nancy, in record 20.

3. After all matches are made, swap attributes on matched records. The adjusted microdata file after these attributes are swapped appears in Table 14.

4. Use the swapped data file directly to produce tables. See Table 15.

Applying the set of data swapping procedures has a great advantage in that multidimensional tables can be prepared easily and the disclosure protection applied will always be consistent.

Table 14: Fictitious Microdata

Delinquent Children After Swapping -- Only County Alpha Shown

Number	Child	County	HH education	HH income	Race	Sex
1	John	Alpha	Very high	201	B	M
2	Jacob	Alpha	High	103	W	M
3	Sue	Alpha	High	75	B	F
4*	**Alfonso**	**Alpha**	**Very high**	**61**	**W**	M
5	Ramesh	Alpha	Medium	72	W	M
6	Dante	Alpha	Low	103	W	M
7	Larry	Alpha	Low	91	B	M
8	Marilyn	Alpha	Low	84	W	F
9	Steve	Alpha	Low	75	W	M
10	Paul	Alpha	Low	62	B	M
11	Renee	Alpha	Low	58	W	F
12*	**June**	**Alpha**	**High**	**56**	**B**	F
13	Mary	Alpha	Low	54	B	F
14	Laura	Alpha	Low	52	W	F
15	Tom	Alpha	Low	55	B	M
16	Al	Alpha	Low	48	W	M
17*	**George**	**Alpha**	**Medium**	**48**	**W**	M
18	Phil	Alpha	Low	41	B	M
19	Brian	Alpha	Low	44	B	M
20*	**Heather**	**Alpha**	**Low**	**37**	**W**	F

Data: first name and education level swapped in fictitious microdata file from another county.
NOTES: HH indicates head of household. Income is shown in thousands of dollars.

Table 15: Table Protected By Data Swapping

Number of Delinquent Children by County and Education Level of Household Head

County	Low	Medium	High	Very High	Total
Alpha	13	2	3	2	20
Beta	18	12	8	17	55
Gamma	5	9	11	0	25
Delta	14	12	8	1	35
Total	50	35	30	20	135

SOURCE: Fictitious microdata.

F.2.d. Blank and Impute for Randomly Selected Records.

The blank and impute method involves deleting the values for selected variables for selected respondents from the microdata file and replacing them with values for those same variables from other respondents or through modeling. This technique is illustrated using data shown in Table 16.

Table 16: Fictitious Microdata -- Sampled, Identifiers Removed

Geographic Detail Limited using Blank and Impute - Delinquent Children

Number	County	HH Education	HH Income	Race
1	AlpGam	High	61	W
2	AlpGam	Low	***63***	W
3	AlpGam	Medium	30	B
4	AlpGam	Medium	52	W
5	AlpGam	Very High	117	W
6	Beta	Very High	***52***	B
7	Beta	Very High	103	W
8	Beta	Low	45	W
9	Beta	Medium	62	W
10	Beta	High	85	W
11	Delta	Low	33	B
12	Delta	Medium	59	B
13	Delta	Medium	***49***	W
14	Delta	High	72	B

NOTE: HH means head of household. Income reported in thousands of dollars. County AlpGam means either Alpha or Gamma.

First, one record is selected at random from each publishable county, AlpGam, Beta and Delta. In the selected record the income value is replaced by an imputed value. If the randomly selected records are 2 in county AlpGam, 6 in county Beta and 13 in county Delta, the income value recorded in those records might be replaced by 63, 52 and 49 respectively. These numbers are also fictitious, but you can imagine that imputed values were calculated as the average over all households in the county with the same race and education. Blank and impute was used as part of the confidentiality edit for tables of frequency data from the 1990 Census sample data files (containing information from the long form of the decennial Census).

F.2.e. Blurring

Blurring replaces a reported value by an average. There are many possible ways to implement blurring. Groups of records for averaging may be formed by matching on other variables or by sorting the variable of interest. The number of records in a group (whose data will be averaged) may be fixed or random. The average associated with a particular group may be assigned to all members of a group, or to the "middle" member (as in a moving average.) It may be performed on more than one variable, with different groupings for each variable.

In our example, we illustrate this technique by blurring the income data. In the complete microdata file we might match on important variables such as county, race and two education groups (very high, high) and (medium, low). Then blurring could involve averaging households in each education group, such as two at a time. In county Alpha (see Table 9) this would mean that the household income for the group consisting of John and Sue would be replaced by the average of their incomes (139), the household income for the group consisting of Jim and Pete would be replaced by their average (82), and so on. After blurring, the data file can be subject to sampling, removal of identifiers, and limitation of geographic detail to further reduce the risk of identification.

F.2.f. Targeted Suppression

Although **suppression** is one of the most commonly used ways of protecting sensitive cells in tables, it may also be used on records in microdata files. When a record contains extreme values or unique values that cannot be adequately protected, it may be necessary to delete the single record in its entirety, or suppress the sensitive values for certain variables on the record.

G. Summary

This chapter describes the standard methods of disclosure limitation used by federal statistical agencies to protect both tables and microdata. It relies heavily on simple examples to illustrate the concepts. A consideration when evaluating different methods is that records subject to swapping, blanking and imputation, and blurring methodologies are not distinguished (or flagged) in any way on a file. This means that not only are the adjusted records protected, but a high degree of uncertainty is introduced such that whatever methods are used to isolate any particular record, the user will not be able to determine with certainty that the isolated record contains actual and not swapped, imputed or blurred values. The mathematical underpinnings of applying disclosure limitation methodology in tables and microdata are reported in more detail in Chapters IV and V, respectively. Agency practices in disclosure limitation are described in Chapter III.

CHAPTER III – Current Federal Statistical Agency Practices

This chapter provides an overview of 14 Federal agency policies, practices, and procedures for statistical disclosure limitation. Agencies are authorized or required to protect individually identifiable data by a variety of statutes, regulations or policies. Statistical disclosure limitation methods are applied by the agencies to limit the risk of disclosure of individual information when statistics are disseminated in tabular or microdata formats.

This review of agency practices is based on three sources. The first source is Jabine (1993b), a paper based in part on information provided by the statistical agencies in response to a request in 1990 by the Panel on Confidentiality and Data Access, Committee on National Statistics. Another source of agency practices was from 1991 when each statistical agency was asked to provide a description of its current disclosure practices, standards, and research plans for tabular and microdata. 12 statistical agencies responded to this request.

The third source was from 2004, when each agency was requested by the Confidentiality and Data Access Committee, a subcommittee of the Federal Committee on Statistical Methodology, to review and supplement their responses concerning current disclosure practices and standards, and to comment on any provisions for researcher access. Thus, the material in this chapter is current as of the publication date.

The first section of this chapter summarizes the disclosure limitation practices for 14 Federal statistical agencies as shown in Statistical Programs of the United States Government: Fiscal Year 2004 (Office of Management and Budget). The agency summaries are followed by an overview of the current status of statistical disclosure limitation policies, practices, and procedures based on the available information. Specific methodologies and the state of software being used are discussed to the extent they were included in the individual agencies' responses.

A. Agency Summaries

A.1. Department of Agriculture

A.1.a. Economic Research Service (ERS)

ERS disclosure limitation practices are documented in the statement of "ERS Policy on Dissemination of Statistical Information," dated September 28, 1989. This statement provides that: Estimates will not be published from sample surveys unless: (1) sufficient nonzero reports are received for the items in a given class or data cell to provide statistically valid results which are clearly free of disclosure of information about individual respondents. In all cases at least three observations must be available, although more restrictive rules may be applied to sensitive data, (2) the second condition is an application of the (n, k) concentration rule or dominance rule to insure that the unexpanded data for any one respondent does not equal a specified threshold, For each published cell value, the respondent must represent less than 60 percent of the total that is being published, except when written permission is obtained from that respondent. In this instance (n, k) = (1, 0.6). Both conditions are applied to magnitude data while the first condition also applies to counts.

Within ERS, access to unpublished, confidential data is controlled by the appropriate branch chief. Authorized users must sign confidentiality certification forms. Restrictions require that data be summarized so individual reports are not revealed.

ERS does not release public-use microdata files. ERS provides access to microdata via its "remote data center" software to authorized users. ERS will share data for statistical purposes with government agencies, universities, and other entities under cooperative agreements as described below for the National Agricultural Statistics Service (NASS). Requests of entities under cooperative agreements with ERS for tabulations of data that were originally collected by NASS are subject to NASS review.

A.1.b. National Agricultural Statistics Service (NASS)

NASS maintains a series of Policy and Standards Memoranda (PSM) which document the policies and standards established for all of the Agency's programs. PSM 12 governs the rules of attribute and inferential disclosure along with provisions for handling special cases. PSM 7 documents NASS policy on the release of unpublished summary data and estimates and access to microdata files. PSM 6 covers the use of the list sampling frame including identity disclosure. PSM 4 presents NASS's legal obligation to protect confidential information and specifies the procedures for confidentiality certification of employees and special agents.

The Agricultural Estimates program includes crop, livestock, environmental, and economic reports that NASS regularly produces through the Agricultural Statistics Board. The Agricultural Estimates program determines primary suppressions using a threshold rule of three and the (n, k) dominance rule. The values of n and k are administratively determined and, with a few exceptions, are consistent across all publications. NASS statisticians are responsible for identifying primary suppressions and their complements, and ensuring that the suppression patterns are consistent over time. Suppressions may be presented individually or as aggregates. PSM 12 allows for the use of informed consent (waivers) for the Agricultural Estimates program if it is determined to be in the interest of the industry. All parties at risk must agree to allow the estimates to be published and have the right to revoke their consent. Agreements are renewed every five years.

For the Census of Agriculture, the Puerto Rico Census of Agriculture, the census follow-on programs including the Farm and Ranch Irrigation Survey, and the Census of Aquaculture, NASS uses the p-percent rule to identify sensitive data cells at risk of disclosure. The threshold rule is also applied to all magnitude data to ensure that a minimum number of farms are represented in each published cell. All magnitude data associated with cells with less than three farms are also suppressed. Complementary suppressions are chosen using network flow methodology. Frequency count data are not considered sensitive and not subject to suppression. Also, NASS does not allow the use of informed consent from respondents for the Census of Agriculture and its follow-on programs.

While it is NASS policy not to release microdata files, NASS operates a Data Lab within its Washington headquarters. Individual researchers may submit a research proposal and request

permission to run specialized models or tabulations on certain microdata files within the lab. Requests are addressed and approved or disapproved on a case-by-case basis by the Associate Administrator. NASS staff monitors the lab and all materials leaving the lab are subject to disclosure review. Individuals using the data lab sign confidentiality forms as NASS agents and are bound by the statutes restricting unlawful use and disclosure of data. NASS will arrange for a data lab in any of its 46 field offices, when needed. Data users may also request special tabulations through the Data Lab. These tabulations are performed by NASS staff and eliminate the need for access to microdata files. The results of each tabulation are considered public domain and are available to any data user.

NASS and the Economic Research Service cooperatively provide an interactive web tool with built-in disclosure review and filtering, that allows individual researchers to run tabulations and special analysis against microdata from the Agricultural Resource Management Survey. Access procedures mirror those of the Data Lab. Individual researchers may submit a research proposal and request an authenticated access ID. Data confidentiality is protected by applying a noise-based approach to the underlying microdata before the tabular data are generated. The parameters used for the noise creation are kept confidential. The p-percent rule is also applied to the aggregates to test a table cell for dominance from a single establishment.

NASS conducts a number of reimbursable surveys for government or academic organizations, and has developed special confidentiality procedures for these surveys. In these situations, NASS will clearly identify the sponsoring organization and purpose of the survey to respondents prior to collecting their voluntary responses. In these situations NASS may provide a microdata file, stripped of identifiers, to the sponsoring organization for their analyses. The microdata file must reside in a physically secure site under security measures approved by NASS. All individuals who will have access to the file must sign confidentiality forms as NASS agents and are bound by the statutes restricting unlawful use and disclosure of data.

In February 1993, USDA's Office of the General Counsel (OGC) reviewed the laws and regulations pertaining to the disclosure of confidential NASS data. In summary, OGC's interpretation of the statutes allows data sharing to other agencies, universities, and private entities as long as it enhances the mission of USDA and is through a cooperative agreement, cost-reimbursement agreement, contract, or memorandum of understanding. Such entities or individuals receiving the data are also bound by the statutes restricting unlawful use and disclosure of the data. NASS's current policy is that data sharing for statistical purposes will occur on a case-by-case basis, as needed, to address an approved specified USDA or public need, and under the specialized situations described above.

To the extent future uses of data are known at the time of data collection, they are explained to the respondent and permission is requested to permit the data to be shared among various users. This permission is requested in writing with a release form signed by each respondent

A.2. Department of Commerce
A.2.a. Bureau of Economic Analysis (BEA)

BEA's disclosure limitation activities pertain mainly to data that it collects on international direct investment and trade in services. These data are collected from U.S. business enterprises—both U.S.-owned and foreign-owned—in mandatory surveys conducted under authority of the International Investment and Trade in Services Survey Act (P.L. 94-472, as amended). Surveys of trade in financial services also are authorized by the Omnibus Trade and Competitiveness Act of 1988. As required by the Survey Act, the data collected are held confidential and are published in a manner that precludes the identification of individual responses. Disclosure limitation activities also are conducted for certain data on regional economic activity that are obtained from the Bureau of Labor Statistics. BLS conducts the disclosure limitation activities for its own purposes and provides a copy of the results to BEA.

With regard to the data on direct investment and trade in services, the general rule for primary suppression involves looking at the data for the top reporter, the second reporter, and all other reporters in a given cell. If the data for all but the top two reporters add up to no more than a certain percent of the top reporter's data, the cell is a primary suppression. This is an application of the p-percent rule.

This rule protects the top reporter from the second reporter, protects the second reporter from the top reporter, and automatically suppresses information in any cell with only one or two reporters. On very rare occasions, respondents may, upon request by BEA, grant a waiver of confidentiality.

When applying the general rule, absolute values are used if the data item can be negative (for example, net income). If a reporter has more than one data record in the same cell, these records are aggregated and suppression is done at the reporter level.

In addition to applying the general rule, several special rules may be applied covering rounded estimates, country and industry aggregates, and "key item" suppression (looking at a set of related items as a group and suppressing all items if the key item is suppressed).

Complementary suppression is done partly by computer and partly by human intervention. The computer programs used include routines that examine different combinations of cells to ensure that suppressions cannot be uncovered through the computation of linear combinations of rows and columns.

Some tables are published on numbers of companies, such as the number of foreign affiliates of U.S. companies in different countries or industries. These number counts are not considered sensitive and are not analyzed for disclosure or suppressed.

Under the International Investment and Trade in Services Survey Act, limited sharing of data with other Federal agencies, and with consultants and contractors of BEA, is permitted, but only for statistical purposes and only to perform specific functions under the Act. Included among these are "Special Sworn Employees", who are allowed on-site access to company-level

microdata for research purposes and who are sworn to uphold the confidentiality of the data on the same basis as regular BEA employees. Certain types of data sharing with other Federal agencies also are authorized by the Foreign Direct Investment and International Financial Data Improvements Act of 1990 and by the Confidential Information Protection and Statistical Efficiency Act of 2002. This data sharing is for statistical purposes only, and any staff of these agencies who must view BEA's unsuppressed data in connection with these activities are required to obtain BEA Special Sworn Employee status.

In another program area, BEA's Regional Economic Measurement Division publishes estimates of local area personal income by major source, based on county-level data on wages and salaries that it obtains from the Federal/state ES-202 Program of the Bureau of Labor Statistics (BLS). BEA is required to follow statistical disclosure limitation rules that satisfy BLS requirements. To prevent either the direct or the indirect disclosure of the confidential information, BEA uses the BLS state and county nondisclosure file to protect the confidential information in the ES-202 data that has been supplied to BEA. The nondisclosure file identifies the sensitive cells that must be protected to avoid release of confidential information.

BEA uses as many BLS nondisclosure cells as possible, but cannot use some of them for various reasons. The most important reasons are that the industry or geographic structure published by BEA does not exactly match the industry or geographic detail provided by BLS and that BEA does not use ES–202 data for the farm sector. For these cases, BEA must select additional cells to prevent the disclosure of confidential information. In order to determine which estimates should be suppressed, the total wages and salaries file and the wages–and–salaries–nondisclosure file are used to prepare a multidimensional matrix. This matrix is tested, and the estimates that should be suppressed are selected. Complementary suppressions, if necessary, are generated by computer and checked to ensure that they are adequate.

A.2.b. Bureau of the Census (BOC)

The Census Bureau conducts its statistical programs under government-wide legislation such as the Privacy Act, the Freedom of Information Act (FOIA), and the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002; and agency-specific legislation such as Title 13, United States Code, of 1954.

Title 13, U.S.C, defines the basis for the Census Bureau standards for confidentiality. Data that identify individuals, businesses, and other organizations must not be shared with anyone unless that person has taken an oath to maintain Census confidentiality and has a business need to know. The Census Bureau protects confidential data through the use of technological safeguards, statistical data protection, and through restricted access. Methods used include encryption software, special dedicated lines, as well as password and firewall techniques.

The Census Bureau has legislative authority to conduct surveys for other agencies under either Title 13 or Title 15 U.S.C. A sponsoring agency with a reimbursable agreement under Title 13 can use samples and sampling frames developed for the various Title 13 surveys and censuses. This would save the sponsor the extra expense that might be incurred if it had to develop its own

sampling frame. However, the data released to an agency that sponsors a reimbursable survey under Title 13 are subject to the confidentiality provisions of any Census Bureau public-use microdata file or tables; for example, the Census Bureau will not release either identifiable microdata or small area data. The situation under Title 15 is quite different. In conducting surveys under Title 15, the Census Bureau may release identifiable information, as well as small area data, to sponsors. However, sources other than surveys and censuses covered by Title 13 must be used to draw the samples. When the sponsoring agency furnishes the frame, the data are collected under Title 15, and the sponsoring agency's confidentiality rules apply.

A Disclosure Review Board (DRB) reviews specifications and proposals relating to each Title 13 data release intended for public use. The DRB ensures adherence to guidelines of the "Census Bureau DRB checklist" and any other criteria previously established by the DRB. It communicates disclosure limitation policy to program managers, Census Bureau officials, data users, prospective sponsors and the general public. The DRB initiates and coordinates research on the disclosure potential in microdata, tabular data, and other statistical outputs; and on the effectiveness of disclosure avoidance techniques as applied to such outputs. Members of the Disclosure Avoidance Research Group in the Statistical Research Division conduct research into the most suitable data protection methods for the materials published.

Some mechanisms exist to provide access to more detailed information on a restricted basis. These include Research Data Centers for approved researchers with Special Sworn Status, as well as remote on-line access in State Data Centers and Census Information Centers via the Advanced Query System for user-defined tables from Census 2000. The latter system allows users to request certain types of tables and then automatically reviews the tables to avoid disclosing confidential information. Users receive only the tables that have passed disclosure review.

Some microdata are accessible to approved researchers at the Census Bureau's Research Data Centers (RDCs). The objective of the Center for Economic Studies (CES) and the RDCs is to increase the utility and quality of Census Bureau data products. Use of microdata can address important policy questions without the need for additional data collections. In addition, it is the best means by which the Census Bureau can check on the quality of the data it collects, edits, and tabulates. These secure research facilities are located at various sites across the country. Access is strictly limited to researchers and staff authorized by the Bureau of the Census. All analysis must be performed within the secure RDC research facility. Ensuring security at RDCs has several aspects: project oversight, a physically secure facility, personnel security, a secure computing environment, an on-site Census employee, and application of disclosure avoidance rules to the analytical results presented to the public.

For the every-fifth-year economic census and associated surveys, the Census Bureau uses the p% rule to identify sensitive cells in tables but does not publish the value of p. Sensitive cells are suppressed and complementary suppressions are identified using the technique of network flow (which may be viewed as a special case of linear programming) which is computationally very fast, or linear programming which is slower. Network flow is ideal for 2-dimensional tables. It has also been applied to 3D tables although for such tables, linear programming is the preferred

method from a theoretical point of view; i.e. full protection of sensitive cells is guaranteed, obviating the need to run a disclosure audit program to check the extent of protection achieved.

For the 2002 Economic Census, network flow was used for all 2-dimensional tables and the larger 3-dimensional tables. Suppression programs based on linear programming were used for smaller 3-dimensional tables. Certain surveys have 4-dimensional or 5-dimensional data, and linear programming based programs may be used for these tables if runtimes are not excessive. Auditing programs are used when necessary.

For non-census demographic data, the Census Bureau primarily uses a combination of geographic thresholds, population thresholds and coarsening. Microdata cannot show geography below a population of 100,000. For the most detailed microdata, that threshold is raised to 250,000 or higher. Some surveys tabulate only at state, region or Census division. For data products that fall outside the main publications, a threshold may be applied at the cell level or to the population. Multi-dimensional tabular data on specific populations must met a minimum of unweighted cases, usually 50. The cell threshold minimum most frequently used is 3 unweighted individuals from 3 distinct households. Coarsening is used to avoid the application of thresholds. For small populations or rare characteristics noise may be added to identifying variables, data may be swapped, or an imputation applied to the characteristic. Census data, which lacks the component of protection provided by sampling, employs targeted swapping in addition to the combination of table design and thresholds described above.

Most of the Census Bureau's current statistical disclosure limitation practices and research are summarized in three papers Zayatz (2002), Zayatz, Massell, and Steel (1999) Hawala, Zayatz, and Rowland (2004). Other references are found in these three papers.

A.3. Department of Education: National Center for Education Statistics (NCES)

The National Center for Education Statistics (NCES) has strong legislation that requires the agency to protect the confidentiality of its data collections. First under the 1988 Hawkins-Stafford Elementary and Secondary School Improvement Amendments, and then under the 1994 National Education Statistics Act, NCES was required to maintain confidentiality of all individually identifiable data about individuals (e.g., principal, teacher or student data). Although the law did not explicitly protect institutional data, protecting data about individuals within institutions frequently resulted in the protection of data about educational institutions as well. The Education Sciences Reform Act of 2002 explicitly requires NCES to protect the confidentiality of all individually identifiable data about students, their families and their schools. Related to these laws, NCES has a statistical standard on maintaining confidentiality (NCES Statistical Standard 4-2 http://nces.ed.gov/statprog/2002/std4_2.asp). That standard summarizes the relevant laws, identifies employee and contractor responsibilities when handling confidential data, describes alternative methods that may be used to protect NCES data from disclosure, and includes the consent notice to be placed on NCES public use data files. In addition, the NCES Disclosure Review Board (DRB) reviews disclosure analysis plans and proposed public-use data releases to protect the confidentiality of the individual reported values.

Most NCES data collections include some institution data, but additionally include data from any combination of institution heads, teachers, librarians, students or student's parents. It's the individual's data that must be protected. These datasets can be made publicly available through either a public-use file or a data analysis system (DAS) after applying a DRB approved disclosure analysis and resolving any observed disclosure risks. This process is described below.

A public-use file is a file or series of linked files that: 1) contain individuals' responses about themselves, and 2) have gone through a DRB approved disclosure analysis. All direct individually identifiable information (e.g., school name, individual name, addresses) is stripped from the public-use file. Continuous variables are top and bottom coded to protect against identification of outliers. After this has be done, the only way a casual data intruder can identify an individual respondent is by first identifying the sampled institution for the individual.

To prevent identification of the sampled institution, all known publicly available lists of education institutions that contain institutions' names and addresses are gathered. Each list is matched with the sample file using all common variables between the two files. If an institution can be identified to within 2 other institutions, using an appropriate distance measure, then that is a disclosure risk and must be resolved before releasing the data.

If too many disclosure risks are obtained then a common variable(s) may be dropped from the public-use file, or the variable(s) may be coarsened. If there are only a few identified disclosure risks found then the appropriate action is to selectively perturb a set of the common variables until all disclosure risks are resolved. This analysis is repeated sequentially for each list file until it can be repeated for each list file without identifying any disclosure risks.

The matching analysis described above is designed to prevent the casual data snooper from determining survey respondents. It is assumed that if the institution cannot be identified then individuals within that institution also cannot be identified. However, data intruders with detailed knowledge about a sampled institution may be able to identify an institution; thereby, increasing the likelihood of identifying an individual. To reduce the likelihood of correctly doing this, additional disclosure edits are required.

Whenever institution head, teacher, student, or parent data are clustered, a subsampling of respondents is required. Data from respondents selected in this sub-sample, are reviewed using an additional disclosure edit. The edit is either: 1) a blanking and imputing, or data swapping of a sampling of sensitive items collected; or 2) a data swapping of the key identification variable of the respondent or institution. The amount of editing is set at a level high enough to protect the confidentiality of the respondent, while not compromising the analytic usefulness of the data file.

The important aspect of this edit is that all respondents have a chance of selection. Usually respondents at greater risk are given a larger selection probability. Should someone think that they have identified a respondent, they cannot be sure that the data is really for that respondent.

Another way NCES distributes data is through a Disclosure Avoidance System (DAS). A DAS is a table generator program that can generate proportions, means, or correlation coefficients with the corresponding standard errors that have been calculated taking into account the complex

sampling procedures used in the NCES surveys. The DAS is linked to a data file, but all data elements are masked so that the file itself is unreadable to anything or anyone other than the table generator program. The data are also protected through the survey sampling process (i.e., any unit selected is likely to have many other similar units in the universe). However, since there is little control on the type and number of tables generated, further disclosure protections are applied through data perturbation (e.g., data swapping) and data coarsening.

In order for a DAS to be released, the underlying data file must include a series of DRB confidentiality edits: either a blanking and imputing, or data swapping of a sampling of sensitive items collected; or a data swapping of the key identification variable of the respondent or institution.

All NCES tables use either a perturbation technique (i.e. a confidentiality edit approach), or a process of collapsing cells until all cells contain values associated with at least three respondents. The confidentiality edit approach is applied to the restricted-use microdata file. The table can then be prepared with no additional disclosure limitation method applied.

A.4. Department of Energy: Energy Information Administration (EIA)

EIA has established statistical standards (http://www.eia.doe.gov/smg/Standard.pdf) including standards for data protection, accessibility, and nondisclosure. Standard 2002-22, "Nondisclosure of Company Identifiable Data in Aggregate Cells," contains the procedures and policies to ensure that sensitive data cell values are suppressed (i.e., withheld from public release) for the protection of confidential survey data. EIA also requires additional confidentiality training for those who have access to data protected under CIPSEA.

EIA's primary method for ensuring confidentiality protection is the application of the pq rule or a combination rule. Regardless of the parameters chosen, the rule assures that nonzero value data cells must be based on three or more respondents. The combination rule is the pq rule in conjunction with some other subadditive linear suppression rule. The value of the pq sensitivity parameter represents the maximum permissible gain in information when one company uses the published cell total and its own value to create better estimates of its competitors' values. The values of the pq parameter that are selected for specific surveys are not published and are considered confidential. Complementary suppression is applied to other cells to assure that the sensitive value cannot be reconstructed from published data. For information collected under a pledge of confidentiality, EIA does not publicly release names or other identifiers of survey respondents linked to their submitted data.

