
**Statistical Policy**
**Working Paper 22**

# Report on Statistical Disclosure Limitation Methodology

**Prepared by**
**Subcommittee on Disclosure Limitation Methodology**
**Federal Committee on Statistical Methodology**

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

**May 1994**

**MEMBERS OF THE FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY**

**(May 1994)**

**Maria E. Gonzalez, Chair**
**Office of Management and Budget**

Yvonne M. Bishop
Energy Information Administration

Warren L. Buckler
Social Security Administration

Cynthia Z.F. Clark
National Agricultural Statistics Service

Steven Cohen
Administration for Health Policy and Research

Zahava D. Doering
Smithsonian Institution

Roger A. Herriot
National Center for Education Statistics

C. Terry Ireland
National Computer Security Center

Charles G. Jones
Bureau of the Census

Daniel Kasprzyk
National Center for Education Statistics

Nancy Kirkendall
Energy Information Administration

Daniel Melnick
Substance Abuse and Mental Health Services Administration

Robert P. Parker
Bureau of Economic Analysis

Charles P. Pautler, Jr.
Bureau of the Census

David A. Pierce
Federal Reserve Board

Thomas J. Plewes
Bureau of Labor Statistics

Wesley L. Schaible
Bureau of Labor Statistics

Fritz J. Scheuren
Internal Revenue Service

Monroe G. Sirken
National Center for Health Statistics

Robert D. Tortora
Bureau of the Census

Alan R. Tupek
National Science Foundation

## PREFACE

The Federal Committee on Statistical Methodology was organized by 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. The subcommittees are open by invitation to Federal employees who wish to participate. Working papers are prepared by the subcommittee members and reflect only their individual and collective ideas.

The Subcommittee on Disclosure Limitation Methodology was formed in 1992 to update the work presented in Statistical Policy Working Paper 2, Report on Statistical Disclosure and Disclosure-Avoidance Techniques published in 1978. 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 report includes a tutorial, guidelines, and recommendations for good practice; recommendations for further research; and an annotated bibliography. The Subcommittee plans to organize seminars and workshops in order to facilitate further communication concerning disclosure limitation.

The Subcommittee on Disclosure Limitation Methodology was chaired by Nancy Kirkendall of the Energy Information Administration, Department of Energy.

**Members of the**
**Subcommittee on Disclosure Limitation Methodology**

Nancy J. Kirkendall, Chairperson
Energy Information Administration
Department of Energy

William L. Arends
National Agricultural Statistics Service
Department of Agriculture

Lawrence H. Cox
Environmental Protection Agency

Virginia de Wolf
Bureau of Labor Statistics
Department of Labor

Arnold Gilbert
Bureau of Economic Analysis
Department of Commerce

Thomas B. Jabine
Committee on National Statistics
National Research Council
National Academy of Sciences

Mel Kollander
Environmental Protection Agency

Donald G. Marks
Department of Defense

Barry Nussbaum
Environmental Protection Agency

Laura V. Zayatz
Bureau of the Census
Department of Commerce

**Acknowledgements**

In early 1992, an ad hoc interagency committee on Disclosure Risk Analysis was organized by Hermann Habermann, Office of Management and Budget. A subcommittee was formed to look at methodological issues and to analyze results of an informal survey of agency practices. That subcommittee became a Subcommittee of the Federal Committee on Statistical Methodology (FCSM) in early 1993. The Subcommittee would like to thank Hermann Habermann for getting us started, and Maria Gonzalez and the FCSM for adopting us and providing an audience for our paper.

Special thanks to Subcommittee member Laura Zayatz for her participation during the last two years. She helped to organize the review of papers, contributed extensively to the annotated bibliography and wrote the chapters on microdata and research issues in this working paper. In addition, she provided considerable input to the discussion of disclosure limitation methods in tables. Her knowledge of both theoretical and practical issues in disclosure limitation were invaluable.

Special thanks also go to Subcommittee member Bill Arends for his participation during the last two years. He helped in the review of papers, analyzed the results of the informal survey of agency practices, contacted agencies to get more detailed information and wrote the chapter on agency practices. He and Mary Ann Higgs pulled together information from all authors and prepared three drafts and the final version of this working paper, making them all look polished and professional. He also arranged for printing the draft reports.

Tom Jabine, Ginny deWolf and Larry Cox are relative newcomers to the subcommittee. Ginny joined in the fall of 1992, Tom and Larry in early 1993. Tom and Larry both participated in the work of the 1978 Subcommittee that prepared Statistical Policy Working Paper 2, providing the current Subcommittee with valuable continuity. Tom, Ginny and Larry all contributed extensively to the introductory and recommended practices chapters, and Tom provided thorough and thoughtful review and comment on all chapters. Larry provided particularly helpful insights on the research chapter.

Special thanks to Tore Dalenius, another participant in the preparation of Statistical Policy Working Paper 2, for his careful review of this paper. Thanks also to FCSM members Daniel Kasprzyk and Cynthia Clark for their thorough reviews of multiple drafts.

The Subcommittee would like to acknowledge three people who contributed to the annotated bibliography: Dorothy Wellington, who retired from the Environmental Protection Agency; Russell Hudson, Social Security Administration; and Bob Burton, National Center for Education Statistics.

Finally, the Subcommittee owes a debt of gratitude to Mary Ann A. Higgs of the National Agricultural Statistics Service for her efforts in preparing the report.

Nancy Kirkendall chaired the subcommittee and wrote the primer and tables chapters.

## TABLE OF CONTENTS

**TABLE OF CONTENTS (Continued)**

Page

**TABLE OF CONTENTS (Continued)**

Page

**TABLE OF CONTENTS (Continued)**

Page

**TABLE OF CONTENTS (Continued)**

Page

**Appendices**

# 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.

In early 1992 the Statistical Policy Office of the Office of Management and Budget convened an **ad hoc** interagency committee to review and evaluate statistical disclosure limitation methods used by federal statistical agencies and to develop recommendations for their improvement. Subsequently, the **ad hoc** committee became the Subcommittee on Disclosure Limitation Methodology, operating under the auspices of the Federal Committee on Statistical Methodology. This is the final report of the Subcommittee.

The Subcommittee's goals in preparing this report were to:

- update a predecessor subcommittee's report on the same topic (Federal Committee on Statistical Methodology, 1978);
- 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;
- encourage the development, sharing and use of software for the applications of disclosure limitation methods; and
- encourage research to develop improved statistical disclosure limitation methods, especially for public-use microdata files.

The Subcommittee believes that every agency or unit within an agency that releases statistical data should have the ability to select and apply 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.

To this end, our 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. We believe that the report will also be of interest to employees with similar responsibilities in other organizations that release statistical data, and to data users, who may find that it helps them to understand and use disclosure-limited 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.

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. Our concern is with methods designed to comply with the first element of the definition, in other words, to minimize the risk of **disclosure** (public identification) of the identity of individual units and information about them.

The release of statistical data inevitably reveals some information about individual data subjects. Disclosure occurs when information that is meant to be treated as confidential is revealed. Sometimes disclosure can occur based on the released data alone; sometimes disclosure results from combination of the released data with publicly available information; and sometimes disclosure is possible only through combination of the released data with detailed external data sources that may or may not be available to the general public. At a minimum, each statistical agency must assure that the risk of disclosure from the released data alone is very low.

Several different definitions of disclosure and of different types of disclosure have been proposed (see Duncan and Lambert, 1987 for a review of definitions of disclosure associated with the release of microdata). Duncan et al. (1993: 23-24) provide 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**).

In the above definition, the word "data" could have been substituted for "file", because each type of disclosure can occur in connection with the release of 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 and Microdata

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 or microdata files. 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 percents.

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 explicit identifiers, like name, address or Social Security number. Removing any such identifiers is an obvious first step in preparing for the release of a file for which the confidentiality of individual information must be protected.

### B.3. Restricted Data and Restricted Access

The confidentiality of individual information can be protected by restricting the amount of information 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 disclosure limitation methods described in this report provide confidentiality protection by restricting the data.

**Public-use** data products are released by statistical agencies to anyone without restrictions on use or other conditions, except for payment of fees to purchase publications or data files in electronic form. Agencies require that the disclosure risks for public-use data products be very low. The application of disclosure limitation methods to meet this requirement sometimes calls for substantial restriction of data content, to the point where the data may no longer be of much value for some purposes. In such circumstances, it may be appropriate to use procedures that allow some users to have access to more detailed data, subject to restrictions on who may have access, at what locations and for what purposes. Such restricted access arrangements normally require written agreements between agency and users, and the latter are subject to penalties for improper disclosure of individual information and other violations of the agreed conditions of use.

The fact that this report deals only with disclosure limitation procedures that restrict 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 the report of the Panel on Confidentiality and Data Access (see below) and in Jabine (1993b).

## C. Report of the Panel on Confidentiality and Data Access

In October 1993, while the Subcommittee was developing this report, the Panel on Confidentiality and Data Access, which was jointly sponsored by the Committee on National Statistics (CNSTAT) of the National Research Council and the Social Science Research Council, released its final report (Duncan et al., 1993). The scope of the CNSTAT report is much broader than this one: disclosure limitation methodology was only one of many topics covered and it was treated in much less detail than it is here. The CNSTAT panel's recommendations on statistical disclosure limitation methods (6.1 to 6.4) are less detailed than the guidelines and recommendations presented in this report. However, we believe that the recommendations in the two reports are entirely consistent with and complement each other. Indeed, the development and publication of this report is directly responsive to the CNSTAT Panel's Recommendation 6.1, which says, in part, that "The Office of Management and Budget's Statistical Policy Office should continue to coordinate research work on statistical disclosure analysis and should disseminate the results of this work broadly among statistical agencies."

## D. Organization of the Report

Chapter II, "Statistical Disclosure Limitation Methods: A Primer", provides a simple description and examples of disclosure limitation techniques that are commonly used to limit the risk of disclosure in releasing tables and microdata. Readers already familiar with the basics of disclosure limitation methods may want to skip over this chapter.

Chapter III describes disclosure limitation methods used by twelve 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 provides a systematic and detailed description and evaluation of statistical disclosure limitation methods for tables of frequency and magnitude data. Chapter V fulfills the same function for microdata. 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. Readers with more general interests may want to skip these chapters and proceed to Chapters VI and VII.

Due in part to the stimulus provided by our predecessor subcommittee's report (which we will identify in this report as Working Paper 2), improved methods of disclosure limitation have been developed and used by some agencies over the past 15 years. Based on its review of these methods, the Subcommittee has developed guidelines for good practice for all agencies. With separate sections for tables and microdata, Chapter VI presents guidelines for recommended practices.

Chapter VII presents an agenda for research on disclosure limitation methods. Because statistical disclosure limitation procedures for tabular data are more fully developed than those for microdata, the research agenda focuses more on the latter. The Subcommittee believed that a high priority should be assigned to research on how the quality and usefulness of data are affected by the application of disclosure limitation procedures.

Two 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 an annotated bibliography of articles about statistical disclosure limitation published since the publication of Working Paper 2.

## E. Underlying Themes of the Report

Five principal themes underlie the guidelines in Chapter VI and the research agenda in Chapter VII:

- There are legitimate differences between the disclosure limitation requirements of different agencies. Nevertheless, agencies 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 25 years. The time has come to make the best technology available to the entire federal statistical system. The Subcommittee believes that methods which have been shown to provide adequate protection against disclosure should be documented clearly in simple formats. The documentation and the corresponding software should then be shared among federal agencies.

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

- Several agencies have formed review panels 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.

- New research should focus on disclosure limitation methods for microdata and on how the methods used affect the usefulness and ease of use of data products.

# Statistical Disclosure Limitation: A Primer

This chapter provides a basic introduction to the disclosure limitation techniques which are used to protect statistical tables and microdata. It uses simple examples to illustrate the techniques. 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 them and provide statistical summaries to the public. Some of the data collected are considered proprietary by respondents. Agencies are authorized or required to protect individually identifiable data by a variety of statutes, regulations or policies. Cecil (1993) summarizes the laws that apply to all agencies and describes the statutes that apply specifically to the Census Bureau, the National Center for Education Statistics, and the National Center for Health Statistics. Regardless of the basis used to protect confidentiality, federal statistical agencies must balance two objectives: to provide useful statistical information to data users, and to assure that the responses of individuals are protected.

Not all data collected and published by the government are subject to disclosure limitation techniques. Some data on businesses collected for regulatory purposes are considered public. 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 to be publicly available. Methods of protecting data by restricting access are alternatives to statistical disclosure limitation. They are not discussed in this paper. See Jabine (1993) for a discussion of restricted access methods. 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.

The historical method of providing data to the public is via statistical tables. With the advent of the computer age in the early 1960's agencies also started releasing **microdata files**. 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, there are no identifiers on 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. A new method of releasing data has been introduced by the National Center for Education Statistics (NCES) in the 1990's. Data are provided on diskette or CD-ROM in a secure data base system with access programs which allow

users to create special tabulations. The NCES disclosure limitation and data accuracy standards are automatically applied to the requested tables before they are displayed to the user.

This chapter provides a simple description of the disclosure limitation techniques which are commonly used to limit the possibility of disclosing identifying information about respondents in tables and microdata. The techniques are illustrated with examples. The tables or microdata 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 the sections and chapters that follow: included are a discussion of the distinction between tables of frequency data and tables of magnitude data, a definition of table dimensionality, 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.

## 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 category 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}. A table that displays value of construction work done during a particular period in the state of Maryland by county and by 4-digit Standard Industrial Code (SIC) groups is comprised of cells such as the cell {SIC 1521, Prince George's County}.

### 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" over all 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. For practical purposes, entirely rigorous definitions are not necessary. The statistical disclosure limitation techniques used for magnitude data can be used for frequency data. However, for tables of frequency data other options are also available.

## 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 {SIC 1521, 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 {SIC 1521, Prince George's County, 1990}. The first two-dimensions are said to be presented in rows and columns, the third variable in "layers".

## B.3. What is Disclosure?

The definition of disclosure given in Chapter I, and discussed further below is very broad. Because this report documents the methodology used to limit disclosure, the focus is on practical situations. Hence, the concern is only with the disclosure of confidential information through the public release of data products.

As stated in Lambert (1993), "disclosure is a difficult topic. People even disagree about what constitutes a disclosure." 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 the likelihood of identification without revealing additional information is minuscule. 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 may occur when confidential information is revealed exactly or when it can be closely estimated. Thus, attribute disclosure comprises identification of the subject and divulging confidential information pertaining to the subject.

Attribute disclosure is the form of disclosure of primary concern to statistical agencies releasing 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 prevent attribute disclosure.

The third type of disclosure, **inferential disclosure**, occurs when 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. In general, statistical agencies are not concerned with inferential disclosure, for two reasons. 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, no data could be released. Second, inferences are designed to predict aggregate behavior, not individual attributes, and thus are often poor predictors of individual data values.

**Table 1: Example Without Disclosure**

**Number of Households by Heated Floorspace and Family Income**
**(Million U.S. Households)**

| | 1990 Family income | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Heated Floor Space sq ft | Total | Less than $5000 | $5000 to $9999 | $10000 to $14999 | $15000 to $24999 | $25000 to $34999 | $35000 to $49999 | $50000 or more |
| Fewer than 600 | 8.0 | 1.5 | 1.9 | 1.6 | 1.5 | .8 | .5 | .3 |
| 600 to 999 | 22.5 | 2.0 | 3.7 | 4.1 | 5.5 | 3.4 | 2.7 | 1.2 |
| 1000 to 1599 | 26.5 | 1.1 | 3.2 | 3.2 | 5.2 | 5.1 | 5.5 | 3.3 |
| 1600 to 1999 | 12.6 | .3 | 1.0 | 1.1 | 2.2 | 2.3 | 2.6 | 3.1 |
| 2000 to 2399 | 9.0 | Q | .5 | .6 | 1.3 | 1.3 | 2.3 | 2.8 |
| 2400 to 2999 | 7.8 | .2 | .3 | .5 | 1.0 | 1.4 | 1.7 | 2.7 |
| 3000 or more | 7.4 | Q | .2 | .3 | .7 | 1.0 | 1.3 | 3.8 |

NOTE: Q -- Data withheld because relative standard error exceeds 50%.
SOURCE: "Housing Characteristics 1990", Residential Energy Consumption Survey, Energy Information Administration, DOE/EIA-0314(90), page 54.

## C. 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 on the previous page as an example.

### C.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 made by multiplying an individual respondent's data by a sampling weight before they are aggregated. If sampling weights are not published, this weighting helps to make an individual respondent's data less identifiable from published totals. Because the weighted numbers represent all households in the United States, the counts in this table 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.

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. For a 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.

When tables of counts or frequencies are based directly on data from all units in the population (for example the 100-percent items in the decennial Census) then disclosure limitation procedures must be applied. 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. Such rules differ from agency to agency and from table to table. The special rules are generally designed to provide protection to data considered particularly sensitive by the agency. The second class is more general: a cell is defined to be sensitive if the number of respondents is less than some specified threshold (the threshold rule). Examples of both classes of disclosure limitation techniques are given in Sections II.C.2 and II.C.3.

### C.2. Special Rules

Special rules impose restrictions on the level of detail that can be provided in a table. For example, Social Security Administration (SSA) rules prohibit tabulations in which a detail cell 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: 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: Working Paper 2.

Table 2 is a two-dimensional table showing the number of beneficiaries by county and size of benefit. This table would not be publishable because the data shown for counties B and D violate Social Security's disclosure rules. For county D, there is only one non-empty detail cell, and a beneficiary in this county is known to be receiving benefits between $40 and $59 per month. This violates two rules. First the detail cell is equal to the cell 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 non-empty cells, but the range of possible benefits is from $40 to $79 per month, an interval of less than the required $50.