For many EIA surveys that use the pq rule, complementary suppressions are selected manually. One survey system that publishes complex price and volume tables for crude oil and refined petroleum products uses software to select complementary suppressions. It assures that there are at least two suppressed cells in each dimension, zero value cells are excluded as candidates for suppression, and that the cells selected are those of lesser importance to data users.

Standard 2002-22 also includes separate supplementary materials with guidelines for understanding and implementing the pq rule. Guidelines are included for situations where all

values are negative; some data are imputed; published values are net values (the difference between positive numbers); and the published values are weighted averages (such as volume weighted prices). Much of the same information is provided in Appendix A of this report.

In selected program areas, EIA does not use disclosure limitation methods on statistical data. For certain energy supply data, the number of companies providing information is relatively small and/or the distribution of energy supply companies is highly skewed with a relatively small number of large companies. Statistical data for sub-United States geographical areas (e.g., States, Petroleum Administration for Defense Districts, Refining Districts) typically include some values that are sensitive and would not be published if disclosure limitation methods were applied. If disclosure limitation methods using primary and complementary suppression were applied, the result would be a significant amount of information loss. This loss of information to data users would seriously erode the value of the information for public and private understanding and analysis of energy supply.

In these program areas, EIA uses a Federal Register notice to announce a proposed policy of not using disclosure limitation methods and requests public comments. After considering public comments, EIA decides whether to formalize its policy. If the policy is to not use such methods, EIA explains the policy at the time an information collection undergoes the Office of Management and Budget approval process and when the survey materials are provided to potential respondents at the time information is requested. The explanation states that disclosure limitation procedures are not applied to the statistical data published from that survey's information. The explanation goes on to state that there may be some resulting statistics that are based on data from fewer than three respondents, or that are dominated by data from one or two large respondents. In these cases, it may be possible for a knowledgeable person to estimate the information reported by a specific respondent.

EIA does not have a standard to address tables of frequency data. However, there are only two primary publications of frequency data in EIA tables. Those publications are the Household Characteristics publication of the Residential Energy Consumption Survey (RECS) and the Building Characteristics publication of the Commercial Building Energy Consumption Survey (CBECS). In both publications, cells are suppressed for accuracy reasons, not for disclosure reasons. For the first publication, cell values are suppressed if there are fewer than 10 respondents or the Relative Standard Errors (RSE's) are 50 percent or greater. For the second publication, cell values are suppressed if there are fewer than 20 respondents or the RSE's are 50 percent or greater. No complementary suppression is used.

EIA does not have a standard for statistical disclosure limitation techniques for microdata files. The only microdata files for confidential data released by EIA are for RECS and CBECS. In these files, various standard statistical disclosure limitation procedures are used to protect the confidentiality of data for individual households and buildings. These procedures include: eliminating identifiers, limiting geographic detail, omitting or collapsing data items, top-coding, bottom-coding, interval-coding, rounding, substituting weighted average numbers (blurring), and introducing noise through a data adjustment method which randomly adjusts respondent level data within a controlled maximum percentage level around the actual published estimate. After applying the randomized adjustment method to the data, the mean values for broad population

groups based on the adjusted data are the same as the mean values generated from the unadjusted data.

A.5. Department of Health and Human Services

A.5.a. Agency for Healthcare Research & Quality (AHRQ)

The disclosure limitation procedures used by AHRQ are similar to those of NCHS. The Medical Expenditure Panel Survey (MEPS) conducted by AHRQ utilizes the National Health Interview Survey as its sampling frame. Therefore, the disclosure limitation procedures used by AHRQ for MEPS public use data files follow the procedures used by NCHS for the MEPS. All public use data file releases are required to be reviewed and approved by the NCHS Disclosure Review Board before they are released. AHRQ also reviews and cross clears release of public use files from the NHIS.

AHRQ has established an on-site data center within the Center for Financing, Access, and Cost Trends (CFACT) to facilitate researcher access to selected non-public use MEPS data.

The CFACT Data Center is a physical space at AHRQ located in Rockville, Maryland where researchers, with approved projects are allowed access to data files not available for public dissemination. These data are classified as "restricted" and contain information that are not released to the public. These data sets may contain geographic variables at a lower level than released for public use, more detailed condition information, or may consist of unedited data base segments not yet prepared for public release. These restricted data sets do not contain information that directly identifies a respondent (name, social security number, street address).

Researchers are allowed access only to the information required to complete their project. No researcher can remove any materials from Data Center until the materials have been reviewed by specific CFACT staff for disclosure avoidance. Only summary output (tables, equations) may be removed from the Data Center. No microdata files are permitted to be removed from the Data Center.

All materials to be removed from the data center are subject to disclosure review. CFACT staff is responsible for insuring the confidentiality of data being used in the data center. In the case of onsite users, CFACT staff reviews output or tables prior to the material leaving the Data Center. In the case of researchers using the Data Center remotely, CFACT staff will conduct a disclosure review of material before forwarding output to the researcher. The development of formal criteria for review of tabular materials is an ongoing process.

For users, the Manager of the CFACT Data Center is the point of contact for arbitration of confidentiality review. Every attempt will be made to work with the researcher to develop specifications for tabulations that will "pass" a confidentiality review. Projects with continuing confidentiality issues will be discussed with CFACT senior staff before a final decision is rendered.

Any output that could potentially identify respondents or small geographic areas, either directly or inferentially cannot be removed from the data center. Tables with geographic areas as one of the tabs (except for those identified on public use files) cannot be removed, nor can tables containing cells with less than 100 observations. Data Center Users are never given access to files with direct identifiers such as name or address. Users may be given access to files with dummy codes for places. However, since data center users have no need to discern the identity of the places, they will not be given the key that would allow the association of a place name with the code. Upon request the entire file can be pre-coded into categories (i.e. residing in a state with high/middle/low Medicaid generosity). Models using geographic area as the dependent variable cannot be removed from the Data Center. The identity of sampling units, which could assist in the identity of the data subject, cannot be removed. In general, any direct or inferential identities not revealed on public use data files cannot be removed from the Data Center.

A.5.b. National Center for Health Statistics (NCHS)

NCHS is the principal federal agency that releases health statistics. It is part of the Department of Health and Human Services Centers for Disease Control and Prevention (CDC). CDC's NCHS statistical disclosure limitation techniques are presented in the NCHS Staff Manual on Confidentiality (September, 2004), Section 9 "Avoiding Inadvertent Disclosures Through Release of Microdata " and Section 10 "Avoiding Inadvertent Disclosures in Tabular Data". No magnitude data figures should be based on fewer than five cases and an (n, k) rule is used. Commenting on an earlier edition of the NCHS Manual, Jabine (1993b) states that "the guidelines allow analysts to take into account the sensitivity and the external availability of the data to be published, as well as the effects of nonresponse and response errors and small sampling fractions in making it more difficult to identify individuals." In almost all survey reports, no low level geographic data are shown, substantially reducing the chance of inadvertent disclosure. The NCHS staff manual states that for tables of frequency data a) "in no table should all cases of any line or column be found in a single cell"; and b) "in no case should the total figure for a line or column of a cross-tabulation be less than 5". One acceptable way to solve the problem (for either tables of frequency data or tables of magnitude data) is to combine rows or columns, or to use cell suppression (plus complementary suppression). Other approaches are in development.

It is NCHS policy to make microdata files available to the scientific community so that additional analyses can be made for the country's benefit. Such files are reviewed for approval by the NCHS Disclosure Review Board following guidance and principles contained in the Staff Manual and the NCHS Checklist for the Release of Micro Data Files. These guidelines require that detailed information that could be used to identify individuals (for example, date of birth) should not be included in microdata files. The identities of geographic places and characteristics of areas with less than 100,000 people are never to be identified and it may be necessary to set this minimum at a higher number if research or other considerations so indicate. Information on the drawing of the sample that could identify data subjects should not be included.

All new microdata sets must be reviewed for confidentiality issues and approved for release by the NCHS Confidentiality Officer who consults with the NCHS Disclosure Review Board in making agency decisions.

Upon successful application to the NCHS Research Data Center, researchers may be provided access to special files that do not permit the identification of individual respondents. This may take place on site at NCHS offices or remotely over secure electronic lines. While information concerning named geographic entities cannot be accessed, data ordered by such units can be analyzed at a level not possible with public use data.

Prospective researchers must submit a research proposal that is reviewed and approved by a committee whose judgment is based upon the availability of RDC resources, consistent with the mission of NCHS, general scientific soundness, and the feasibility of the project. Although researchers sign confidentiality agreements, strict confidentiality protocols require that researchers with approved projects complete their work using the facilities located within the RDC. Researchers can supply their own data to be merged with NCHS data sets. Completed by the RDC staff, the merged files are only available to the originating researcher unless written permission is given to allow access to others. Further details on NCHS' Research Data Center are available at http://www.cdc.gov/nchs/r&d/rdc.htm.

Areas under current investigation include software for balancing data quality and statistical disclosure limitation (SDL) in tabular data and enhanced procedures for SDL and disclosure risk assessment in microdata.

A.6. Department of Justice: Bureau of Justice Statistics (BJS)

The same requirements under Title 13 of the U.S.C. that cover the Census Bureau are followed by BJS for those data collected for BJS by the Census Bureau. For tabular data, cells with fewer than 10 observations are not displayed in published tables. Published tables may further limit identifiability by presenting quantifiable classification variables (such as age and years of education) in aggregated ranges. Cell and marginal entries may also be restricted to rates, percentages, and weighted counts. Standards for microdata protection are incorporated in BJS enabling legislation. Individual identifiers are routinely stripped from all microdata files before they are released for public use.

A.7. Department of Labor: Bureau of Labor Statistics (BLS)

Commissioner's Order 3-04, "The Confidential Nature of BLS Records," dated October 4, 2004, contains the BLS' policy on the confidential data it collects. One of the requirements is that:

> "Publications shall be prepared in such a way that they will not reveal the identity of any specific respondent and, to the knowledge of the preparer, will not allow information concerning the respondent to be reasonably inferred by either direct or indirect means."

A subsequent provision allows for exceptions under conditions of informed consent and requires prior authorization of the Commissioner before such an informed consent provision is used.

The statistical methods used to limit disclosure vary by program. For tables, the most commonly used procedure has two steps--the threshold rule, followed by a concentration rule. BLS programs use the p percent rule or the (n, k) rule to assess concentration depending upon program. The value of the parameters used for thresholds and various concentration rules used by BLS is not released to the public. Current practice at BLS is to replace use of the (n, k) concentration rule by the p percent rule.

For example, the Quarterly Census of Employment and Wages (QCEW), a census of monthly employment and quarterly wage information from Unemployment Insurance filings, uses a threshold rule and the p percent rule for calendar year (CY) 2002 data and beyond. Prior to CY 2002, QCEW used a threshold rule and a concentration rule of (n, k). In a few cases, a two-step rule is used--an (n, k) rule for a single establishment is followed by an (n, k) rule for two establishments. The Survey of Occupational Injuries and Illnesses is using a threshold rule and the p percent rule for the CY 2003 data replacing the threshold rule used in conjunction with a concentration rule of (n, k).

The National Compensation Survey uses an approach that combines two threshold rules and an (n, k) rule. The threshold rules require that each estimate be comprised of establishments from at least m companies (unweighted) and that there are at least t distinct occupational selections (unweighted). It also uses an (n, k) concentration rule, which requires that the weighted employment among all establishments contributing to the estimate that are part of n companies cannot exceed k percent of the weighted employment of all establishments contributing to the estimate.

The Consumer Price Index Program uses a combination of a threshold rule and a minimum number of quotes from distinct sample units. The Producer Price Index uses a threshold rule on units and quotes in conjunction with the (n, k) rule.

BLS releases very few public-use microdata files. Most of these microdata files contain data collected by the Bureau of the Census under an interagency agreement and Census' Title 13 authority. For these surveys (Current Population Survey, Consumer Expenditure Survey, and four of the five surveys in the family of National Longitudinal Surveys) the Bureau of the Census determines the statistical disclosure limitation procedures that are used. Disclosure limitation methods used for the public-use microdata files containing data from the National Longitudinal Survey of Youth, collected under contract by Ohio State University and the National Opinion Research Center at the University of Chicago, are similar to those used by the Bureau of the Census.

The Bureau of Labor Statistics (BLS) has opportunities available on a limited basis for researchers from colleges and universities, government, and eligible nonprofit organizations to obtain access to confidential BLS data files for exclusively statistical purposes. These data files are derived from BLS surveys and administrative databases for which no public-use version is available. These confidential BLS data are available for research that is exclusively statistical,

with appropriate controls to protect the data from unauthorized disclosure. BLS confidential data files are available for use only at the BLS National Office in Washington, D.C., on statistical research projects approved by the BLS. Researchers granted access to the confidential data sign agreements stating that they are responsible for adhering to the confidentiality policies of the BLS.

The BLS considers applications for research proposals four times a year. Research proposals should be between 5 and 10 pages and should contain detailed information about the research project, including a literature review and an indication of how the proposed research contributes to the literature, the hypotheses to be tested, the data set and variables to be used in the analysis, the empirical methods to be used, and the specific data outputs that will result from the project.

A.8. Department of the Transportation: Bureau of Transportation Statistics (BTS)

The Bureau of Transportation Statistics (BTS) collects transportation-related data. BTS' confidentiality statutes and a set of comprehensive confidentiality procedures protect these data. The BTS *Confidentiality Procedures Manual* documents the confidentiality procedures for the agency.

BTS' confidentiality officer (CO) is responsible for the day-to-day operations of the confidentiality program. The CO also chairs the BTS' disclosure review board (DRB), which is responsible for reviewing microdata, tabular data and other information products for disclosure risks prior to public release. BTS staff and contractors are required to have annual confidentiality training, and to sign non-disclosure agreements when they enter or leave service with BTS.

BTS confidentiality program objectives guide the data review process for whether disclosure limitation methods should be applied. These objectives seek to:

- Protect confidential data while increasing access to data,
- Apply statistical disclosure limitation (SDL) methods on a case-by-case basis, and
- Take into account data user opinions on applications of SDL methods.

For most microdata and tabular data products, BTS program managers are required to complete a checklist identifying potential disclosure risks and outline any steps taken to mitigate such risk. The BTS' DRB reviews the data product and checklist and makes a final determination on disclosure risk. The DRB can recommend application of SDL methods prior to public dissemination.

BTS uses various microdata SDL methods based on the disclosure review findings and the unique characteristics of the data files. Some SDL procedures used include data suppression and modification. Data modification includes recoding continuous variables into categorical variables, collapsing categories, top and bottom coding, introduction of noise, and data swapping. BTS program managers must also identify any external data that could be matched to BTS datasets and take steps to minimize the ability to match.

The DRB conducts disclosure review of tabular data products when they are developed from microdata files that are not released to the public. BTS also uses tabular data SDL methods based on the disclosure review findings and on the characteristics of the tables.

A.9. Department of the Treasury: Internal Revenue Service, Statistics of Income Division (IRS, SOI)

The Statistics of Income (SOI) function within the larger organization Research, Analysis, and Statistics (RAS) is to establish and implement IRS guidance rules for the public release of tax data in tables and public-use microdata files. This role is primarily necessitated by sections 6108(c) and 6103j(4) of the Internal Revenue Code (IRC), which require that the data in statistical publications produced by IRS and authorized recipient agencies be anonymous.

The administrative rules are found in Chapter VI of the SOI Division Operating Manual (January, 1985), and require that at or above the state level each cell in a publicly released tabulation be based on at least three observations. Below the state level the requirement is at least ten observations. Data cells not meeting these thresholds are suppressed or combined with other cells. Combined or deleted data are included in the corresponding column totals. These rules also apply for secondary disclosure in which taxpayer identities might be revealed by subtraction of associated cells within a table or between tables, and even indirectly through similar data in other publications.

SOI documents disclosure procedures in its own publications. For example, disclosure limitations are discussed in "SOI Sampling Methodology and Data Limitations" in the Appendix to the quarterly SOI Bulletins and online at http://www.irs.gov/taxstats.

SOI produces one annual public-use microdata file, known as the SOI "tax model", containing a sample of data based on the Form 1040 series of individual tax returns. The disclosure protection procedures applied to this file include: (1) subsampling certainty records at a 33% rate; (2) removing certain records having extreme values; (3) suppressing certain fields from all records and geographical fields from high income records; (4) top coding and modifying some fields; (5) blurring some fields of high income records by locally averaging across records; and (6) rounding amount fields to four significant digits. To help ensure that taxpayer privacy is protected in the SOI tax model file, SOI has periodically contracted with experts who employ so-called "professional intruder" techniques to both verify that confidentiality is protected and to inform techniques to be applied to future releases of the SOI tax model file. For additional details on the disclosure avoidance techniques used to produce SOI public-use files see: Sailer, P., Weber, M. and Wong, W., (2001);

In addition to its own role in producing tax statistics, SOI is also responsible for coordinating the provision of tax data for statistical purposes to authorized recipients under section 6103j of the IRC. This function includes ensuring that authorized recipients of tax data also follow the rules of 3/10 described above or an equivalent methodology approved by SOI, as stipulated in the IRS Publication 1075, *Tax Information Security Guidelines for Federal, State, and Local Agencies (June 2000)*. Because of the considerable onus this requirement can entail for both SOI and agencies using alternative disclosure protection methodologies, recent efforts have begun to

establish inter-agency agreements with experienced users, such as the US Census Bureau, in which responsibility for alternative tabular protection methodologies is accepted by the recipient agency. The IRS-Census agreement for this purpose was effective June 2, 2003. Because the challenges of protecting public-use microdata files are considered unique and such data are deemed more sensitive to disclosure risk, public-use microdata files are excluded. That is, under these agreements, IRS approval would still be needed before an outside agency could release a public-use microdata file based on tax data.

Currently, the IRS Office of Research within RAS is working with Census to ensure that all data in a proposed Census public-use file based on tax data [earnings] linked to Census' Survey of Income and Program Participation (SIPP) will be anonymous. The proposed SIPP/earnings public-use file methodology is exploring using "synthetic data" to produce public-use files tailored for particular users, as opposed to a "one size fits all" approach.

A.10. National Science Foundation (NSF)

The National Science Foundation (NSF), Division of Science Resources Statistics (SRS), balances the requirement to guard the confidentiality of its respondents against the desire of the research community to access data collected using taxpayer dollars. NSF applies either the (n, k) dominance rule or p-percent rule, or sometimes both rules in conjunction with each other depending upon the survey. When it is possible to create a microdata file that is useful to a broad group of researchers while protecting respondent confidentiality, SRS releases public use data files consistent with these dual objectives. When releasing public-use microdata files, individual identifiers are removed from all records and other high risk variables that contain distinguishing characteristics are modified to prevent identification of survey respondents and their responses. Top-codes and bottom-codes are employed for numeric fields to avoid showing extreme field values on a data record. Values beyond the top-code or bottom-code are replaced either by the average of the values in excess of the respective top-code or bottom-code or through the application of various imputation methodologies.

When the researcher demonstrates that available SRS public use data files do not meet research needs and in keeping with SRS's mission to help provide the statistical information about the US science and engineering enterprise, it is sometimes possible to accommodate the request by providing access to restricted data files. One method for access is a recently created on-site secure analysis area for visiting researchers. Another method of access is off-site licensing.

Under the Office of the Director, SRS, the Chief Statistician coordinates a restricted-use data-licensing program. To acquire restricted-use files, the researcher and the researcher's institution indicates their knowledge of confidentiality issues and willingness to ensure protection of the data by completing a formal legal contract, the license agreement, that details the use of the data, promises to prevent disclosure of confidential data, agrees to a prepublication review by SRS, and stipulates the return of the data to SRS upon expiration of the license. Research conducted by licensees often is found in scientific journals as well as highly cited in policy forums.

A.11. Social Security Administration (SSA)

The Office of Research, Evaluation, and Statistics (ORES), the statistical office of the Social Security Administration, reviews and establishes methodology and procedures for protecting the confidentiality of data. For the release of statistical tables, ORES uses a strategy combining both suppression and rounding to prevent the release of identifiable information.

Statistical tables for Social Security beneficiaries and benefits consist of frequency counts for beneficiaries and summary benefit amounts. Detailed beneficiary information is suppressed when the marginal total is less than a cut-off value and only the marginal value is shown. For the rows in which only the marginal counts are shown, dollar amounts are suppressed when the number of cases contributing to the total is less than a cutoff. Detailed frequency counts are suppressed when all details for a marginal total are in a single category. When suppressions are introduced to prevent disclosure in an individual cell, complementary suppressions are employed to prevent the inference of a suppressed value. Controlled rounding is also used as a disclosure avoidance method in statistical tables for frequency counts.

Publications that include earnings and employment information conform to IRS rules when presenting tables (See section A.9 of this chapter). In particular, table cells with fewer than 3 persons at the state level and 10 persons at the county level are suppressed and the corresponding summary income is also not shown. Whenever data cells are suppressed, complementary suppressions are introduced to prevent inferring a suppressed value. All dollar amounts are shown in thousands of dollars. Earnings and employment statistics are derived from a sample of IRS records rather than a 100-percent file of earnings and employment information.

When releasing public-use microdata files, individual identifiers are removed from all records and other distinguishing characteristics are modified to prevent identification of persons to whom a record pertains. Records are sequenced in random order to avoid revealing information due to the ordering of records on the file. Top-codes and bottom-codes are employed for numeric fields to avoid showing extreme field values on a data record. Values beyond the top-code or bottom-code are replaced by the average of the values in excess of the respective top-code or bottom-code. Top-code and bottom-code values are derived at the national level and the replacement values are derived and applied at the state level when appropriate. Values shown for some categorical fields are combined into broader groupings than those present on the internal file and dollar amounts are rounded. Top-code and bottom-code values, replacement values, and related information are provided to users as part of file documentation.

A Disclosure Review Board (DRB) reviews proposed public-use microdata files prior to their release. The DRB consists of staff from ORES who are familiar with the underlying data files, their uses, and confidentiality requirements. In addition, confidentiality specialists from other federal agencies may serve on the DRB to provide further perspective and additional confidentiality expertise. Staff who are responsible for file creation complete the *Checklist on Disclosure Potential of Proposed Data Releases,* prepared by the Interagency Confidentiality and Data Access Committee, and the Checklist is included in the DRB review.

B. Summary

Most of the 14 agencies covered in this chapter have standards, guidelines, or formal review mechanisms that are designed to ensure that adequate disclosure analyses are performed and appropriate statistical disclosure limitation techniques are applied prior to release of tabulations and microdata. The agency standards and guidelines exhibit a wide range of specificity: Some contain only one or two simple rules while others are much more detailed. Some agencies publish the parameter values they use, while others feel withholding the values provides additional protection to the data. Obviously, there is great diversity in policies, procedures, and practices among Federal agencies to appropriately protect the wide variations in the content and format of information released.

B.1. Magnitude and Frequency Data

Most standards or guidelines provide for minimum cell sizes and some type of concentration rule. Some agencies (for example, ERS, NASS, and NCHS) publish the values of the parameters they use in (n, k) concentration rules, whereas others, such as Census and BLS, do not. Minimum cell sizes of 3 are routinely used, because each member of a cell of size 2 could derive a specific value for the other member. Some agencies cited accuracy standards as guidelines for releasing certain tabular data. **Accuracy standards** refer to specific rules that an agency applies to the data that relate to some measure of data quality such as a threshold level for relative standard error or coefficient of variation estimates.

Most of the agencies that published their parameter values for concentration rules used a single set, with n = 1. Values of k ranged from 0.5 to 0.8. The most elaborate rule included in standards or guidelines were EIA's pq rule and BEA's and Census Bureau's related p-percent rules. All these rules have the property of subadditivity. The p percent and pq rule give the disclosure analyst flexibility to specify how much gain in information about its competitors by an individual company is acceptable.

One possible method for dealing with data cells that are dominated by one or two large respondents is to ask those respondents for permission to publish the cells, even though the cell would be suppressed or masked under the agency's normal statistical disclosure limitation procedures. Agencies including NASS, EIA, the Census Bureau, and some of the state agencies that cooperate with BLS in its Federal-state statistical programs, use this type of procedure for some surveys to allow publication of those sensitive cell values. Another disclosure limitation method used by two agencies is to apply noise to the underlying micro data before aggregating the reported values.

B.2. Microdata

The agencies that release public use microdata files have established statistical disclosure limitation procedures for releasing microdata. Some agencies noted that the disclosure limitation procedures for surveys they sponsored were set by the Census Bureau's Disclosure Review Board, because the surveys had been conducted for them under the Census Bureau's authority (Title 13). Major releasers of public-use microdata--Census, NCHS and NCES--have all

established formal procedures through Disclosure Review Boards for review and approval of new microdata sets. As Jabine (1993b) wrote, "In general these procedures do not rely on parameter-driven rules like those used for tabulations. Instead, they require judgments by reviewers that take into account factors such as: the availability of external files with comparable data, the resources that might be needed by an `attacker' to identify individual units, the sensitivity of individual data items, the expected number of unique records in the file, the proportion of the study population included in the sample, the expected amount of error in the data, and the age of the data."

Geography is an important factor. Census and NCHS specify that no geographic codes for areas with a sampling frame of less than 100,000 persons can be included in public-use data sets. If a file contains large numbers of variables, a higher cutoff may be used. The inclusion of local area characteristics, such as the mean income, population density and percent minority population of a census tract, is also limited by this requirement because if enough variables of this type are included, the local area can be uniquely identified. An interesting example of this latter problem was provided by EIA's Residential Energy Consumption Surveys, where the local weather information included in the microdata sets had to be masked to prevent disclosure of the geographic location of households included in the survey.

Top-coding is commonly used to prevent disclosure of individuals or other units with extreme values in a distribution. Dollar cutoffs are established for items like income and assets and exact values are not given for units exceeding these cutoffs. Blurring, swapping, blank and impute, noise introduction, recoding, threshold rules, and rounding are other methods commonly used to prevent disclosure.

Summary of Agency Practices

Agency	Magnitude Data	Frequency Data	Microdata	Waivers	Restricted Access Allowed for Researchers
ERS	(n, k), (1,.6) 3+	Threshold Rule 3+	No	Yes	Yes
NASS	(n, k), p-percent Parameters Confidential	1+ Not Sensitive for Est. Surveys	No	Yes	Yes
BEA	p-percent c=1	1+ Not Sensitive for Est. Surveys	No	No	Yes
CENSUS	p-percent Parameters Confidential Noise addition	Data Swapping, Access Query System rules, Threshold Rule	Yes -- Disclosure Review Board	Yes	Yes
NCES	Data Swapping Data Coarsening Accuracy	Data Swapping Data Coarsening Accuracy	Yes – Disclosure Review	No	Yes

Agency	Magnitude Data	Frequency Data	Microdata	Waivers	Restricted Access Allowed for Researchers
	Standards/Threshold Rule 3+	Standards/Threshold Rule 3+	Board		
EIA	(n, k), pq, Parameters Confidential	Threshold Rule Accuracy Standards	Yes – Office Review	Yes	No
NCHS	(n, k), (1,.6)	Threshold Rule 4+	Yes – Disclosure Review Board	No	Yes
AHRQ	N/A	Threshold Rule 4+	Yes – Disclosure Review Board	Yes – Disclosure Review Board	Yes
SSA	Threshold Rule 3+	Threshold Rule, 5+ Marginals, 3+ cells	Yes - Agency Review	No	No
BJS	N/A	Threshold Rule 10+, Accuracy Standards	Yes - Legislatively Controlled Agency Review	No	No
BLS	(n, k), p% rule, Parameters vary by survey and data element	Minimum Number varies by survey	BOC Collects Title 13	Yes	Yes
IRS	Threshold Rule 3+	Threshold Rule 3+	Yes - Legislatively Controlled	No	No
BTS	Varies by data	Threshold Rule 3+	Yes – Disclosure Review Board	No	No
NSF	(n, k) and/or p as appropriate	Varies by risk	Yes – Meet or exceed Census public use products which are merged	Yes	Yes

Notes: Details of specific methodologies being used are shown in this table and discussed in the text to the extent they were included in the individual agencies' responses. Rules shown in the various table cells (p-percent, (n, k), for example) are explained in the text.

The following page contains a brief explanation of the key terms used in the table.

The Threshold Rule: With the threshold rule, a cell in a table of frequencies is defined to be **sensitive** if the number of respondents is less than some specified number. Some agencies require at least 5 respondents in a cell, others require 3. Sometimes, the threshold rule is applied to the universe of a table. For example, a minimum size may be needed to publish values in all cells of a table. An agency may restructure tables and combine categories or use cell suppression, random rounding, or controlled rounding. The "+" notation (3+ for example) means at least that many non-zero observations must be present for the cell to be published. (See Section II.C.3)

Data Swapping is the procedure that was used by the U.S. Census Bureau to provide protection in data tables prepared from the 2000 Census. The technique applies statistical disclosure avoidance to the microdata records before they are used to prepare tables. The adjusted microdata files are not released, they are used only to prepare tables. For both the 100 percent data file and the sample, a small sample of households were selected and matched with households in other geographic regions that had identical characteristics on a set of selected key variables. Most variables in the matched records were interchanged. This technique is called swapping. The key variables used for matching were selected to assure that Census aggregates mandated by law would be unchanged by applying this procedure. NCES recommends using data swapping and coarsening for all internal and external microdata records. If these techniques are not used, NCES prohibits the publication of any cells with fewer than three cases and prohibits the use of cell suppression. Tabulations must be reconfigured until there are no remaining cells with fewer than 3 cases

The p-Percent Rule: Approximate disclosure of magnitude data occurs if the user can estimate the reported value of some respondent too accurately. Such disclosure occurs, and the table cell is declared sensitive, if upper or lower estimates for the respondent's value are closer to the reported value than a pre-specified percentage, p. This method assumes that before data are published a user can estimate the true value to within plus or minus 100%. This rule is referred to as the "p-percent estimation equivocation level" in Statistical Policy Working Paper 2, but it is more generally referred to as the **p-percent rule**. (See Section IV.B.1.a)

The pq Rule: The pq rule is similar to the p% rule, but assumes that before data are published the general public can estimate a company's data to within q% (where q<100). Hence, an agency can specify how much prior knowledge there is by assigning a value q which represents how accurately respondents can estimate another respondent's value before any data are published ($p < q < 100$). (See Section IV.B.1.b)

The (n, k) Rule: The **(n, k) rule**, or dominance rule was described as follows in Statistical Policy Working Paper 2. "Regardless of the number of respondents in a cell, if a small number (n or fewer) of these respondents contribute a large percentage (k percent or more) of the total cell value, then the so-called **n respondent, k percent rule** of cell dominance defines this cell as

sensitive." Many people consider this to be an intuitively appealing rule, because, for example, if a cell is dominated by one respondent then the published total alone is a natural upper estimate for the largest respondent's value. (See Section IV.B.1.c)

CHAPTER IV – Methods for Tabular Data

Chapter II presented examples of disclosure limitation techniques used to protect tables and microdata. Chapter III described agency practices in disclosure limitation. This chapter presents more detail concerning methodological issues regarding confidentiality protection in tables.

As mentioned earlier, tables are classified into two categories for purposes of disclosure risk analysis: tables of frequency (or count) data and tables of magnitude data. Tables containing frequency data show the percent of the population that have certain characteristics, or equivalently, the number in the population which have certain characteristics. If a cell has only a few respondents and the characteristics are sufficiently distinctive, then it may be possible for a knowledgeable user to identify the individuals in the population. For tables of frequency data disclosure limitation methods are applied to cells with fewer than a specified **threshold number** of respondents to minimize the risk that individuals can be identified from their data. Disclosure limitation methods applied after tabulation include random rounding, controlled rounding, cell suppression, and controlled tabular adjustment. Disclosure limitation methods applied before tabulation include micordata protection techniques such as data perturbation and data swapping.

Tables of magnitude data typically present the results of surveys of organizations or establishments, where the items published are aggregates of nonnegative reported values. For such surveys the values reported by respondents may vary widely, with some extremely large values and some small values. The confidentiality problem relates to assuring that a person cannot use the published total and other publicly available data to estimate an individual respondent's value too closely. Disclosure limitation methods are applied to cells for which a **linear sensitivity measure** indicates that some respondent's data may be estimated too closely. For tables of magnitude data cell suppression is the most widely used method. Controlled tabular adjustment offers another alternative. Both methods are applied after tabulation. Disclosure limitation methods applied before tabulation include microdata protection techniques such as adding noise.

Tables of frequency data are discussed in Section A. The major methodological areas of interest are in controlled rounding and the use of microdata methods such as data swapping. Tables of magnitude data are discussed in Section B. This section provides some detail concerning linear sensitivity measures, auditing of proposed suppression patterns and automated cell suppression methodologies.

A. Tables of Frequency Data

Tables of frequency data may relate to people or establishments. Frequency data for establishments are generally not considered sensitive because so much information about an establishment is publicly available. Disclosure limitation techniques are generally applied to tables of frequencies based on demographic data. As discussed earlier, the most commonly used **primary disclosure rule** for deciding whether a cell in a table of frequency data reveals too

much information is the "threshold rule". A cell is defined to be sensitive when the number of respondents is less than some predetermined threshold. If there are cells that are identified as being sensitive, steps must be taken to protect them. The methods of preventing disclosure in tables of counts or frequencies were illustrated in II.C.2. Included are: combining cells, random rounding, controlled rounding, cell suppression, controlled tabular adjustment, and microdata techniques. Combining cells is generally a judgmental activity, performed by the survey manager. There are no methodological issues to discuss. Selection of cells for complementary suppression is the same problem for both tables of frequencies and tables of magnitude data. Complementary suppression will be discussed in Section B.2 of this Chapter. Controlled tabular adjustment is most valuable for establishment level data and is also discussed in Section B2. Microdata techniques have been used to publish data from the decennial census since 1990. These techniques were illustrated in Chapter II, and the technical issues are described in Chapter V.

Perturbation methods include random rounding and controlled rounding as special cases. Controlled rounding is a special case of random rounding. Controlled rounding is the most desirable of the perturbation methods, because it sets a condition that the cell values must add to the published row and column totals. It results in an additive table (sums of row, column and layer entries are equal to the published marginal total). Controlled Rounding can always be solved for two-dimensional tables, and can generally be solved for three-dimensional tables. Section A.1 provides more detail on the methodology used in controlled rounding.

A.1. Controlled Rounding

Controlled rounding was developed to overcome the shortcomings of conventional and random rounding and to combine their desirable features. Examples of random rounding and controlled rounding were given in II.C.2. Like random rounding, controlled rounding replaces an original two-dimensional table by an array whose entries are rounded values that are adjacent to the corresponding original values. However, the rounded array is guaranteed to be additive and can be chosen to minimize any of a class of standard measures of deviation between the original and the rounded tables.

A solution to the controlled rounding problem in two-dimensional tables was found in the early 1980's (Cox and Ernst, 1982). With this solution the table structure is described as a mathematical network, a linear programming method that takes advantage of the special structures in a system of data tables. The network method can also be used to solve controlled rounding for sets of two-dimensional tables that are related hierarchically along one dimension (Cox and George, 1989).

For three-dimensional tables an exact network solution does not exist (Cox and Ernst, 1982). Current methods make use of an iterative approximate solution using a sequence of two-dimensional networks. The exact solutions for two-dimensional tables and the approximate solutions for three-dimensional tables are both fast and accurate. Although solutions to the controlled rounding problem are available, controlled rounding is not a common practice among

U.S. government agencies.

B. Tables of Magnitude Data

For tables of magnitude data the values reported by respondents are aggregated in the cells of a table. Examples of magnitude data are income for individuals and sales volumes and revenues for establishments. Particularly for establishments, these reported values are typically highly skewed with a few very large reported values that might easily be associated with a particular respondent by a knowledgeable user. As a result, a more mathematical definition of a **sensitive cell** is needed for tables of magnitude data. For tables of frequency data each respondent contributes equally to each cell, leading to the simple threshold definition of a sensitive cell.

Mathematical definitions of sensitive cells are discussed in Section B.1 below. After the tables have been created and the sensitive cells are identified, a decision must be made as to how to prevent disclosure from occurring. For tables of magnitude data the possibilities include combining cells and rolling up categories, cell suppression, and controlled tabular adjustment. All were summarized and illustrated in Chapter II.

In the combination method tables are redesigned (categories rolled-up) so there are fewer sensitive cells. Table redesign methods are useful exercises, particularly with tables from a new survey or where portions of a table contain many sensitive cells because the population is sparse. However, it is not generally possible to eliminate all sensitive cells by collapsing tables, and rigorous automated procedures for collapsing in general remain to be developed.

The historical method of protecting sensitive cells in tables of magnitude data is cell suppression. Sensitive cells are not published (they are suppressed). These sensitive suppressed cells are called **primary suppressions**. To make sure the primary suppressions cannot be derived by subtraction from published marginal totals, additional cells are selected for **complementary suppression**. Complementary suppressions are sometimes called **secondary suppressions**.

For small tables, it is possible to manually select cells for complementary suppression, and to apply audit procedures (see Section 2.a) to guarantee that the selected cells adequately protect the sensitive cells. For large-scale survey publications with many related tables, the selection of a set of complementary suppression cells that are "optimal" in some sense is an extremely complex problem. Complementary suppression is discussed in Section B.2.

Controlled tabular adjustment is also illustrated in Chapter II. Some of the technical details are discussed in IV.B.3. Finally, microdata methods are increasingly being used to protect tabular data prior to tabulation. For establishment level data, noise addition is the technique that has been applied to date. This is summarized in Chapter II, and discussed in more detail in IV.B.4.

B.1. Definition of Sensitive Cells – Linear Sensitivity Rules

The definitions and mathematical properties of linear sensitivity measures and their relationship to the identification of sensitive cells in tables were formalized by Cox (1981). Although the common linear sensitivity rules were known in 1978 and were used to identify sensitive cells, their mathematical properties had not been formally demonstrated. The important definitions and properties are given below.

For a given cell, *X*, with *N* respondents the respondent level data contributing to that cell can be arranged in order from large to small: $x_1 \geq x_2 \geq \ldots \geq x_N \geq 0$. Then, an **upper linear sensitivity measure**, $S(X)$, is a linear combination

$$S(X) = \sum_{i=1}^{N} w_i x_i$$

defined for each cell or cell union *X* and its respondent data $\{x_i\}$. The sequence of constants, $\{w_i\}$, is called the sequence of weights of $S(X)$. These weights may be positive or negative. A cell or cell union X is **sensitive** if $S(X) > 0$. Note that multiplying a linear sensitivity measure by a constant yields another (equivalent) linear sensitivity measure. The linear sensitivity measures described in this section are all normalized so that the weight multiplying x_1 is equal to 1. This normalization makes it easier to compare them. If a respondent contributes to two cells, X and Y, then it remains a single respondent to the union of X and Y, with value equal to the sum of its X and Y contributions.

One of the properties which assists in the search for complementary cells is **subadditivity**, which guarantees that the union of disjoint cells which are not sensitive is also not sensitive. Cox shows that a linear sensitivity measure is subadditive if the sequence of weights is nonincreasing, i.e. if $w_1 \geq w_2 \geq \ldots \geq w_N$. Subadditivity is an important property because it means that aggregates of cells which are not sensitive are not sensitive and do not need to be tested. Valid complementary cells have the property that their union with the sensitive cell(s) in a row, column or layer where marginal totals are published is not sensitive according to the linear sensitivity measure. A simple result is that zero cells are not valid candidates for complementary suppression as the union of a sensitive cell and a zero cell is equal to the sensitive cell, and is therefore still sensitive. Complementary suppressions may not be needed if marginal totals are not published.

The commonly used primary suppression rules are described Sections a, b, and c below. They are compared in Section d. Each of these rules involves parameters that determine the values taken by the weights, $w_1 \ldots w_n$. Although agencies may reveal the primary suppression rule they use, they should not disclose parameter values, as knowledge of the rule and its parameters enables a respondent to make better inferences concerning the values reported by other respondents. An example is presented in Section 3.

There are three linear sensitivity measures that are discussed in the literature and used in practical applications. These are the p-percent rule, the pq rule and the (n, k) rule. They are described below. All are subadditive, as can be seen by the fact that the weights in the equations defining $S(x)$ are non-increasing. The p-percent and pq rule classify cells of count data as sensitive if n<3.

B.1.a. The p-Percent Rule

Approximate disclosure of magnitude data occurs if the user can estimate the reported value of some respondent too accurately. Such disclosure occurs, and the table cell is declared sensitive, if upper and lower estimates for the respondent's value are closer to the reported value than a pre-specified percentage, p. This is referred to as the "p-percent estimation equivocation level" in Statistical Policy Working Paper 2. It is more generally referred to as the **p-percent rule**, and has linear sensitivity measure,

$$S^{p\%}(X) = x_1 - \frac{100}{p}\sum_{i=c+2}^{N} x_i .$$

Here, c is the size of a coalition, a group of respondents who pool their data in an attempt to estimate the largest reported value. The cell is sensitive if $S^{p\%}(X) > 0$. Note that if there are less than 3 respondents $(N < 3)$ in cell X, then $S^{p\%}(X) = x_1 > 0$ and the cell is sensitive for all values of p and c.

The p-percent rule is derived as follows. Assume that from general knowledge any respondent can estimate the contribution of another respondent to within 100-percent of its value. This means that the estimating respondent knows that the other respondents' values are nonnegative and less than two times the actual value. For the p-percent rule, it is desired that after the data are published no respondent's value should be estimable more accurately than within p percent (where $p < 100$).

It can be shown that the coalition including the second largest respondent is in a position to estimate the value of x_1 most accurately, and that if x_1 is protected, so are all the smaller respondents. Thus, it suffices to provide the protection to the largest respondent, and to assume that the estimating party is a coalition of the second largest respondent and the next largest $c-1$ respondents. As the coalition respondents may estimate each of $x_{c+2},\ldots,x_N$ to within 100 percent, they have an estimate for the sum of these smallest respondents, E, such that

$$\left| \sum_{i=c+2}^{N} x_i - E \right| \le \sum_{i=c+2}^{N} x_i .$$

They can estimate the value of x_1 by subtracting the value they reported to the survey $\left(\sum_{i=2}^{c+1} x_i\right)$ and their estimate of the smaller respondent's total, E, from the published total. The error in this estimate will be equal to the error in estimating E, which is less than or equal to $\sum_{i=c+2}^{N} x_i$.

The requirement that this estimate be no closer than p-percent of the value of x_1 $(p < 100)$ implies that

$$\sum_{i=c+2}^{N} x_i \geq \frac{p}{100} x_1.$$

This can be rewritten as the linear sensitivity rule above. A simpler version of the p percent rule that assumes coalitions of size c can be written as follows:

$S = x_1 - 100/p * (T - T_c - x_1)$

Where T is the cell total of all respondents, T_c is the sum of the respondent values in the coalition, and x_1 is the largest value. Using this formula the cell is sensitive if S is positive. In the simple case where $T_c = x_2$ (i.e., the coalition is only a size of one), then $T - T_c - x_1 = T - x_2 - x_1$ which means the remaining cell value is the sum of all the smallest companies in the cell with the exception of the two largest. $T - T_c - x_1$ will equal zero only if the coalition (T_c) includes all the respondents in the cell except the largest company.

B.1.b. The pq Rule

In the derivation for the p-percent rule, we assumed that there was limited prior knowledge about respondent's values. Some people believe that agencies should not make this assumption. In the pq rule, agencies can specify how much prior knowledge there is by assigning a value *q* which represents how accurately respondents can estimate another respondent's value before any data are published $(p < q < 100)$. Thus, there is an improved estimate, E_2, of $\sum_{i=c+2}^{N} x_i$ with the property that

$$\left| \sum_{i=c+2}^{N} x_i - E_2 \right| \leq \frac{q}{100} \sum_{i=c+2}^{N} x_i.$$

This leads directly to a more accurate estimate for the largest respondent's value, x_1. The requirement that this estimate be no closer than *p*-percent of the value of x_1 implies that

$$\frac{q}{100} \sum_{i=c+2}^{N} x_i \geq \frac{p}{100} x_1.$$

$$S^{pq}(X) = x_1 - \frac{q}{p}\sum_{i=c+2}^{N} x_i .$$

This can be rewritten as the linear sensitivity rule.
Note that the pq rule (sometimes called a prior-posterior ambiguity rule) and the p-percent rule are identical if the ratio q/p, the "information gain", is equal to $100/p$. In the table below we use the ratio q/p as a single parameter for the pq rule. If users fix a value for p and a value for $q < 100$, the pq rule is more conservative (will suppress more cells) than the p-percent rule using the same value of p.

Note that if there are fewer than 3 respondents $(N < 3)$, then $S^{pq} = x_1 > 0$ and cell X is sensitive for all values of c and q/p.

Most frequently the pq rule is given with the size of a coalition equal to 1. In this case the linear sensitivity rule is given by

$$S^{pq}(X) = x_1 - \frac{q}{p}\sum_{i=3}^{N} x_i .$$

B.1.c. The (n, k) Rule

The **(n, k) rule**, or dominance rule was described as follows in Statistical Policy Working Paper 2. "Regardless of the number of respondents in a cell, if a small number (*n* or fewer) of these respondents contribute a large percentage (*k* percent or more) of the total cell value, then the so-called **n respondent, k percent rule** of cell dominance defines this cell as sensitive." Many people consider this to be an intuitively appealing rule, because, for example, if a cell is dominated by one respondent then the published total alone is a natural upper estimate for the largest respondent's value. Although coalitions are not specifically discussed in the derivation of the (n, k) rule, agencies select the value of *n* to be larger than the number of any suspected coalitions. Many agencies use an (n, k) rule with $n = 1$ or 2.

The linear sensitivity measure for the (n, k) rule is given by

$$S^{(n,k)}(X) = \sum_{i=1}^{n} x_i - \frac{k}{100-k}\sum_{i=n+1}^{N} x_i .$$

If $N \le n$, $S^{(n,k)} = \sum_{i=1}^{N} x_i > 0$ and cell X is sensitive for all values of k.

B.1.d. The Relationship Between (n, k) and p-Percent or pq Rules

Table 1 is designed to assist users in selecting a value of the parameter p for use with the p-percent rule with coalitions of size 1 (or for the value of the ratio, q / p, for the pq rule with coalitions of size 1) when they are used to thinking in terms of the (n, k) rule. For various values of $p\%$ (q / p), the table shows the value of k_1 and k_2 such that if the linear sensitivity rule for $(1, k_1)$ or $(2, k_2)$ is positive then the linear sensitivity rule for the p-percent (pq) rule will be positive. With this formulation, the p-percent (pq) rule is more conservative. It will suppress more cells than will either of the two (n, k) rules individually, and also more than the combination rule based on the two (n, k) rules. The derivation of the inequalities used in Table 1 is presented in the Technical Notes at the end of this Chapter. Additionally, the sensitivity regions for (n, k), p-percent, and pq rules are illustrated graphically in the Technical Notes. See Robertson, D. A. (1993) for theoretical analysis comparing disclosure rules.

To illustrate the use of Table 1, if the analyst wants to make sure that a cell where the largest respondent contributes more than 75 percent of the total is suppressed, and that a cell where the largest two respondents exceed 85 percent of the total is suppressed, he/she could approximately accomplish this by using the p-percent rule with p equal to 33.3 percent, or the pq rule with information gain, $q / p = 3$.

The p-percent, pq and (n, k) rules as well as the combination rule,

$$S^{comb} = \max(S^a(X), S^b(X))$$

are subadditive linear sensitivity rules. (Here $S^a(X)$ and $S^b(X)$ denote any two subadditive linear sensitivity measures.) Any of these rules is acceptable from a mathematical point of view. However, the p-percent or pq rule is preferred for two major reasons. First, the tolerance interval concept directly parallels methods currently used for complementary suppression, (see section B.2.a.iii). Second, as illustrated in the table above and the example in the Technical Notes, the p-percent (pq) rule provides more consistent protection areas than a single version of the (n, k) rule.

TABLE 1 Relationship Between Suppression Regions for p-Percent or (pq) Rule and (1,k), (2,k) Rules

		$S^{p\%}(X) > 0$ and	Sensitive when cell
P	q/p	x_1/T exceeds:	$(x_1 + x_2)/T$ exceeds:
50.0%	2	66.7%	80.0%
33.3%	3	75.0%	85.7%
16.7%	6	85.7%	92.3%
11.1%	9	90.0%	94.7%

NOTE: $T = \sum_{i=1}^{N} x_i$ is the cell total.

B.1.e. Information in Parameter Values

Agencies may publish their suppression rules, however, they should keep the parameter values they use confidential. Knowledge of both the rule and the parameter values enables the user to make better inferences concerning the value of suppressed cells, and may defeat the purpose of suppression.

For example, assume that an agency uses the p-percent rule with p=20 percent, and that the same value of p is used to determine the protection regions for complementary suppression. We assume that a cell total is 100 and that the cell is sensitive according to the p-percent rule. That cell will be suppressed along with other suitable complementary cells. For this cell (as with any suppressed cell), any user can use a linear programming package to calculate upper and lower bounds for the cell total based on the published row and column equations. Assume that this leads to the following inequality:

$$80 = \text{lower bound} \leq \text{cell total} \leq \text{upper bound} = 120.$$

In this case, the protection region used in selecting cells for complementary suppression assures that the cell total cannot be estimated more closely than plus or minus 20 percent of the cell value, or plus or minus 20 in this case. A knowledgable user has thus uniquely determined that the value of the suppressed cell total must be 100. Once the total for one suppressed cell has been uniquely determined, it is likely that other cell values can easily be derived by subtraction from published marginal totals.

B.2. Complementary Suppression

Once sensitive cells are identified by a primary suppression rule, other nonsensitive cells must be selected for suppression to assure that the respondent level data in sensitive cells cannot be estimated too accurately. This is the only requirement for a proposed set of complementary cells for tables of magnitude data and is generally considered to mean that a respondent's data cannot be estimated more closely than plus or minus some percentage.

There are two ways respondent level data can be compromised. First, implicitly published unions of suppressed cells may be sensitive according to the linear sensitivity measure. This depends on the characteristics of the respondent level data in the cell union, and tends to be a problem only where the same respondents contribute to both cells. Second, the row and column equations represented by the published table may be solved, and the value for a suppressed cell estimated too accurately. Automated methods of **auditing** a proposed suppression pattern may be needed to assure that the primary suppressions are sufficiently well protected (see Section B.2.a).

Any set of cells proposed for complementary suppression is acceptable as long as the sensitive cells are protected. For small tables this means that selection of complementary cells may be done manually. Typically the data analyst knows which cells are of greatest interest to users (and should not be used for complementary suppression if possible), and which are of less interest to users (and therefore likely candidates for complementary suppression). Manual selection of complementary cells is acceptable as long as the resultant table provides sufficient protection to the sensitive cells. An automated audit should be applied to assure this is true.

For large systems of tables, for example, those based on an Economic Census, the selection of complementary cells is a major effort. Manual selection of cells may mean that a sensitive cell is inadvertently left unprotected or that consistency is not achieved from one table to another in a publication. (Cox, 1980). Inconsistency in the suppression patterns in a publication increases the likelihood of inadvertent disclosure. For this reason linear programming techniques have been applied to the selection of cells for complementary suppression by statistical agencies for many years. (Cox, 1995). As an additional benefit, agencies expect automated selection of the complementary cells will result in less information lost through suppression. Examples of the theory and methods for automatic selection of cells for complementary suppression are discussed in Section B.2.b.

B.2.a. Audits of Proposed Complementary Suppressions

B.2.a.i. Implicitly Published Unions of Suppressed Cells Are Sensitive

If sensitive cells are protected by suppressing other internal table cells when publishing the marginal totals, the implicit result is that the unions of the suppressed cells in rows, columns and layers are revealed by subtracting from the total. Thus, one way to audit the protection supplied by the suppression pattern is to apply the linear sensitivity rule to those unions to assure that they are not sensitive. While this type of audit is a simple matter for small tables, Cox (1980) points

out that for large tables it may be computationally intractable unless a systematic approach is used. This type of audit is not included in standard audit software because of its dependence on respondent level data.

Clearly a table for which suppression patterns have been selected manually requires an audit to assure that the pattern is acceptable. Early versions of complementary suppression software used approximation arguments to select cells for complementary suppression (individual respondent data were not used.) These methods guaranteed that unions of suppressed cells were not sensitive as long as different respondents contributed to each cell. However, if the same respondents contributed to multiple cells in a cell union, then an audit was needed.

B.2.a.ii. Row, Column and/or Layer Equations Can Be Solved for Suppressed Cells

A two-dimensional table with row and column subtotals and a three-dimensional table with row, column and layer subtotals can be viewed as a large system of linear equations. The suppressed cells represent unknown values in the equations. It is possible that the equations can be manipulated and the suppressed values estimated too accurately. Audits for this type of disclosure require the use of linear programming techniques. The output of this type of audit is the maximum and the minimum value each suppressed cell can take given the other information in the table. When the maximum and the minimum are equal, the value of the cell is disclosed exactly. To assure that cells cannot be estimated too accurately the analyst makes sure the maximum and the minimum value for the suppressed cell are no closer to the true value than some specified percentage protection.

It is well known that a minimal suppression pattern where marginal totals are presented will have at least two suppressed cells in every row, column and layer requiring suppression. This is not sufficient, however, as was illustrated in Chapter 2 Section C.2.a.