To protect confidentiality, Table 2 could be restructured and rows or columns combined (sometimes referred to as "rolling-up categories"). 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: Working Paper 2.

### C.3. 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. An agency may restructure tables and combine categories (as illustrated above), or use cell suppression, random rounding, controlled rounding or the confidentiality edit. Cell suppression, random rounding, controlled rounding and the confidentiality edit are described and illustrated below.

Table 4 is a fictitious example of a table with disclosures. The fictitious data set consists of information concerning delinquent children. We define a cell with fewer than 5 respondents to be sensitive. Sensitive cells are shown with an asterisk.

### C.3.a. Suppression

One of the most commonly used ways of protecting sensitive cells is via **suppression**. It is obvious that 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 cells must also be suppressed. These are referred to as **complementary** suppressions. While it is possible to select cells for complementary suppression manually, it is difficult to guarantee that the result provides adequate protection.

**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, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column headings are fictitious.

Table 5 shows an example of a system of suppressed cells for Table 4 which has at least two suppressed cells in each row and column. This table appears to offer protection to the sensitive cells. But does it?

**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 | $D_1$ | $D_2$ | $D_3$ | 20 |
| Beta | 20 | $D_4$ | $D_5$ | 15 | 55 |
| Gamma | $D_6$ | 10 | 10 | $D_7$ | 25 |
| Delta | $D_8$ | 14 | 7 | $D_9$ | 35 |
| Total | 50 | 35 | 30 | 20 | 135 |

NOTE: D indicates data withheld to limit disclosure.
SOURCE: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column headings are fictitious.

The answer is no. 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

$$(15 + D_1 + D_2 + D_3) + (20 + D_4 + D_5 + 15) - (D_1 + D_4 + 10 + 14) - (D_2 + D_5 + 10 + 7) = 20 + 55 - 35 - 30.$$

This reduces to $D_3 = 1$.

This example shows that selection of cells for complementary suppression is more complicated than it would appear at first. Mathematical methods of linear programming are used to automatically select cells for complementary suppression and also to **audit** a proposed suppression pattern (eg. 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, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column headings are fictitious.

### C.3.b. Random Rounding

In order to reduce the amount of data loss which occurs with suppression, the U.S. Census Bureau has investigated alternative methods to protect sensitive cells in tables of frequencies. Perturbation methods such as random rounding and controlled rounding are examples of such alternatives. 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.

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.

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 in 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. The New Zealand Department of Statistics has used random rounding in its publications and this is one of the criticisms it has heard (George and Penny, 1987).

**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, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column headings are fictitious.

### C.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. Linear programming methods are used to identify a controlled rounding for a table. There was considerable research into controlled rounding in the late 1970's and early 1980's and controlled rounding was proposed for use with data from the 1990 Census, (Greenberg, 1986). However, to date it has not been used by any federal statistical agency. Table 8 illustrates controlled rounding.

One disadvantage of controlled rounding is that it requires the use of specialized computer programs. At present these programs are not widely available. Another disadvantage is that controlled rounding solutions may not always exist for complex tables. These issues are discussed further in Chapters IV and VI.

### C.3.d. Confidentiality Edit

The **confidentiality edit** is a new procedure developed by the U.S. Census Bureau to provide protection in data tables prepared from the 1990 Census (Griffin, Navarro, and Flores-Baez, 1989). There are two different approaches: one was used for the regular decennial Census data (the 100 percent data file); the other was used for the long-form of the Census which was filed by a sample of the population (the sample data file). Both techniques apply statistical disclosure limitation techniques to the microdata files before they are used to prepare tables. The adjusted files themselves are not released, they are used only to prepare tables.

**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, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column headings are fictitious.

First, for the 100 percent microdata file, the confidentiality edit involves "data swapping" or "switching" (Dalenius and Reiss, 1982; Navarro, Flores-Baez, and Thompson, 1988). The confidentiality edit proceeds as follows. First, take a sample of records from the microdata file. Second, find a match for these records in some other geographic region, matching on a specified set of important attributes. Third, swap all attributes on the matched records. For small blocks, the Census Bureau increases the sampling fraction to provide additional protection. After the microdata file has been treated in this way it can be used directly to prepare tables and no further disclosure analysis is needed.

Second, the sample data file already consists of data from only a sample of the population, and as noted previously, sampling provides confidentiality protection. Studies showed that this protection was sufficient except in small geographic regions. To provide additional protection in small geographic regions, one household was randomly selected and a sample of its data fields were blanked. These fields were replaced by imputed values. After the microdata file has been treated in this way it is used directly to prepare tables and no further disclosure analysis is needed.

To illustrate the confidentiality edit as 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 9 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 confidentiality edit 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 confidentiality edit.

**Table 9: Fictitious Microdata**

**All Records in County Alpha Shown**
**Delinquent Children**

| Number | Child | County | HH education | HH income | Race |
|---|---|---|---|---|---|
| 1 | John | Alpha | Very high | 201 | B |
| 2 | Jim | Alpha | High | 103 | W |
| 3 | Sue | Alpha | High | 77 | B |
| 4 | Pete | Alpha | High | 61 | W |
| 5 | Ramesh | Alpha | Medium | 72 | W |
| 6 | Dante | Alpha | Low | 103 | W |
| 7 | Virgil | Alpha | Low | 91 | B |
| 8 | Wanda | Alpha | Low | 84 | W |
| 9 | Stan | Alpha | Low | 75 | W |
| 10 | Irmi | Alpha | Low | 62 | B |
| 11 | Renee | Alpha | Low | 58 | W |
| 12 | Virginia | Alpha | Low | 56 | B |
| 13 | Mary | Alpha | Low | 54 | B |
| 14 | Kim | Alpha | Low | 52 | W |
| 15 | Tom | Alpha | Low | 55 | B |
| 16 | Ken | Alpha | Low | 48 | W |
| 17 | Mike | Alpha | Low | 48 | W |
| 18 | Joe | Alpha | Low | 41 | B |
| 19 | Jeff | Alpha | Low | 44 | B |
| 20 | Nancy | Alpha | Low | 37 | W |

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

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

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 A. 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 10.

**Table 10: Fictitious Microdata**

**Delinquent Children -- After Swapping**
**Only County Alpha Shown**

| Number | Child | County | HH education | HH income | Race |
|---|---|---|---|---|---|
| 1 | John | Alpha | Very high | 201 | B |
| 2 | Jim | Alpha | High | 103 | W |
| 3 | Sue | Alpha | High | 75 | B |
| **4*** | **Alfonso** | **Alpha** | **Very high** | **61** | **W** |
| 5 | Ramesh | Alpha | Medium | 72 | W |
| 6 | Dante | Alpha | Low | 103 | W |
| 7 | Virgil | Alpha | Low | 91 | B |
| 8 | Wanda | Alpha | Low | 84 | W |
| 9 | Stan | Alpha | Low | 75 | W |
| 10 | Irmi | Alpha | Low | 62 | B |
| 11 | Renee | Alpha | Low | 58 | W |
| **12*** | **June** | **Alpha** | **High** | **56** | **B** |
| 13 | Mary | Alpha | Low | 54 | B |
| 14 | Kim | Alpha | Low | 52 | W |
| 15 | Tom | Alpha | Low | 55 | B |
| 16 | Ken | Alpha | Low | 48 | W |
| **17*** | **George** | **Alpha** | **Medium** | **48** | **W** |
| 18 | Joe | Alpha | Low | 41 | B |
| 19 | Jeff | Alpha | Low | 44 | B |
| **20*** | **Heather** | **Alpha** | **Low** | **37** | **W** |

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

4. Use the swapped data file directly to produce tables, see Table 11.

The confidentiality edit has a great advantage in that multidimensional tables can be prepared easily and the disclosure protection applied will always be consistent. A disadvantage is that it does not look as if disclosure protection has been applied.

**Table 11: Example -- Without Disclosure, Protected by Confidentiality Edit**

**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 | 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. Data only for County Alpha shown in Table 10.

## D. 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 the values reported by the largest, most highly visible respondents too closely. By protecting the largest values, we, in effect, protect all values.

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 which are selected with certainty. Thus, the units which 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. Such a cell is called a **sensitive** cell, and cannot be published.

The primary suppression rules most commonly used to identify sensitive cells by government agencies are the (n,k) rule, the p-percent rule and the pq rule. 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. Primary suppression rules are discussed in more detail in Section VI.B.1.

Once sensitive cells have been identified, there are only two options: restructure the table and collapse cells until no sensitive cells remain, or cell suppression. 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 II.C.3.a. The same problems exist for tables of magnitude data.

An administrative way to avoid cell suppression is used by a number of agencies. They obtain written permission 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. In this case, respondents are willing to accept the possibility that their data might be estimated closely from the published cell total.

## E. Microdata

Information collected about establishments is primarily magnitude data. These data are likely to be highly skewed, and there are likely to be high visibility respondents that could easily be identified via other publicly available information. As a result there are virtually no public use microdata files released for establishment data. Exceptions are a microdata file consisting of survey data from the Commercial Building Energy Consumption Survey, which is provided by the Energy Information Administration and two files from the 1987 Census of Agriculture provided by the Census Bureau. Disclosure protection is provided using the techniques described below.

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. (This is a topic for which research is needed, as discussed in Chapter VII). 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, virtually all 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 of variables on the file.

Additional methods used to disguise high visibility variables include:

1. 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.

### E.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. Part of the population (County A) was shown in Table 9. Second: remove obvious identifiers. 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 12. These manipulations result in the fictitious microdata file shown in Table 12.

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 could be identified by a knowledgeable person. Limiting the number of variables to 5 makes such identification less likely.

### E.2. High Visibility Variables

It may be that information available to others in the population could be used with the income data shown in Table 12 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. Such variables are called **high visibility** variables and require additional protection.

### E.2.a. Top-coding, Bottom-coding, Recoding into Intervals

Large income values are **top-coded** by showing only that the income is greater than 100 thousand dollars per year. Small income values are **bottom-coded** by showing only that the income is less than 40 thousand dollars per year. Finally, income values are **recoded** by presenting income in 10 thousand dollar intervals. The result of these manipulations yields the fictitious public use data file in Table 13. Top-coding, bottom-coding and recoding into intervals are among the most commonly used methods to protect high visibility variables in microdata files.

**Table 12: 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 | 51 | 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.

**Table 13: Fictitious Microdata -- Sampled, Identifiers Removed**

**Geographic Detail Limited, Income Top, Bottom and Recoded**
**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.

### E.2.b. Adding Random Noise

An alternative method of disguising high visibility 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 14. To produce this table, 14 random numbers were selected from the specified normal distribution, and were added to the income data in Table 12.

**Table 14: 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.

### E.2.c. Swapping or Rank Swapping

Swapping involves selecting a sample of the records, finding a match in the data base on a set of predetermined variables and swapping all other variables. Swapping (or switching) was illustrated as part of the confidentiality edit for tables of frequency data. In that example records were identified from different counties which 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.

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 which 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.

### E.2.d. Blank and Impute for Randomly Selected Records

The blank and impute method involves selecting a few records from the microdata file, blanking out selected variables and replacing them by imputed values. This technique is illustrated using data shown in Table 12. 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 Census sample data files (containing information from the long form of the decennial Census).

### E.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 group, say 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 would be subject to sampling, removal of identifiers, and limitation of geographic detail.

## F. Summary

This chapter has described the standard methods of disclosure limitation used by federal statistical agencies to protect both tables and microdata. It has relied heavily on simple examples to illustrate the concepts. The mathematical underpinnings of disclosure limitation 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.

# Current Federal Statistical Agency Practices

This chapter provides an overview of Federal agency policies, practices, and procedures for statistical disclosure limitation. 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. Some of the statistical agencies conduct or support research on statistical disclosure limitation methods. Information on recent and current research is included in Chapter VII.

This review of agency practices is based on two 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. Additional information for the Jabine paper was taken from an appendix to Working Paper 2.

The second source for this summary of agency practices was a late 1991 request by Hermann Habermann, Office of Management and Budget, to Heads of Statistical Agencies. Each agency was asked to provide, for use by a proposed ad hoc Committee on Disclosure Risk Analysis, a description of its current disclosure practices, standards, and research plans for tabular and microdata. Responses were received from 12 statistical agencies. Prior to publication, the agencies were asked to review this chapter and update any of their practices. 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 each of the 12 largest Federal statistical agencies as shown in Statistical Programs of the United States Government: Fiscal Year 1993 (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 unexpanded data for any one respondent must represent less than 60 percent of the total that is being published, except when written permission is obtained from that respondent ...

The second condition is an application of the (n,k) concentration rule. 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. 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)

Policy and Standards Memorandum (PSM) 12-89, dated July 12, 1989, outlines NASS policy for suppressing estimates and summary data to preserve confidentiality. PSM 7-90 (March 28, 1990) documents NASS policy on the release of unpublished summary data and estimates. In general, summary data and estimates may not be published if a nonzero value is based on information from fewer than three respondents or if the data for one respondent represents more than 60 percent of the published value. Thus NASS and ERS follow the same basic (n,k) concentration rule.

Suppressed data may be aggregated to a higher level, but steps are defined to ensure that the suppressed data cannot be reconstructed from the published materials. This is particularly important when the same data are published at various time intervals such as monthly, quarterly, and yearly. These rules often mean that geographic subdivisions must be combined to avoid revealing information about individual operations. Data for many counties cannot be published for some crop and livestock items and State level data must be suppressed in other situations.

NASS uses a procedure for obtaining waivers from respondents which permits publication of values that otherwise would be suppressed. Written approval must be obtained and updated periodically. If waivers cannot be obtained, data are not published or cells are combined to limit disclosure.

NASS generally publishes magnitude data only, but the same requirement of three respondents is applied when tables of counts are generated by special request or for reimbursable surveys done for other agencies.

NASS does not release public-use microdata. PSM 4-90 (Confidentiality of Information), PSM 5-89 (Privacy Act of 1974), and PSM 6-90 (Access to Lists and Individual Reports) cover NASS policies for microdata protection. Almost all NASS surveys depend upon voluntary reporting by farmers and business firms. This cooperation is secured by a statutory pledge that individual reports will be kept confidential and used only for statistical purposes.

While it is NASS policy to not release microdata files, NASS and ERS have developed an arrangement for sharing individual farm data from the annual Farm Costs and Returns Survey which protects confidentiality while permitting some limited access by outside researchers. The data reside in an ERS data base under security measures approved by NASS. All ERS employees with access to the data base operate under the same confidentiality regulations as NASS employees. Researchers wishing access to this data base must have their requests approved by NASS and come to the ERS offices to access the data under confidentiality and security regulations.

USDA's Office of the General Counsel (OGC) has recently (February 1993) 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 contract, cooperative agreement, cost-reimbursement agreement, 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.

To the extent future uses of data are known at the time of data collection, they can be explained to the respondent and permission 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.

NASS will also work with researchers and others to provide as much data for analysis as possible. Some data requests do not require individual reports and NASS can often publish additional summary data which are a benefit to the agricultural sector.

### A.2. Department of Commerce

#### A.2.a. Bureau of Economic Analysis (BEA)

BEA standards for disclosure limitation for tabular data are determined by its individual divisions. The International Investment Division is one of the few--and the major--division in BEA that collects data directly from U.S. business enterprises. It collects data on USDIA (U.S. Direct Investment Abroad), FDIUS (Foreign Direct Investment in the United States), and international services trade by means of statistical surveys. The surveys are mandatory and the

data in them are held strictly confidential under the International Investment and Trade in Services Survey Act (P.L. 94-472, as amended).

A standards statement, "International Investment Division Primary Suppression Rules," covers the Division's statistical disclosure limitation procedures for aggregate data from its surveys. This statement provides that:

> 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 some given percent of the top reporter's data, the cell is a primary suppression.

This is an application of the p-percent rule with no coalitions (c=1). This rule protects the top reporter from the second reporter, protects the second reporter from the top reporter, and automatically suppresses any cell with only one or two reporters. The value of that percent and certain other details of the procedures are not published "because information on the exact form of the suppression rules can allow users to deduce suppressed information for cells in published tables."

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 primary suppression, only reported data are counted in obtaining totals for the top two reporters; data estimated for any reason are not treated as confidential.

The statement includes several "special rules" covering rounded estimates, country and industry aggregates, key item suppression (looking at a set of related items as a group and suppressing all items if the key item is suppressed), and the treatment of time series data.

Complementary suppression is done partly by computer and partly by human intervention. All tables are checked by computer to see if the complementary suppression is adequate. Limited applications of linear programming techniques have been used to refine the secondary suppression methods and help redesign tables to lessen the potential of disclosure.

The International Investment Division publishes some tables of counts. These are counts pertaining to establishments and are not considered sensitive.

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. Beyond this limited sharing, BEA does not make its microdata on international investment and services available to outsiders. Confidentiality practices and procedures with respect to the data are clearly specified and strictly upheld.

According to Jabine (1993b), "BEA's Regional Measurement Division publishes estimates of local area personal income by major source. Quarterly data on wages and salaries paid by county are obtained from BLS's Federal/state ES-202 Program and BEA is obliged to follow statistical disclosure limitation rules that satisfy BLS requirements." Statistical disclosure limitation procedures used are a combination of suppression and combining data (such as, for two or more counties or industries).

Primary cell suppressions are identified by combining a systematic roll up of three types of payments to earnings and a dominant-cell suppression test of wages as a specified percentage of earnings. Two additional types of complementary cell suppressions are necessary to prevent the derivation (indirect disclosure) of primary disclosure cells. The first type is the suppression of additional industry cells to prevent indirect disclosure of the primary disclosure cells through subtraction from higher level industry totals. The second type is the suppression of additional geographic units for the same industry that are suppressed to prevent indirect disclosure through subtraction from higher level geographic totals. These suppressions are determined using computer programs to impose a set of rules and priorities on a multi-dimensional matrix consisting of industry and county cells for each state and region.