B.2.a.iii. Software For Auditing A Suppression Pattern

Automated methods of auditing a suppression pattern have been available since the mid 1970's at the U.S. Census Bureau, and at Statistics Canada. Modern versions of audit software set up the linear programming problem and use commercially available linear programming packages. All audit systems produce upper and lower estimates for the value of each suppressed cell based on linear combinations of the published cells. A suppression audit can uncover three types of problems for tables cells: 1) the upper and lower limits may be the same; 2) the upper and lower limits may be too close together; 3) the upper and/or lower limits may be too close to the cell value. The data analyst uses the output from the audit to determine whether the protection provided to the sensitive cells by the proposed complementary cells is sufficient. The user should know the type and extent of the rounding of cell values in a table that is being audited to avoid misleading evaluations of data protection. Depending upon whether suppression was

applied to the rounded or unrounded data can result in over- or under-suppression of cells in a table. These audit methods are applicable to tables of both magnitude and frequency.

Linear programming is the most common procedure used for auditing suppression patterns in a table because it can be used for higher-dimensional tables. (Zayatz 1992a). Network procedures have been shown to provide fast solutions for two-dimensional tables. The network flow method for cell suppression is self-auditing only for two-dimensional tables in which there is a hierarchy in one dimension. The network flow method is not-self auditing for two-dimensional tables with a hierarchical variable structure in both the row and column, and it is not self-auditing for three dimensional or higher dimensional tables that contain a hierarchical structure. (Massell 2002).

At the U. S. Census Bureau both types of audits are subsumed into the algorithm that selects cells for complementary suppression. The company level contributions for a cell are used in selecting a protection level or tolerance interval for each cell that provides protection to all respondents in the cell. The algorithm that selects cells for complementary suppression provides that the primary cells cannot be estimated more accurately than that specified tolerance interval. The complementary suppressions selected by applying the algorithm do not require additional audits.

Audit software was developed by the Confidentiality and Data Access Committee, with support from a number of statistical agencies, and is available with documentation at http://www.fcsm.gov/committees/cdac/resources.html. This software is written in SAS® and checks the lower and upper bounds around suppressed cells in a table that contains non-additive, independently rounded cells. The program requires a specific format for the ASCII input file. The program also checks for whether independent rounded cells exist in the table and adjusts the cell values to preserve additivity within the row and columns at the same time it is performing the import function. The user has the option of specifying a protection range based on a plus/minus percent basis or absolute value basis. The software is not limited by the number of dimensions in a table and the linear programming methodology provides for two types of optimizers.

B.2.b. Automatic Selection of Cells for Complementary Suppression

Software that automatically selects complementary cells for suppression has been available since the 1970's at Statistics Canada and at the U.S. Census Bureau. These programs typically use linear programming methods implemented by accessing general purpose linear programming routines which make use of special structures in the data. Due to refinements in linear programming algorithms, these routines run much faster now than in the 1980's. Network flow methods may be viewed as a special case of linear programming. They work best for two dimensional tables, with at most one level of hierarchy (in either rows or columns). Routines based on network flow methods are typically much faster than linear programming routines. Cell suppression programs can be used for both magnitude data tables and frequency data tables.

At the U.S. Census Bureau these programs have been used mainly for business survey magnitude data. Robert Jewett (Jewett, 1993) wrote a set of cell suppression programs for this purpose. They include much beyond the basic cell suppression model. For example, the program can be used to identify sensitive cells from a given input microdata file by using the p% rule. The problem of "common respondents" is handled by defining a table of capacities for each primary. It is constructed just before complementary suppressions are selected to protect a given primary. The "common respondents" problem arises frequently with business survey data since many companies have more than one establishment and often these establishments are contributors to different cells of the same table. The U.S. Census Bureau must protect not only each establishment's contribution but all sums of an establishment's facilities, including the company's total contribution. These programs are also able to handle tables that are linked and interrelated to cells in two or more of the tables. It uses the method of backtracking to check that a given suppressed cell has the same degree of uncertainty in each table in which it appears.

The software, tau-Argus, developed from the Computational Aspects of Statistical Confidentiality (CASC) European project offers methods to identify sensitive cells, a choice of algorithms to select secondary suppressions, an suppression audit program to compute interval bounds for suppressed cells, and a module to generate synthetic values to replace suppressed original ones in a publication. The documentation and software for operating tau-Argus are available at http://neon.vb.cbs.nl/casc.

In the straightforward implementation of linear programming, sensitive cells are treated sequentially beginning with the most sensitive. At each step (i.e. for each sensitive cell) the set of complementary cells that minimizes a cost function (usually the sum of the suppressed values) is identified. Zayatz (1992a) describes the formulation for two-dimensional tables. Zayatz (1992b) gives the parallel formulation for three-dimensional tables. As above, these are implemented by using a commercially available linear programming package. The disadvantage of the straightforward linear programming approach is the computer time it requires. For large problems, the run time of the Central Processing Unit of a personal computer increases significantly with 3 or more dimensions.

Another linear programming approach is based on describing the table structure as a mathematical network, and using that framework and the required tolerance intervals for each cell to balance the table. The network methods are favored because they give the same result as the straightforward linear programming methods, but the solution requires much less computer time. The network method is directly applicable to two-dimensional tables and to two-dimensional tables with subtotal constraints in one dimension (Cox, 1995). Subtotal constraints occur when data in one dimension have a hierarchical additive structure such as the North American Industry Classification System (NAICS) coding system. In the past 20 years, there was considerable research in developing faster and more efficient procedures for both two-dimensional and three-dimensional tables. Research has involved using methods based on integer programming, network flow theory, and neural networks.

Complementary suppression and controlled rounding can both be solved using network theory. Ernst (1989) demonstrated the impossibility of representing a general three or higher dimension table as a network. For this reason, complementary suppression for three-dimensional tables currently uses linear programming as the main approach. (Zayatz, 1992b). The straightforward linear programming methods can be used for small three-dimensional tables. However, for large three-dimensional tables, an iterative approximate approach based on a sequence of two-dimensional networks is used. The complementary suppression pattern identified by this approximate approach must still be audited to assure that an individual sensitive cell cannot be estimated too accurately.

As mentioned above, one possible objective or cost function for automated procedures is to minimize the sum of the suppressed values. With this objective function, automated procedures tend to suppress many small cells, a result not generally considered "optimal" by the analyst. Other possible cost functions include minimizing the total number of suppressed cells in a table or minimizing the suppression for specific data series in a table. Further research is needed into the identification of cost functions for use in selecting the "optimal" complementary suppressions. Possibilities here include research into a cost function for use in a single run of the software, as well as cost functions for use in multiple runs of the software. An example is the development of a cost function that is used during a second pass through the software to remove superfluous suppressions (Zayatz, 1992b).

Another reason the complementary cells selected by automated methods do not provide the "optimal" set for the table as a whole is that all current implementations protect sensitive cells sequentially. For any given sensitive cell, the complementary cells selected to protect it will be optimal according to the objective function, conditional on all suppressions selected for previously considered sensitive cells. The sequential nature of the approach leads to over-suppression.

In spite of the lack of "optimality" of the result, the automated complementary cell suppression procedures identify useful sets of complementary suppressions. However, work is often needed to fine tune, reduce over-suppression, and assure that the analysts' nonmathematical definition of an "optimal" solution is more closely realized.

B.3. Controlled Tabular Adjustment

Controlled Tabular Adjustment is a useful methodology for protecting tables of magnitude data as well as count data. It is discussed with an example in Chapter 2 Section D.3.d. Each sensitive original value in a table is replaced 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. Another appropriate optimization function may be to minimize the sum of the absolute values of the data adjustments. The reciprocal of the cell value 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 nonsensitive cells 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 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 either type of 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. (Dandekar, 2004). 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, accepts all necessary adjustments in marginal table cell values to restore additive table structure.

B.4. Adding Noise to Microdata Prior to Tabulating Data

Adding noise to the underlying microdata is a method that has been used to protect magnitude tabular data. It is different from the noise procedures used to protect public use microdata files. The noise addition method adjusts each value by a small amount (the exact percent to remain confidential within the statistical agency). Each establishment reporting in the sample or survey is assigned a multiplier, or noise factor. A company may have several different stores or establishments. In this case, each establishment may be assigned a slightly different multiplier as

long as the overall distribution of the multipliers across all establishments within a company average the specified percent for adjusting that company's reported values. (Evans, 1998).

For example, if an establishment's data is adjusted by 10%, then its data would be multiplied by a number that is close to either 1.1 or 0.9. Any type of distribution can be used to choose the multipliers for each establishment. In this example, whatever distribution is used to generate a multiplier of 1.1, it is important that the same distribution shape, or its "mirror image," be used to generate the multipliers near 0.9 to adjust data in the opposite direction. The two distributions of multipliers should produce a joint distribution of multipliers that is symmetrical and approximates 1.

The direction of adding the noise to each responding company is randomly assigned. Using the example of 10% as the base for perturbation, this is equivalent to determining if all establishments in a company have multipliers close to 1.1 or close to 0.9. The next step in the process is to randomly assign a multiplier to each establishment within a company. The multipliers would be generated from that half of the overall distribution of the multipliers that corresponds to the direction of perturbation assigned to that company. An example of assigning multipliers to a set of respondents is as follows:

Example 1:

Company	Establishment	Direction	Multiplier
Company A		1.1	
	Establishment A1		1.12
	Establishment A2		1.09
	Establishment A3		1.10
	Establishment A4		1.11
Company B		0.9	
	Establishment B1		0.89
	Establishment B2		0.93
Company C		1.1	
	Establishment C1		1.08

In this example, the expected value of the amount of noise added in any cell value is zero because of the symmetry of the distribution of the multipliers and the random assignment of both the direction of perturbation and the multipliers within each company. The probability that a company's establishments will be perturbed in a positive direction is equal to the probability that they will be perturbed in a negative direction. The distribution of the multipliers is symmetric about 1. The expected value of any given multiplier is 1, hence the expected value of the *amount* of noise in any given establishment is 0, and the amount of noise in any cell value is simply the sum of the noise in its component establishments.

Noise addition differs from Controlled Tabular Adjustment because noise addition adjusts the reported values prior to any tabulations. Controlled Tabular Adjustment adjusts the cells after

the data have been tabulated on a cell by cell basis. Noise addition relies on the random assignment of the multiplier to control the effects of adding noise to different types of cells.

C. Online Data Query Systems

Most online query systems that were developed by the federal agencies allow access to summary files with matrices of aggregated data. These query systems allow users to design queries to generate customized tabulations. Special disclosure limitation methods should be considered when users access microdata files to produce customized tabulations.

One example of an online query system that allows users to access microdata files is the "Advanced Query System" (AQS) which is part of the Census Bureau's "American Fact Finder" online data dissemination system. The microdata files in the AQS contain information on individuals and households. To ensure that tabulations from these microdata files do not reveal the identities of respondents, the Census Bureau uses data recoding and data swapping techniques in addition to other microdata techniques.

Variables such as geography, detailed race, age, occupation, industry, Hispanic origin, and group quarters are re-coded and/or collapsed. All continuous variables, such as income, fuel and utility costs, property taxes, rent, and mortgage payments are top coded to mask the outlying values in the tails of the distributions of each continuous variable. The re-coded variables are added to the files used by AQS. An external user is diverted to the re-coded variables and geographic area when submitting a query.

In addition to recoding, a swapping technique is also applied to the records in the microdata files. The technique consists of swapping pairs of household records selected as having the highest disclosure risk based upon a predetermined set of key variables. In the AQS system, records are selected for swapping with a probability inversely proportional to block size.

Any request submitted by a user passes through two filters; the Query filter and the Statistical Results filter. The purpose of the Query filter is to detect those queries that will not pass disclosure limitation before the query is submitted for execution, such as the geographic variable must meet a minimum threshold. The Statistical Results Filter checks the final values in the cells of the resulting table. If a table does not pass the filters, the entire table is suppressed and the user does not receive the table. The AQS system does not perform any cell suppression. A message is sent to a user that the table is suppressed for confidentiality reasons and the user may then try requesting a table with less detail.

The disclosure protection procedures applied by the Agriculture Resource Management Survey (ARMS) uses a different approach than the AQS system. The ARMS on line query system allows users to select across survey data sets and build customized reports. There are three stages to the disclosure protection procedures used in the ARMS system. In the first step, noise is added to the weights for underlying microdata in a unique way to protect large establishments that may dominate a cell. The second step is to develop minimum expanded farm counts in a

cell and to test the sensitivity of that cell using the p-percent rule. The third step applies primary cell suppression without any complementary suppression. No complementary suppression is necessary because of the noise that was initially added to the microdata provides the necessary protection to the aggregates. Cells in the outputted files are suppressed if they fail any of the three criteria: 1) if the ratio of the cell value with noise to the cell value without noise is outside a set range, then the cell is suppressed; 2) if the weighted farm count for a cell is small then the cell is suppressed; 3) if the cell fails the p-percent rule and has insufficient noise to protect the actual value, then the cell is also suppressed. The approach used in ARMS avoids the need for complementary suppression and simplifies the computational problems associated with disclosure protection in an on-line query system.

D. Technical Notes: Relationships Between Common Linear Sensitivity Measures

This section illustrates the relationship between the p-percent, pq and (n, k) rules described in the text by using plots of regions of cell sensitivity. To simplify this presentation we make a few assumptions. First, for the p-percent rule we assume there are no coalitions $(c=1)$ and for the (n, k) rules we consider only $n=1$ and $n=2$. Second, replace $\sum_{i=3}^{N} x_i$ by $(T-x_1-x_2)$. Third, divide each sensitivity rule through by the cell total, *T*, and multiply by 100. Finally, set $z_i = 100x_i / T$, the percent contributed to the cell total by company *i*. The sensitivity rules can be written

$$S^{p\%}(X) = \left(1+\frac{100}{p}\right)z_1 + \frac{100}{p}z_2 - \frac{100}{p}100,$$

$$S^{pq}(X) = \left(1+\frac{q}{p}\right)z_1 + \frac{q}{p}z_2 - \frac{q}{p}100,$$

$$S^{(1,k_1)}(X) = \left(1+\frac{k_1}{100-k_1}\right)z_1 - \frac{k_1}{100-k_1}100$$

$$S^{(2,k_2)} = \left(1+\frac{k_2}{100-k_2}\right)z_1 + \left(1+\frac{k_2}{100-k_2}\right)z_2 - \frac{k_2}{100-k_2}100$$

The regions where these sensitivity rules are positive (i.e. where the cells are sensitive) are shown in Figure 1. The horizontal axis represents the percent contributed by the largest unit, z_1 and the vertical axis represents the percent contributed by the second largest unit, z_2. Since $z_1 \geq z_2$ and $z_1 + z_2 \leq 1$ (the sum of the two largest is less than or equal to the cell total), the only possible values in a table cell will be in the lower triangular region bounded from below by the line $z_2 = 0$, from above by the line $z_1 = z_2$ and to the right by the line $z_1 + z_2 = 1$.

The $(1,k_1)$ and $(2,k_2)$ rules are particularly simple to illustrate graphically. The inequality $(1,k_1)$ rule simplifies, and a cell is classified as sensitive if $z_1 > k_1$. The dividing line between sensitive and nonsensitive region is given by a vertical line through the point $(0,k_1)$. Similarly, the inequality for the $(2,k_2)$ rule simplifies and a cell is classified as sensitive if $(z_1 + z_2) > k_2$ (z1 + z2) > k2. The dividing line between the sensitive and nonsensitive regions is the line through the points $(0,k_2)$ and $(k_2,0)$. This line intersects $z_1 = z_2$ at the point $(k_2/2, k_2/2)$. In all cases the sensitive region is the area to the right of the dividing line. The sensitivity regions for the (1,75) and (2,85) rules are illustrated in Figure 1A.

For the p-percent rule the inequality above yields the boundary line for sensitive cells as the line joining the points (0,100) and $\left(\frac{100}{\frac{p}{100}+1},0\right)$. This line intersects $z_1 = z_2$ at the point $\left(\frac{100}{\frac{p}{100}+2}, \frac{100}{\frac{p}{100}+2}\right)$. The pq rule is the same, with q/p = 100/p.

FIGURE 1A
EXAMPLES OF SUPPRESSION REGIONS
THE (N,K) RULE WITH N=1 AND K=75, N=2 AND K=85

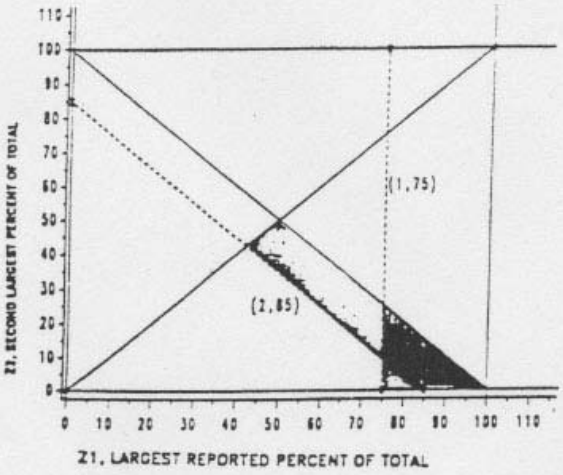


FIGURE 1B
EXAMPLES OF SUPPRESSION REGIONS
THE P-PERCENT RULE WITH P=17.65 PERCENT, AND P=35.3 PER

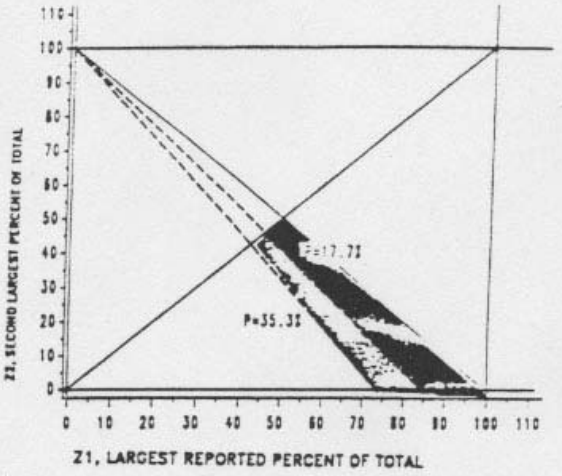


FIGURE 1C
P-PERCENT LESS CONSERVATIVE THAN (2,85)
P = 17.7 PERCENT

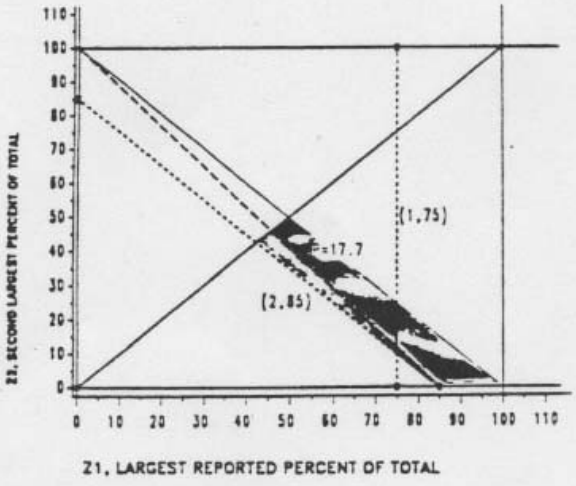
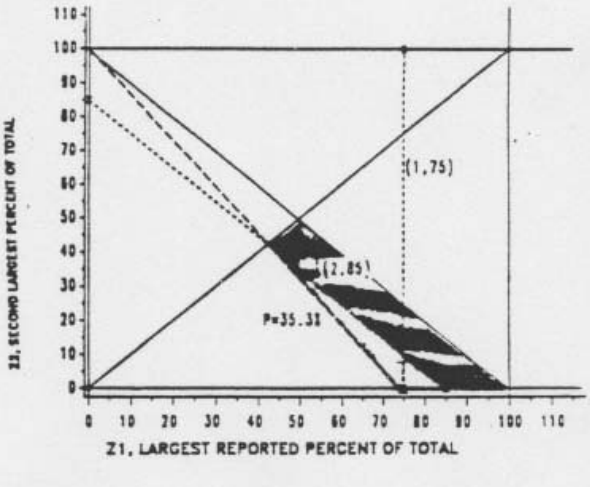


FIGURE 1D
P-PERCENT MORE CONSERVATIVE THAN (2,85) AND (1,
P = 35.3 PERCENT



Note: Values corresponding to cells in a table are in the triangle bounded by the lines z1 = z2, z2 = 0, and z1 + z2 = 1. Values which correspond to sensitive cells are shaded.

Figure 1B shows the sensitivity regions for the p-percent rule with $p = 17.65$ and $p = 35.29$. The selection of these values of p will be discussed below. Note that if $p = 0$, the sensitivity line falls on top of the line $z_1 + z_2 = 1$. At that point there are no sensitive cells. Similarly if p is negative, there are no sensitive cells.

To Find p so that $S^{p\%}(X) \leq S^{(n,k)}(X)$ for all cells, X.

Consider the case where the (n, k) rule is being used and there is also a requirement that no respondent's contribution be estimable to within p-percent of its value. We would like to find the value of p so that the p-percent rule is closest to the (n, k) rule with $S^{(n,k)}(X) \geq S^{p\%}(X)$. Thus, there may be cells classified as sensitive by the (n, k) rule which would not be sensitive by the p-percent rule, but all cells classified as sensitive by the p-percent rule would be classified as sensitive by the (n, k) rule. Consider the (2, 85) rule illustrated in Figure 1A. The p-percent rule, closest to the (2, 85) rule, which would satisfy this requirement would be the one which intersects the line $z_2 = 0$ at the same point as the (2, 85) rule. Thus, for a given value of k_2 we must have

$$S^{(n,k)}(X) \geq S^{p\%}(X)$$

Similarly, if we were first given the value of p for the p-percent rule, we must have

$$S^{(n,k)}(X) \geq S^{p\%}(X)$$

For the (2, 85) rule, $p/100 = 15/85 = .1765$, so that $p = 17.65$ percent. Figure 1C shows the (2,85) sensitivity region along with the less conservative $p = 17.65$ percent region.

For the $(1, k_1)$ rule, the p-percent rule closest to the (1, 75) rule satisfying this requirement would be the one intersecting the line $z_1 = z_2$ at the point (75, 75). For a given value of k1 we must have

$$\frac{p}{100} = \frac{100}{k_1} - 2.$$

Similarly, if we were first given the value of p,

$$k_1 = \frac{100}{\frac{p}{100} + 2}.$$

With $k_1 = 75$, the less conservative p-percent rule would have $p = -66.7$, which would result in no cell suppression. For $p = 17.65\%$, we would need $k_1 = 45.94$, a very restrictive rule.

To find parameter p so that $S^{p\%}(X) \geq S^{(n,k)}(X)$ for all X.

We would like to find the value of p so that the p-percent rule is closest to the (n, k) rule with $S^{k)(n,}(X) \leq S^{p\%}(X)$. Thus, there may be cells classified as sensitive by the p-percent rule which would not be sensitive by the (n, k) rule, but all cells classified as sensitive by the (n, k) rule would be classified as sensitive by the p-percent rule. Again, we consider the (2, 85) rule as illustrated in Figure 1A. In this case the most conservative p-percent rule needed would be the one that intersects the line $z_1 = z_2$ at the same point as the (2, 85) rule. Given the value of k_2 this leads to

$$\frac{p}{100} = \frac{200}{k_2} - 2.$$

If we were first given the value of p, we would need

$$k_2 = \frac{200}{\frac{p}{100} + 2}.$$

For $k_2 = 85$, this gives $p/100 = 200/85 - 2 = .3529$. Figure 1D shows the (2,85) sensitivity region along with the $p = 35.29$ percent region.

To find the most conservative p% rule needed to include the sensitivity region of the $(1, k_1)$ rule, we need the p-percent rule which intersects the line $z_2 = 0$ at the same point as the $(1, k_1)$ rule. Given the value of k_1, this leads to

$$\frac{p}{100} = \frac{100}{k_1} - 1.$$

If we were first given the value of p, we would need

$$k_1 = \frac{100}{\frac{p}{100} + 1}.$$

For the (1,75) rule, this leads to $p/100 = 25/75 = .3333$.

To find the $(1, k_1)$ rule going through the same point as the (2,85) rule and the p-percent rule with $p = 35.29\%$, substitute the desired value of p into the above equation and find $k_1 = 73.91$.

In this case since we started with the (2,85) rule, which lead to $p = 35.29$, a consistently less conservative $(1, k_1)$ rule is the one that has $k_1 = 73.91$. Thus the p-percent rule with $p = 35.29$ provides slightly more protection than either the (2,85) rule or the (1,73.91) rule. Table 1 in the text summarizes these results for selected values of p, or equivalently for selected values of q/p.

Example

Consider the three cells below. Let x_1^k represent the largest value reported by a respondent in cell k; x_2^k the second largest value reported by a respondent in cell k; and so on. Here we assume that respondents report in only one of the cells 1, 2 or 3. Cell membership is denoted by the superscript k. Superscript T represents the total.

	Cell 1	Cell 2	Cell 3	Total
	$x_1^1 = 100$	$x_1^2 = 1$	$x_1^3 = 100$	$x_1^T = 100$
		$x_2^2 = 1$		$x_2^T = 100$
		$x_3^2 = 1$		$x_3^T = 100$
		.		.
		.		.
		.		.
		$x_{20}^2 = 1$		$x_{20}^T = 100$
SUM	100	20	100	220

Assume that we are using the (n, k) rule with $n = 2$ and $k = 85$ percent. As described above, the related rules are the p-percent rule with $p = 17.65$ (more conservative), the p-percent rule with $p = 35.29$ (less conservative) and the (1,73.91) rule.

Using any of these rules, Cell 1 and Cell 3 are clearly sensitive ($N = 1$, so $S(X) > 0$). It is also easy to verify that using any sensible rule Cell 2 is not sensitive. We consider two cells, the union of Cell 1 and Cell 2 and the Total.

The cell sensitivities for these rules are

$$
\begin{aligned}
S^{(2,85)}(\text{Cell}\,1 \cup \text{Cell}\,2) &= 100 + 1 - 5.667 \cdot 19 = -6.67 \\
S^{(17.6\%)}(\text{Cell}\,1 \cup \text{Cell}\,2) &= 100 \quad - 5.667 \cdot 19 = -7.67 \\
S^{(1,73.91)}(\text{Cell}\,1 \cup \text{Cell}\,2) &= 100 \quad - 2.833 \cdot 20 = -43.34 \\
S^{(35.29\%)}(\text{Cell}\,1 \cup \text{Cell}\,2) &= 100 \quad - 2.834 \cdot 19 = 46.16
\end{aligned}
$$

$$
\begin{aligned}
S^{(2,85)}\ (\text{Total}) &= 100 + 100 - 5.667 \cdot 20 = 86.66 \\
S^{(17.6\%)}\ (\text{Total}) &= 100 \qquad - 5.667 \cdot 20 = -13.34 \\
S^{(1,73.91)}\ (\text{Total}) &= 100 \qquad - 2.833 \cdot 120 = -239.96 \\
S^{(35.29\%)}\ (\text{Total}) &= 100 \qquad - 2.834 \cdot 20 = 43.32
\end{aligned}
$$

The union of Cell 1 and Cell 2 is not sensitive according to the (2, 85) rule and the 17.65% rule. However, both the (1, 75) and the 33.3% rule classify the cell as sensitive. Looking at the respondent level data, it is intuitively reasonable that the union of Cell 1 and Cell 2 is sensitive, even though the rule of choice for this example was to protect only against dominance by the 2 largest respondents. This cell corresponds to the point (83.3, .008) on Figure 1.