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

According to Jabine (1993b):

> "The Census Bureau's past and current practices in the application of statistical disclosure limitation techniques and its research and development work in this area cover a long period and are well documented. As a pioneer in the release of public-use microdata sets, Census had to develop suitable statistical disclosure limitation techniques for this mode of data release. It would probably be fair to say that the Census Bureau's practices have provided a model for other statistical agencies as the latter have become more aware of the need to protect the confidentiality of individually identifiable information when releasing tabulations and microdata sets."

The Census Bureau's current and recent statistical disclosure limitation practices and research are summarized in two papers by Greenberg (1990a, 1990b). Disclosure limitation procedures for frequency count tables from the 1990 Census of Population are described by Griffin, Navarro and Flores-Baez (1989). Earlier perspectives on the Census Bureau's statistical disclosure limitation practices are provided by Cox et al. (1985) and Barabba and Kaplan (1975). Many other references will be found in these five papers.

For tabular data from the 1992 Census of Agriculture, the Census Bureau will use the p-percent rule and will not publish the value of p. For other economic censuses, the Census Bureau uses the (n,k) rule and will not publish the values of n or k. Sensitive cells are suppressed and complementary suppressions are identified by using network flow methodology for two-dimensional tables (see Chapter IV). For the three-dimensional tables from the 1992 Economic Censuses, the Bureau will be using an iterative approach based on a series of two-dimensional

networks, primarily because the alternatives (linear programming methods) are too slow for the large amount of data involved.

For all demographic tabular data, other than data from the decennial census, disclosure analysis is not needed because of 1) very small sampling fractions; 2) weighted counts; and 3) very large categories (geographic and other). For economic magnitude data most surveys do not need disclosure analysis for the above reasons. For the economic censuses, data suppression is used. However, even if some magnitude data are suppressed, all counts are published, even for cells of 1 and 2 units.

Microdata files are standard products with unrestricted use from all Census Bureau demographic surveys. In February 1981, the Census Bureau established a formal Microdata Review Panel, being the first agency to do so. (For more details on methods used by the panel, see Greenberg (1985)). Approval of the Panel is required for each release of a microdata file (even files released every year must be approved). In February 1994, the Census Bureau added two outside advisory members to the Panel, a privacy representative and a data user representative. One criterion used by the Panel is that geographic codes included in microdata sets should not identify areas with less than 100,000 persons in the sampling frame, except for SIPP data (Survey of Income and Program Participation) for which 250,000 is used. This cutoff was adopted in 1981; previously a figure of 250,000 had been used for all data. Where businesses are concerned, the presence of dominant establishments on the files virtually precludes the release of any useful microdata.

The Census Bureau has legislative authority to conduct surveys for other agencies under either Title 13 or Title 15 U.S.C. Title 13 is the statute that describes the statistical mission of the Census Bureau. This statute also contains the strict confidentiality provisions that pertain to the collection of data from the decennial census of housing and population as well as the quinquennial censuses of agriculture, etc. 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; for example, the Census Bureau will not release identifiable microdata nor 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, samples must be drawn from sources other than the surveys and censuses covered by Title 13. If the sponsoring agency furnishes the frame, then the data are collected under Title 15 and the sponsoring agency's confidentiality rules apply.

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

As stated in NCES standard IV-01-91, Standard for Maintaining Confidentiality: " In reporting on surveys and preparing public-use data tapes, the goal is to have an acceptably low probability of identifying individual respondents." The standard recognizes that it is not possible to reduce this probability to zero.

The specific requirement for reports is that publication cells be based on at least three unweighted observations and subsequent tabulations (such as crosstabulations) must not provide additional information which would disclose individual identities. For percentages, there must be three observations in the numerator. However, in fact the issue is largely moot at NCES since all published tables for which disclosure problems might exist are typically based on sample data. For this situation the rule of three or more is superseded by the rule of thirty or more; that is, the minimum cell size is driven by statistical (variance) considerations.

For public-use microdata tapes, consideration is given to any proposed variables that are unusual (such as very high salaries) and data sources that may be available in the public or private sectors for matching purposes. Further details are documented in NCES's Policies and Procedures for Public Release Data.

Public-use microdata tapes must undergo a disclosure analysis. A Disclosure Review Board was established in 1989 following passage of the 1988 Hawkins-Stafford Amendment which emphasized the need for NCES to follow disclosure limitation practices for tabulations and microdata files. The Board reviews all disclosure analyses and makes recommendations to the Commissioner of NCES concerning public release of microdata. The Board is required to "...take into consideration information such as resources needed in order to disclose individually identifiable information, age of the data, accessibility of external files, detail and specificity of the data, and reliability and completeness of any external files."

The NCES has pioneered in the release of a new data product: a data base system combined with a spreadsheet program. The user may request tables to be constructed from many variables. The data base system accesses the respondent level data (which are stored without identifiers in a protected format and result from sample surveys) to construct these custom tables. The only access to the respondent level data is through the spreadsheet program. The user does not have a password or other special device to unlock the hidden respondent-level data. The software presents only weighted totals in tables and automatically tests to assure that no fewer than 30 respondents contribute to a cell (an NCES standard for data availability.)

The first release of the protected data base product was for the NCES National Survey of Postsecondary Faculty, which was made available to users on diskette. In 1994 a number of NCES sample surveys are being made available in a CD-ROM data base system. This is an updated version of the original diskette system mentioned above. The CD-ROM implementation is more secure, faster and easier to use.

The NCES Microdata Review Board evaluated the data protection capabilities of these products and determined that they provided the required protection. They believed that the danger of identification of a respondent's data via multiple queries of the data base was minimal because only weighted data are presented in the tables, and no fewer than 30 respondents contribute to a published cell total.

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

EIA standard 88-05-06 "Nondisclosure of Company Identifiable Data in Aggregate Cells" appears in the Energy Information Administration Standards Manual (April 1989). Nonzero value data cells must be based on three or more respondents. Primary suppression rule is the pq rule alone or in conjunction with some other subadditive rule. Values of pq (an input sensitivity parameter representing 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) selected for specific surveys are not published and are considered confidential. Complementary suppression is also applied to other cells to assure that the sensitive value cannot be reconstructed from published data. The Standards Manual includes a separate section with guidelines for implementation of 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). These guidelines have been augmented by other agencies' practices and appear as a Technical Note to this chapter.

An alternative approach pursued by managers of a number of EIA surveys from which data were published without disclosure limitation protection for many years was to use a Federal Register Notice to announce EIA's intention to continue to publish these tables without disclosure limitation protection. The Notice pointed out that the result might be that a knowledgeable user could estimate an individual respondent's data.

For most EIA surveys that use the pq rule, complementary suppressions are selected manually. One survey system that publishes complex tables makes use of software designed particularly for that survey to select complementary suppressions. It assures that there are at least two suppressed cells in each dimension, and that the cells selected are those of lesser importance to data users.

EIA does not have a standard to address tables of frequency data. However, it appears that there are only two routine publications of frequency data in EIA tables, 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 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 from 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.

## A.5. Department of Health and Human Services

### A.5.a. National Center for Health Statistics (NCHS)

NCHS statistical disclosure limitation techniques are presented in the NCHS Staff Manual on Confidentiality (September 1984), Section 10 "Avoiding Inadvertent Disclosures in Published Data" and Section 11 "Avoiding Inadvertent Disclosures Through Release of Microdata Tapes." No magnitude data figures should be based on fewer than three cases and a (1, 0.6) (n,k) rule is used. Jabine (1993b) points out 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 3". The acceptable ways to solve the problem (for either tables of frequency data or tables of magnitude data) are to combine rows or columns, or to use cell suppression (plus complementary suppression).

The above rules apply only for census surveys. For their other data, which come from sample surveys, the general policy is that "the usual rules precluding publication of sample estimates that do not have a reasonably small relative standard error should prevent any disclosures from occurring in tabulations from sample data."

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. The manual contains rules that apply to all microdata tapes released which contain any information about individuals or establishments, except where the data supplier was told prior to providing the information that the data would be made public. Detailed information that could identify individuals (for example, date of birth) should not be included. Geographic places and characteristics of areas with less than 100,000 people are not to be identified. Information on the drawing of the sample which could identify data subjects should not be included. All new microdata sets must be reviewed for confidentiality issues and approved for release by the Director, Deputy Director, or Assistant to the Director, NCHS.

### A.5.b. Social Security Administration (SSA)

SSA basic rules are from a 1977 document "Guidelines for Preventing Disclosure in Tabulations of Program Data," published in Working Paper 2. A threshold rule is used in many cases. In general, the rule is 5 or more respondents for a marginal cell. For more sensitive data, 3 or more respondents for all cells may be required. IRS rules are applied for publications based on IRS data. The SSA guidelines established in 1977 are:

a) No tabulation should be released showing distributions by age, earnings or benefits in which the individuals (or beneficiary units, where applicable) in any group can be identified to

   (1) an age interval of 5 years or less.
   (2) an earnings interval of less than $1000.
   (3) a benefit interval of less than $50.

b) For distribution by variables other than age, earnings and benefits, no tabulation should be released in which a group total is equal to one of its detail cells. Some exceptions to this rule may be made on a case-by-case basis when the detail cell in question includes individuals in more than one broad category.

c) The basic rule does not prohibit empty cells as long as there are 2 or more non-empty cells corresponding to a marginal total, nor does it prohibit detail cells with only one person. However, additional restrictions (see below) should be applied whenever the detailed classifications are based on sensitive information. The same restrictions should be applied to non-sensitive data if it can be readily done and does not place serious limitations on the uses of the tabulations. Additional restrictions may include one or more of the following:

   (1) No empty cells. An empty cell tells the user that an individual included in the marginal total is not in the class represented by the empty cell.

   (2) No cells with one person. An individual included in a one-person cell will know that no one else included in the marginal is a member of that cell.

SSA mentions ways of avoiding disclosure to include a) suppression and grouping of data and b) introduction of error (for example, random rounding). In 1978 the agency tested a program for random rounding of individual tabulation cells in their semi-annual tabulations of Supplemental Security Income State and County data. Although SSA considered random rounding and/or controlled rounding they decided not to use it. SSA did not think that it provided sufficient protection, and feared that the data were less useful than with suppression or combining data. Thus, their typical method of dealing with cells that represent disclosure is through suppression and grouping of data.

One example of their practices is from "Earnings and Employment Data for Wage and Salary Workers Covered Under Social Security by State an County, 1985", in which SSA states that they do not show table cells with fewer than 3 sample cases at the State level and fewer than 10 sample cases at the county level to protect the privacy of the worker. These are IRS rules and are applied because the data come from IRS.

Standards for microdata protection are documented in an article by Alexander and Jabine (1978). SSA's basic policy is to make microdata without identifiers as widely available as possible, subject only to necessary legal and operational constraints. SSA has adopted a two-tier system for the release of microdata files with identifiers removed. Designated as public-use files are those microdata files for which, in SSA's judgment, virtually no chance exists that users will be able to identify specific individuals and obtain additional information about them from the records on the file. No restrictions are made on the uses of such files. Typically the public-use files are based on national samples with small sampling fractions and the files contain no geographic codes or at most regional and/or size of place identifiers. Those microdata files considered as carrying a disclosure risk greater than is acceptable for a public-use file are released only under restricted use conditions set forth in user agreements, including the purposes to be made of the data.

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

Cells with fewer than 10 observations are not displayed in published tables. Display of geographic data is limited by Census Bureau Title 13 restrictions for those data collected for BJS by the Census Bureau. 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. In addition to BJS statutes, the release of all data collected by the Census Bureau for BJS is further restricted by Title 13 microdata restrictions. Individual identifiers are routinely stripped from all other microdata files before they are released for public use.

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

Commissioner's Order 3-93, "The Confidential Nature of BLS Records," dated August 18, 1993, contains BLS's policy on the confidential data it collects. One of the requirements is that:

> 9e. 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 the data of any specific respondent to be imputed from the published information.

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 (for two programs this authority is delegated to specific Associate Commissioners).

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 the (n,k) concentration rule. For example, the BLS collective bargaining program, a census of all collective bargaining agreements covering 1,000 workers or more, requires that (1) each cell must have three or more units and (2) no unit can account for more than 50 percent of the total employment for that cell. The ES-202 program, a census of monthly employment and quarterly wage information from Unemployment Insurance filings, uses a threshold rule that requires three or more establishments and a concentration rule of (1,0.80). In general, the values of k range from 0.5 to 0.8. In a few cases, a two-step rule used--an (n,k) rule for a single establishment is followed by an (n,k) rule for two establishments.

Several wage and compensation statistics programs use a more complex approach that combines disclosure limitation methods and a certain level of reliability before the estimate can be published. For instance, one such approach uses a threshold rule requiring that each estimate be comprised of at least three establishments (unweighted) and at least six employees (weighted). It then uses a (1,0.60) concentration rule where n can be either a single establishment or a multi-establishment organization. Lastly, the reliability of the estimate is determined and if the estimate meets a certain criterion, then it can be published.

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. 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, are similar to those used by the Bureau of the Census.

### A.8. Department of the Treasury: Internal Revenue Service, Statistics of Income Division (IRS, SOI)

Chapter VI of the SOI Division Operating Manual (January 1985) specifies that "no cell in a tabulation at or above the state level will have a frequency of less than three or an amount based on a frequency of less than three." Data cells for areas below the state level, for example counties, require at least ten observations. Data cells considered sensitive are suppressed or combined with other cells. Combined or deleted data are included in the corresponding column totals. SOI also documents its disclosure procedures in its publications, "Individual Income Tax Returns, 1989" and "Corporation Income Tax Returns, 1989."

One example given (Individual Income Tax Returns, 1989) states that if a weighted frequency (the weighting frequency is obtained by dividing the population count of returns in a sample stratum by the number of sample returns for that stratum) is less than 3, the estimate and its corresponding amount are combined or deleted in order to avoid disclosure.

SOI makes available to the public a microdata file of a sample of individual taxpayers' returns (the Tax Model). The data must be issued in a form that protects the confidentiality of individual taxpayers. Several procedural changes were made in 1984 including: removing some data fields and codes, altering some codes, reducing the size of subgroups used for the blurring process, and subsampling high-income returns.

Jabine points out that "the SOI Division has sponsored research on statistical disclosure limitation techniques, notably the work by Nancy Spruill (1982, 1983) in the early 1980's, which was directed at the evaluation of masking procedures for business microdata. On the basis of her findings, the SOI released some microdata files for unincorporated businesses." Except for this and a few other instances, "the statistical agencies have not issued public-use microdata sets of establishment or company data, presumably because they judge that application of the statistical disclosure limitation procedures necessary to meet legal and ethical requirements would produce files of relatively little value to researchers. Therefore, access to such files continues to be almost entirely on a restricted basis."

### A.9. Environmental Protection Agency (EPA)

EPA program offices are responsible for their own data collections. The types and subjects of data collections are required by statutes and regulations and the need to conduct studies. Data confidentiality policies and procedures are required by specific Acts or are determined on a case-by-case basis. Individual program offices are responsible for data confidentiality and disclosure as described in the following examples.

The Office of Prevention, Pesticides and Toxic Substances (OPPT) collects confidential business information (CBI) for which there are disclosure avoidance requirements. These requirements come under the Toxic Substance Control Act (TSCA). Procedures are described in the CBI security manual.

An OPPT Branch that conducts surveys does not have a formal policy in respect to disclosure avoidance for non-CBI data. The primary issue regarding confidentiality for most of their data collection projects is protection of respondent name and other personal identification characteristics. Data collection contractors develop a coding scheme to ensure confidentiality of these data elements and all raw data remain in the possession of the contractor. Summary statistics are reported in final reports. If individual responses are listed in an appendix to a final report identities are protected by using the contractor's coding scheme.

In the Pesticides Program, certain submitted or collected data are covered by the provisions of the Federal Insecticide, Fungicide and Rodenticide Act (FIFRA). The Act addresses the protection of CBI and even includes a provision for exemption from Freedom of Information Act disclosure for information that is accorded protection.

Two large scale surveys of EPA employees have taken place in the past five years under the aegis of intra-program task groups. In each survey, all employees of EPA in the Washington, D.C. area were surveyed. In each instance, a contractor was responsible for data collection,

analysis and final report. Data disclosure avoidance procedures were in place to ensure that the identification and responses of individuals and specific small groups of individuals could not occur.

All returned questionnaires remained in the possession of the contractor. The data file was produced by the contractor and permanently remained in the contractor's possession. Each record was assigned a serial number and the employee name file was permanently separated from the survey data file.

The final reports contained summary statistics and cross-tabulations. A minimum cell size standard was adopted to avoid the possibility of disclosure. Individual responses were not shown in the Appendix of the reports. A public-use data tape was produced for one of the surveys it included a wide array of tabulations and cross-tabulations. Again, a minimum cell-size standard was used.

## B. Summary

Most of the 12 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. 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.

### 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, NCHS, and BLS) publish the values of the parameters they use in (n,k) concentration rules, whereas others do not. Minimum cell sizes of 3 are almost invariably used, because each member of a cell of size 2 could derive a specific value for the other member.

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. BLS uses the lower value of k in one of its programs and the upper value in another. 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. They both have the property of subadditivity, and they give the disclosure analyst flexibility to specify how much gain in information about its competitors by an individual company is acceptable. Also, they provide a somewhat more satisfying rationale for what is being done than does the arbitrary selection of parameters for a (n,k) concentration rule.

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.

### B.2. Microdata

Only about half of the agencies included in this review have established statistical disclosure limitation procedures for microdata. Some agencies pointed out that the procedures for surveys they sponsored were set by the Census Bureau's Microdata 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 more recently NCES--have all established formal procedures 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, noise introduction, and rounding are other methods used to prevent disclosure.

## Summary of Agency Practices

| Agency | Magnitude Data | Frequency Data | Microdata | Waivers |
|---|---|---|---|---|
| ERS | (n,k), (1,.6)<br>3+ | Threshold Rule<br>3+ | No | Yes |
| NASS | (n,k), (1,.6)<br>3+ | Threshold Rule<br>3+ | No | Yes |
| BEA | p-percent<br>c=1 | 1+ Not Sensitive for Est. Surveys | No | No |
| BOC | (n,k),<br>p-percent (Ag Census),<br>Parameters Confidential | 1+ (Economic Census),<br>Confidentiality Edit (Demographic Census),<br>Accuracy Requirements (Demographic Surveys) | Yes --<br>Microdata Review Panel | Yes |
| NCES | 3+ Accuracy Standards | 3+ Accuracy Standards | Yes --<br>Disclosure Review Board<br>"Protected" Data File | No |
| EIA | pq, Parameters Confidential | Accuracy Requirements | Yes --<br>Agency Review | Yes |
| NCHS | (n,k), (1,.6) | 3+ | Yes -- Review by Director or Deputy | No |
| SSA | 3+ | Threshold Rule<br>5+ Marginals<br>3+ Cells | Yes --<br>Agency Review | No |
| BJS | N/A | 10+, Accuracy Requirements | Yes -- Legislatively Controlled,<br>Agency Review | No |
| BLS | (n,k)<br>Parameters<br>Vary by Survey | Minimum Number<br>Varies by Data Collection | BOC Collects<br>Title 13 | Yes |
| IRS | 3+ | 3+ | Yes -- Legislatively Controlled | No |
| EPA | Minimum Number<br>Varies by Data Collection | Minimum Number<br>Varies by Data Collection | Yes --<br>Agency Review | No |

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. An agency may restructure tables and combine categories or use cell suppression, random rounding, controlled rounding or the confidentiality edit. 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)

**The Confidentiality Edit:** The **confidentiality edit** is a new procedure developed by the U.S. Census Bureau to provide protection in data tables prepared from the 1990 Census. There are two different approaches: one was used for the regular decennial Census data (the 100 percent data file); the other was used for the long-form of the Census which was filed by a sample of the population (the sample data file). Both techniques apply statistical disclosure limitation techniques to the microdata files before they are used to prepare tables. The adjusted files themselves are not released, they are used only to prepare tables. For the basic decennial Census data (the 100 percent data file) 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. All variables in the matched records were interchanged. This technique is called switching. The key variables used for matching were selected to assure that Census aggregates mandated by law would be unchanged by the confidentiality edit. For the sample file, consisting of the data collected on the long form, the sampling was shown to provide adequate protection except in small geographic regions (blocks). In these regions one record was selected and a sample of the variables on the record were blanked and replaced by imputed data. This procedure is called "blank and impute". Both "blank and impute" and "switching" have been suggested as methods to provide disclosure limitation to microdata files. (See Sections II.C.3.d and IV.A.2)

**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 prespecified percentage, p. This is referred to as the "p-percent estimation equivocation level" in Working Paper 2, but it is more generally referred to as the **p-percent rule**. For this rule the parameter c is the size of a coalition, a group of respondents who pool their data in an attempt to estimate the largest reported value. (See Section IV.B.1.a)

**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 ($p < q < 100$). (See Section IV.B.1.b)

**The (n,k) Rule:** The **(n,k) rule**, or dominance rule was described as follows in 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)

# 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 confidentiality purposes: tables of frequency (or count) data and tables of magnitude data. Tables of frequency data show the percent of the population which 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. 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 include cell suppression, data perturbation methods and the confidentiality edit.

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 only disclosure limitation method used.

Tables of frequency data are discussed in Section A. The major methodological areas of interest are in controlled rounding and the confidentiality edit. 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 which 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, cell suppression, perturbation methods, random rounding, controlled rounding and the confidentiality edit. 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.

Perturbation methods include random rounding and controlled rounding as special cases, and controlled rounding is a special case of random rounding. Controlled rounding is the most desirable of the perturbation methods, because it results in an additive table (sums of row, column and layer entries are equal to the published marginal total), can always be solved for two-dimensional tables, and can generally be solved for three-dimensional tables. Section 1 provides more detail on the methodology used in controlled rounding. The confidentiality edit is a relatively new technique and was used by the Census Bureau to publish data from the 1990 Census. The confidentiality edit is discussed in Section 2.

### 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 which 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 which 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 which are related hierarchically along one dimension (Cox and George, 1989). For three-dimensional tables an exact network solution does not exist. 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.

Current research focuses on refinements to the two-dimensional problem and solutions to the three-dimensional problem. These are described in Greenberg (1988a); Fagan, Greenberg and Hemmig (1988); Kelly, Assad and Golden (1990); Kelly, Golden, Assad and Baker (1990); and Kelly, Golden and Assad (1990c).

Although solutions to the controlled rounding problem are available, controlled rounding has not been used by U.S. government agencies.

### A.2. The Confidentiality Edit

The newest approach to protection of tables of frequency data is the confidentiality edit, (Griffin, Navarro and Flores-Baez, 1989), which was illustrated in II.C.2.d. 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. In 1990 the confidentiality edit used different procedures to protect tables based on these two systems of data. For the basic decennial Census data (the 100 percent data file) 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. All variables in the matched records were interchanged. This technique is called switching. The key variables used for matching were selected to assure that Census aggregates mandated by law would be unchanged by the confidentiality edit.

The effectiveness of the data switching procedure was investigated by simulation (Navarro, Flores-Baez and Thompson, 1988). It was found that switching provides adequate protection except in areas with small populations (blocks). The solution used by the Census Bureau was to use higher sampling fractions to select households for switching in such areas.

For the sample file, consisting of the data collected on the long form, the sampling was shown to provide adequate protection except in small geographic regions (blocks). In these regions one record was selected and a sample of the variables on the record were blanked and replaced by imputed data. This procedure is called "blank and impute". Both "blank and impute" and "switching" have been suggested as methods to provide disclosure limitation to microdata files.

Once the data for the sampled households were switched in the 100% microdata file and blank and impute was done in the sample file, the files were used directly to prepare all Census tabulations. The advantage of the confidentiality edit is that it maximizes the information that can be provided in tables. Additionally, all tables are protected in a consistent way.

## 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 which 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. Once 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, and cell suppression. 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 which are "optimal" in some sense is an extremely complex problem. Complementary suppression is discussed in Section 2.b.

Instead of suppressing data, some agencies ask respondents for permission to publish cells even though they are sensitive. This is referred to as the waiver approach. Waivers are signed records of the respondents permission to publish. 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 believe that the data are not particularly sensitive before they will sign waivers.

### B.1. Definition of Sensitive Cells

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). This is one of the important advancements since Working Paper 2. 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: $x_1 \geq x_2 \geq \ldots x_N \geq 0$. 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 $w_1 \geq w_2 \geq ... \geq w_N$. 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 which determine the values taken by the weights, $w_1$ ..., $w_N$. 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 which have been 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.

### 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 prespecified percentage, p. This is referred to as the "p-percent estimation equivocation level" in 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 $x_{c+2}$, ..., $x_N$ to within 100 percent, they have an estimate for the sum of these smallest respondents, E, such that

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

They can estimate the value of $x_1$ by subtracting the value they reported to the survey ($\sum_{i=2}^{c+1} x_i$) 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.

### 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

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

This can be rewritten as the linear sensitivity rule

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

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 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 $N \leq n$, $S^{(n,k)} = \sum_{i=1}^{N} x_i > 0$ 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 (p/q) 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 are 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.

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.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. 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. 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 but publishing the marginal totals, the implicit result is that the unions of the suppressed cells in rows, columns and layers are published. 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 did guarantee 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, 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 a suppression. This is not sufficient, however, as was illustrated in II.C.2.a.

### B.2.a.iii. Software

Automated methods of auditing a suppression pattern for the second problem have been available since the mid 1970's at the U.S. Census Bureau, (Cox, 1980) and at Statistics Canada, (Sande, 1984). Modern versions of audit software set up the linear programming problem as described in Zayatz (1992a) 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. 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. These audit methods are applicable to tables of both magnitude and frequency.

In more recent formulations of the complementary suppression problem 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 which will provide protection to all respondents in the cell, and the algorithm which selects cells for complementary suppression now assures that the primary cells cannot be estimated more accurately than that specified tolerance interval. The complementary suppressions selected by such computer systems do not require additional audits.

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

Automatic methods,of selecting cells for complementary suppression have also been available since the late 1970's at Statistics Canada, (Sande, 1984), and at the U. S. Census Bureau, (Cox, 1980). These programs typically rely on linear programming methods, either using standard approaches or approaches which make use of special structures in the data, such as network theory. The Statistics Canada software, CONFID, has been made available to U. S. Government agencies, where it is currently being implemented and evaluated. Complementary suppression software is applicable to tables of both frequency and magnitude.

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 which minimizes a cost function (usually the sum of the suppressed values) is identified. Zayatz (1992b) describes the formulation for two-dimensional tables. Zayatz (1992a) 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, it is essentially impossible to use.

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 (Cox, 1992). 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 (Cox and Ernst, 1982; Cox, 1987b) and to two-dimension}l tables with subtotal constraints in one dimension (Cox and George, 1989). Subtotal constraints occur when data in one dimension have a hierarchical additive structure. One common example of this structure occurs when one variable is the Standard Industrial Classification (SIC) code. An interior table cell might relate to a specific 4 digit SIC code, with subtotals given by 3-digit SIC codes, and the marginal total given by the appropriate 2-digit code. Sullivan (1992b) describes how to represent tables with this hierarchical structure in a network.

Complementary suppression and controlled rounding can both be solved using network theory. The parallelism between the two problems was demonstrated in Cox, Fagan, Greenberg and Hemmig (1986). 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 one of two approaches, (Zayatz, 1992a). 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.

There is continuing research in developing faster and more efficient procedures for both two-dimensional and three-dimensional tables, (see Greenberg, 1986; Kelly, 1990; Kelly, Golden and Assad, 1990a and 1990b; Kumar, Golden and Assad, 1992; Desilets, Golden, Kumar and Wang, 1992; Lougee-Heimer, 1989; and Wang, Sun and Golden, 1991). Mathematical approaches mentioned in current research include methods based on integer programming, network flow theory, and neural networks.

As mentioned above, one possible objective 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. As observed by Cox (1992) "what data analysts want to see coming out of the complementary suppression process isn't always minimum number of suppressions and isn't always minimum value suppressed, but rather sometimes one and sometimes the other and, in general, a suppression pattern that somehow balances these two objectives to avoid worst-case scenarios."

Further research is needed into the identification of cost functions for use in selecting the "optimal" complementary suppressions. Possibilities here include both research into a cost function to be used for a single run of the software, as well as cost functions for use in multiple runs of the software. An example is development of a cost function to be used during a second pass through the software to remove superfluous suppressions. Rowe (1991) and Zayatz (1992b) provide examples of current research into cost functions.

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. 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.

## C. 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

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

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

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

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

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 $(z_1 + z_2) > k_2$. 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 $(\frac{100}{\frac{p}{100} + 1}, 0)$. This line intersects $z_1=z_2$ at the point $(\frac{100}{\frac{p}{100} + 2}, \frac{100}{\frac{p}{100} + 2})$ 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

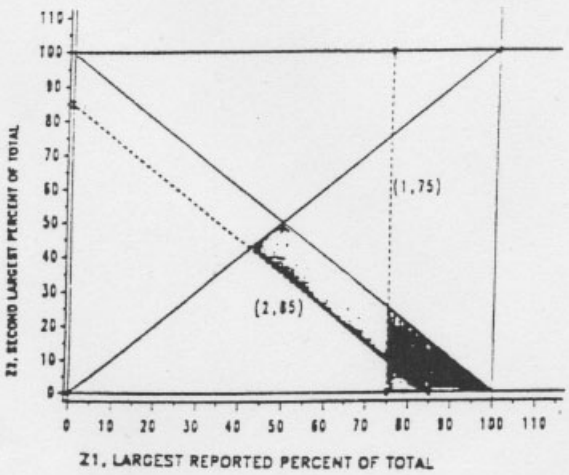



FIGURE 1B
EXAMPLES OF SUPPRESSION REGIONS
THE P-PERCENT RULE WITH P=17.65 PERCENT, AND P=35.3 PERCENT

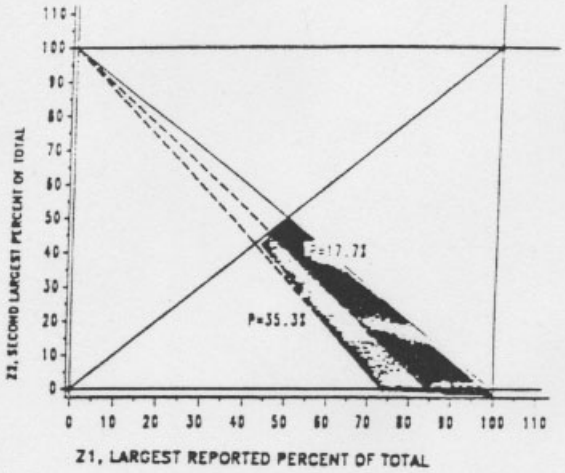



FIGURE 1C
P-PERCENT LESS CONSERVATIVE THAN (2,85)
P = 17.7 PERCENT

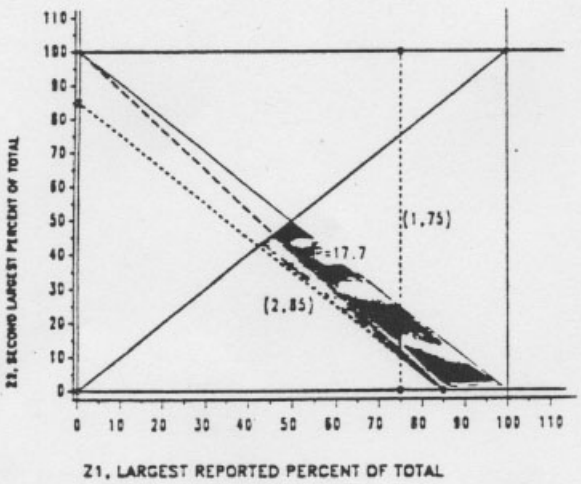
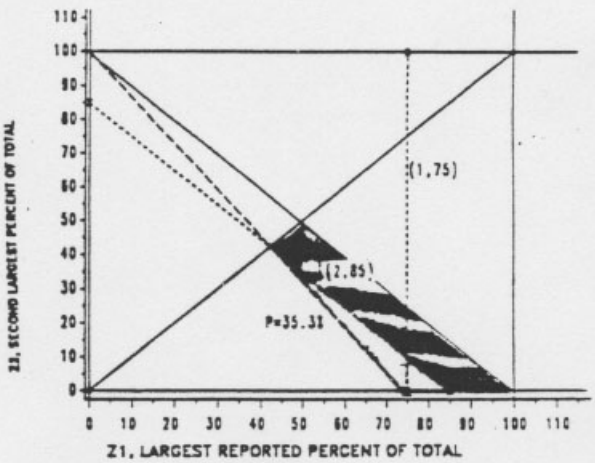



FIGURE 1D
P-PERCENT MORE CONSERVATIVE THAN (2,85) AND (1,75)
P = 35.3 PERCENT




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

$$\frac{p}{100} = \frac{100}{k_2} - 1.$$

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

$$k_2 = \frac{100}{\frac{p}{100} + 1}.$$

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 $k_1$ 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 $S^{(n,k)}(X) \leq S^{p\%}(X)$. 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 |
|---|---|---|---|---|
| | $x_1^1 = 100$ | $x_1^2 = 1$ | $x_1^3 = 100$ | $x_1^T = 100$ |
| | | $x_2^2 = 1$ | | $x_2^T = 100$ |
| | | $x_3^2 = 1$ | | $x_3^T = 1$ |
| | | . | | . |
| | | . | | . |
| | | . | | . |
| | | $x_{20}^2 = 1$ | | $x_{22}^T = 1$ |
| 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

$S^{(2,85)}$ (Cell 1 $\cup$ Cell 2) = 100 + 1 - 5.667*19 = -6.67
$S^{17.6\%}$ (Cell 1 $\cup$ Cell 2) = 100 - 5.667*19 = -7.67
$S^{(1,73.91)}$(Cell 1 $\cup$ Cell 2) = 100 - 2.833*20 = 43.34
$S^{35.29\%}$ (Cell 1 $\cup$ Cell 2) = 100 - 2.834*19 = 46.16

$S^{(2,85)}$ (Total) = 100 + 100 - 5.667*20 = 86.66
$S^{17.6\%}$ (Total) = 100 - 5.667*20 = -13.34
$S^{(1,73.91)}$ (Total) = 100 - 2.833*120 = -239.96
$S^{35.29\%}$ (Total) = 100 - 2.834*20 = 43.32

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.

# 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 II.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. We explain the reasons for this in Section C.4 of this chapter.

This report concerns **public-use** microdata files that are available at a fee to anyone who wishes to purchase them. 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 (Jabine, 1993a).

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. A statistical agency releasing a microdata file containing confidential data must do its best to ensure that an outside data user cannot 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.