The Total is sensitive for the (2, 85) rule and the p-percent rule with p=35.3%. It is not sensitive for the (1, 73.9) rule or the p-percent rule with p=17.6%. This point corresponds with the point (45.5, 45.5) on Figure 1.

Consider the inconsistency in using the (2, 85) rule alone. In the above example, if the union of cell 1 and cell 2 (not sensitive by the (2, 85) rule,) is published, then the largest respondent knows that the other respondents' values sum to 20, and each of other respondents knows that the other respondents' values sum to 119. If the total (sensitive by the (2, 85) rule) is published then the largest two respondents each knows that the sum of the remaining respondents' values is 120, and each of the small respondents knows that the sum of the others' values is 219.

Intuitively, it would seem that more information about respondent's data is released by publishing the nonsensitive union of cell 1 and cell 2 than by publishing the sensitive total. The inconsistency can be resolved by using a combination of (n, k) rules, such as the (1, 73.91) and (2, 85), or by using a single p-percent rule with p = 35.29 or a pq-rule with q/p = 2.83. These changes result in additional, but more consistent suppressions.

Proponents of the simple (2, 85) rule claim that more protection is needed when respondents have competitors with values close to their own. Proponents of the simple (1, 75) rule claim that more protection is needed if the cell is dominated by a single respondent. These people argue that the use of a simple (n, k) rule allows them to determine which rules are needed for their special situations without the additional suppressions which would result from a more consistent approach.

CHAPTER V – Methods for Public-Use Microdata Files

One method of publishing the information collected in a census or survey is to release a **public-use** microdata file (see Section 2.D). A microdata file consists of records at the respondent level where each record on the file represents one respondent. Each record consists of values of characteristic variables for that respondent. Typical variables for a demographic microdata file are age, race, and sex of the responding person. Typical variables for an establishment microdata file are Standard Industrial Classification (SIC) code, employment size, and value of shipments of the responding business or industry. Most public-use microdata files contain only demographic microdata. The disclosure risk for most kinds of establishment microdata is much higher than for demographic microdata. The reasons for this are explained in Section C.4 of this chapter.

This chapter concerns microdata files that are publicly available, that are **public-use** microdata files. In addition to or instead of public-use files, some agencies offer **restricted-use** microdata files. Access to these files is restricted to certain users at certain locations and is governed by a restricted use agreement.

To protect the confidentiality of microdata, agencies remove all obvious identifiers of respondents, such as name and address, from microdata files. However, there is still a concern that the release of microdata files could lead to a disclosure. Some people and some businesses and industries in the country have characteristics or combinations of characteristics that would make them stand out from other respondents on a microdata file. Public use microdata files contain some measure of risk of disclosing confidential information. A statistical agency releasing a microdata file containing confidential data must do its best to minimize the risk that an outside data user can correctly link a respondent to a record on the file. Aside from not releasing any microdata, there is no way of removing all disclosure risk from a file; however, agencies must make reasonable efforts to minimize this risk and still release as much useful information as possible.

Several Federal agencies including the Census Bureau, National Center for Education Statistics, National Center for Health Statistics, Centers for Medicare and Medicaid Services, Energy Information Administration, Social Security Administration, Bureau of Transportation Statistics, and Internal Revenue Service release microdata files. This chapter describes the disclosure risk associated with microdata files, mathematical frameworks for addressing the problem, and necessary and stringent methods of limiting disclosure risk.

A. Disclosure Risk of Microdata

Statistical agencies are concerned with a specific type of disclosure of personal information that relates to a respondent, and there are several factors that play a role in the disclosure risk of a microdata file. A record is at risk of being identified if a respondent is unique in the database with respect to a set of identifying variables and if the intruder knows that the respondent is on

the file. Data providers that are subject to the privacy rule under the Health Insurance Portability and Protection Act (HIPAA) and/or the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) must take affirmative steps to protect the confidentiality of the reported values before the database is released as a public-use file.

A.1. Disclosure Risk and Intruders

Most national statistical agencies collect data under a pledge of confidentiality. Any violation of this pledge is a disclosure. An outside user who attempts to link a respondent to a microdata record is called an **intruder**. The disclosure risk of a microdata file greatly depends on the motive of the intruder. If the intruder is hunting for the records of specific individuals or firms, chances are that those individuals or firms are not even represented on the file that possesses information about a small sample of the population. In this case, the disclosure risk of the file is very small. The risk is much greater, on the other hand, if the intruder is attempting to match *any* respondent with their record to an external file. We can measure disclosure risk only against a specific compromising technique that we assume the intruder to be using (Keller-McNulty, McNulty, and Unger, 1989).

A.2. Factors Contributing to Risk

There are two main sources of the disclosure risk of a microdata file. One source of risk is the existence of high-risk records. Some records on the file may represent respondents with unique characteristics such as very unusual jobs (e.g. movie star, Federal judge) or very large incomes (e.g. over one million dollars). An agency must decrease the visibility of such records. Another type of high-risk records includes those cases where multiple records in a data file are known to belong to the same cluster (for example, household or school). In this case, there is a greater risk that either one may be identified (even if no information about the cluster per se is provided). A third type of high-risk records can occur when one dimension of the data are released in too fine a level of detail. In this case, if data are released for small areas, such as school districts, variables that would not create disclosure problem at a higher level of aggregation, such as a state or region, may result in an increased risk of disclosure. An example might be, teacher's income by race/ethnicity and age.

The second source of disclosure risk is the possibility of matching the microdata file with external files. There may be individuals or firms in the population that possess a unique combination of the characteristic variables on the microdata file. If some of those individuals or firms happen to be chosen in the sample of the population represented on that file, there is a disclosure risk. Intruders potentially could use external files that possess the same characteristic variables and identifiers to link these unique respondents to their records on the microdata file.

Knowledge of which individuals participated in a survey, or even which areas were in the sample, can greatly help an intruder to identify individuals on a microdata file from that survey. Advising respondents to use discretion when telling others about their past participation in surveys is appropriate but may make respondents wary of participating in the survey. The

disclosure risk of a microdata file is greatly increased if it contains administrative data or any other type of data from an outside source linked to survey data. Those providing the administrative data could use those data to link respondents to their records on the file. This is not to imply that providers of administrative data would attempt to link files, however, it is a possibility and precautions should be taken. In addition, in some cases, the administrative data may be already released as a public use file, so any intruder could use the information to try to identify an individual. The potential for linking files (and thus the disclosure risk) increases as the number of variables common to both files increases, as the accuracy or resolution of the data increases, and as the number and availability of external files increases, not all of which may be known to the agency releasing the microdata file.

Longitudinal and panel surveys create a special case of disclosure risk that may be associated with linked files. In this case, the disclosure risk of a microdata file increases if some records on the file are released on another file with more detailed or overlapping recodes (categorizations) of the same variables. Likewise, risk increases if some records on the file are released on another file containing some of the same variables and some additional variables.

As a corollary, there is greater risk when the statistical agency explicitly links a new microdata file on a set of respondents with published data for those same respondents at an earlier point in time. This occurs in longitudinal surveys, such as the Census Bureau's Survey of Income and Program Participation, where the same respondents are surveyed several times and the NCES high school longitudinal surveys where students are followed for 10 to 12 years through high school, postsecondary education, and into the labor force and/or parenthood. The amount of risk is increased when the data from the different time periods can be linked for each respondent. Changes that an intruder may or may not see in a respondent's record (such as a change in occupation or marital status or a large change in income) over time could lead to the disclosure of the respondent's identity.

In general, the disclosure risk of a file increases as the structure of the data becomes more complex - whether it is through the addition of linked data from an external source, or through the addition of linked data for a set of respondents across time, the effect is the same. More complex variable structure also leads to an increase in the likelihood of unique streams of data responses, and thus an increase in the likelihood of disclosure.

A.3. Factors that Naturally Decrease Risk

Sampling is an important factor in decreasing risk of disclosure in microdata files. As we stated previously, if an intruder possesses such a microdata file and is looking for the record of a specific individual or firm, chances are that that individual or firm is not even represented on the file. Also, records on such a file that are unique compared with all other records on the file may not represent respondents with unique characteristics in the population. There may be several other individuals or firms in the population with those same characteristics that did not get chosen in the sample. This creates a problem for an intruder attempting to link files.

The disclosure risk of the file can be decreased even further if only a subsample of the sampled population is represented on the file. Then, even if an intruder knew that an individual or firm participated in the survey, he or she still would not know if that respondent appeared on the file. Data users, however, generally want the whole sample.

Another naturally occurring factor that decreases the risk of disclosure is the age of the data on microdata files and any potentially matchable external files. When an agency publishes a microdata file, the data on the file are usually at least one to two years old. The characteristics of individuals and firms can change considerably in this length of time. Also, the age of data on potentially matchable files is probably different from the age of the data on the microdata file. One caveat is that the difference in age of the data between files may not complicate the job of linking older files if an intruder has access to an external file that corresponds in time to the data collection.

The naturally occurring noise in the microdata file and in potentially matchable files decreases the ability to link files. All such data files will reflect reporting variability, non-response, and various edit and imputation techniques.

Many potentially matchable files have few variables in common. Even if two files possess the "same" characteristic variables, often the variables are defined slightly differently depending on the purpose for collecting the data. Sometimes the variables on different files are recoded differently. The definitions of any variables that are common to both files should be checked to verify that the definitions are the same, otherwise, the variables may actually be measuring different activity. Differences in variable definitions and recodes can make an intruder's job more difficult.

The final factors that decrease risk are the time, effort, and money needed to link files, although, as computer technology advances, these factors are diminished.

A.4 Disclosure Risks Associated with Regression Models

The question of whether disclosure risks exist in regression-type models has become more important over the past decade as federal agencies expand access to their micro data. The risks associated with public use files have increased due to increased computing power coupled with the development of sophisticated data matching software and the increasing availability of electronic databases on the Internet. At the same time, demand for access to microdata files has increased as the researcher community has recognized the value of the files and increased computing power has made analyzing the files much easier. In response to these developments, agencies have developed several modes of restricted access to data: the U.S. Census Bureau has taken the lead on establishing Research Data Centers (RDCs); NCES has made use of licensing agreements; and NCHS has developed remote access systems for users to access micro data files.

The U.S. National Science Foundation and NCES have jointly funded work by the U.S. National Institutes of Statistical Sciences (NISS) to study issues in developing "model servers," which will

allow researchers to estimate models from databases of confidential microdata without having direct access to the microdata. The NISS researchers have investigated how to release useful results (e.g., regression parameter estimates and model diagnostics) while not compromising confidential information (Gomatam et al, 2005). They have also investigated how to estimate regressions using a combination of confidential data from several sources; e.g., several statistical agencies (Karr et al, 2005).

Disclosure risks may arise from the use of regression models, particularly in the standard linear regression model estimated using Ordinary Least Squares methods as well as in logit and probit models (which use binary (0,1) dependent variables) and other Generalized Linear Models (Reznek 2003, Reznek and Riggs, 2004). The risks in regression models that contain continuous variables on the right-hand side are small if the overall sample is large enough to pass tabular disclosure analysis. However, risks may exist in models that contain dummy variables as independent variables. Coefficients of models that contain only fully-interacted (saturated) sets of dummy variables on the right-hand sides can be used to obtain entries in cross-tabulations of the dependent variable, where the cross-tabulation categories are defined by the dummy variables. The same types of cross-tabulations can also arise from correlation and covariance matrices of the variables, and from variance-covariance matrices of model coefficients, if these matrices include dummy variables. These research outputs present disclosure risks if the cross-tabulations present disclosure risks.

B. Mathematical Methods of Addressing the Problem

Although several mathematical measures of risk have been proposed, none has been widely accepted. Techniques that reduce the disclosure risk of microdata include methods that either reduce the amount of information provided to data users or methods that slightly distort the information provided to data users. Several mathematical measures of the usefulness of disclosure-limited data sets have been proposed to evaluate the trade off between protection and usefulness. Again, none has been widely accepted. More research is necessary to identify the best disclosure limitation methodology sufficient for both data users and suppliers of confidential microdata.

Before describing these mathematical methods of addressing the problem of disclosure risk, we must mention several mathematical and computer science problems that in some way relate to this problem. For example, various mathematical methods of matching a microdata file to an outside file can be found in literature concerning record linkage methodology at http://www.fcsm.gov/working-papers/RLT_1997.html. Record Linkage Techniques, 1997 -- Proceedings of An International Record Linkage Workshop and Exposition presents reprints of the major background papers in record linkage as well as discussions of current work.

B.1. Proposed Measures of Risk

Measuring the disclosure risk of a public use microdata file involves measuring the probability that an intruder is able to identify a record. Most research has considered some or all of the following factors:

- the probability that the respondent for whom an intruder is looking is represented on both the microdata file and some matchable file,
- the probability that the matching variables are recorded identically on the microdata file and on the matchable file,
- the probability that the respondent for whom the intruder is looking is unique in the population for the matchable variables, and
- the degree of confidence of the intruder that he or she has correctly identified a unique respondent.

A model for measuring disclosure risk should reflect certain a prior assumptions about the intruder. The level of risk varies depending upon whether the intruder wishes to disclose the reported values of a particular respondent, or the reported values of any respondent, or a group of respondents. (See Steel, 2004). The validity of the measures of risk depend upon the accuracy of the file preparer's designation of the key variable list. This is a set of variables on the microdata file that may be used to identify unique records in the file and that also exist on data that is in the public domain (or could be held privately from some outside commercial source). A frequency count of the records in the microdata file is usually generated using the key variable list. The most common rule applied in preparing public microdata files is the Threshold rule, or sometimes referred to as the k-anonymity rule. This rule requires a minimum number of records, of at least k records, (usually k=3), that are identical with respect to the specified set of key variables. This is also used as a risk measure in mu-ARGUS, a software product developed by Statistics Netherlands and the Computational Aspects of Statistical Confidentiality (CASC) project. (See Websites in Appendix B for further information on CASC).

The percent of records representing respondents who are unique in the population plays a major role in the disclosure risk of a microdata file. These records are often called **population uniques**. The records that represent respondents who are unique compared with everyone else in the sample are called **sample uniques**. Every population unique is a sample unique, however, not every sample unique is a population unique. There may be other persons in the population who were not chosen in the sample and whom have the same characteristics as a person represented by a sample unique. Statistical Policy Working Paper 2 states that "uniqueness in the population is the real question, and this cannot be determined without a census or administrative file exhausting the population." This corollary remains true for each individual record on a sample microdata file. Several methods of estimating the percent of population uniques on a sample microdata file have been developed. These methods are based on subsampling techniques, the equivalence class structure of the sample together with the hypergeometric distribution, and modeling the distribution of equivalence class sizes (Bethlehem, Keller, and Pannekoek, 1990; Steel, 2004; Winkler, 2004).

A measure of relative risk for two versions of the same microdata file has been developed using the classic entropy function on the distribution of equivalence class sizes (Greenberg and Zayatz, 1992). For example, one version of a microdata file may have few variables with a lot of detail on those variables while another version may have many variables with little detail on those variables. Entropy, used as a measure of relative risk, can point out which of the two versions of the file has a higher risk of disclosure.

B.1.a. MASSC.

Another measure of risk used in the Micro Agglomeration, Substitution, Subsampling, and Calibration (MASSC) disclosure limitation method (discussed later in Section B.3.d) 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. 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. MASSC computes four measures of risk to generate an upper bound measure of disclosure risk for a target record, stratum, or file. A measure of disclosure risk is calculated based on whether the target looks like a unique, a non-unique double, a non-unique triple, or a non-unique-four-plus, i.e., a non-unique cluster size of four records or more. An overall measure of the target is generated by taking a weighted average of the four disclosure risk measures where the weights are the relative proportion of each type of record in the adjusted database. By collapsing over the strata, a disclosure risk can be calculated for an entire database as well as an individual record.

B.1.b. R-U Confidentiality Map.

This approach attempts to measure the simultaneous impact on disclosure risk and data utility of applying a specific disclosure limitation technique and can serve as a tool by a data provider for choosing the appropriate parameter value. 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, the gains, if any, in 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 Map can be constructed for different disclosure limitation techniques and serve as a useful tool in applying a specific disclosure limitation methodology. (Duncan, McNulty, and Stokes, 2001)

B.2. Methods of Reducing Risk by Reducing the Amount of Information Released

Recoding variables into categories is one commonly used way of reducing the disclosure risk of a microdata file (Skinner, 1992). The resulting information in the file is no less accurate, but it is less precise. This reduction in precision reduces the ability of an intruder to correctly link a respondent to a record because it decreases the percent of population uniques on the file. Recoding variables can also reduce the high risk of some records. For example, if occupation is on the file in great detail, a record showing an occupation of United States Senator in combination with a geographic identifier of Delaware points to one of two people. Other variables on the file would probably lead to the identification of that respondent. Occupation could be recoded into fewer, less discriminatory categories to alleviate this problem.

If an agency is particularly worried about an outside, potentially matchable file, the agency may recode the variables common to both files so that there are no unique variable combinations on the microdata file, thus preventing one-to-one matches. For example, rather than release the complete date of birth, an agency might publish only year of birth. Rounding values, such as rounding income to the nearest one thousand dollars, is also a form of recoding.

Another commonly used way of reducing the disclosure risk of a file is through setting top-codes and/or bottom-codes on continuous variables (see Section II.D.2). A **top-code** for a variable is an upper limit on all published values of that variable. Any value greater than this upper limit is not published on the microdata file. In its place is some type of flag that tells the user what the top-code is and that this value exceeds it. For example, rather than publishing a record showing an income of $2,000,000, the record may only show that the income is > $150,000. Similarly, a **bottom-code** is a lower limit on all published values for a variable. Top- and bottom-coding reduce the high risk of some records. Examples of top-coded variables might be income and age for demographic microdata files and value of shipments for establishment microdata files. If an agency published these variables on a microdata file with no top-coding, there would probably be a disclosure of confidential information. Examples of bottom-coded variables might be year of birth or year built for some particular structure.

Recoding and top-coding obviously reduce the usefulness of the data. However, agencies could provide means, medians, and variances of the values in each category and of all top-coded values to data users to compensate somewhat for the loss of information. Also, recoding and top-coding can cause problems for users of time series data when top-codes or interval boundaries are changed from one period to the next.

B.3. Methods of Reducing Risk by Disturbing Microdata

Since Statistical Policy Working Paper 2 was published, researchers have proposed and evaluated several methods for disturbing microdata in order to limit disclosure risk. These techniques, described in Chapter II, slightly alter the data in a manner that hinders an intruder who is trying to match files.

Probably the most basic form of disturbing continuous variables is the addition of, or multiplication by, random numbers with a given distribution This **noise** may be added to the data records in their original form or to some transformation of the data depending on the intended use of the file. Probability distributions can be used to add error to a small percent of categorical values. An agency must decide whether or not to publish the distribution(s) used to add noise to the data. Publishing the distribution(s) could aid data users in their statistical analyses of the data but might also increase disclosure risk of the data. Another proposed method of disturbing microdata is to randomly choose a small percent of records and blank out a few of the values on the records (see Section II.D.5). Imputation techniques are then used to impute for the values that were blanked.

B.3.a. Data Swapping

Swapping (or **switching**) and **rank swapping** are two proposed methods of disturbing microdata. The purpose of any swapping methodology is to introduce uncertainty so that the data user doesn't know whether real data values correspond to certain records. Records with a high risk of disclosure are usually selected for swapping. In the swapping procedure, a small percent of records are matched with other records in the same file, perhaps in different geographic regions, on a set of predetermined variables that are used as swapping attributes. The values of variables used as swapping attributes in the file are then swapped between the two records. In the rank swapping procedure, values of continuous variables are sorted and values that are close in rank are then swapped between pairs of records. As the percentage of swapped records increases, the greater the losses in data utility of the microdata file. Although swapping does not change the marginal distribution of any variable in a file, it does distort joint distributions involving both swapped and unswapped variables.

B.3.b. Data Shuffling

Data Shuffling is another data masking procedure that has been successfully 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. This method preserves the rank order correlation between the confidential and non-confidential attributes, thereby maintaining monotonic relationships between attributes.

Before the data are perturbed, the non-confidential variables (**S**) and confidential variables (**X**) on the file are identified. The conditional distribution of $f(\mathbf{X}|\mathbf{S} = \mathbf{s_i})$ between the confidential and non-confidential variables is then derived. For i = 1 to n, generate a vector $\mathbf{y_i}$ from $f(\mathbf{X}|\mathbf{S} = \mathbf{s_i})$. The perturbed values of **Y** are the collection of the values $\mathbf{y_i}$ (i = 1, 2, …n).

The shuffling of data records occurs after the values for the confidential variable have been perturbed and ranked. For each confidential variable let $Y = (y_1, y_2, \ldots, y_n)$ represent the perturbed values of the confidential variable $X = (x_1, x_2, \ldots, x_n)$. Let $X^j = (x^1, x^2, \ldots, x^n)$ represent the rank ordered values of X.

For i = 1 to n: Find the rank of y_i. Let this rank be k. Replace the value of y_i by x^k.
In the example below, notice that the rank of the first perturbed observation is 17. The value of the 17^{th} ordered value of X (x^{17}) = 42.79. Hence, the first perturbed observation is replaced by 42.79. Similarly, the rank of observation y_2 is 16 and is replaced by x^{16} = 41.74, the value of X at the 16^{th} rank of X. The process is repeated for every perturbed observation until the all of the perturbed values are replaced with original values from the confidential variable.

Example Data Set

ID#	S	X	Rank of X	Perturbed Y	Rank of Perturbed Y	Shuffled Y
1	41	54.24	27	43.8024	17	42.79
2	53	52.98	25	43.7608	16	41.74
3	40	33.77	4	31.2382	3	32.54
4	51	43.15	18	41.6440	13	40.41
5	37	48.70	22	36.3746	8	36.94
6	41	41.74	16	43.6570	15	40.77
7	24	36.00	7	46.5293	20	46.80
8	57	48.06	21	51.1033	23	48.76
9	52	57.69	29	54.3518	28	55.21
10	27	34.14	5	42.1101	14	40.72
11	39	32.54	3	40.6861	11	38.79
12	54	55.21	28	48.5196	22	48.70
13	52	40.77	15	53.7893	26	53.19
14	47	48.76	23	41.5140	12	39.50
15	41	27.52	1	44.6543	19	45.35
16	52	50.36	24	40.2965	10	38.68
17	20	42.79	17	34.6577	6	35.43
18	42	39.50	12	40.1456	9	38.05
19	52	53.19	26	51.5981	24	50.36
20	45	40.72	14	32.4994	4	33.77
21	52	38.68	10	47.7596	21	48.06
22	42	46.80	20	32.9835	5	34.14
23	50	59.08	30	44.4699	18	43.15
24	48	32.28	2	51.8446	25	52.98
25	33	36.94	8	35.7985	7	36.00
26	50	38.05	9	54.5523	29	57.69
27	46	40.41	13	25.2914	1	27.52
28	43	38.79	11	54.1997	27	54.24
29	56	45.35	19	54.7677	30	59.08
30	41	35.43	6	29.0405	2	32.28

The marginal distribution of the masked (Shuffled Y) variable is the same as that of the original variable X and the product moment correlation (linear relationships) and rank order correlation (non-linear monotonic relationships) are not disturbed. In the example provided the correlation between (S and X) is 0.4507 and that between (Shuffled Y and S) is 0.4474. The rank order correlation between (S and X) is 0.52 and that between (Shuffled Y and S) is 0.54. These estimates will approach each other as the size of the data set increases.

B.3.c. Data Blurring and Microaggregation

Blurring involves aggregating values across small sets of respondents for selected variables and replacing a reported value (or values) by the aggregate. Different groups of respondents may be formed for different data variables by matching on other variables or by sorting the variable of interest (see Section II.D.6). Records are placed in groups of size k, where k is commonly set between 3 and 10 and the original values associated with sensitive variables are replaced with the aggregate value. Data may be aggregated across a fixed number of records, a randomly chosen number of records, or a number determined by (n, k) or p-percent type rules as used for aggregate data. For a definition of the (n, k) and p-percent rules, see Chapter IV. The aggregate associated with a group may be assigned to all members of the group or to the "middle" member (as in a moving average). Aggregating over groups of 3 records or less may not be sufficient for reducing the risk of disclosure, especially if the blurring is performed on only one or two variables in a file. As the size of the group of records increases, the chance of re-identification is reduced. If the grouping is larger than 10 records, there may be greater distortion introduced into the microdata file which may lead to inaccurate published data. **Microaggregation** is a form of data blurring where records are grouped based on a proximity measure of all variables of interest, and the same groups of records are used in calculating aggregates for those variables. Blurring and microaggregation may be done in a way to preserve variable means. However, single variable data blurring or microaggregation may lead to re-identification and therefore should be combined with other disclosure limitation techniques to provide adequate data protection.

Another proposed disturbance technique involves super and subsampling (Cox and Kim, 1991). The original data are sampled with replacement to create a file larger than the intended microdata file. Differential probabilities of selection are used for the unique records in the original data set, and record weights are adjusted. This larger file is then subsampled to create the final microdata file. This procedure confuses the idea of sample uniqueness. Some unique records are eliminated through non-selection, and some no longer appear to be unique due to duplication. Some non-unique records appear to be unique due to nonselection of their clones (records with the same combination of values). Biases introduced by this method could be computed and perhaps released to users as a file adjunct.

B.3.d. Micro Agglomeration, Substitution, Subsampling, and Calibration (MASSC)

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 in preparation for the level of modification to the data to reduce the risk of disclosure. Some recoding of variables is done during this phase. 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 records using certain key variables to preserve the domain counts from the original dataset. The calibration step reduces bias due to the substitution and it reduces variance due to the subsampling step. In the methodology, every record in the database is subject to modification or swapping, however, when applying this methodology, only a small random portion of the records are actually modified. (Yu, Dunteman, Dai, and Wilson, 2004).

B.4. Methods of Reducing Risk by Using Simulated Microdata

B.4.a. Latin Hypercube Sampling.

Latin Hypercube Sampling (LHS) is another technique that involves creating a replacement file containing replacement values for the sensitive variables in the microdata file. The LHS method ensures that the synthetic data set has nearly the same univariate statistical characteristics of the original data such as mean, standard deviation and coefficient of skew ness. LHS can be used to generate a synthetic data set for a group of uncorrelated variables. In the case where 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. (Dandekar, Cohen, Kirkendall, 2001). Latin Hypercube Sampling provides one method of using multiple imputation techniques to produce a set of pseudo-data with the same specified statistical properties as the true microdata.

B.4.b. Inference-Valid Synthetic Data.

Another variation in the use of synthetic data for releasing public use data files is by drawing samples from the posterior predictive distribution of the adjusted confidential data. In this approach, the actual confidential variable(s) in the micro data 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 data can then 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's plus the residuals for the relevant confidential variable, i.e., the posterior predictive distribution, various micro datasets can be created and the statistical inferences from the synthetic data are valid 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 to different groups of users. The inference valid synthetic data methodology was applied to the Survey of Income and Program Participation (SIPP) data after the SIPP data was linked to earnings data from the Social Security Administration. (Abowd and Lane, 2003).