In 1962, the Social Security Administration's Office of Research and Statistics began releasing microdata files on tape from their Continuous Work History Sample to other Federal and State agencies. There were essentially no restrictions on these files, and they were later used extensively by non-government researchers. The first broad release of a public-use microdata file occurred in 1963 when the Census Bureau released a file consisting of a 1 in 1,000 sample from the 1960 Census of Population and Housing. A few years later, the Census Bureau publicly released a microdata file from the Current Population Survey. Currently, unrestricted microdata files are standard products of all Census Bureau demographic surveys. They are available to any purchaser, and researchers use them extensively (Greenberg and Zayatz, 1991). Several other Federal agencies including the National Center for Education Statistics, National Center for Health Statistics, Energy Information Administration, and Internal Revenue Service currently 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, and there are several factors that play a role in the disclosure risk of a microdata 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 which 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 simply as a challenge or in order to discredit the agency that published the 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).

These issues as well as the contents of the file should be considered when an agency discusses the potential release of a proposed microdata file.

### 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 visibility records. Some records on the file may represent respondents with unique characteristics such as very unusual jobs (movie star, Federal judge) or very large incomes (over one million dollars). An agency must decrease the visibility of such records.

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 outside 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 sample, can greatly help an intruder to identify individuals on a microdata file from that survey. Warning survey respondents not to tell others about their participation in the survey might alleviate this problem, but it also might 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 that 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 theoretical possibility and precautions should be taken.

The potential for linking files (and thus the disclosure risk) is increased as the number of variables common to both files increases, as the accuracy or resolution of the data increases, and as the number of outside files, not all of which may be known to the agency releasing the microdata file, increases. Also, as computer technology advances, it becomes quicker, easier, and less costly to link files.

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 the microdata file to another published microdata file. This happens in the case of longitudinal surveys, such as the Census Bureau's Survey of Income and Program Participation, 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. 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.

The disclosure risk of a file increases as the structure of the data becomes more complex. If two records are known to belong to the same cluster (for example, household), there is a greater risk that either one may be identified (even if no information about the cluster per se is provided).

### A.3. Factors that Naturally Decrease Risk

Probably the factor with the biggest role in decreasing risk is the fact that almost all microdata files contain records that represent only a sample of the population. 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 need 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. 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. This difference in age complicates the job of linking files.

The naturally occurring noise in the microdata file and in potentially matchable files decreases the ability to link files (Mugge, 1983b). 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, and often the variables on different files will be recoded differently. These differences in variable definitions and recodes 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 mentioned previously, as computer technology advances, these factors are diminished.

**B. Mathematical Methods of Addressing the Problem**

Although several mathematical measures of risk have been proposed, none has been widely accepted. Techniques for reducing the disclosure risk of microdata include methods that reduce the amount of information provided to data users and 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. However, none has been widely accepted. More research is necessary to identify the best disclosure limitation methodology sufficient for both 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. Record Linkage Techniques, 1985 -- Proceedings of the Workshop on Exact Matching Methodologies presents reprints of the major background papers in record linkage as well as discussions of current work. Another related problem concerns computer science methods of addressing disclosure risk that involve storing the confidential data in a sequential database and monitoring and restricting access to the data (Lunt, 1990). The danger that this method seeks to avoid is that a data user could gain information about an individual respondent through multiple queries of the database. The National Center for Education Statistics has released compressed data with access controls along with software that allows users to obtain weighted cross tabulations of the data without being able to examine individual data records. Due to constraints of time and space, we will not describe these techniques in detail, though we will discuss some suggested research concerning

databases in Section VII.C.1. Readers interested in these techniques are encouraged to consult the references.

### B.1. Proposed Measures of Risk

Several researchers have proposed mathematical measures of the disclosure risk of a microdata file (Spruill, 1982; Duncan and Lambert, 1987; Paass, 1988; Mokken, Pannekoek, and Willenborg, 1990; Skinner, Marsh, Openshaw, and Wymer, 1990; Cox and Kim, 1991). Most include calculations of:

- 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.

More research into defining a computable measure of risk is necessary (see Section VII.A.1).

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 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 who have the same characteristics as a person represented by a sample unique. 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 remains true for each individual record on a sample microdata file. However, since then, researchers have developed and tested several methods of estimating the percent of population uniques on a sample microdata file (Skinner and Holmes, 1992). 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; Greenberg and Zayatz, 1991).

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, 1991). 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.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. 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 month and year of birth or only year of birth. Rounding values, such as rounding income to the nearest one thousand dollars, is a form of recoding. Recoding variables can also reduce the high visibility 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.

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 visibility 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 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 (McGuckin and Nguyen, 1988;

Sullivan and Fuller, 1989; Kim, 1990a; Skinner, 1992). 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 (Kim, 1986). 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. See (Kim, 1990a) for a description of one process that involved the addition of random noise to a microdata file.

**Swapping** (or **switching**) and **rank swapping** are two proposed methods of disturbing microdata. 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 (Dalenius and Reiss, 1982; Dalenius, 1988; Griffin, Navarro, and Flores-Baez, 1989). The values of all other variables on 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.

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 (Griffin, Navarro, and Flores-Baez, 1989).

**Blurring** involves aggregating values across small sets of respondents for selected variables and replacing a reported value (or values) by the aggregate (Spruill, 1983). 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). 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). See (Strudler, Oh, and Scheuren, 1986) for an application of blurring. In **microaggregation**, 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 (Govoni and Waite, 1985; Wolf, 1988). Blurring and microaggregation may be done in a way to preserve variable means.

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 nonselection, 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.

Two other procedures have been suggested that have similar objectives, but differ from the disturbance procedures described above in that they are not applied to a set of true data records before their release. Randomized response is a technique used to collect sensitive information from individuals in such a way that survey interviewers and those who process the data do not know which of two alternative questions the respondent has answered (Kim, 1986; Dalenius, 1988). Rubin has proposed the use of multiple imputation techniques to produce a set of pseudo-data with the same specified statistical properties as the true microdata (Rubin, 1993).

### B.4. 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 (Wolf, 1988).

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 (Voshell, 1990). 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, almost no public-use microdata files are disturbed, and agencies, instead, 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 procedures 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.

## C.2. Limiting Geographic Detail

Geographic location is a characteristic that appears on all 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.

In Working Paper 2, it was stated that the Census Bureau will not identify on a microdata file any geographic region with less than 250,000 persons in the sampling frame. After Working Paper 2 was published, however, the Census Bureau determined that this geographic cut-off size was excessive for most surveys. Currently, the Census Bureau will 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 (Greenberg and Voshell, 1990). This is easier said than done. Decisions of this kind are often based on precedents and judgement 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 of Continuous High Visibility Variables

The variables on microdata files that contribute to the high visibility of certain respondents are called **high visibility variables**. Examples of continuous high visibility 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.

For 10 years following the 1980 Census of Population and Housing, the Census Bureau used a top-code of $100,000 on all types of income variables. Following the 1990 Census of Population and Housing, the Bureau's Microdata Review Panel members raised the top-code for some types of income that are usually high and lowered the top-code for income variables that are usually low. The Panel often requests that a given percentage of values be top-coded for a variable.

The percentage may depend on the sensitivity of the variable. The Bureau will be providing the medians of top-coded values on microdata files from the 1990 Census of Population and Housing. Appropriate top-codes (and/or bottom-codes in some cases) should be set for all of the continuous high visibility 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.

## C.4. Precautions for Certain Types of Microdata

### C.4.a. Establishment Microdata

Almost all microdata files now 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 very skew, the size of the establishment universe is small, and there are many high visibility variables on potential establishment microdata files. Also, there are a large number of subject matter experts and many possible motives for attempting to identify respondents on 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. One needs to consider the entire data collection program 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 will reduce the possibility that an intruder can link the microdata file to the administrative file. Many feel 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's Microdata Review Panel recently began using methods that estimate the number of population uniques on a microdata file to determine if that file was matchable to an outside file on a certain set of key variables. The National Center for Education Statistics has matched microdata files under consideration for release to commercially available school files looking for unique matches.

## 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. 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 (Dalenius and Denning, 1982). 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 (Spruill, 1982). 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 (Burton, 1990).

Microaggregation 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. For a definition of the (n,k) and p-percent rules, see Chapter IV.

Microaggregation and other methods of disturbance that hinder file matching, however, result in inaccurate 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 which, 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).

# Recommended Practices

## A. Introduction

Based on its review of current agency practices and relevant research, the Subcommittee developed a set of recommendations for disclosure limitation practices. The Subcommittee believes that 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.

The first set of recommendations (Section B.1 is general and pertains to both tables and microdata. There are five general recommendations. First, federal agencies should consult with both respondents and data users, the former to obtain their perceptions of data sensitivity and the latter to gather their views on the ease of use of disclosure-limited data products. Second, each agency should centralize the review of its disclosure-limited products. Next, agencies should move further toward sharing of statistical disclosure limitation software and methods. Fourth, interagency cooperation is needed when identical or similar data are released by different agencies or groups within agencies. Finally, each agency should use consistent disclosure limitation procedures across data sets and across time.

Section B.2 describes Subcommittee recommendations for tables of frequency data. Several methods of protecting data in tables of frequency data have been developed. These include cell suppression, controlled rounding and the confidentiality edit. The Subcommittee was unable to recommend one method in preference to the others. Instead, it recommended that further research concentrate on comparisons of the protection provided and on the usefulness of the end product.

Recommendations 7 to 11 in Section B.3 pertain to tables of magnitude data. Effective procedures, grounded in theory, have been developed for tabular data. Hence, the Subcommittee is comfortable in recommending that for tables of magnitude data agencies should 1) use only subadditive rules for identifying sensitive cells, 2) move toward using the p-percent or pq-ambiguity rule rather than the (n,k) rule, 3) do not reveal parameter values used in suppression rules, 4) use cell suppression or combining rows and\or columns in tables to protect sensitive cells and 5) audit tables using suppression to assure that cells are adequately protected.

Lastly, Recommendation 12 in Section B.4 pertains to microdata. Many decisions concerning the disclosure limitation methods used for microdata are based solely on precedents and judgment calls. The only recommendation presented here is to remove all directly identifying information, such as name, and limit the amount of information that is reported from other identifying variables. There is clearly a need for further research. Hence, Chapter VII "Research Agenda," focuses on 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 how respondents feel about disclosure and related issues. For example, whether they consider some data items more sensitive than others. The research done by Eleanor Singer for the Census Bureau provides one model (see Singer and Miller, 1993).

Similarly, agencies should consult data users on issues relating disclosure limitation methods to the utility and ease of use of disclosure-limited data products. For instance, whether rounded frequency counts are preferable to suppressed counts, or whether categorized or collapsed microdata are preferable to microdata that have been altered by techniques such as blurring or swapping.

**Recommendation 2: Centralize Agency Review of Disclosure-Limited Data Products.** The Subcommittee believes that it is most important that disclosure limitation policies and procedures of individual agencies be internally consistent. Results of disclosure limitation procedures should be reviewed. Agencies should centralize responsibility for this review.

Because microdata represent a relatively greater disclosure risk, the Subcommittee recommends that agencies that release microdata and tables first focus their attention on the review of microdata releases. 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 from a statistical disclosure limitation point of view. Review panels should be as broadly representative of agency programs as is practicable, should be knowledgeable about disclosure limitation methods for microdata, should be prepared to recommend and facilitate the use of disclosure limitation strategies by program managers, and should be empowered to verify that disclosure limitation techniques have been properly applied.

The Subcommittee believes that tabular data products of agencies should also be reviewed. Disclosure limitation should be an auditable, replicable process. As discussed below, for tabular data this can be achieved through the use of software based on self-auditing mathematical methods. (Disclosure limitation for microdata is not currently at the stage where a similar approach is feasible.) Depending upon institutional size, programs and culture, an agency may be able to combine the review of microdata and tables in a single review panel or office.

As needed, the Statistical Policy Office, Office of Management and Budget should draw upon agencies experienced with microdata release and review panels to provide advice and assistance to agencies getting started in microdata release and review.

**Recommendation 3: Share Software and Methodology Across the Government.** Federal agencies should share software products for disclosure limitation, as well as methodological and technical advances. Based on its long standing commitment to research in disclosure limitation, the Census Bureau is in a unique position to provide software implementations of many disclosure limitation methods and documentation of that software to other federal agencies. This software would not be the large-scale data processing systems used for specific Census Bureau data products, but the basic, simple prototype programs that have been used for testing purposes. In this way, other agencies could evaluate both the software and the methodology in light of their own needs. Specifically, software for the implementation of the following methods should be documented and made available:

- network-based cell suppression: The inputs are table dimensions, table values, identification of disclosure cells and minimal upper- and lower-protection bounds for suppressed cells. The outputs are complementary suppressions and achieved protection ranges for all suppressed cells.

- linear programming: The inputs are table dimensions, table values and identification of suppressed cells. The output is the achieved protection ranges for all suppressed cells.

Once this software is modified and documented so that it can be made available, individual agencies or programs may be interested in more extensive capabilities of the Census Bureau, such as disclosure processing of sets of two-dimensional tables related hierarchically along one dimension (such as SIC).

In addition, the National Center for Education Statistics should share information about the new system of releasing data that it introduced in the 1990's. This system, containing compressed data on a diskette or CD-ROM with access controls and software that allow users to create special tabulations without being able to examine individual data records, might prove to be very useful to other agencies.

Lastly, as advances are made in software for statistical disclosure limitation they should be made broadly available to members of the federal statistical community. Software for other methods should also be made available. For example, Statistics Canada has developed a disclosure limitation software package, CONFID (Robertson, 1993). This should be evaluated and the evaluation shared within the Federal statistical community. An interagency subcommittee of the Federal Committee on Statistical Methodology should coordinate the evaluation of CONFID and other software, such as the relatively new system of the National Center for Education Statistics.

**Recommendation 4: Interagency Cooperation is Needed for Overlapping Data Sets.** An emerging problem is the publication or release of identical or similar data by different agencies or groups within agencies (either from identical or similar data sets). There is a potential for similar problems with linked data sets. In such cases, disclosure may be possible if agencies do not use the same disclosure limitation rules and procedures. Interagency cooperation on

overlapping data sets and the use of identical disclosure limitation procedures seems to be the best approach.

**Recommendation 5: Use Consistent Practices.** Agencies should strive to limit the number of disclosure limitation practices they use, and to employ disclosure limitation methods in standard ways. Insofar as possible, agencies should be consistent in defining categories in different data products and over time. Such practices will make disclosure-limited tables and microdata more user-friendly. Examples include using consistent schemes for combining categories, establishing standardized practices for similar data such as categorizing or top-coding items like age or income, and moving towards standardized application of geographic size limitations. 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 Methods.** There has been considerable research into disclosure limitation methods for tables of frequency data. The most commonly used method at present is suppression. Besides suppression, other well-developed methods include controlled rounding and the confidentiality edit. The Subcommittee was unable to recommend one preferred method. Instead, we recommend that a research project be undertaken to compare these three methods in terms of data protection and usefulness of the data product. (Further discussion of this recommendation can be found in Chapter VII.)

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.** Disclosure occurs in statistical tables when published cell values divulge or permit narrow estimation of confidential data. For example, a count of 1 or 2 in a frequency count table of race by income may divulge the income category of one respondent to another respondent or data user. Or, a cell representing total sales within an industry for a particular county may allow narrow estimation of the sales of a single company. Such cells are called **primary disclosure cells** and must be subjected to disclosure limitation.

Agencies develop operational rules to identify primary disclosure cells. Research has shown that sensible and operationally tractable disclosure rules enjoy the mathematical property of **subadditivity** which assures that a cell formed by the combination of two disjoint nondisclosure cells remains a nondisclosure cell. Agencies should employ only subadditive primary disclosure rules. The p-percent, pq and (n,k) rules are all subadditive.

**Recommendation 8: The p-Percent or pq-Ambiguity Rules are Preferred.** The p-percent and pq-ambiguity rule 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 feels 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, respondents may be more comfortable with a combination of (n,k) rules with different values of n. An example of a combination rule is (1,75) and (2,85). With a combination rule 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 Section IV.B.3 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: Use Cell Suppression or Combine Rows and/or Columns.** There are two methods of limiting disclosure in tables of magnitude data. For single tables or sets of tables that are not related hierarchically, agencies may limit disclosure by combining rows and/or columns. Agencies should verify that the cells in the resulting table do not fail the primary suppression rule. For more complicated tables, cell suppression should be used to limit disclosure. 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. Suppression methods should provide protection with minimum data loss as measured by an appropriate criterion, for example minimum number of suppressed cells or minimum total value suppressed. These recommended practices also apply if suppression is used for tables of frequency count data.

**Recommendation 11: Auditing of Tabular Data is a Necessity.** Tables for which suppression is used to protect sensitive cells should be audited to assure that the values in suppressed cells cannot be derived by manipulating row and column equations. If the complementary suppressions were derived via network methods there is no need for a separate audit, because network methods are self-auditing. Self-auditing means that the protection provided is measured and compared to prespecified levels, thereby ensuring automatically that sufficient protection is achieved. Where self-auditing methods are not used to select cells for complementary suppression, linear programming methods should be used to audit the table with its proposed pattern of suppressions. This recommendation applies to both tables of frequency data and tables of magnitude data.

## B.4. Microdata Files

**Recommendation 12: Remove Direct Identifiers and Limit Other Identifying Information.** The challenge of applying statistical disclosure methods to microdata is to thwart identification of a respondent from data appearing on a record while allowing release of the maximum amount of data. The first step to protect the respondent's confidentiality is to remove from the microdata all **directly 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. Circumstances can vary widely between agencies or between microdata files.

Agencies should 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.

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 for data users.

There is clearly a need for further research. Hence, the next chapter, entitled "Research Agenda," focuses on microdata.

# Research Agenda

Although much research and development on disclosure limitation have been done and should be disseminated, many topics worthy of research remain. The Subcommittee has focused on fourteen topics broadly useful to Federal agencies. The Subcommittee organized these topics into three **research areas** (microdata, tabular data, and other data products) and associated with each topic a development activity designed to facilitate implementation and use.