B.4.c. The FRITZ Algorithm for Disclosure Limitation.

The Federal Reserve Imputation Technique Zeta (FRITZ) system is used for both missing value imputation and disclosure limitation in the Survey of Consumer Finances (SCF). The FRITZ model reviews the data along a sequential predetermined 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 cases are selected for protection. All dollar values in the SCF are set to missing and the FRITZ algorithm is applied to generate imputed values. The subsequent analysis of this methodology indicates that while the imputations provided the protection to the sensitive individual records, it had only minimal effects on the distributional characteristics of the file (Kennickell, 1998).

B.5. Methods of Analyzing Disturbed Microdata to Determine Usefulness

There are several statistical tests that can be performed to determine the effects of disturbance on the statistical properties of the data. These include the Kolmogorov-Smirnov 2-sample test, Fischer's z-transformation of the Pearson Correlations, and the Chi-Square approximation statistic to the likelihood ratio test for the homogeneity of the covariance matrices.

These procedures are mainly conducted to see if the means and the variance-covariance and correlational structure of the data remain the same after disturbance. Even if these tests come out favorably, disturbance can still have adverse effects on statistical properties such as means and correlational structure of subsets and on time series analyses of longitudinal data. If an agency knows how the file will be used, it can disturb the data in such a way that the statistical properties pertinent to that application are maintained. However, public-use files are available to the entire public, and they are used in many ways. Levels of disturbance needed to protect the data from disclosure may render the final product useless for many applications. For this reason, agencies limit the amount of modification to the data in the microdata file, or attempt to limit disclosure risk by limiting the amount of information in the microdata files. Disturbance may be necessary, however, when potentially linkable files are available to users, and recoding efforts do not eliminate population uniques.

C. Necessary Procedures for Releasing Microdata Files

Before publicly releasing a microdata file, a statistical agency must attempt to preserve the usefulness of the data, reduce the visibility of respondents with unique characteristics, and ensure that the file cannot be linked to any outside files with identifiers. While there is no method of completely eliminating the disclosure risk of a microdata file, agencies should perform the following steps before releasing a microdata file to limit the file's potential for disclosure. Statistical agencies have used most of these methods for many years. They continue to be important.

C.1. Removal of Identifiers

Obviously, an agency must purge a microdata file of all direct personal and institutional identifiers such as name, address, Social Security number, and Employer Identification number. An internal file with the names or other direct identifiers removed may still be at risk of **indirect disclosure**, if sufficient data are left on the file with which to match with information from an external source that *also contains names or other direct identifiers.* In such a case, the identity, as well as all information in the file associated with that person or establishment will be disclosed if the file is released without further modifications.

C.2. Limiting Geographic Detail

The match does not need to be exact. An intruder could link the characteristics of all respondents with the same sample unit with similar information from an external source of data. Other variables on a file may cause an indirect disclosure problem if they could be used to distinguish a small geographic unit on the basis of certain socioeconomic characteristics. Once an individual or establishment’s records are associated with a small geographic area, the possibility of identification is greatly increased. Geographic location is a characteristic that appears on most microdata files. Agencies should give geographic detail special consideration before releasing a microdata file because it is much easier for an intruder to link a respondent to the respondent's record if the intruder knows the respondent's city, for example, rather than if he or she only knows the respondent's state.

Based on these considerations, the Census Bureau does not identify any geographic region with less than 100,000 persons in the sampling frame. A higher cut-off is used for surveys with a presumed higher disclosure risk. Microdata files from the Survey of Income and Program Participation, for example, still have a geographic cut-off of 250,000 persons per identified region. Agencies releasing microdata files should set geographic cut-offs that are simply lower bounds on the size of the sampled population of each geographic region identified on microdata files. This is easier said than done. Decisions of this kind are often based on precedents and judgment calls. More research is needed to provide a scientific basis for such decisions (Zayatz, 1992a).

Some microdata files contain contextual variables. Contextual variables are variables that describe the area in which a respondent or establishment resides but do not identify that area. In general, the areas described are smaller than areas normally identified on microdata files. Care must be taken to ensure that the contextual variables do not identify areas that do not meet the desired geographic cut-off. An example of a contextual variable that could lead to disclosure is average temperature of an area. The Energy Information Administration adds random noise to temperature data (because temperature data are widely available) and provides an equation so the user can calculate approximate heating degree-days and cooling degree-days (important for regression analysis of energy consumption).

C.3. Top-Coding High Risk Variables That Are Continuous.

The variables on microdata files that contribute to the high risk of certain respondents are called **high-risk variables**. Examples of continuous high-risk variables are income and age for demographic microdata files and value of shipments for establishment microdata files. As stated previously, if an agency published these variables on a microdata file with no top-coding, there would probably be a disclosure of confidential information. For example, intruders could probably correctly identify respondents who are over the age of 100 or who have incomes of over one million dollars.

Appropriate top-codes (and/or bottom-codes in some cases) should be set for all of the continuous high-risk variables on a microdata file. Top-coded records should then only show a representative value for the upper tail of the distribution, such as the cut-off value for the tail or the mean or median value for the tail, depending on user preference. Angle (2003) developed a methodology for estimating the distribution of top coded values using a distribution more general than the traditional Pareto, and illustrates it using annual wage and salary income. The model's estimate of the right tail truncated by top-coding has been shown to have many of the dynamics of the right tails of empirical annual wage and salary income distributions. This methodology uses a probability density function model for generating the right tail of an income distribution that has been truncated by top-coding. The model's parameters are estimated in the fit of the model to data below the cutoff for top-coding. The model's right tail is used in the estimation of statistics of the whole distribution. The model is able to generate the distribution of top coded values even after lowering the threshold level for minimum top-codeable annual wage and salary income well below the 99^{th} percentile. (Angle, 2003).

C.4. Precautions for Certain Types of Microdata

There are certain types of microdata that may raise the risk of disclosure when reviewing a file for release.

C.4.a. Establishment Microdata

Most microdata files that are publicly released contain demographic microdata. It is presumed that the disclosure risk for establishment microdata is higher than that for demographic microdata. Establishment data are typically skewed, the size of the establishment universe may be small, and there are many high-risk variables on potential establishment microdata files. Industry publications and trade associations may also exist and function as outside sources of information for a data user. Publicly available administrative databases may also be available for matching to the establishment microdata files and create additional disclosure risks. Also, there are a large number of subject matter experts and many possible motives for attempting to identify respondents on some types of establishment microdata files. For example, there may be financial incentives associated with learning something about the competition. Agencies should take into account all of these factors when considering the release of an establishment microdata file.

C.4.b. Longitudinal Microdata

There is greater risk when the microdata on a file are from a longitudinal survey where the same respondents are surveyed several times. Risk is increased when the data from the different time periods can be linked for each respondent because there are much more data for each respondent and because changes that may or may not occur in a respondent's record over time could lead to the disclosure of the respondent's identity. Agencies should take this into account when considering the release of such a file. One piece of advice is to plan ahead. Releasing a first cross-sectional file without giving any thought to future plans for longitudinal files can cause unnecessary problems when it comes to releasing the latter. The entire data collection program should be considered in making judgments on the release of public-use microdata.

C.4.c. Microdata Containing Administrative Data

The disclosure risk of a microdata file is increased if it contains administrative data or any other type of data from an outside source linked to the survey data. Those providing the administrative data could use that data to link respondents to their records. This is not to imply that providers of administrative data would attempt to link files, however, it is a theoretical possibility and precautions should be taken. At the very least, some type of disturbance should be performed on the administrative data or the administrative data should be categorized so there exists no unique combination of administrative variables. This reduces the possibility that an intruder can link the microdata file to the administrative file. There are concerns that agencies should not release such microdata at all or should release it only under a restricted access agreement.

C.4.d. Consideration of Potentially Matchable Files and Population Uniques

Statistical agencies must attempt to identify outside files that are potentially matchable to the microdata file in question. Comparability of all such files with the file in question must be examined. The Census Bureau uses re-identification and record linkage experiments to

determine if their files are matchable to outside files on a certain set of key variables. The National Center for Education Statistics matches microdata files under consideration for release to commercially available school files to identify unique matches. Re-identification of microdata refers to the ability to use public available information to attach names, addresses, and other partially unique identifiers to individual records in a public-use file. An identifier is partially unique if it can be used in conjunction with other variables to re-identify a record even though it may not exactly identify a linkage between two records by itself. Record linkage software has been developed to handle a large variety of both minor and major spelling variations and errors in the variables used in the matching process.

Another measure of the risk of re-identification for a file is the number or proportion of population uniques, where consideration is restricted to those variables thought to be available on external files. Statistical models have been developed that relate the distribution of the sample uniques in a file to the distribution of the population uniques. However, these models only provide an estimate for the percentage of sample uniques that are true population uniques. This estimate tends to have a high variance and estimating the percentage doesn't provide any guide to determining which uniques are artifacts of sampling and which are population uniques. Record linkage experiments also can provide a measure of re-identification risk, but are heavily dependent on acquiring or modeling external data sources (Winkler 2004). A record linkage experiment may identify some population uniques, but should not be considered as an assurance that all risky records have been discovered.

D. Stringent Methods of Limiting Disclosure Risk

There are a few procedures that can be performed on microdata files prior to release that severely limit the disclosure risk of the files such as data swapping and data coarsening. One must keep in mind, however, that the usefulness of the resulting published data will also be extremely limited. The resulting files will contain either much less information or information that is inaccurate to a degree that depends on the file and its contents.

D.1. Do Not Release the Microdata

One obvious way of eliminating the disclosure risk of microdata is to not release the microdata records. The statistical agency could release only the variance-covariance matrix of the data or perhaps a specified set of low-order finite moments of the data. This greatly reduces the usefulness of the data because the user receives much less information and data analyses are restricted.

D.2. Recode Data to Eliminate Uniques

Recoding the data in such a way that no sample uniques remain in the microdata file is generally considered a sufficient method of limiting the disclosure risk of the file. A milder procedure allowing for broader categorization--recoding such that there are no population uniques--would suffice. Recoding the data to eliminate either sample or population uniques would likely result in

very limited published information.

D.3. Disturb Data to Prevent Matching to External Files

Showing that a file containing disturbed microdata cannot be successfully matched to the original data file or to another file with comparable variables is generally considered sufficient evidence of adequate protection. Several proximity measures should be used when attempting to link the two files. An alternative demonstration of adequate protection is that no exact match is correct or that the correct match for each record on a comparable file is not among the K closest matches. Microaggregation or data shuffling could be used to protect data, perhaps using (n, k) or p-percent type rules as used for tables. In this way, no individual data are provided, and intruders would be prevented from matching the data to external files. See Chapter IV for a definition of the (n, k) and p-percent rules. Microaggregation, data blurring, and other methods of disturbance that hinder file matching, however, may cause distortions in published data. Taken to a degree that would absolutely prevent matching, the methods would usually result in greatly distorted published information.

E. Conclusion

Public-use microdata files are used for a variety of purposes. Any disclosure of confidential data on microdata files may constitute a violation of the law or of an agency's policy and could hinder an agency's ability to collect data in the future. Short of releasing no information at all, there is no way to completely eliminate disclosure risk. However, there are techniques that, if performed on the data prior to release, should sufficiently limit the disclosure risk of the microdata file. Research is needed to understand better the effects of those techniques on the disclosure risk and on the usefulness of resulting data files (see Section VI.A.2).

CHAPTER VI – Recommended Practices For Federal Agencies

A. Introduction

Based on its review of current agency practices and relevant research, the Confidentiality and Data Access Committee (CDAC), a subcommittee of the FCSM, developed a set of recommendations for disclosure limitation practices. The implementation of these practices by federal agencies will result in an overall increase in disclosure protection and will improve the understanding and ease of use of federal disclosure-limited data products. Sometimes the methods used to reduce the risk of disclosure make the data unsuitable for statistical analysis (for example, as mentioned in Chapter V, recoding can cause problems for users of time series data when top-codes are changed from one period to the next). In deciding what statistical procedures to use, agencies also need to consider the usefulness of the resulting data product to data users.

The first set of recommendations in Section B.1 is general and pertains to both tables and microdata. Section B.2 describes CDAC recommendations for tables of frequency data. Recommendations 7 to 11 in Section B.3 pertain to tables of magnitude data. Lastly, Recommendations 12 and 13 in Section B.4 pertain to microdata.

B. Recommendations

B.1. General Recommendations for Tables and Microdata

Recommendation 1: Seek Advice from Respondents and Data Users. In order to plan and evaluate disclosure limitation policies and procedures, agencies should consult with both respondents and data users. Agencies should seek a better understanding of how respondents feel about data disclosure risks, data sharing across agencies, the availability of matching to external administrative data files, and data protections under CIPSEA and non-CIPSEA surveys.

Similarly, 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.

Recommendation 2: Standardize and Centralize Agency Review of Disclosure-Limited Data Products. It is important that disclosure limitation policies and procedures of individual agencies be internally consistent. Results of disclosure limitation procedures should be reviewed. Agencies should standardize the review process by adopting standards and/or guidelines on protecting data confidentiality. The "Checklist on Disclosure Potential of Proposed Data Releases" available at http://www.fcsm.gov/committees/cdac/ should be used as a guide for this review process. The checklist should be modified to suit the agency's data release policy and procedures. Agencies should also centralize responsibility for this review in the organizational

structure through mechanisms such as disclosure review boards (permanent or ad hoc), or a confidentiality officer, review panel, or group of staff knowledgeable and experienced in the area of disclosure limitation procedures and data confidentiality protection.

CDAC recommends that agencies become familiar with external databases that are available to users for matching to agency data products. They should evaluate any proposed data release both in terms of disclosure risks internal to the variables and values inside the file and in terms of external risks of disclosure from potential matching to external files. In agencies with small or single programs for microdata release, this may be assigned to a single individual knowledgeable in statistical disclosure limitation methods and agency confidentiality policy. In agencies with multiple or large programs, a review panel should be formed with responsibility to review each microdata file proposed for release and determine whether it is suitable for release. Review panels should be: as broadly representative of agency programs as is practicable; knowledgeable about disclosure limitation methods for microdata; prepared to recommend and facilitate the use of disclosure limitation methodologies by program managers, and should be empowered by their agency to verify that disclosure limitation techniques have been properly applied.

Tabular data products of agencies should also be reviewed. Disclosure limitation and suppression should be an auditable and replicable process. (Disclosure limitation for microdata is not currently at the stage where a similar approach is feasible.) There are administrative efficiencies for centralizing the review of both micro data files and table files. Depending upon institutional size, programs, and culture, an agency should combine the review of microdata and tables in a single individual, review panel or office.

Recommendation 3: Share Software and Methodology Across the Government.
Federal agencies should share software products for disclosure limitation and record linkage, as well as methodological and technical advances. Specifically, CDAC should continue to make software for disclosure limitation methodologies and documentation available from its website to the federal agencies and public for their use. Software should be written in a common processing language that is easily modifiable with clear documentation.

As advances are made in software for statistical disclosure limitation and record linkage by academia, government, and private businesses, CDAC should evaluate these new methodologies and software, and provide guidance to the federal agencies on the practical and appropriate applications for their use. CDAC has available on its website at http://www.fcsm.gov/committees/cdac/ software which performs primary and complementary suppression, and suppression auditing software which reviews and generates a report indicating the extent of the protection applied from the suppression pattern used for a table.

Recommendation 4: Formal Interagency Cooperation is Needed for Data Sharing. Sharing data files between agencies requires formalized agreements between agencies in order to safeguard data confidentiality protections and meet an agency's legal obligations for collecting and publishing information. 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 other factors that contribute to the need for interagency cooperation. Interagency panels or teams may be needed to plan and review data sharing activities between agencies. Interagency cooperation on reviewing overlapping data sets and the use of identical disclosure limitation procedures is encouraged. Agencies should expand the shared use of research data centers as a method for increasing access to confidential data by researchers. Agencies may also consider requesting representatives from other agencies that have more experience with releasing public use micro data files to serve on disclosure review boards so that knowledge and experience across agencies may be shared.

Recommendation 5: Use Consistent Practices. Agencies should strive to employ disclosure limitation methods in standard ways and be consistent in defining categories in different data products and over time. They should standardize variable definitions internally to the extent it meets the agency's program needs and common definitions between agencies should be developed where possible. Such practices will improve data access by the public and make it easier to implement disclosure limitation methodologies. Examples include using consistent schemes for combining categories, establishing standardized practices for similar data such as categorizing or top-coding variables like age or income, and moving towards standardized application of minimum geographic size limitations for household data. Software should be developed, made broadly available and used to implement these methods to assure both consistency and correct implementation.

B.2. Tables of Frequency Count Data

Recommendation 6: Research is Needed to Compare and Evaluate Methods. There has been considerable research into disclosure limitation methods for tables of frequency data. The most common method used is suppression. Besides suppression, other well-developed methods that are available include controlled rounding, controlled tabular adjustment, and applying data perturbation methods prior to tabulation. Additional research is needed to apply these methods to different types of data and compare and evaluate these different methods in terms of data protection and usefulness of the resulting data product. If suppression is used, the guidelines listed in Recommendations 9 and 10 also apply to tables of frequency data.

B.3. Tables of Magnitude Data

Recommendation 7: Use Only Subadditive Disclosure Rules For Identifying Sensitive Cells. Agencies should develop and apply operational linear sensitivity rules (See Chapter 4) to identify and then protect primary disclosure cells. Disclosure rules that have the mathematical property of **subadditivity** provide assurance that a cell formed by the combination of two non-sensitive cells remains non-sensitive. Agencies should employ only subadditive primary disclosure rules. The p-percent, pq, N, and (n, k) rules are all subadditive. **Primary disclosure cells** must be protected using disclosure limitation techniques.

Recommendation 8: The p-Percent or pq-Ambiguity Rules are Preferred. The p-percent and pq-ambiguity rules are recommended because the use of a single (n, k) rule is inconsistent in the amount of information allowed to be derived about respondents (see Chapter IV). The p-percent and pq rules do provide consistent protection to all respondents. In particular, the pq rule should be used if an agency can quantify the extent that data users already know something about respondent values. If, however, an agency feels that respondents need additional protection from close competitors within the same cells, they might use the p-percent or pq rule in conjunction with an (n, k) rule. When using only the (n, k) rule, a sequence of (n, k) rules is better than a single set of parameters. An example of a sequence of (n, k) rules is (1,75) and (2,85). When a combination of (n, k) rules is applied, a cell is sensitive if it violates either rule.

Recommendation 9: Do Not Reveal Suppression Parameters. To facilitate releasing as much information as possible at acceptable levels of disclosure risk, agencies are encouraged to make public the kind of rule they are using (e.g. a p-percent rule) but they should not make public the specific value(s) of the disclosure limitation rule (e.g., the precise value of "p" in the p-percent rule) since such knowledge can reduce disclosure protection. (See Chapter 4 Section B.4 for an illustration of how knowledge of both the rule and the parameter value can enable the user to infer the value of the suppressed cell.) The value of the parameters used for statistical disclosure limitation can depend on programmatic considerations such as the sensitivity of the data to be released.

Recommendation 10: Redesign Tables, Apply Cell Suppression, Controlled Tabular Adjustment, or Perturbation Methods to Microdata Prior to Tabulation There are four methods of limiting disclosure in tables of magnitude data. First, for single tables or sets of tables that are not related hierarchically, agencies may limit disclosure by combining rows and/or columns. Second, for more complicated tables, cell suppression may be used to limit disclosure. Third, controlled tabular adjustment may be applied to protect sensitive cells after tabulation. Fourth, sensitive cells may be protected prior to tabulation by applying some perturbation method that adds noise to the underlying microdata.

Suppression is widely used by the federal agencies. Cell suppression removes from publication (suppresses) all cells that represent disclosure, together with other, nondisclosure cells that could be used to recalculate or narrowly estimate the primary, sensitive disclosure cells. Zero cells are often easily identified and should not be used as complementary suppressions. The suppression patterns should be audited to check whether the algorithms that select the complementary suppression pattern permit estimation of the suppressed cell values within “too close” of a range. Suppression methods should provide protection with minimum data loss as measured by an appropriate criterion such as minimum number of suppressed cells or minimum total value suppressed. If the information loss from cell suppression undermines the utility of the data, other methods may more useful.

Controlled tabular adjustment applied to tables and perturbation methods applied to microdata prior to tabulation eliminate the information loss associated with suppression. One cautionary note is that both methodologies may not provide sufficient protection to a cell that has one

respondent or a cell that is dominated by one respondent. There may also be some inferential loss in information from changing the data. The interrelationship between tables also needs to be checked to minimize any adjustments to cells in other tables or set of tables should be reviewed to check if any of the table(s)' analytical properties have been distorted or limited. These recommended practices also apply if suppression is used for tables of frequency count data.

Recommendation 11: If Applying Cell Suppression, Auditing of Tabular Data is a Necessity. Tables where suppression is applied to protect sensitive cells should be audited to assure that the values in suppressed cells may not be derived by manipulating row and column equations. This recommendation applies to both tables of frequency data and magnitude data.

B.4. Microdata

Recommendation 12: Remove Direct Identifiers and Limit Other Identifying Information From Microdata Files. The challenge of applying statistical disclosure methods to microdata is to thwart the identification of a respondent from data appearing on a record while allowing release of the maximum amount of data. The ability to match variables from external files generates additional disclosure risks that expand the list of variables on a file that need to be reviewed. The first step to protect the respondent's confidentiality is to remove from the microdata file all **direct identifying information** such as name, social security number, exact address, or date of birth. Certain univariate information such as occupation or precise geographic location can also be identifying. Other univariate information such as a very high income or presence of a rare disease can serve both to identify a respondent and disclose confidential data. These data should also be removed or protected. Agencies should also continue to identify univariate data that tend to facilitate identification or represent disclosure, and set limits on how this information is reported. For example, the Census Bureau presents geographic information only for areas of 100,000 or more persons. Income and other information may be top-coded to a predetermined value such as the 99th percentile of the distribution. Lastly, appropriate distributions and cross tabulations should be examined to ensure that individuals are not directly identified. Circumstances can vary widely between agencies or within an agency between microdata files.

After direct identifiers have been removed, a file may still remain identifiable, if sufficient data are left on the file with which to match with information from an external source that also contains names or other direct identifiers. For this reason, agencies should perform re-identification studies and attempt to match variables on the released files to external files outside of the agency.

Recommendation 13: Agencies Need to Share Information on Assessing Disclosure Risks. Agencies need to share information on what external files that 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.

GLOSSARY

Attribute disclosure – A disclosure that reveals sensitive information about a data subject.

Audit – Check a proposed suppression pattern to make sure sensitive cells are adequately protected.

Bottom-coded – Replacing values below a certain number or percentile ranking with the same value.

Complementary suppression – Withholding non-sensitive cells from release in order to protect other sensitive cells from disclosing confidential information.

Confidential Information – information reported under an expectation that the information will not be released in a manner that allows public identification of the respondent or causes some harm to a respondent.

Disclosure – revealing information that relates to the identity of a data subject, or some sensitive information about a data subject through the release of either tables or microdata.

Frequency count data – Data that show the number of units of analysis in a cell.

Hierarchy – A series of items organized or classified according to rank or order; especially a ranked classification schema used to structure a table or microdata file such as NAICS codes.

High risk – information that has a high probability of being used to either identify a respondent or reveal confidential information about the respondent.

Identifiable form – Any representation of information that permits the identity of the respondent to whom the information applies to the reasonably inferred by either indirect or indirect means.

Inferential disclosure – A disclosure that makes it possible to determine the value of some characteristic of any individual more accurately than otherwise would have been possible.

Identity disclosure – A disclosure that identifies a data subject.

Informed consent – Written permission from a respondent to publish sensitive cell values. It is has the effect of acting as a waiver of the promise to protect sensitive cells and specific authorization or consent to the agency for public releasing he confidential information.

Intruder - An outside user who attempts to link a respondent to a microdata record.

Linear sensitivity measure – A rule that indicates how close a respondent's data may be estimated from a published cell value.

Magnitude data – Data that show the aggregate of a "quantity of interest" that applies to units of analysis in the cell.

Primary suppression rules – A linear combination of respondent level data that is used to determine whether a given table cell could reveal individual respondent information.

Primary suppression – Withholding from publication any cells that are identified as being by a primary suppression rule.

Public-use – Data products that are released by statistical agencies to anyone without restrictions on use or other conditions, except for payment of fees to purchase data in electronic form.

Restricted Data – Adjusting data in released tables and microdata files or limiting the amount of information released.

Restricted Access – Imposing terms and conditions on users' access to the data products.

Sample – A set of records or data elements drawn from a population and used to estimate the characteristics of a population.

Sensitive – A classification of a cell value established by using a primary suppression rule.

Suppression – Withholding information in selected table cells from release.

Subaddivity – The property that the union of two non-sensitive cells is also non-sensitive.

Tabular Data – Data presented in tables.

Three-dimensional table – A table containing aggregate cell values over three variables.

Top-coded – Replacing values above a certain percentile ranking with the same value.

Two-dimensional table – A table containing aggregate cell values over two variables.

APPENDIX A – Technical Notes: Extending Primary Suppression Rules To Other Common Situations

This appendix contains practices the statistical agencies have found useful when applying disclosure limitation to tables in common situations. The primary and complementary suppression procedures for tables of magnitude data discussed in Chapter IV are based on the assumption that the reported data are strictly positive, and that the published number is the simple sum of the data from all respondents. In some situations published data are not simple sums, and it is not clear how to apply primary and complementary suppression methodology. For example, in this appendix we extend primary suppression rules used for tabular data to tables containing imputed data.

Further, the methods discussed in this paper are implicitly to be applied to every published variable. In practice, simplifying assumptions have been made to reduce the workload associated with disclosure limitation and to improve the consistency of published tables over time.

Section 2 presents the disclosure limitation practices that have been used where there may be some question as to how to apply the standard procedures. Section 3 presents the simplifying assumptions that have been found useful by federal statistical agencies. Both sections are intended as a reference for other agencies facing similar situations.

1. Background

The (n, k), pq-ambiguity and p-percent rules described in Chapter IV can all be written in the following form:

$$S(X) = \sum_{i=1}^{n} x_i - c\left(T - \sum_{i=1}^{s} x_i\right)$$

where the values of n, c and s depend on the specific rule and the parameters chosen, T is the total to be published, x_1 is the largest reported value, x_2 is the second largest reported value, and so on. In this framework, the x_i are all nonnegative.

2. Extension of Disclosure Limitation Practices

2.a. Sample Survey Data

The equation above assumes that all data are reported (as in a census). How can this rule be applied to data from a sample survey? One way of handling this is to let the values of the largest respondents, the xi, be specified by the unweighted reported values, but to let T be the weighted total to be published. (Note: this is a consistent way of stating that there is no disclosure with data from a sample survey when no units are selected with certainty and the sampling fractions

are small.)