Each Subcommittee member prioritized the fourteen research topics, and we combined these rankings to achieve the prioritization seen in Table 1 at the end of this chapter. The members' rankings varied considerably. This is not surprising. Various agencies will have different priorities because of the different types of data products they release and because of differences in the methodologies and technologies they currently use. These differences even occur within agencies causing people from different areas within an agency to have different priorities. Also, users of the data may rank these research topics differently than would suppliers of the data.

Table 1 lists only the three topics with the highest priority. In combining our rankings, we found that all Subcommittee members feel that these three topics are of great importance. The rankings of the other eleven topics listed in Table 2 were consistently lower. In general, most Subcommittee members feel that good methodology already exists for tabular data while many decisions concerning the disclosure limitation of microdata are based solely on precedents and judgement calls. Thus the Subcommittee feels that research concerning the disclosure limitation of microdata takes priority over research concerning tabular data. Also, Subcommittee members feel that research focusing on the effects of disclosure limitation on data quality and usefulness for both microdata and tabular data is of great importance.

## A. Microdata

While the disclosure risk of microdata is higher than that of tabular data, the usefulness of microdata is also correspondingly higher (Cox and Zayatz, 1993). The following research topics are aimed at increasing the ability of statistical agencies to release microdata subject to confidentiality constraints.

### A.1. Defining Disclosure

Primary disclosure rules for statistical tabulations are fairly well established. New research is not a high priority, particularly if agencies follow the Recommended Practices in Chapter VI. However, the problem of defining disclosure in microdata is far from solved.

The definition and the assessment of disclosure risk in microdata need to be put on a sound statistical footing. Probability theory provides an intuitively appealing framework for defining

disclosure in microdata in which we relate disclosure to the probability of reidentification. On this basis, we could measure disclosure limitation quantitatively and we could easily incorporate extensions, such as analysis based on prior knowledge. Without a measure of disclosure risk, decisions concerning the disclosure limitation of microdata files must be based on precedents and judgement calls. Research into probability-based definitions of disclosure in microdata should have high priority.

One part of this research involves developing a method of estimating the percent of records on a sample microdata file that represent unique persons or establishments in the population (Zayatz, 1991b). Another part involves developing a measure of marginal disclosure risk for each variable on a microdata file and analyzing changes in overall risk that result from changes in detail of each variable. After a measure of risk is developed, agencies may choose different maximum allowable levels of disclosure risk for different public-use microdata files.

### A.2. Effects of Disclosure Limitation on Data Quality and Usefulness

#### A.2.a. Disturbing Data

Due to advances in computer technology and an increase in the number of available, potentially linkable data files, it may become necessary to disturb microdata prior to release in order to make matching more difficult. Some agencies have used disturbance techniques (see Section V.B.3) such as addition of random noise or data swapping, and more research is needed to investigate the protection provided by the various disturbance techniques and the usefulness of the resulting microdata (Spruill, 1983).

#### A.2.b. More Information about Recoded Values

Users should be consulted as to the benefit of releasing means, medians, and/or variances of all values that have been top-coded or bottom-coded and of all values in each category of a recoded variable. A minimum size requirement for each category would be necessary.

### A.3. Reidentification Issues

The principal risk in releasing microdata is that a third party could match microrecords to another file containing identifiable information with reasonable accuracy. However, due to differences in sample, in responses (reporting variability), in age of data, in edit and imputation techniques, in nonresponse, and in definition and recoding of variables, linking records from two different files may be difficult. Agencies should conduct realistic attempts to match files with overlapping information, **reidentification experiments**, in order to better understand and reduce disclosure risk.

A controversial research proposal involves hiring an "intruder" to attempt to link respondents to their corresponding records on a microdata file. It would be useful to see how an intruder might approach the problem, whether or not any correct matches were made, and if correct matches were made, the amount of time and work that were required. A potential problem with

this research activity could arise if the intruder was successful in making one or more correct matches. Even if the hired intruder was an agency employee, people outside the agency could find out the results of this research under the Freedom of Information Act. This could harm the agency's reputation for maintaining confidentiality.

### A.4. Economic Microdata

The feasibility of releasing microdata from economic censuses and surveys should be further investigated. Models for release or administrative alternatives need to be proposed (McGuckin and Nguyen, 1990; McGuckin, 1992). Some agencies have released very limited establishment-based microdata files, for example the 1987 Census of Agriculture files released by the Bureau of the Census. However, as stated in Chapter V, the disclosure risk for establishment microdata files is much greater than for demographic files, and data content of currently released establishment-based files is so limited that many users consider them useless.

### A.5. Longitudinal Microdata

Longitudinal information increases both the identifiability of individual respondents and the amount of confidential information at risk of disclosure (see Section V.C.4.b). When linkable files are released on a flow basis, the disclosure risk of a given file depends on what information was released in previous files. The advice given in Chapter V was to plan ahead, considering disclosure limitation techniques that will be used for future files, but this is difficult because the choice of variables and their detail and variable sensitivity may change over time. Research is needed to assess the level of disclosure risk in longitudinal microdata files and to develop appropriate disclosure limitation policies and procedures.

### A.6. Contextual Variable Data

Social scientists are interested in obtaining microdata files with contextual variables. **Contextual variables** are variables that describe the area in which a respondent resides (such as average income of all residents in the county) but do not identify that area. In general, the areas described are much smaller than areas explicitly identified on microdata files. Contextual variables are often costly to compute, and they can increase the disclosure risk of a microdata file because a detailed description of an area can lead to the identification of that area. Identification of such small areas is undesirable in terms of disclosure risk. Further study is needed which will identify an affordable method of generating contextual variables that will not lead to the identification of small areas (Saalfeld, Zayatz, and Hoel, 1992).

### A.7. Implementation Issues for Microdata

Standard software for disclosure limitation in microdata would be of considerable benefit within and across agencies, and would serve as a quality assurance tool for Review Panels (Cox and Zayatz, 1993). This software could perform disclosure limitation techniques such as top-coding, recoding, and adding noise and could provide Panels with distributions and cross tabulations for review. Systematic development of this software should facilitate further research on improved

disclosure limitation methods for microdata and analysis of their effects on data quality and utility.

A potentially useful tool for this research and development is **matrix masking**--the representation of disclosure limitation methods in terms of matrix algebra to facilitate their implementation and analysis (Cox, 1991; Cox, 1993b). Matrix masking should be explored as a format for representing, implementing and comparing microdata disclosure limitation methods.

## B. Tabular Data

The research topics below are designed to improve the efficiency of disclosure limitation for tabular data in terms of the amount of work required and the amount of information sacrificed to achieve protection.

### B.1. Effects of Disclosure Limitation on Data Quality and Usefulness

#### B.1.a. Frequency Count Data

Research should be conducted to find out which protection method data users prefer for frequency count data (see Section IV.A). The options include controlled rounding, controlled perturbation, cell suppression, and perhaps the confidentiality edit used by the Census Bureau for the 1990 Census of Population and Housing publications.

#### B.1.b. Magnitude Data

Agencies normally use cell suppression for protection of confidential tabular data. Agencies should find out if data users prefer the collapsing (or rolling up) of categories to cell suppression.

Users should be consulted as to the benefit of publishing ranges for suppressed cells. Some publishing of ranges is already being done. For example, the Census Bureau publishes feasibility ranges for suppressed cells containing the number of employees in its County Business Patterns publications. These ranges, however, are larger than the feasibility ranges of those same cells that a data user could determine with a linear programming package. If users would like agencies to publish the smaller ranges, a feasibility study should be done to see if users could assimilate and manipulate information provided in this form and if a large volume of data would preclude this action due to the time needed for determining the ranges.

Users should also be consulted as to the benefit of releasing means, medians, and/or variances of all values in each suppressed cell in a table. A minimum number of respondents in each of these cells would be necessary.

### B.2. Near-Optimal Cell Suppression in Two-Dimensional Tables

Network flow methods work well in two-dimensional tables. However, other methods also offer desirable features. It would be useful to specify the characteristics of and develop an optimal method for disclosure limitation in two-dimensional tables and, by combining desirable features of existing methods, to create a new method that improves upon each of the original methods. Such a method would be advantageous, as Federal statistical agencies analyze many thousands of tables each year. It would also improve the handling of three- and higher dimensional tables and interrelated sets of tables using methods based on combining two-dimensional procedures. The list of original methods includes network flow, general linear programming, integer programming, and neural networks. These methods are discussed in Chapter IV. The question is which method (or combination) is best with respect to well-defined criteria such as cost, computational efficiency, transportability, extension to higher dimensions, ease of implementation and maintenance, and data use.

### B.3. Evaluating CONFID

Statistics Canada has a set of programs called CONFID which performs cell suppression on tabular data. CONFID has been made available to U.S. Federal statistical agencies. It would be worthwhile to extensively test and evaluate this system of programs and to compare it to the current cell suppression systems used at the Census Bureau and elsewhere.

### B.4. Faster Software

Federal agencies would benefit from locating and purchasing the fastest network flow package available. Current cell suppression methodology uses network flow methodology when applying complementary suppressions (see Section IV.2.b). In addition, one proposed **filter technique** also uses network flow methodology to locate (and remove) superfluous complementary suppressions.

Agencies would also benefit from locating and purchasing the fastest linear programming package available. The network flow-plus-heuristic technique currently used at the Census Bureau to find complementary suppression patterns in three-dimensional tables yields non-optimal solutions. Any technique for finding optimal solutions to the cell suppression problem is currently impractical due to computer time constraints. However, there exists a linear programming technique which yields better solutions for three-dimensional tables than the currently used network-based procedure and which agencies could use if a faster linear programming package was available.

Auditing programs (see Section IV.B.2.a) also use linear programming packages. Auditing programs are programs that check to see if all primary suppressions in a table are indeed sufficiently protected after complementary suppressions have been applied. So, a faster linear programming package would lead to faster auditing programs.

### B.5. Reducing Over-suppression

One problem with the currently used cell suppression methodology is that it applies complementary suppressions to only one primary suppression at a time. The system has no way of considering all of the primary suppressions at once, and this leads to the application of too many suppressions (**over-suppression**). There may be one or more procedures which could be applied prior to using current techniques and which would reduce the amount of over-suppression caused by this one-primary-at-a-time approach. These procedures should be developed and tested.

To reduce over-suppression, the Census Bureau is currently investigating the possibility of incorporating one small integer programming tool into the current cell suppression procedures to obtain better results. The purpose of this particular integer program is to eliminate superfluous complementary suppressions (Sullivan and Rowe, 1992). Other ways of using integer programming procedures to reduce over-suppression should be examined.

There may be one or more procedures that could be applied after the current cell suppression techniques to reduce the amount of over-suppression caused by the one-primary-at-a-time approach. These techniques, often called **filter** or **clean-up techniques**, examine the table including the complementary suppressions and attempt to locate cells which were chosen as complementary suppressions but which could be published without a loss of protection of the primary suppressions. One procedure involves using network flow methodology repetitively to locate the superfluous complementary suppressions. The practicality of using this filter technique, which increases the computer time needed to perform disclosure analysis considerably, should be investigated. Other filter techniques should be developed and tested.

Another major drawback of the currently used cell suppression methodology is that, often, due to computer storage and methodology constraints, not all data values in all additive relationships can be considered simultaneously. This is particularly true for the large amount of tabular data published by the Census Bureau. Thus the problem must be broken down into pieces (sets of data) that are processed separately. Unfortunately, some data values are necessarily in more than one piece. While trying to ensure that the various pieces, when considered as a whole, do not result in a disclosure of confidential information, it is necessary to reprocess many of the pieces several times. This reprocessing of sets of data due to the inability to process all data simultaneously is called **backtracking**. Backtracking is time consuming and results in over-suppression. Research that could reduce the amount of backtracking needed and the over-suppression caused by backtracking would be beneficial.

## C. Data Products Other Than Microdata and Tabular Data

Previously, disclosure limitation research focused on either microdata or tabular data. However, agencies are also releasing information in the form of database systems and analytical reports. Research is needed to analyze disclosure risk and develop and evaluate disclosure limitation techniques for these types of data products. The next section describing database systems is

lengthy because we provided no background information on this subject in previous chapters. The length of the section does not reflect the importance of the research topic.

### C.1. Database Systems

More research is needed to analyze the feasibility of storing and allowing access to microdata in a database system with security controls and inferential disclosure limiting techniques (Keller-McNulty and Unger, 1993). Research in database systems assumes quite a different approach than we have discussed so far. Rather than releasing known subsets of data, it is possible to keep all data on-line in a **database management system** (DBMS) that dynamically enforces the controls. Normally this is a relational database that organizes the data in a series of tables. These data tables are similar to microdata files, although they might also contain sensitive data such as salaries, which are releasable only as an aggregated value such as average salary. Users may request whatever subset of information they need, and the database may return either an exact or approximate answer, or even refuse to answer if the data would disclose individual identities. The relational database management system, therefore, simplifies retrieving data, but does not actually improve availability, i.e., exactly the same data are still available. References considering such an approach to statistical databases are (Adam and Wortman, 1989) and (Michalewicz, 1991).

If such a system were successful, it would allow the greatest possible use of the data, while still maintaining the confidentiality of the individual. Such a system could allow for use of microdata containing administrative data. The system would handle special requests without any manual intervention. Reports, special files, tapes, etc. would not be generated unless actually requested. Present research in this type of system has discovered many problems, and proposed a few solutions, but no one technique is presently accepted as a standard. Implicit in such proposals is the requirement that the data can only be accessed via the database, not directly through the file system or by transmission protocols. Encryption of the data by the database is one method of protecting the data during either storage or transmission.

Implementation of dynamic controls on data access could possibly be accomplished using some developing concepts from the area of database security. For this purpose, we can assume that some information is more sensitive than other information. The sensitive information is classified "High", while the less sensitive information is classified "Low". For example, the entire survey could be available on-line, but the individual's name (and other identifying information) could be classified as "High". Names would be available only to survey personnel. Such a database would have to incorporate various protection mechanisms common to secure database systems. Of most significance to statistical database management systems would be the assurance that no process or subroutine has been surreptitiously inserted into the system to simply write protected information into a "Low" area (such a process is called a "Trojan Horse"). If users are limited to reading data then the disclosure problem becomes identical to the **inference problem** of interest to database security researchers. Can it be guaranteed that "Low" users cannot infer the information from the other, authorized, Low data that they retrieve? This is an open problem that has been solved only for a few specialized types of inferences. Perhaps the most promising approach is to define a set of constraints, or rules, to be checked for each query.

Such constraints could enforce query set size, overlap, or complexity restrictions. Constraints could dynamically simulate cell suppression techniques. They could also correlate recent queries to insure that the particular user cannot combine these answers to infer unauthorized information. Research into such inference issues would be especially valuable since it addresses the inference links that exist between separate tables (or relations), or even users. It, therefore, addresses the question of whether the data in one table can be used to compromise the data in another table. This is an increasingly common issue due to the growing popularity of relational databases. A good reference for inference control is (Qian, Stickel, Karp, Lunt, and Garvey, 1993).

The ability to calculate aggregates (such as averages, sums, counts, etc.) over sensitive data leads to many problems in maintaining secrecy of the data. Successive queries over varying subsets of data may be sufficient to determine specific values of the sensitive data. This is known as a **tracker attack** (Denning, 1982). Prevention of tracker attacks using query histories requires additional computation steps (Michalewicz, 1991), but using an algorithm that does not keep a history seriously restricts the amount of data available to the user.

These problems have taken on increased importance in recent years because of research in the field of **knowledge discovery** which concerns artificial intelligence based programs whose purpose is to discover relations between data items. Unauthorized disclosures may therefore be automated, and the probability of such disclosure is dramatically increased. These methods will change the way that disclosure risk must be calculated. Rather than selecting a target individual, the program will search the statistical database for *any* specific individual, and then match the one found to the outside database. For example, if average income is released by county, and some county has only one doctor, then that doctor's income may be determined. Although the probability of being able to determine the income for any arbitrary doctor is quite small, a knowledge discovery program will check the entire database to find the one doctor whose income can be determined. Such techniques have proven quite useful to direct marketing companies.

The database techniques currently used tend to emulate the paper world, and are adequate for a certain balance of protection and accessibility. If we wish to provide more accessibility, while maintaining the same level of protection, we will require techniques that are only available through automated computer programs. These will tend to require extensive computations. In fact, some protection proposals suggest that all possible combinations of characteristics be checked to insure that individual names cannot be determined from the data. Such computations may currently be suitable for small or simple datasets that can be verified once and then widely distributed. The cost of computing power must decline further before such methods are feasible for complicated databases, however.

The National Center for Education Statistics has released compressed data with access controls along with software that allows users to obtain weighted, minimum cell count cross tabulations of the data without being able to examine individual data records. Research should be done which investigates the disclosure risk in releasing results of other types of statistical analysis of the data.

## C.2. Disclosure Risk in Analytic Reports

The Center for Economic Studies (CES) at the Census Bureau releases output from statistical models, such as econometric equations, estimated using confidential data. Some agencies, such as the Bureau of Labor Statistics, have fellowship programs and some, such as the Census Bureau, can "swear in" researchers as special employees to allow use of confidential data for analytical purposes. Fellows and other researchers would like to publicly release the results of their statistical analyses.

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 recover exact input data from a regression equation if the number of coefficients is greater than or equal to the number of observations, regression output generally poses little disclosure risk because normally the number of observations is much larger than the number of coefficients. One question to be addressed, however, is if the number of observations is only slightly larger than the number of coefficients, how closely can a user estimate the input data?

Also, sometimes researchers use dummy (0,1) variables in statistical models to capture certain effects, and these dummy variables may take on values for only a small number of observations. Currently, CES treats these dummy variables as though they were cells in a table and performs disclosure analysis on the observations for which the dummy variables take on the value of 1. CES applies the n,k rule to these "cells" based on total value of shipments in deciding whether or not to release their corresponding regression coefficients. Research is needed to determine if this technique leads to withholding too much information.

**Table 1**
**Prioritization of the Three**
**Most Important Research Topics**

| Priority | Research Topic | Data Type |
|---|---|---|
| 1 | Defining Disclosure | Microdata |
| 2 | Data Quality and Usefulness | Microdata |
| 3 | Data Quality and Usefulness | Tabular Data |

**Table 2**
**Other Research Topics**

| Research Topic | Data Type |
|---|---|
| Reidentification Issues | Microdata |
| Economic Microdata | Microdata |
| Longitudinal Microdata | Microdata |
| Contextual Variable Microdata | Microdata |
| Implementation Issues | Microdata |
| Near-Optimal Cell Suppression in Two-Dimensional Tables | Tabular Data |
| Evaluating CONFID | Tabular Data |
| Faster Software | Tabular Data |
| Reducing Over-Suppression | Tabular Data |
| Database Systems | Other |
| Analytic Reports | Other |

# 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 which have been used where there may be some question as to how to apply the standard procedures. Section 3 presents the simplifying assumptions which 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:

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

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 $x_i$, 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.

# Government References

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. (April 1989). Energy Information Administration, U.S. Department of Energy. Washington, DC.

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 1992). National Center for Education Statistics, U.S. Department of Education. Washington, DC.

6. NCES Policies and Procedures for Public Release Data. National Center for Education Statistics, U.S. Department of Education. Washington, DC.

7. NCHS Staff Manual on Confidentiality. (September 1984). National Center for Health Statistics, U.S. Department of Health and Human Services. Washington, DC.

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.

# Bibliography

The purpose of this bibliography is to update the references on disclosure limitation methodology that were cited in **Working Paper 2**. Much has been written since **Working Paper 2** was published in 1978. Subcommittee members reviewed papers dealing with methodological issues which appeared after 1980 and prepared the abstracts in this bibliography. **An asterisk (*) indicates that the document has been specifically referenced in this report.**

In the Federal statistical system the Bureau of Census 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 and has been well represented on the Subcommittee. For these reasons the statistical disclosure limitation research that has been sponsored by the Bureau of the Census is thoroughly and adequately covered in this bibliography. In addition the Subcommittee tried to include important papers which 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.

Within the past two years statistical disclosure limitation research has been highlighted in publications from Western Europe. An international Seminar on Statistical Confidentiality was held in September, 1992, in Dublin, Ireland. The seminar was organized by Eurostat (Statistical Office of the European Community) and ISI (International Statistical Institute). The papers were published in Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute. In addition, a special issue of Statistica Neerlandica, Vol. 46, No. 1, 1992 was dedicated to disclosure limitation. Selected papers from these sources are included in the annotated bibliography.

In 1993, a special issue of the Journal of Official Statistics, Vol. 9, No. 2 was dedicated to disclosure limitation. That issue contains the papers which were presented at a workshop sponsored by the Panel on Confidentiality and Data Access, of the Committee on National Statistics. The panel report was published later in 1993. It is entitled Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, by Duncan et. al., and was published by the Committee on National Statistics and the Social Science Research Council, National Academy Press, Washington, DC. The panel report and selected papers from the special issue of the Journal of Official Statistics are included in the bibliography.

Areas of potential applicability which are not covered in this bibliography include mathematical methods of matching a microdata file to an outside file. A good summary of the state-of-the-art in exact matching as of 1985 can be found in "Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methodologies", Dept of Treasury, IRS, SOI, Publication

1299 (2-86). A more recent reference is a special issue of Survey Methodology, Volume 19, Number 1, 1993.

The Subcommittee on Disclosure Limitation Methodology would like to thank the following people who contributed to the annotated bibliography: Robert Burton, National Center for Education Statistics; Russell Hudson, Social Security Administration; Dorothy Wellington, Environmental Protection Agency.

## Bibliography on Methodology for Disclosure Limitation

*Adam, N. R. and Wortmann, J. C., "Security-control Methods for Statistical Databases: A Comparative Study," ACM Computing Surveys, Vol 21, No. 4, pp. 515-556 (Dec. 1989).

The authors carefully define and explain the various problems inherent in disclosure control of on-line systems. Proposed solutions, along with their strengths and weaknesses are discussed. This paper is written on a tutorial level, the purpose being to educate the reader in the current methods. Security control methods are classified into four approaches: conceptual, query restriction, data perturbation and output perturbation. Methods based on these approaches are compared. Promising methods for protecting dynamic-online statistical databases are presented.

*Alexander, L. B., and Jabine, T. B. (1978), "Access to Social Security Microdata Files for Research and Statistical Purposes," Social Security Bulletin, Vol. 41, No. 8, pp. 3-17.

This article focuses on the characteristics of SSA microdata files and on the development of a disclosure policy aimed at serving the public interest while protecting the privacy of individuals and the confidentiality of research and statistical information. Several dimensions of the disclosure question are explored. The factors controlling the decision whether or not to release microdata are also discussed. Some particular practices are described to illustrate application of present policy principles.

*Barabba, V. P. and Kaplan, D. L. (1975), "U. S. Census Bureau Statistical Techniques to Prevent Disclosure -- The Right to Privacy vs. the Need to Know," paper read at the 40th session of the International Statistical Institute, Warsaw.

*Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), "Disclosure Control of Microdata," Journal of the American Statistical Association, Vol. 85, pp. 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, micro aggregation, 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.

Blien, U., Wirth, H., and Muller, M. (1992), "Disclosure Risk for Microdata Stemming from Official Statistics," Statistica Neerlandica, Vol. 46, No. 1, pp. 69-82.

Empirical results from matching a portion of a microdata file on individuals against a reference file are presented. Two matching strategies are considered: a simple (or exact) procedure, and a procedure developed by Paass and based on the use of discriminant function analysis. It is concluded that exact matching is a poor strategy and would not be

used by an astute data snooper. It is further concluded that Paass' procedure is also deficient, on the grounds that, among the "identifications" that it yields, only a small proportion are true identifications, and the snooper does not know which matches are correct. One criticism of this is that the authors do not consider the possibility that a "small proportion" of correct matches may be too many. This weakens their overall conclusion that disclosure in microdata files may be a less serious problem than it is typically thought to be.

Bowden, R. J. and Sim, A. B. (1992), "The Privacy Bootstrap," Journal of Business and Economic Statistics, Vol. 10, No. 3, pp. 337-345.

The authors describe a method of masking microdata by adding noise. The noise is generated by bootstrapping from the original empirical distribution of the data. The technique is analyzed in terms of protection gained and statistical efficiency lost, and it is compared with the technique of adding random noise.

Burton, R. (1990), "Strategies to Ensure the Confidentiality of Data Collected by the National Center for Education Statistics," unpublished manuscript presented to the Washington Statistical Society.

The author discusses the confidentiality problems that arise with microdata files containing information from complex national surveys, the steps taken at NCES to minimize disclosure risk from these files, and the testing procedures that are used to assess risk. Since education data bases are typically hierarchical, disclosure risk centers on identification of schools, rather than on direct identification of individuals. It is therefore necessary to recode school data, which, at NCES, takes the form of converting continuous data to categorical.

The recoded files are tested by matching against publicly available reference files, using the "nearest neighbor" concept, which is defined in terms of Euclidean distance. The author discusses the difficulty of developing a strict criterion of "sufficiently safe" in this context, and presents the criteria that NCES has used.

Caudill, C. (1990), "A Federal Agency Looks for Answers to Data Sharing/Confidentiality Issues," presented at the annual meeting of the American Agricultural Economics Association, Vancouver, British Columbia.

NASS has a clear set of publication rules. No reports can be released which may reveal information about individual operations. NASS will only issue information which is based on reports of three or more operations. Also, data will not be released if one operation accounts for 60 percent or more of a total. These publication rules often mean that geographic subdivisions must be combined to avoid revealing information about individual operations. Data for many counties cannot be published for some crop and livestock items and State level data must be suppressed in other situations.

If only a few operations are present in a particular universe, such as hatcheries in a State or firms holding cold storage supplies of specific commodities, it may not be possible to publish totals at all. NASS can publish data in such cases only if a signed waiver is received under which an operation accounting for more than 60 percent of a total agrees to allow data to be published. If waivers cannot be obtained, data are not published. If waivers are obtained, the waivers are reviewed and signed periodically to be sure that cooperation has not changed.

Causey, B., Cox, L. H., and Ernst, L. R. (1985), "Application of Transportation Theory to Statistical Problems," Journal of the American Statistical Association, 80, 392, pp. 903-909.

This paper demonstrates that the transportation theory that solves the two-dimensional (zero-restricted) controlled rounding problem, Cox and Ernst (1982), can be used for other statistical problems: 1) general statistical problems which involve replacing nonintegers by integers in tabular arrays (eg. iterative proportional fitting or raking); 2) controlled selection for a sample; and 3) sample selection to maximize the overlap between old and new primary sampling units after a sample redesign. The paper mentions that the controlled rounding of a two-way table can be used to prevent statistical disclosure in a microdata release (by replacing the true value by an appropriately rounded value). The paper also provides a simple example that shows that the three-way controlled rounding problem does not always have a solution.

*Cecil, J. S. (1993), "Confidentiality Legislation and the United States Federal Statistical System," Journal of Official Statistics, Vol. 9, No. 2, pp. 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.

Cigrang, M. and Rainwater, L. (1990), "Balancing Data Access and Data Protection: the Luxembourg Income Study Experience" Proceedings of the Statistical Computing Section, American Statistical Association, Alexandria, VA, pp. 24-26.

Details of a computer system allowing access to multi-national microdata files are presented. Data protection is achieved through use of a security system based on user identification, passwords, output control, operating system safeguards, and review of both job requests and output. The authors do not address analytical issues.

Cox, L. H. (1979), "Confidentiality Problems in Microdata Release," Proceedings of the Third Annual Symposium on Computer Applications in Medical Care, IEEE Computer Society, pp. 397-402.

The examples of disclosure avoidance techniques given and views expressed were drawn from Statistical Policy Working Paper 2: Report on Disclosure-Avoidance and Disclosure Avoidance Techniques. This was followed by Cox's observations. He points out that there are no generally accepted and quantifiable notions of the degree of disclosure or the degree of protection. Thus, there is no concept of sensitivity of microdata upon which the necessary protective techniques must be defined. It is difficult to accurately measure the degree to which a technique reduces the sensitivity of a microdata set without first quantifying the notion of sensitive data. He suggested empirical research into quantifying the concept of sensitivity, simulating likely privacy invasion tactics and engaging in related cost-benefit analyses. For theoretical research he suggested casting the problem in terms of data base theory, which he claims includes data base security, multidimensional transformation and data swapping.

One interesting thing about this paper is that although some of the research projects have been tried, the same questions remain and many of the same protection techniques are still used.

Cox, L. H. (1980), "Suppression Methodology and Statistical Disclosure Control," Journal of the American Statistical Association, Vol. 75, pp. 377-385.

This article highlights the interrelationships between the processes of disclosure definitions, subproblem 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. This method formed the basis of an automated system for disclosure control used by the Census Bureau in the 1977 and 1982 Economic Censuses.

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 subproblems. 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. This approach was implemented in a data base environment in an automated disclosure control system for the 1977 U.S. Economic Censuses.

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, pp. 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. (1983), "Some Mathematical Problems Arising from Confidentiality Concerns," Essays in Honour of Tore E. Dalenius Statistical Review, Vol. 21, Number 5, Statistics Sweden, pp. 179-189.

This is a nice summary of disclosure problems in tables, both of frequency data and magnitude data. Included are discussions of the definition of a sensitive cell and the mathematics involved in selection of cells for complementary suppression and controlled rounding.

Cox, L. H. (1984), "Disclosure Control Methods for Frequency Count Data," presented at the Census Advisory Committee Meeting of the American Statistical Association, Bureau of the Census.

Four methods for controlling statistical disclosure in frequency count data are discussed along with their pros and cons: cell suppression, random perturbation, random rounding and controlled rounding. Each method is viable for single 2-way tables. With some effort, cell suppression can be applied consistently between sets of tables but creates problems for data use. Because the other methods distort every value somewhat, they avoid abbreviating detail and do not produce seemingly arbitrary and possible cumbersome patterns of data suppression. Among these other methods, only controlled rounding meets all of the following objectives: additivity, unbiasedness and reducing data distortion. The paper recommends research concerning the extent to which various methods, particularly controlled rounding can be applied consistently between tables.

This paper defines disclosure in frequency count data to occur when one can infer with certainty that the number of respondents is less than a predetermined threshold. Most other references say that disclosure occurs when the number of respondents is less than the predetermined threshold.

Cox, L. H. (1987a), "New Results in Disclosure Avoidance for Tabulations," International Statistical Institute-Proceedings of the 46th Session: Contributed Papers, Tokyo, pp. 83-84.

For two-way tables, this paper considers the three standard disclosure avoidance procedures, suppression, perturbation and rounding in a single mathematical framework. The two unifying formulations mentioned are the use of alternating cycles and network optimization models. Alternating cycles are described in more detail in Cox (1987b). Network optimization models are described in more detail in Cox (1992).

Cox, L. H. (1987b), "A Constructive Procedure for Unbiased Controlled Rounding," Journal of the American Statistical Association, Vol. 82, June 1987, pp. 520-524.

A constructive algorithm for achieving zero-restricted unbiased controlled rounding, simple enough to be implemented by hand is presented. The procedure is based on adjustments in alternating cycles of cells in an array. Gives a counterexample to the existence of unbiased controlled rounding in three dimensional tables. Cox's solution also allows one to perform random data perturbation in a way that assures additivity.

Cox, L. H. (1991), "Comment," a comment on Duncan, G. and Pearson, R., "Enhancing Access to Microdata while protecting Confidentiality: Prospects for the Future," Statistical Science, No. 6, pp. 232-234.

This is an initial formulation of matrix masking, which is described more completely in Cox (1993b).

Cox, L. H. (1992), "Solving Confidentiality Protection Problems in Tabulations Using Network Optimization: A Network Model for Cell Suppression in U.S. Economic Censuses." Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, Dublin, pp. 229-245.

Mathematical issues in confidentiality protection for aggregate economic statistics are discussed. A network model of the minimum-cell cell suppression algorithm of Cox (1980) is presented and the development of network models that combine minimum-cell and minimum-value-suppressed criteria are discussed.

Cox, L. H. (1993a), "Network Models for Complementary Cell Suppression", unpublished manuscript.

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. (1993b), "Matrix Masking methods for Disclosure Limitation in Microdata," unpublished manuscript.

The statistical literature contains many methods for disclosure limitation in microdata. However, their use and understanding of their properties and effects has been limited. For purposes of furthering education, research, and use of these methods, and facilitating their evaluation, comparison, implementation and quality assurance, it would be desirable to formulate them within a single framework. A framework called "matrix masking"-- based on ordinary matrix arithmetic--is presented, and explicit matrix mask formulations are given for the principal microdata disclosure limitation methods in current use. This enables improved understanding and implementation of these methods by statistical agencies and other practitioners.

Cox, L. H. (1994), "Protecting Confidentiality in Establishment Surveys," in Survey Methods for Businesses, Farms and Institutions, Brenda Cox (ed.), John Wiley and Sons, NY.

This paper focuses on the issue of disclosure limitation in tables of establishment data, namely cell suppression and some of the mathematical issues associated with determining "optimal" patterns of complementary cell suppressions. The methods used by the U. S. Census Bureau and Statistics Canada, current research by the Census Bureau, and problems associated with microdata files for establishment data are described.

Cox, L. H. and Ernst, L. R. (1982), "Controlled Rounding," INFOR, Vol 20, No. 4, pp. 423-432. Reprinted: Some Recent Advances in the Theory, Computation and Application of Network Flow Methods, University of Toronto Press, 1983, pp. 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., Fagan, J. T., Greenberg, B., and Hemmig, R. (1986), "Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data, "Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 388-393.

Results obtained by the Census Bureau's confidentiality staff in its research in disclosure avoidance methods for publicly released tabular data are described. The paper reports new procedures (based on network theory) developed for rounding, perturbation, and cell suppression in two-dimensional tables, with a focus on their common underlying structure. The goal is to develop unbiased procedures which maintain additivity and alter marginals as infrequently as feasible.

The common underlying structure considered in this paper is using circuits in a graph, referred to as "alternating cycles" in Cox (1987b). This paper describes the approach and illustrates its use for unbiased controlled rounding, unbiased controlled perturbation, unbiased restricted controlled perturbation, auditing protection in a suppression scheme, and selecting cells for complementary suppression.

Cox, L. H. and George, J. A. (1989), "Controlled Rounding for Tables with Subtotals," Annals of Operations Research, 20 (1989) pp. 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., Johnson, B., McDonald, S., Nelson, D. and Vazquez, V. (1985), "Confidentiality Issues at the Census Bureau," Proceeding of the Bureau of the Census First Annual Research Conference, Bureau of the Census, Washington D. C., pp. 199-218. (Revised: Cox, L. H., McDonald, S. K. and Nelson, D. (1986), "Confidentiality Issues of the U.S. Bureau of the Census," Journal of Official Statistics, 2, 2, pp. 135-160.)

This paper summarizes confidentiality issues and presents a fair amount of detail in selected areas such as methods applied to tables of frequency counts. U. S. Census Bureau studies in the early 1980's pointed out the need for an integrated program of research and development work to contribute to key decisions on important policy issues. This paper presents the major ideas that were raised in these studies and merit further attention as opportunities for research, and highlights research in progress.