2.b. Tables Containing Imputed Data

If some data are imputed, disclosure potential depends on the method of imputation.

a) Imputation for a sample survey is done by adjusting weights: In this case, method 2.a applies (the adjusted weights are used to calculate the weighted total, T).

b) Imputed values may be based on other respondent's data, as in "hot decking": In this case, the imputed value should not constitute a disclosure about the nonrespondent, so the imputed value (weighted, if appropriate) is included in the estimated total, T. The imputed value is counted as an individual reported value for purposes of identifying the largest respondents only for the donor respondent.

c) Imputed values may be based on past data from the nonrespondent: If the imputed value were revealed, it could constitute disclosure about the nonrespondent (for example, if the imputed value is based on data submitted by the same respondent in a different time period). The imputed value is included in the estimated total, T, and is also treated as submitted data for purposes of identifying the largest respondents.

2.c. Tables that Report Negative Values

If all reported values are negative, suppression rules can be applied directly by taking the absolute value of the reported data.

2.d. Tables Where Differences Between Positive Values Are Reported

If the published item is the difference between two positive quantities reported for the same time period (e.g. net production equals gross production minus inputs), then apply the primary suppression rule as follows:

a) If the resultant difference is generally positive, apply the suppression procedure to the first item (gross production in the above example).

b) If the resultant difference is generally negative, apply the suppression procedure to the second item (inputs in the above example.)

c) If the resultant difference can be either positive or negative and is not dominated by either, there are two approaches. One method is to set a threshold for the minimum number of respondents in a cell. A very conservative approach is to take the absolute value of the difference before applying the primary suppression rule.

2.e. Tables Reporting Net Changes (that is, Difference Between Values Reported at Different Times)

If either of the values used to calculate net change were suppressed in the original publication, then net change must also be suppressed.

2.f. Tables Reporting Weighted Averages

If a published item is the weighted average of two positive reported quantities, such as volume weighted price, apply the suppression procedure to the weighting variable (volume in this example).

2.g. Output from Statistical Models

Output from statistical models, such as econometric equations estimated using confidential data, may pose a disclosure risk. Often the resulting output from the statistical analyses takes the form of parameter coefficients in various types of regression equations or systems of equations. Since it is only possible to exactly recover input data from a regression equation if the number of coefficients is equal to the number of observations, regression output generally poses no disclose risk. However, sometimes dummy (0,1) variables are used in the model to capture certain effects, and these dummy variables may take on values for only a small number of observations.

One way of handling this situation is provided by the Center for Economic Studies of the Census Bureau. They treat the dummy variables as though they were cells in a table. Using the (n, k) rule, disclosure analysis is performed on the observations for which the dummy variable takes on the value 1.

3. Simplifying Procedures

3.a. Key Item Suppression

In several economic censuses, the Census Bureau employs key item suppression: performing primary disclosure analysis and complementary suppression on certain key data items only, and applying the same suppression pattern to other related items. Under key item suppression, fewer agency resources are devoted to disclosure limitation and data products are more uniform across data items. Key and related items are identified by expert judgment. They should remain stable over time.

3.b. Preliminary and Final Data

For magnitude data released in both preliminary and final form, the suppression pattern identified and used for the preliminary data should be carried forward to the final publication. The final data tables are then subjected to an audit to assure that there are no new disclosures. This conservative approach reduces the risk that a third party will identify a respondent's data from the changes in suppression patterns between preliminary and final publication.

3.c. Time Series Data

For routine monthly or quarterly publications of magnitude data, a standard suppression pattern (primary and complementary) can be developed based on the previous year's monthly data. This suppression pattern, after auditing to assure no new disclosures, would be used in the regular monthly publication.

APPENDIX B – Government References and Websites

1. Report on Statistical and Disclosure-Avoidance Techniques. Statistical Policy Working Paper 2 (May 1978). Washington, DC: U.S. Department of Commerce, Office of Policy and Federal Statistical Standards. This report is available from the National Technical Information Service: NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161; 703-487-4650. The NTIS document number is PB86-211539/AS.

2. Energy Information Administration Standards Manual. (September 2002). Energy Information Administration, U.S. Department of Energy. Washington, DC. http://www.eia.doe.gov/smg/Standard.pdf

3. Federal Statistics: Report of the President's Commission on Federal Statistics, Vol. 1. President's Commission on Federal Statistics. Washington, DC: U.S. Government Printing Office.

4. NASS Policy and Standards Memoranda. National Agricultural Statistics Service, U.S. Department of Agriculture. Washington, DC.

5. NCES Statistical Standards. (June 2003). National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/statprog/2002/stdtoc.asp

6. NCES Standard on "Maintaining Confidentiality" National Center for Education Statistics, U.S. Department of Education. Washington, DC. http://nces.ed.gov/statprog/2002/std4_2.asp

7. NCHS Staff Manual on Confidentiality. (September 2004). National Center for Health Statistics, U.S. Department of Health and Human Services. Washington, DC. http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

8. Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methodologies. Publication 1299 (February, 1986). Statistics of Income Division, Internal revenue Service, U.S. Department of Treasury. Washington, DC.

9. SOI Division Operating Manual. (January 1985). Statistics of Income Division, Internal revenue Service, U.S. Department of Treasury. Washington, DC.

WEBSITES FOR ADDITIONAL SOURCES

1) http://www.fcsm.gov/committees/cdac/ Website for the Confidentiality and Data Access Committee. This site provides useful links to resources for disclosure avoidance methodologies and related data access issues.

2) http://www.amstat.org/comm/cmtepc/index.cfm Website for the American Statistical Association's Privacy, Confidentiality, and Data Security. This site provides comprehensive information and references for the methodological, legal, ethical, and technical issues that arise out of protecting and using statistical data

3) www.census.gov/srd/sdc/index.html This site provides links and conventional references for research sponsored by the U.S. Census Bureau in the areas of statistical disclosure control, confidentiality, and disclosure limitation

4) http://aspe.os.dhhs.gov/datacncl/privcmte.htm U.S. Dept. of Health and Human Services Privacy Committee's website.

5) http://neon.vb.cbs.nl/casc/ Website for Computational Aspects of Statistical Confidentiality (CASC) (managed by the Netherlands Statistical Bureau). This site provides links for downloading Mu-Argus and Tau-Argus for applying disclosure avoidance rules to either microdata or tabular data; There are other useful links to books, papers, and presentations.

APPENDIX C – References

The purpose of this listing is to update the references on disclosure limitation methodology that were cited in Statistical Policy Working Paper 2 and the original version of Statistical Policy Working Paper 22. Several papers have been written since both these Statistical Policy Working Papers were published in 1978 and 1994, respectively.

In the Federal statistical system the Census Bureau has been the leading agency for conducting research into statistical disclosure limitation methods. The Census Bureau staff has been very active in publishing the results of their research through their website shown in Appendix B. For these reasons the statistical disclosure limitation research sponsored by the Bureau of the Census is thoroughly and adequately covered in this bibliography. In addition, important papers that either describe new methodology or summarize important research questions in the areas of disclosure limitation for tables of magnitude data, tables of frequency data and microdata are also included.

The "books" listed below in alphabetical order refer to traditional technical books written by a single author or a few co-authors, special collections of papers by many different authors, special issues of journals devoted to disclosure, and various online sources (e.g., references, manuals).

Books

"Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies"; edited by Pat Doyle, Julia I. Lane, Jules J.M. Theeuwes, Laura V. Zayatz. Published in 2001 by Elsevier Science B.V., Amsterdam, The Netherlands.
This volume has sixteen chapters written by leading researchers in a wide variety of disclosure topics. A description and list of articles appears at:

www.elsevier.com/wps/find/bookdescription.cws_home/622129/description#description
Chapter 1 is available online: www.census.gov/srd/sdc/ConfidentialityCH1.pdf

"Elements of Statistical Disclosure Control" by Leon Willenborg and Ton de Waal. Published by Springer in 2001. Lecture Notes in Statistics, volume 155. This volume is more theoretical than the earlier volume by these authors and goes into depth on many important methods. It has chapters on (i) disclosure risk (ii) information loss (iii) non-perturbative techniques (iv) perturbative techniques first for microdata and then for tabular data. There are 119 literature references presented at the end of the volume.

"For the Record, Protecting Electronic Health Information," by the National Academy of Sciences and National Research Council. Published in 1997 by the National Academy Press, Washington, D. C. In 1996, the Computer Science and Telecommunications Board (CSTB) formed a 15 member Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure. The committee addressed threats to healthcare information, adequacy of existing privacy and security measures, and best practices. The results of the committee's work were published in this book.

"Improving Access to and Confidentiality of Research Data", Committee on National Statistics", National Research Council, edited by Christopher Mackie and Norman Bradburn; published by National Research Council, National Academy Press, Washington, D.C., 2000. Summary of a workshop convened by CNSTAT to promote discussion about methods for advancing the often conflicting goals of exploiting the research potential of microdata and maintaining acceptable levels of confidentiality.

"Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics," edited by George T. Duncan, Thomas B. Jabine, Virginia A. de Wolf; published by the Committee on National Statistics and the Social Science Research Council, National Academy Press, Washington, D.C., 1993. This short (23 pages) but important volume consists of the executive summary and recommendations of the Panel on Confidentiality and Data Access. This panel was organized by CNSTAT and the Social Science Research Council to develop recommendations that could aid federal statistical agencies in their stewardship of data for policy decisions and research.

"Record Linkage and Privacy, Issues in Creating New Federal Research and Statistical Information," (GAO-01-126SP). This book provides a summary of various methodologies and matching techniques for matching a microdata file to an outside file. It updates a previous summary of mathematical methods used for matching found in "Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methodologies", Dept of Treasury, IRS, SOI, Publication 1299 (2-86).

"Statistical Disclosure Control in Practice" by Leon Willenborg and Ton de Waal. Published by Springer in 1996. Lecture Notes in Statistics, Volume 111. This book aims to discuss various aspects associated with disseminating personal or business data collected in censuses or surveys or copied from administrative sources. There are two detailed chapters on statistical disclosure control discussing the protection issues for microdata and several techniques that have been developed and used at various agencies. These are similar chapters for tabular data. There are 79 literature references presented at the end of the volume.

Reports Of Conferences and Workshops

Workshop on statistical data confidentiality (Skopje, Macedonia, March 2001). Sponsored by United Nations Economic Commission for Europe (UNECE).
Proceedings are available at http://192.91.247.58/stats/documents/2001.03.confidentiality.htm. This site also provides useful links to the papers and other statistical methodology materials.

"Inference Control in Statistical Databases: From Theory to Practice" (conference in Luxemburg, December 2001). Edited by Josep Domingo-Ferrer. Published by Springer in 2002 in Lecture Notes in Computer Science series, LNCS #2316. The list of articles with brief abstracts are available at: http://www.springerlink.com/app/home/search-articles-results.asp?wasp=5n5d6ynmwn0vwp8d4gfy&referrer=searchmainxml&backto=journal,1,1;linkingpublicationresults,1:105633,1

Workshops sponsored or co-sponsored by Eurostat. "Privacy in Statistical Databases", proceedings of Barcelona, June 2004 conference). Edited by Jose Domingo-Ferrer and Vicenc Torra. Published by Springer in 2004 in Lecture Notes in Computer Science series, #3050. The list of articles with brief abstracts may be found at: http://www.springerlink.com/app/home/search-articles-results.asp?wasp=3l93gmuvtj7yuk32wmf0&referrer=searchmainxml&backto=journal,1,1;linkingpublicationresults,1:105633,1

"Monographs of Official Statistics: Work session on statistical data confidentiality" (Proceedings of Luxembourg conference, April 2003). Published by Eurostat in 2004. The following three online .pdf documents form the entire proceedings.
http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-CR-03-004-1/EN/KS-CR-03-004-1-EN.PDF
http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-CR-03-004-2/EN/KS-CR-03-004-2-EN.PDF
http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-CR-03-004-3/EN/KS-CR-03-004-3-EN.PDF

Special Issues of Journals

Journal of Official Statistics: Special Issue on Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data, Vol. 19, No. 4, December 1998. Edited by Stephen E. Fienberg and Leon C.R.J. Willenborg (This journal is published by Statistics Sweden)
For list of articles see: http://www.jos.nu/Contents/issue.asp?vol=14&no=4

Journal of Official Statistics: Special Issue on Confidentiality and Data Access, Vol.9, No. 2., June 1993. For list of articles see: http://www.jos.nu/Contents/issue.asp?vol=9&no=2

The journal "Of Significance", published by the Association of Public Data Users, had a special issue on Confidentiality in 2000. It is volume 2, number 1 and is available online at: www.apdu.org/resources/docs/OfSignificance_v2n1.pdf

Netherlands Official Statistics: Special issue on Statistical Disclosure Control, vol. 14, Spring 1999. www.cbs.nl/nl/publicaties/publicaties/algemeen/a-125/1999/nos-99-1.pdf

Online References

An annotated list of references in contained in the article by John M. Abowd and Simon D. Woodcock in the volume, "Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies." This list is also available online at http://www.census.gov/srd/sdc/abowd-woodcock2001-appendix-only.pdf

A list of Microdata Confidentiality References compiled by William E. Winkler in March 2004 may also be found at www.census.gov/srd/sdc.

Websites dedicated to disclosure issues and/or references:
www.fcsm.gov/committees/cdac/cdac.html
www.census.gov/srd/sdc.

Manual

Checklist on Disclosure Potential of Proposed Data Releases (prepared by Confidentiality and Data Access Committee (CDAC) of the Federal Committee on Statistical Methodology (FCSM). http://www.fcsm.gov/committees/cdac/cdac.html

Report of the Task Force on Disclosure: GSS Methodology Series, no. 4, Government Statistical Service. Dec 1995, Office of National Statistics, London. This report is available online: http://www.statistics.gov.uk/downloads/theme_other/GSSMethodology_No_04_v2.pdf

National Center for Health Statistics Staff Manual on Confidentiality
http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

Articles

Aboud, J. M. and Lane, J. I., "Synthetic Data and Confidentiality Protection," (September, 2003). Technical Paper No. TP-2003-10, U.S. Census Bureau. The authors describe a method of creating multiple public use files from a single database where the actual values are replaced with scientifically valid estimates. The analytical value of the selected confidential variables is preserved while providing disclosure protection to the file.

Angle, John. (2003). "Imitating the Salamander: Reproduction of the Truncated Right Tail of an Income Distribution." This paper proposes a method to estimate the right tail of an income distribution using knowledge of the left and center portion of the variable's distribution and provides insight in applying top coding to a microdata file.
http://www.fcsm.gov/03papers/Angle_Final.pdf.

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), "Disclosure Control of Microdata," Journal of the American Statistical Association, Vol. 85, p. 38-45. A general overview of disclosure risk in the release of microdata is presented. Topics discussed are population uniqueness, sample uniqueness, subpopulation uniqueness and disclosure protection procedures such as adding noise, data swapping, microaggregation, rounding and collapsing. One conclusion reached by the authors is that it is very difficult to protect a data set from disclosure because of the possible use of matching procedures. Their view is that the data should be released to users with legal restrictions which preclude the use of matching.

Cecil, J. S. (1993), "Confidentiality Legislation and the United States Federal Statistical System," Journal of Official Statistics, Vol. 9, No. 2, p. 519-535. Access to records, both statistical and administrative, maintained by federal agencies in the United States is governed by a complex web of federal statutes. The author provides some detail concerning the Privacy Act

of 1974, which applies to all agencies, and the laws which apply specifically to the U. S. Bureau of Census, the National Center for Education Statistics and the National Center for Health Statistics. The author also describes ways these agencies have made data available to researchers.

Cox, L. H., (1980) "Suppression Methodology and Statistical Disclosure Control," Journal of the American Statistical Association, Vol. 75, No. 370, p. 377-385. This article highlights the interrelationships between the processes of disclosure definitions, sub-problem construction, complementary cell suppression, and validation of the results. It introduces the application of linear programming (transportation theory) to complementary suppression analysis and validation. It presents a mathematical algorithm for minimizing the total number of complementary suppressions along rows and columns in two-dimensional statistical tables. In a census or major survey, the typically large number of tabulation cells and linear relations between them necessitate partitioning a single disclosure problem into a well-defined sequence of inter-related sub-problems. Over suppression can be minimized and processing efficiency maintained if the cell suppression and validation processes are first performed on the highest level aggregations and successively on the lower level aggregates. The paper gives an example of a table with 2 or more suppressed cells in each row and column, where the value of the sensitive cell can be determined exactly, as an example of the need for validation.

Cox, L. H. (1981), "Linear Sensitivity Measures in Statistical Disclosure Control," Journal of Statistical Planning and Inference, Vol. 5, p. 153-164. Through analysis of important sensitivity criteria such as concentration rules, linear sensitivity measures are seen to arise naturally from practical definitions of statistical disclosure. This paper provides a quantitative condition for determining whether a particular linear sensitivity measure is subadditive. This is a basis on which to accept or reject proposed disclosure definitions. Restricting attention to subadditive linear sensitivity measures leads to well-defined techniques of complementary suppression. This paper presents the mathematical basis for claiming that any linear suppression rule used for disclose rule must be "subadditive". It gives as examples the n-k rule, the pq rule, and the p percent rule and discusses the question of sensitivity of cell unions. It provides bounding arguments for evaluating (in special cases) whether a candidate complementary cell might protect a sensitive cell.

Cox, L. H. and Ernst, L. R. (1982), "Controlled Rounding," INFOR, Canadian Journal of Operation Research and Information Processing, Vol. 20, No. 4, p. 423-432. Reprinted: Some Recent Advances in the Theory, Computation and Application of Network Flow Methods, University of Toronto Press, 1983, p. 139-148.) This paper demonstrates that a solution to the (zero-restricted) controlled rounding problem in two-way tables always exists. The solution is based on a capacitated transportation problem.

Cox, L. H., S.K. McDonald and D.W. Nelson, (1986). "Confidentiality Issues at the U.S. Bureau of the Census," Journal of Official Statistics Vol. 2, No. 2, p. 135 –160. This paper describes the policies and procedures of the U.S. Census Bureau following a major review and research program in data confidentiality protection during the mid-1980's.
http://www.jos.nu/Contents/jos_online.asp

Cox, L. H. (1987), "A Constructive Procedure for Unbiased Controlled Rounding," Journal of the American Statistical Association, Vol. 82, p. 520-524. Unbiased controlled rounding in a table involves rounding to an integer base, preserving additive structure, and assuring that the expected value of the rounded entry equals the original entry. This paper provides an easy-to-implement algorithm for achieving unbiased controlled rounding in a 2-dimensional table. The method also solves the two-way stratification problem in survey sampling and can be used to assure integer sample counts in an unbiased manner following, e.g., iterative proportional fitting (raking).

Cox, L. H. and George, J. A. (1989), "Controlled Rounding for Tables with Subtotals," Annals of Operations Research, 20 (1989) p. 141-157. Controlled rounding in two-way tables, Cox and Ernst (1982), is extended to two-way tables with subtotal constraints. The paper notes that these methods can be viewed as providing unbiased solutions. The method used is a capacitated network (transshipment) formulation. The solution is exact with row or column subtotals. It is demonstrated that the network solution with both row and column subtotal constraints is additive, but that it may fail zero-restricted constraints and may leave grand-totals of the subtables uncontrolled for the adjacency condition. An example is given of a table for which no zero-restricted controlled rounding exists.

Cox, L. H. (1995), "Network Models for Complementary Cell Suppression," Journal of the American Statistical Association, Vol. 90, No. 432, pp. 1453-1462. Complementary cell suppression is a method for protecting data pertaining to individual respondents from statistical disclosure when the data are presented in statistical tables. Several mathematical methods to perform complementary cell suppression have been proposed in the statistical literature, some of which have been implemented in large-scale statistical data processing environments. Each proposed method has limitations either theoretically or computationally. This paper presents solutions to the complementary cell suppression problem based on linear optimization over a mathematical network. These methods are shown to be optimal for certain problems and to offer several theoretical and practical advantages, including tractability and computational efficiency.

Cox, L. H. (1996), "Protecting Confidentiality in Small Population Health and Environmental Statistics," Statistics in Medicine, Vol. 15, p. 1895-1905. This paper discusses confidentiality problems in small domains and suggests the use of subsampling and supersampling for disclosure limitation in microdata files.

Cox, L. H. (2002), "Bounds on Entries in 3-Dimensional Contingency Tables Subject to Given Marginal Totals," in: Inference Control in Statistical Databases—From Theory to Practice, Lecture Notes in Computer Science 2316 (J. Domingo-Ferrer, ed.), New York: Springer, p. 21-33. This paper examines the problem of determining exact bounds for suppressed entries in 3-dimensional contingency tables given specified marginal totals and flaws in previous approaches, and compares several methods analytically.

Cox, L. H. (2003), "On Properties of Multi-Dimensional Statistical Tables," Journal of Statistical Planning and Inference, Vol. 117, 251-273. This paper examines mathematical properties of multi-dimensional statistical tables, including problems and procedures for assuring the existence of a feasible table given specified marginal tables, failure of linear programming to produce

integer solutions given integer constraints, and conditions under which integral solutions are assured based on network structure and network linear programming.

Cox, L. H. and Dandekar, R. A. (2004), "A New Disclosure Limitation Method for Tabular Data that Preserves Data Accuracy and Ease of Use," Proceedings of the 2002 FCSM Statistical Policy Seminar, Statistical Policy Working Paper 35, Federal Committee on Statistical Methodology, Washington, DC: U.S. Office of Management and Budget, p. 15-30. http://www.fcsm.gov/working-papers/spwp35.html
This paper introduces controlled tabular adjustment to the federal statistical community, focusing on its potential to improve data quality.

Cox, L. H., Kelly J., Patil, R. (2004). "Balancing Quality and Confidentiality for Multi-Variate Tabular Data. This paper proposes the use of certain linear and non-linear models subject to specific constraints that may be used to adjust tabular data in order to preserve additivity, covariance, correlation, and regression coefficients and other data relationships from the original table are preserved.

Cox, L. H., James P. Kelly, and Rahul J. Patil. (2005). "Computational Aspects of Controlled Tabular Adjustment: Algorithm and Analysis" in the book "The Next Wave in Computing, Optimization, and Decision Technologies", ed. B. Golden, S. Raghavan, E. Wasil, published by Springer. This paper presents a cutting plane algorithm for speeding controlled tabular adjustment.

Dandekar, R., Cohen, M., and Kirkendall, N. (2002). "Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique. Lecture Notes in Computer Science, vol. 2316, pp. 117-125, Apr. 2002. ISSN 0302-9743. Vol. Inference Control in Statistical Databases, ed. Josep Domingo-Ferrer, Berlin:Springer-Verlag. This paper discusses a methodology for creating synthetic micro data that can be used in place of actual reported data or to create either additive or multiplicative noise which when merged with the original data can provide disclosure protection while reproducing many of the essential quality of the original micro data file.
Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique
<http://taz/smg/papers/BARCEL.pdf

Dandekar Ramesh A., (2004) "Cost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems", (2004), p. 428-434, Monographs of Official Statistics, Luxembourg: Eurostat. The paper describes a simplified procedure as an alternative to the linear programming based controlled tabular adjustment (CTA) methodology to generate synthetic tabular data to protect data containing sensitive information. The simplified CTA procedure is a low cost approach that allows statistical agencies to use conventional readily available software tools to generate synthetic tabular data.

Dandekar, Ramesh, (2004). "Maximum Utility-Minimum Information Loss Table Server Design for Statistical Disclosure Control of Tabular Data." Lecture Notes in Computer Science, Springer-Verlag Heidelberg, ISSN: 0302-9743, Vol. 3050 p. 121-135. The paper discusses a simplified version of the CTA and applies it to categorical and magnitude test data. It also

provides a comparative evaluation of this simplified CTA approach and LP-based CTA using magnitude test data. For these test data, the simplified CTA is able to protect the tables with many fewer adjustments to cell values than the LP-based CTA requires.

De Loera, J., Ohn, Shmuel, "All Rational Polytopes Are Transportation Polytopes and All Polytopal Integer Sets Are Contingency Tables." IPCO 2004, LNCS 3064, pp. 338–351. This paper shows that any rational polytope is polynomial-time representable as a "slim" $r \times c \times 3$ three-way line-sum transportation polytope. This universality theorem has important consequences for linear and integer programming and for confidential statistical data disclosure. It provides polynomial-time embedding of arbitrary linear programs and integer programs in such slim transportation programs and in bipartite bi-flow programs. It resolves several standing problems on 3-way transportation polytopes. It also demonstrates that the range of values an entry can attain in any slim 3-way contingency table with specified 2-margins can contain arbitrary gaps, suggesting that disclosure of k-margins of d-tables for $2 \leq k < d$ is confidential. http://www.opt.math.tu-graz.ac.at/IPCO/prog.10

Dobra, Adrian, Fienberg, Stephen E., (2000), "Bounds for Cell Entries in Contingency Tables Given Marginal Totals and Decomposable Graphs", Proceedings of the National Academy of Sciences Vol. 97 No. 22 p. 1885-1892: Upper and lower bounds on cell counts play an important role in statistical disclosure limitation. This paper provides the theoretical framework and proofs of the exactness of Frechet bounds on decomposable graphical loglinear models. For such models, simple formulae, in lieu of computationally demanding integer programs, yield exact bounds. Some of these models are familiar in statistics, e.g. complete independence models, but overall this entire class of models is relatively small.

Dobra, Adrian, Fienberg, Stephen E. (2001) "Bounds for Cell Entries in Contingency Tables Induced by Fixed Marginal Totals with Applications to Disclosure Limitation." Statistical Journal of the United Nations ECE. Vol. 18, p. 363–371. This paper is a more descriptive version of the results presented in Dobra and Fienberg (2000) on computing exact bounds for decomposable graphical models.

Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2001), "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map," Los Alamos National Laboratory Technical Report, LA-UR-01-6428. Methods are discussed for assessing the disclosure risk of a file and trade offs in data utility as the parameters in various disclosure limitation methodologies are changed. The authors describe a method for calculating separate numerical assessments of the disclosure risk and data utility while allowing different values for the disclosure limitation parameters.

Evans, T., Zayatz, L., Slanta, J., (1998). "Using Noise for Disclosure Limitation Establishment Tabular Data," Journal of Official Statistics, Vol. 14, p. 537-551. This paper discusses the disclosure limitation method for protecting establishment magnitude tabular data by adding noise to the underlying microdata prior to tabulation.

Ernst, L., (1989), "Further Applications of Linear Programming to Sampling Problems," Proceedings of the Survey Research Methods Section, American Statistical Association, p. 625-630. In a previous paper, Cox and Ernst (1982), it was demonstrated that a controlled rounding exists for every two-dimensional additive table. In this paper the author establishes by means of

a counter-example that the natural generalization of their result to three dimensions does not hold.

Fienberg, Stephen E. 1997. "Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research." Carnegie Mellon Department of Statistics Technical Report, Working Paper No. 668. Carnegie Mellon Department of Statistics. Pittsburgh, Pennsylvania. http://www.stat.cmu.edu/tr/tr668/tr668.html. This paper provides an overview of the statistical issues that are related to the evolving area of statistical disclosure limitation methodology.

Fischetti, M and Salazar, JJ (1999), "Models and Algorithms for the 2-Dimensional Cell Suppression Problem in Statistical Disclosure Control," Mathematical Programming, Vol. 84, 283-312. This paper introduces the Fischetti-Salazar method for solving the decision problem associated with complementary cell suppression. Unlike previous methods, it protects all sensitive cells at once rather than sequentially, and can produce optimal results in medium to large problems.

Fischetti, M. and Salazar, JJ (2000), "Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints," Journal of the American Statistical Association, Vol. 95, p. 916-928. Algorithms for complementary cell suppression for tabular data shown to run to optimality in large, but not enormous, problem settings.

Gomatam, S., Karr, A. F., Sanil, A. P. (2005), "Data swapping as a decision problem," Journal of Official Statistics. This paper discusses risk-utility formulation of data swapping for categorical data.

Gomatam, S., Karr, A. F., Reiter, J. P., Sanil, A. P. (2005), "Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers," Statistical Science, Vol. 20, p. 163 - 177. Remote access analysis servers allow users to submit requests for output from statistical models fit using confidential data. The users are not allowed access to the data themselves. Analysis servers, however, are not free from the risk of disclosure, especially in the face of multiple, interacting queries. In this paper, the authors describe these risks and propose quantifiable measures of risk and data utility that can be used to specify which queries can be answered, and with what output. The risk-utility framework is illustrated for regression models.

Gonzalez, JF and Cox, LH (2005), "Software for Tabular Data Protection," Statistics in Medicine, Vol. 24 (4), p. 659-669. This paper describes software for data protection in two-way tables developed for the National Center for Health Statistics: complementary cell suppression, rounding, perturbation and controlled tabular adjustment. The software is available at no charge.

Greenberg, B. and Zayatz, L. (1992), "Strategies for Measuring Risk in Public Use Microdata Files," Statistica Neerlandica, Vol. 46, No. 1, p. 33-48. Methods of reducing the risk of disclosure for microdata files and factors that diminish the ability to link files and to obtain correct matches are described. Two methods of estimating the percent of population uniques on a microdata file are explained. A measure of relative risk for a microdata file based on the notion of entropy is introduced.

Griffin, R. A., Navarro, A., and Flores-Baez, L. (1989), "Disclosure Avoidance for the 1990 Census," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, p. 516-521. This paper presents the 1990 Census disclosure avoidance procedures for 100 percent and sample data and the effects on the data. The Census Bureau's objective is to maximize the level of useful statistical information provided subject to the condition that confidentiality is not violated. Three types of procedures for 100 percent data have been investigated: suppression, controlled rounding, and confidentiality edit. Advantages and disadvantages of each are discussed. Confidentiality Edit is based on selecting a small sample of census households from the internal census data files and interchanging their data with other households that have identical characteristics on a set of selected key variables. For the census sample data, the sampling provides adequate protection except in small blocks. A blanking and imputation-based methodology is proposed to reduce the risk of disclosure in small blocks.

Hawala, S., Zayatz, L., Rowland, S., (2004). "American FactFinder: U.S. Bureau of the Census works towards meeting the needs of users while protecting confidentiality," Journal of Official Statistics, Vol. 20, p. 115-124. This paper discusses the special disclosure limitation techniques that are applied to protect the confidentiality of tabulations generated from an online query of microdata files. http://www.jos.nu/Contents/jos_online.asp

Jabine, T. B. (1993a), "Procedures for Restricted Data Access," Journal of Official Statistics, Vol. 9, No. 2, p. 537-589. Statistical agencies have two main options for protecting the confidentiality of the data they release. One is to restrict the data through the use of statistical disclosure limitation procedures. The other is to impose conditions on who may have access, for what purpose, at what locations, and so forth. For the second option, the term, **restricted access,** is used. This paper is a summary of restricted access procedures that U. S. statistical agencies use to make data available to other statistical agencies and to other organizations and individuals. Included are many examples that illustrate both successful modes and procedures for providing access, and failures to gain the desired access.
http://www.jos.nu/Contents/jos_online.asp

Jabine, T. B. (1993b), "Statistical Disclosure Limitation Practices of United States Statistical Agencies," Journal of Official Statistics, Vol. 9., No. 2, p. 427-454. One of the topics examined by the Panel on Confidentiality and Data Access of the Committee on National Statistics of the National Academy of Sciences was the use of statistical disclosure limitation procedures to limit the risk of disclosure of individual information when data are released by Federal statistical agencies in tabular or microdata formats. To assist the Panel in its review, the author prepared a summary of the disclosure limitation procedures that were being used by the agencies in early 1991. This paper is an updated version of that summary.
http://www.jos.nu/Contents/jos_online.asp

Jewett, R. (1993), "Disclosure Analysis for the 1992 Economic Census," unpublished manuscript, Economic Programming Division, Bureau of Census, Washington, DC. The author describes in detail the network flow methodology used for cell suppression for the 1992

Economic Censuses. The programs used in the disclosure system and their inputs and outputs are also described. http://www.census.gov/srd/sdc/Jewett.disc.econ.1992.pdf

Karr, A. F., Lin, X., Reiter, J. P., Sanil, A. P. (2005). Secure regression on distributed databases. Journal of Computational Graphical Statistics Vol. 14 No. (2) p. 263–279. This article presents several methods for performing linear regression on the union of distributed databases that preserve, to varying degrees, confidentiality of those databases. Such methods can be used by federal or state statistical agencies to share information from their individual databases, or to make such information available to others.

Keller-McNulty, S., McNulty, M. S., and Unger, E. A. (1989), "The Protection of Confidential Data," Proceeding of the 21st Symposium on the Interface, American Statistical Association, Alexandria, VA, pp. 215-219. A broad overview of analytic methods that have been or might be used to protect confidentiality is provided for both microdata files and for tabular releases. Some methods that might be used with microdata, e.g.. "blurring," "slicing," are described. The authors also discuss the need for a standard measure of "control" or protection.

Kennickell, Arthur B. (1998). "Multiple Imputation in the Survey of Consumer Finances," Proceedings of the Joint Statistical Meetings American Statistical Association 1998. This paper describes the FRITZ system of multiple imputation developed for the Survey of Consumer Finances. In addition to describing the application of the system to ordinary problems of imputation of missing data, the paper presents the results of using the system for a set of experiments in data simulation for disclosure avoidance.
http://www.federalreserve.gov/pubs/oss/oss2/papers/impute98.pdf

Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," American Statistical Association, Proceedings of the Section on Survey Research Methods, p. 370-374. Although noise addition is effective in reducing disclosure risk, it has an adverse affect on any data analysis. If one knows how the data are to be used, transformations of the data before and after the addition of noise can maintain the usefulness of the data. The author recommends using linear transformations subject to the constraints that the first and second moments of the new variable are identical to those of the original. He presents the properties of the transformed variable when the variance is known, and when it is estimated. He sets forth the impacts of masking on the regression parameter estimates under different conditions of preserving the first and second moments of the original data.

Kim, J. J., and W.E. Winkler (1995). "Masking Microdata Files," American Statistical Association, Proceedings of the Section on Survey Research Methods, p. 114-119. No single masking scheme so far meets the needs of all data users. This article describes the masking scheme used for a specific case of providing microdata to two users that took into account their analytic needs. Since it was done before Kim (1990b), each group was masked separately. In this example the user planned to construct multiple regression models, with the dependent variable of two types - proportions transformed into logits, and medians. Kim discusses 1) whether to add the noise before or after transformation, 2) what distribution of the noise to use, and 3) whether to add correlated or uncorrelated noise. He presents in clear detail the masking process, the statistical properties of the masked variables, and how they satisfied these users'

needs. Excellent results were obtained for estimates of the mean and variance/covariance, except when considerable censoring accompanied the logit transformation of the proportions.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm," Journal of Official Statistics, Vol. 9, No. 2, p. 313-331. The definition of disclosure depends on the context. Sometimes a disclosure is said to occur even though the information revealed is incorrect. A disclosure may violate a respondent's anonymity and sometimes reveal sensitive information. This paper tries to untangle disclosure issues by differentiating between linking a respondent to a record and learning sensitive information from the linking. The extent to which a released record can be linked to a respondent determines disclosure risk; the information revealed when a respondent is linked to a released record determines disclosure harm. There can be harm even if the wrong record is identified or an incorrect sensitive value inferred. In this paper, measures of disclosure risk and harm that reflect what is learned about a respondent are studied, and some implications for data release policies are given. http://www.jos.nu/Contents/jos_online.asp

Lee, J., Holloman, C., Karr, A. F. and Sanil, A. P. (2001), "Analysis of Aggregated Data in Survey Sampling with Application to Fertilizer/Pesticide Usage Surveys," Research in Official Statistics, Vol. 4, p. 101-116: This paper proposes a Bayesian simulation approach for analysis of data aggregated to protect disclosure.

Massell, Paul B., (2002). "Optimization Models and Programs for Cell Suppression in Statistical Tables," Proceeding of the Joint Statistical Meetings American Statistical Association 2002. This paper compares the different mathematical approaches to applying cell suppression and evaluates the usefulness of the different programs based on the optimization method as well as other practical considerations. Network based programs and extended network based programs are compared with linear programming, integer based, and hypercube based programs. http://www.census.gov/srd/sdc/Massell.JSM2002.v4.pdf

Massell, Paul B., (2004). "Comparing Statistical Disclosure Control Methods for Tables: Identifying the Key Factors", Proceedings of the Joint Statistical Meetings American Statistical Association 2004. This paper describes the key factors involved in deciding how to select a statistical disclosure method that is suitable for protecting a given set of tables.
http://www.census.gov/srd/sdc/Massell.JSM2004.paper.v3.pdf

Michalewicz, Zbigniew (1991). "Security of a Statistical Database," in Statistical and Scientific Data-bases, ed., Ellis Horwood, Ltd. This article discusses statistical database security, also known as inference control or disclosure control. It is assumed that all data is available in an on-line, as in a microdata file. A critique of current methods, both query restriction and perturbation, is included using an abstract model of a statistical database. **Tracker** type attacks are extensively discussed. The balance between security and usability is developed, with usability for query restriction methods being dependent upon the number and ranges of restricted data intervals. Methods of determining these intervals are compared.

Muralidhar, K., Sarathy, R. (May, 2002). "A Data Shuffling Procedure for Masking Data," This paper discusses the methodology and theoretical basis for applying a two-step data swapping procedure for protecting confidential numerical data. Report to the Census Bureau, May, 2002. http://gatton.uky.edu/faculty/muralidhar/maskingpapers.

Paass, G. (1988), "Disclosure Risk and Disclosure Avoidance for Microdata," Journal of Business and Economic Statistics, Vol. 6, p. 487-500. This paper gives estimates for the fraction of identifiable records when specific types of outside information may be available to the investigator, this fraction being dependent primarily on the number of variables in common, and the frequency and distribution of the values of these variables. The author discusses the costs involved. Paass then evaluates the performance of disclosure-avoidance measures such as slicing, microaggregations, and recombinations. In an appendix, he presents the technical details of the proposed methods.

Qian, X., Stickel, M., Karp, P., Lunt, T. and Garvey, T., "Detection and Elimination of Inference Channels in Multilevel Relational Database Systems," IEEE Symposium on Research in Security and Privacy, Oakland, CA, May 24-26, 1993. This paper addresses the problem where information from one table may be used to **infer** information contained in another table. it assumes an on-line, relational database system of several tables. The implied solution to the problem is to classify (and thus to deny access to) appropriate data. The advantage of this approach is that such discoveries are made at the **design** time, not execution time. The disadvantage is that the technique only addresses those situations where inferences always hold, not those cases where the inference is dependant upon specific values of data. The technique needs to be investigated for applicability to the disclosure limitation problem.

Raghunathan, T.E., Reiter, J. P., and Rubin, D.R. (2003), "Multiple Imputation for Statistical Disclosure Limitation," Journal of Official Statistics, Vol. 19, p. 1-16. This article evaluates the use of the multiple imputation framework to protect the confidentiality of respondents' answers in sample surveys. The basic proposal is to simulate multiple copies of the population from which these respondents have been selected and release a random sample from each of these synthetic populations. Users can analyze the synthetic sample data sets with standard complete-data software for simple random samples, then obtain valid inferences by combining the point and variance estimates using the methods in this article.
http://www.jos.nu/Contents/jos_online.asp

Reiter, J.P. (2002), "Satisfying Disclosure Restrictions with Synthetic Data Sets," Journal of Official Statistics, Vol. 18, No. 4, p. 531-543. To avoid disclosures, Rubin proposed creating multiple, synthetic data sets for public release so that (i) no unit in the released data has sensitive data from an actual unit in the population, and (ii) statistical procedures that are valid for the original data are valid for the released data. This paper discusses through the use of simulation studies that valid inferences can be obtained from synthetic data in a variety of settings, including simple random sampling, probability proportional to size sampling, two-stage cluster sampling, and stratified sampling. http://www.jos.nu/Articles/abstract.asp?article=184531

Reznek, A. P., "Disclosure Risks in Cross-Section Regression Models," (2003). This paper describes the disclosure risks associated with certain types of cross section regression models. In

particular, it shows via examples that models with only fully interacted dummy (0,1) variables on the right-hand side allow recovery of entries from a table of means of the left-hand side variable, broken down by the categories of the dummy variables. Proceedings of the Joint Statistical Meetings American Statistical Association 2003

Reznek, Arnold P. and T. Lynn Riggs (2004). "Disclosure Risks in Regression Models: Some Further Results." Proceedings of the Joint Statistical Meetings American Statistical Association 2004. This paper illustrates that correlation matrices and variance-covariance matrices of variables, as well as variance-covariance matrices of model coefficients, can also allow recovery of table entries if the variables include dummy variables.

Robertson, D. A., (1993), "Cell Suppression at Statistics Canada," Proceedings of the Bureau of the 1993 Census Annual Research Conference, Bureau of the Census, Washington, DC, pp. 107-131. Statistics Canada has developed Computer software (CONFID) to ensure respondent confidentiality via cell suppression. It assembles tabulation cells from microdata and identifies confidential cells and then selects complementary suppressions. This paper discusses the design and algorithms used and its performance in the 1991 Canadian Census of Agriculture.

Rubin, D. (1993), "Discussion, Statistical Disclosure Limitation," Journal of Official Statistics, Vol. 9, No. 2, pp. 461-468. Rubin proposes that the government should release only "synthetic data" rather than actual micro-data. The synthetic data would be generated using multiple imputation. They would look like individual reported data and would have the same multivariate statistical properties. However, with this scheme there would be no possibility of disclosure, as no individual data would be released.

Saalfeld, A., Zayatz, L. and Hoel, E. (1992), "Contextual Variables via Geographic Sorting: A Moving Averages Approach," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, p. 691-696. Social scientists would like to perform spatial analysis on microdata. They want to know relative geographic information about each record such as average income of neighboring individuals. Variables providing this type of information are called "contextual variables." This paper introduces a technique which could generate contextual variables which do not comprise the exact location of respondents. The technique is based on taking moving averages of a sorted data set.

Sailer, P., Weber, M., and Wong, W., (2001), "Disclosure-Proofing the 1996 Individual Tax Return Public-use File," Proceedings of the American Statistical Association 2001; This paper provides an overview of the disclosure-proofing techniques applied to the Statistics of Income Individual Tax Return Public-Use File (PUF). It also discusses the results of two tests of these procedures: the matching of a publicly available marketing database to the PUF: and the matching of the IRS Individual Master File to the PUF.

Sanil, A. P., Karr, A. F., Lin. X., Reiter, J. P. (2004), "Privacy preserving regression modeling via distributed computation," Proceedings of the Tenth ACM SIGKDD 2004 International Conference on Knowledge Discovery and Data Mining p. 677–682. This paper discusses secure regression for distributed, vertically partitioned data when the response is shared.

Singer, E. and Miller, E. (1993), "Recent Research on Confidentiality Issues at the Census Bureau," Proceedings of the Bureau of the Census 1993 Annual Research Conference, Bureau of the Census, Washington, DC, p. 99-106. The Census Bureau conducted focus group discussions concerning participants' reactions to the use of administrative records for the Year 2000 Census, their fears concerning confidentiality breaches, their reactions to a set of motivational statements, and ways of reassuring them about the confidentiality of their data. This paper highlights results of these discussions and relates findings from other research in this area.

Steel, Philip M. (2004) "Disclosure Risk Assessment for Microdata," This is an introduction to risk assessment for microdata, for the beginning practioner. It presents some background on legal concepts of identifiablility, discusses risk measurement and its applicability, demonstrates how public data and context can effect risk. There is also an eclectic set of references. http://www.census.gov/srd/sdc/Steel.Disclosure%20Risk%20Assessment%20for%20Microdata.pdf .

Van Den Hout, A., and Van Der Heijden, P. G. M. (2002), "Randomized Response, Statistical Disclosure Control, and Misclassification: A Review." International Statistical Review, Vol. 70 (2), p. 269-288. This paper discusses analysis of categorical data which have been misclassified and where misclassification probabilities are known. Fields where this kind of misclassification occurs are randomized response, statistical disclosure control, and classification with known sensitivity and specificity. Estimates of true frequencies are given, and adjustments to the odds ratio are discussed. Moment estimates and maximum likelihood estimates are compared and it is proved that they are the same in the interior of the parameter space. http://isi.cbs.nl/ISReview/abst01-13.pdf

Winkler, William E. (1998). "Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata," Research in Official Statistics, Vol. 1, p. 87-114. This paper compares several masking methods in terms of their ability to produce analytically valid, confidential microdata. For a public-use microdata file to be analytically valid, it should, for a very small number of uses, yield analytic results that are approximately the same as the original, confidential file that is not distributed. If a microdata file contains a moderate number of variables and is required to meet a single set of analytic needs, then many more records are likely to be re-identified via modern record linkage methods than via the re-identification methods typically used in the confidentiality literature.

Winkler, William E., (2004). "Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems." This paper provides an overview of various methods applied for masking microdata. It also discusses different measures for estimating disclosure risk for a public-use data file. http://www.census.gov/srd/papers/pdf/rrs2004-03.pdf

Yu, Dunteman, Dai, and Wilson (2004). "Measuring the Performance of MASSC Using NCHS-2000 NHIS Public Use File." The paper discusses the Micro Agglomeration, Substitution, Subsampling, and Calibration disclosure limitation method. Work session on data confidentiality. Conference of European Statisticians 2003. http://www.unece.org/stats/documents/2003.04.confidentiality.htm

Zayatz, L. (1992a). "Using Linear Programming Methodology for Disclosure Avoidance Purposes," Statistical Research Division Report Series, Census/SRD/RR-92/02, Bureau of the Census, Statistical Research Division, Washington, DC. This paper presents a linear-programming scheme for finding complementary suppressions for a primary suppression that is applicable to two or three dimensional tables. The method yields good but not optimal results. The paper discusses three ways of improving results: 1) sorting the primary suppressions by the protection they need and finding complementary cells for each primary cell sequentially beginning with the largest; 2) adding an additional run through the linear program with an adjusted cost function to eliminate unnecessary complementary suppressions identified in the first run; and 3) using different cost functions. A general comparison with network flow methodology is also given. The paper also provides an example using the commercially available linear programming package, LINDO.

Zayatz, L. V. (1992b), "Linear Programming Methodology for Disclosure Avoidance Purposes at the Census Bureau." Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, p. 679-684. This paper recommends specific approaches for finding complementary suppressions for two-dimensional tables, small three-dimensional tables and large three-dimensional tables. Network flow procedures are recommended for two-dimensional tables. Linear programming methods are recommended (and described) for small three-dimensional tables. In the case of large three-dimensional tables, the recommended procedure is a sequence of network flow algorithms applied to the two-dimensional sub-tables. The resultant system of suppressions must then be audited to assure that the sensitive cells are protected. A linear programming algorithm for validating a pattern of suppressions is described.

Zayatz, L. (2002). "SDC in the 2000 U.S. Decennial Census," Inference Control in Statistical Databases: From Theory to Practice, Springer, p.193-202. This paper describes the statistical disclosure limitation techniques used for all U.S. Census 2000 data products. It includes procedures for short form tables, long form tables, public use microdata files, and an online query system for tables. The procedures that were used include data swapping, rounding, noise addition, collapsing categories, and applying thresholds.

Zayatz, L., Massell, P., and Steel, P. (1999). "Disclosure limitation practices and research at the U. S. Census Bureau" Netherlands Official Statistics, Spring, 1999, Vol. 14, p. 26-29. This paper discusses disclosure limitation practices in effect at the Census Bureau, as well as current Census Bureau research into alternative disclosure limitation procedures and some analysis of these procedures.

APPENDIX D – Confidentiality and Data Access Committee

In 1995, the Interagency Confidentiality and Data Access Group (ICDAG) was formed to (1) promote and implement the goals and recommendations outlined in Chapter 6 of Statistical Policy Working Paper #22 (2) increase cooperation and sharing of statistical disclosure limitation methods among federal agencies and (3) provide a forum for sharing information and ideas on protecting data confidentiality and improving data access. Its members are employees of Executive Branch federal agencies working on data confidentiality and data access issues expressed the need for a forum to share their knowledge and discuss common issues and concerns. Back in 1995, ICDAG was informally affiliated with the Federal Committee on Statistical Methodology (FCSM).

In 1997, the FCSM formally recognized ICDAG as an "Interest Group" to better facilitate communication and cooperation among agencies. In 2000, the name of the group was changed to the Confidentiality and Data Access Committee (CDAC). Since 1997, CDAC has developed several data products to help centralize agency review of disclosure limited data products, share methodology, software, and information across federal agencies on data confidentiality and data access issues and activities. See http://www.fcsm.gov/committees/cdac/ In addition, its members provide presentations on statistical disclosure methodology to various audiences throughout the year to help expand working knowledge in these areas.

Data products that CDAC has developed include:

Checklist on Disclosure Potential of Proposed Data Releases – This document standardizes the review for disclosure risks associated any proposed data release.

Brochure on "Confidentiality and Data Access Issues Among Federal Agencies – This brochure describes some examples of data protections used by federal agencies - legal sanctions, removal of personal identifiers from data sets, the application of statistical procedures to published information, certificates of confidentiality, institutional and disclosure review boards, and restricted data access (research data centers, remote access, special employee status and data licensing).

Restricted Access Procedures - This paper discusses various methods used by five federal agencies for providing access to statistical data while limiting the risk of disclosure of confidential information. The methods include Research Data Centers (RDCs), remote access and on-line query systems, research fellowships and post-doctoral programs, and licensing agreements.

Identifiability in Microdata Files - This document provides an understanding of what variables and types of data might make individual respondents identifiable in a microdata file.

Disclosure Auditing Software – This PC based SAS software identifies the lower and upper bounds on the values of a withheld (suppressed) cell in a tabular statistical table, and provides other useful measures for auditing the suppression pattern in a table.

Reports Available in the Federal Committee on Statistical Methodology's Statistical Policy Working Paper Series

1. ***Report on Statistics for Allocation of Funds***, 1978 (NTIS PB86-211521/AS)
2. ***Report on Statistical Disclosure and Disclosure-Avoidance Techniques***, 1978 (NTIS PB86-211539/AS)
3. ***An Error Profile: Employment as Measured by the Current Population Survey***, 1978 (NTIS PB86-214269/AS)
4. ***Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics***, 1978 (NTIS PB86-211547/AS)
5. ***Report on Exact and Statistical Matching Techniques***, 1980 (NTIS PB86-215829/AS)
6. ***Report on Statistical Uses of Administrative Records***, 1980 (NTIS PB86-214285/AS)
7. ***An Interagency Review of Time-Series Revision Policies***, 1982 (NTIS PB86-232451/AS)
8. ***Statistical Interagency Agreements***, 1982 (NTIS PB86-230570/AS)
9. ***Contracting for Surveys***, 1983 (NTIS PB83-233148)
10. ***Approaches to Developing Questionnaires***, 1983 (NTIS PB84-105055)
11. ***A Review of Industry Coding Systems***, 1984 (NTIS PB84-135276)
12. ***The Role of Telephone Data Collection in Federal Statistics***, 1984 (NTIS PB85-105971)
13. ***Federal Longitudinal Surveys***, 1986 (NTIS PB86-139730)
14. ***Workshop on Statistical Uses of Microcomputers in Federal Agencies***, 1987 (NTIS PB87-166393)
15. ***Quality in Establishment Surveys***, 1988 (NTIS PB88-232921)
16. ***A Comparative Study of Reporting Units in Selected Employer Data Systems***, 1990 (NTIS PB90-205238)
17. ***Survey Coverage***, 1990 (NTIS PB90-205246)
18. ***Data Editing in Federal Statistical Agencies***, 1990 (NTIS PB90-205253)
19. ***Computer Assisted Survey Information Collection***, 1990 (NTIS PB90-205261)
20. ***Seminar on Quality of Federal Data***, 1991 (NTIS PB91-142414)
21. ***Indirect Estimators in Federal Programs***, 1993 (NTIS PB93-209294)
22. ***Report on Statistical Disclosure Limitation Methodology***, Second version 2005
23. ***Seminar on New Directions in Statistical Methodology***, 1995 (NTIS PB95-182978)
24. ***Electronic Dissemination of Statistical Data***, 1995 (NTIS PB96-121629)
25. ***Data Editing Workshop and Exposition***, 1996 (NTIS PB97-104624)
26. ***Seminar on Statistical Methodology in the Public Service***, 1997 (NTIS PB97-162580)
27. ***Training for the Future: Addressing Tomorrow's Survey Tasks***, 1998 (NTIS PB99-102576)
28. ***Seminar on Interagency Coordination and Cooperation***, 1999 (NTIS PB99-132029)
29. ***Federal Committee on Statistical Methodology Research Conference (Conference Papers)***, 1999 (NTIS PB99-166795)
30. ***1999 Federal Committee on Statistical Methodology Research Conference: Complete Proceedings***, 2000 (NTIS PB2000-105886)
31. ***Measuring and Reporting Sources of Error in Surveys***, 2001 (NTIS PB2001-104329)
32. ***Seminar on Integrating Federal Statistical Information and Processes***, 2001 (NTIS PB2001-104626)
33. ***Seminar on the Funding Opportunity in Survey Research***, 2001 (NTIS PB2001-108851)
34. ***Federal Committee on Statistical Methodology Research Conference (Conference Papers)***, 2001 (NTIS PB2002-100103)
35. ***Seminar on Challenges to the Federal Statistical System in Fostering Access to Statistics.*** 2004.
36. ***Seminar on the Funding Opportunity in Survey and Statistical Research.*** 2004.
37. ***Federal Committee on Statistical Methodology Research Conference (Conference Papers)***, 2003.
38. ***Summary Report of the FCSM-GSS Workshop on Web-Based Data Collection.*** 2004.

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161; telephone: 1-800-553-6847. The Statistical Policy Working Paper series is also available electronically from FCSM's web site <*http://www.fcsm.gov*>.