*Cox, L. H. and Kim, J. (1991), "Concept Paper: Thwarting unique identification in Microdata Files. A Proposal for Research," unpublished manuscript.

This paper proposes a disturbance technique involving super and subsampling. 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.

Cox, L. H. and Zayatz, L. (1993), "Setting an Agenda for Research in the Federal Statistical System: Needs for Statistical Disclosure Limitation Procedures," Proceedings of the Section on Government Statistics, American Statistical Association.

The authors describe the confidentiality protection problem for different types of data, summarize confidentiality protection techniques in current use, and present an agenda for future research in confidentiality protection.

Dalenius, T. (1981), "A Simple Procedure for Controlled Rounding," Statistisk Tidskrift, No. 3, pp. 202-208.

Disclosure control in frequency tables is discussed and available methods are summarized. Dalenius proposes a method of rounding only part of a table, which assures that the rounded table preserve the marginal totals. The cells to be rounded include the sensitive cells and selected other cells to assure that the marginal totals are unchanged. Selecting the additional cells for the "rounding" may be akin to selecting cells for complementary suppression.

Dalenius, T. (1982), "Disclosure Control of Magnitude Data," Statistisk Tidskrift, No. 3, pp. 173-175.

Discusses disclosure control in tables of magnitudes, where cells are determined to be sensitive either because there are too few respondents, or because they fail the (n,k) rule. The approach is similar to that described in Dalenius (1981) in that for one sensitive cell, three additional cells are selected to complete a rectangle. Then random rounding is applied to the counts in four cells, and the magnitude to be published is calculated based on the adjusted number of respondents and the assumption that each respondent has the average volume. The new aggregates are unbiased estimates for the originals.

Dalenius, T. (1986), "Finding a Needle in a Haystack or Identifying Anonymous Census Records," Journal of Official Statistics, Vol. 2, pp. 329-336.

The author presents two variants of a simple method for identifying the unique records in microdata. The first variant involves three different approaches to sorting the data to identify unique records. In the second variant he considers two types of transformation of the data and shows how sorting can identify the unique records under either transformation. A cost function to determine which variant to use is based on number of variables with data in the public domain and computer memory necessary to perform that variant. The identified data can be protected by destroying the unique records in the public domain or by blocking out some but not all of the data, by data perturbation (replacing original data with different values for one or more variables), or by encryption (a reversible transformation of the data.)

Dalenius, T. (1988), Controlling Invasion of Privacy in Surveys, Department of Development and Research, Statistics Sweden.

This book discusses many problems associated with protecting against invasion of privacy in surveys. It was intended as a text for a course on the subject, and includes many examples from Europe and the U.S. Included are chapters on the basic concept and legal framework, safeguards provided as a result of agency codes, professional codes and informed consent. Other chapters discuss safeguards provided by sampling, measurement methods (including randomized response), and transformations. These are followed by a discussion of safeguards taken during the data processing and safeguards for use in the release of statistics (publication of tables and microdata). These chapters on release are the most applicable for this bibliography.

Dalenius defines disclosure and provides examples. He describes disclosure control for tables of counts to involve cell suppression, changing the classification scheme, perturbation and rounding. For tables of magnitude data disclosure control may be based on cell suppression, changing the classification scheme or perturbation. He discuses release of low order moments, release through a data-base and queries of the data base. Finally he describes release of microdata. Protective measures for microdata include deidentification, sampling, placing a restriction on population size, reduction in detail, adding noise to the data, removing well known data subjects, suppression, data-swapping, and transformations.

The book concludes with a discussion of the safeguards involved in the closing operations of a survey, including deidentification of records, file-splitting, taking action on unique vectors, and encryption. The epilogue summarizes what is ahead and includes a discussion of research ideas, most of which do not deal with the methodological issues of release.

Dalenius, T. (1993), "Safeguarding Privacy in Surveys at the Turn of the Century," unpublished manuscript.

This memo discusses the change in public perception concerning invasion of privacy with surveys and considers what survey statisticians can do to counteract concerns by survey respondents. The assumption is made that the sources of public concern in the past may be present in the next few years, but with different forces. This assumption makes it necessary to identify the key sources in the past likely to have a force in the future that cannot be neglected.

Dalenius, T. (1993), "Disclosure Control of Microdata using Data Shifting," unpublished manuscript.

This memo proposes "data shifting" as a way to limit disclosure in microdata. Data shifting is related to, but not the same as, "data swapping".

Dalenius, T. and Denning, D. E. (1982), "A Hybrid Scheme for Release of Statistics," Statistisk Tidskrift, Vol 2, pp. 97-102.

For population survey data this paper proposes for the release of data a scheme that is a hybrid of microstatistics and macrostatistics (tables and summary statistics). A specified set of low-order finite moments of the variables are computed and released, allowing users to compute the low-order statistics corresponding to their needs. They consider the computational feasibility of doing this and discuss the protection implied. They also observe that users would not be able to calculate even simple moments for subgroups of the respondents (eg. all females.) The authors balance the feasibility of giving higher order moments against the increased amount of computation needed as well as the increased risk of disclosure.

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

The data-swapping technique proposed by the authors can be used on categorical data to produce microdata and to release statistical tabulations while protecting confidentiality. The raw data matrix is converted to a new matrix by rearranging entries in such a way that the marginals up to a specified order of cross-tabulation are unaffected and the desired order of statistics is preserved. The authors illustrate mathematically how the data base presented only in terms of t-order statistics is unlikely to be compromised. An appended comment points out that this approach is applicable to data from individual respondents with relatively few categorical responses for each data item. This technique can be used to both produce microdata and release statistical tabulations so that confidentiality is not violated. This is the technique which has been used by the U. S. Census Bureau as part of the Confidentiality Edit. The Confidentiality Edit was used to protect data tables from the 1990 census.

*Denning, D. E. (1982), Cryptography and Data Security, Addison-Wesley, Reading, MA.

This is **the** standard book addressing computer security issues. The book is quite rigorous and very thorough in coverage, including cryptography and transmission issues as well as data-base security. Statistical databases are covered in depth, incorporating much of the author's previous work. The topics covered still provide the basis for understanding more recent work in this area.

DeSilets, L., Golden, B., Kumar, R., and Wang, Q. (1992), "A Neural Network Model for Cell Suppression of Tabular Data," College of Business and Management, University of Maryland, College Park, MD.

For three-dimensional tables, the objective is to select cells for complementary suppression which minimize the total value suppressed but assure that the sensitive cells are protected to within pre-specified tolerance levels. A neural network is trained on the solutions from the heuristic, network based model described in Kelly et al. (1992). Thus, the neural

network can be used on a new problem to quickly identify a good starting solution for the more general optimization method. This paper provides detail on the neural network design and training. Results are promising. Run time of the network is minimal once the network is trained. The trained neural network was able to match about 80% of the cell suppression solution for a new table.

Duncan, G. T. (1990), "Inferential Disclosure-Limited Microdata Dissemination," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 440-445.

Duncan discusses the various forms of disclosure such as complete identification of a record (identity disclosure), obtaining a good approximation of confidential data (attribute disclosure) and the more general inferential disclosure in which the release of the microdata data allows a user to make more accurate estimates of confidential information. This paper also presents a method to measure the risk of disclosure in terms of unauthorized information gained when microdata are released. This method can then be used to measure the effectiveness of data masking techniques.

Duncan, G. T. and Lambert, D. (1986), "Disclosure-limited Data Dissemination" (with comment), Journal of the American Statistical Association, Vol. 81, pp. 10-28.

The authors briefly summarize the legal aspects of maintaining the confidentiality of records, in particular they site various United States laws. The most important part of this paper deals with a general disclosure limiting approach that quantifies the extent of statistical disclosure by means of an uncertainty function applied to predictive distributions.

Duncan, G. T. and Lambert, D. (1987), "The Risk of Disclosure for Microdata," Proceedings of the Bureau of the Census Third Annual Research Conference, Bureau of the Census, Washington, DC.

Various types of disclosure are discussed, including identity and attribute disclosure. The authors then present a model to estimate the risk of disclosure that can take into account the user's prior knowledge and also the type of masking technique that has been used. The model presented uses predictive distributions and loss functions. Using this model they show that sampling and the including of simulated artificial records can reduce the disclosure risk.

Duncan, G. T., Jabine, T. B. and de Wolf, V. A., eds. (1993), Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, Committee on national Statistics and the Social Science Research Council, National Academy Press, Washington, DC.

This is the final report of the Panel on Confidentiality and Data Access, which was jointly sponsored by the Committee on National Statistics of the National Research Council and the Social Science Research Council. The Panel's charge was to develop recommendations that could help federal statistical agencies to protect the confidentiality of data subjects and,

at the same time, facilitate responsible dissemination of data to users. Chapter 6, "Technical and Administrative Procedures," covers statistical disclosure limitation methodology and administrative procedures for restricting access to data. The chapter includes recommendations on both topics.

Duncan, G. T. and Pearson, R. W. (1991), "Enhancing Access to the Microdata While Protecting Confidentiality: Prospects for the Future" (with comment), Statistical Science, Vol. 6, No. 3, pp. 219-239.

Methods on increasing data access while assuring an acceptable level of protection are discussed, including statistical masking, electronic gatekeepers, licensing contracts, punitive damages for improper use of data and researchers code of ethics. The authors also suggest that respondents to data collection procedures should be informed that there is a remote risk of re-identification of their responses.

Ernst, L., (1989), "Further Applications of Linear Programming to Sampling Problems," Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, VA, pp. 625-630.

Cox and Ernst (1982) demonstrated that a controlled rounding exists for every two-dimensional additive table. Here, the author establishes by means of a counter-example, that the natural generalization of their result to three dimensions does not hold. However, a rounding does always exist under less restrictive conditions.

Fagan, J. T., Greenberg, B. V. and Hemmig, R., (1988), "Controlled Rounding of Three Dimensional Tables," Bureau of the Census, SRD Research Report No: Census/SRD/RR-88/02

A heuristic procedure for finding controlled roundings of three dimensional tables is presented. The three-dimensional controlled rounding problem is much more difficult than its two-dimensional counterpart. The solution to the two dimensional problem involves representing the table as a system of linear equations, formulating a network flow problem, modeling the system of equations, finding a saturated flow through the network and interpreting the flow as a controlled rounding of the original table. In three dimensions, the system of linear equations cannot be represented as a single network. The heuristic model discussed in this paper employs a sequence of network flow problems; the solution of each reduces the size of the table to be rounded. The sequence of solutions is then used to attempt to extract a controlled rounding of the original table, if one exists. An alternative approach to the problem is in Kelly, Golden, Assad and Baker (1988).

Fienberg, S. (1993), "Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality," Technical Report #577, Department of Statistics, Carnegie Mellon University. An earlier version of this paper was published in Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, 1992.

This paper examines the conflicts between the two perspectives of data access and confidentiality protection and briefly outlines some of the issues involved from the perspectives of governments, statistical agencies, other large-scale gatherers of data, and individual researchers.

Fuller, W.A. (1991), "Masking Procedures for Disclosure Limitation," Journal of Official Statistics, Vol. 9, No. 2, pp. 383-406.

In this study Fuller focuses on masking through the addition of noise to the characteristics, or to a subset of the characteristics, by data switching or imputation methods. He illustrates how an "intruder" can estimate the characteristics of a particular individual, and by what probability, under different characteristics of the microdata and different masking procedures. Fuller maintains that masking error can be treated as measurement error. He outlines a method Sullivan (1989) developed for adding measurement error to the variables of a data set in which the masking procedure is designed to maintain the marginal distribution functions and the covariance structure of the original data. Fuller then applies measurement error procedures to the masked data to construct consistent estimators of regression parameters and other higher order statistics.

*Gates, G. W. (1988), "Census Bureau Microdata: Providing Useful Research Data while Protecting the Anonymity of Respondents," Presented at the annual meeting of the American Statistical Association, New Orleans, LA.

Gates describes some solutions used by the Census Bureau to provide microdata for public use while controlling disclosure risk. These include: thorough review of the files to evaluate risk of individual identification; research on the methodological evaluation of various masking techniques; microaggregations; and remote access whereby users submit computer programs to be run by authorized staff. He also lists administrative solutions such as surveys that are reimbursable rather than protective, special sworn employees, and programs for which all research must be done on site at the Census Bureau. Gates also lists various legal options for dealing with the problem.

*George, J. A. and Penny, R. N. (1987), "Initial Experience in Implementing Controlled Rounding for Confidentiality Control, "Proceedings of the Bureau of the Census Third Annual Research Conference, Bureau of the Census, Washington DC., pp. 253-262.

The New Zealand Bureau of Statistics has been using random rounding. This paper documents a study of controlled rounding to offset the disadvantages of random rounding: (1) that the published values in the rows and columns of the table to not necessarily add to the published marginal totals, and (2) the procedure would not result in consistent

random roundings if applied to the same table at different times. They use the methodology for controlled rounding in two dimensional tables with subtotal constraints described in the paper by Cox and George (1989).

This paper describes the capacitated transshipment formulation for controlled rounding in terms of a network formed from nodes and arcs, with flows created by the nodes and relayed along the arcs. The network is defined to be capacitated when there are non-zero upper and/or lower limits on the flow along some or all of the arcs. The search for a controlled rounding solution becomes the search for a solution to a network that has an integer flow in every arc. The authors describe their implementation of this search using the SAS/OR Network module, but state that most commercial Mathematical Programming systems will solve the problem. The effect of different types of tables are considered as well as the difficulties encountered the implementation.

The authors point to the need for further theoretical work on controlled rounding for multi-dimensional tables and tables with other complex structures, given the advantage of controlled over random rounding in preserving the additivity of table totals.

Govoni, J. P. and Waite, P. J. (1985), "Development of a Public Use File for Manufacturing," Proceedings of the Section on Business and Economic Statistics, American Statistical Association, Alexandria, VA, pp. 300-302.

A procedure for producing a public use data product for the Longitudinal Establishment Data file (LED) is described. The procedure involves sorting on value of shipments within 4-digit SIC code, and then aggregating 3 or more establishments at a time to form pseudo-establishments. The exact extent of aggregation depends upon the (n,k) rules that would be used in publishing tabular data for the same data set.

Testing led to the conclusion that the resulting file was disclosure-free. There is, however, no description of the testing method, other than the statement that testing involved matching to the original data file. In terms of utility of the public use file, the authors noted that correlations were increased by aggregation, but that relative relationships seemed to be preserved.

Greenberg, B. (1985), "Notes on Confidentiality Issues when Releasing Survey or Census Data," presented at the Conference on Access to Public Data sponsored by the Social Science Research Council.

The author notes that currently there is no measure of disclosure risk for microdata files and explains the need for such a measure. Techniques for reducing the disclosure risk of a microdata file such as top-coding, coding into ranges, and limiting geographic detail are discussed. The majority of the paper describes the role and the review process of the Microdata Review Panel.

*Greenberg, B. (1986), "Designing a Disclosure Avoidance Methodology for the 1990 Decennial Censuses," presented at the 1990 Census Data Products Fall Conference, Arlington, VA.

The Census Bureau's objective data release strategy is to maximize the level of user statistical information provided subject to the condition that pledges of confidentiality are not violated. A Confidentiality Staff has been established at the Census Bureau to develop disclosure avoidance methods for use in Census products, most prominently the 1990 Decennial Censuses data products. This paper describes procedures developed, their impact, and how they relate of the Bureau's goals. The two types of procedures described in the paper for reducing disclosure risk in the release of tabular data are suppression (primary and complementary) and noise introduction (controlled rounding and controlled perturbation). The paper concludes that controlled rounding appears to be the preferred method for use on the 1990 Decennial Census data products. However, it was not used (see Griffin, et. al.).

Greenberg, B. (1988a), "An Alternative Formulation of Controlled Rounding," Statistical Research Division Report Series, Census/SRD/RR-88/01, Bureau of the Census, Washington, DC.

The standard definition of controlled rounding is extended to allow a non-zero multiple of the base to decrease as well as increase. Greenberg compares the formulations of the standard and of this extended version of controlled rounding. He shows that, in their respective solutions, the underlying networks differ only with respect to arcs and to costs. The paper gives step-by-step examples of each procedure and contrasts their performances. The procedures developed by the author minimize a measure of closeness-of-fit to provide solutions to either of the controlled rounding definitions. Greenberg asserts that the new definition and its solution process also can be applied to tables of more than two dimensions and refers to Fagan, Greenberg, and Hemmig (1988).

Greenberg, B. (1988b), "Disclosure Avoidance Research at the Census Bureau," Presented at the Joint Advisory Committee Meeting, April 13-14, Oxon Hill, MD.

Greenberg discusses research in a) improving complementary cell suppression procedures for economic tabular data, b) assessing risk inherent in public use demographic microdata, c) design of data release and masking strategies for demographic public use microdata, and d) design of data release and masking for economic microdata.

Greenberg, B. (1988c), "Disclosure Avoidance Research for Economic Data," Presented at the Joint Advisory Committee Meeting, October 13-14, Oxon Hill, MD.

The primary method of releasing economic data by the Census Bureau is through cross-classified tables of aggregate amounts. This report discusses the basic disclosure avoidance methodology employed for tabular data. Even though economic microdata is not systematically released by the Census Bureau, they also report on research into methods for the design of surrogate economic microdata files. This report outlines the

complementary suppression problem, with examples and briefly discusses the use of network theory.

Greenberg, B. (1990a), "Disclosure Avoidance Research at the Census Bureau," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC, pp. 144-166.

The confidentiality staff at Census is involved in diverse projects, including; (a) improving cell suppression procedures for economic tabular data, (b) assessing risk inherent in public use demographic microdata, (c) development and implementation of data masking schemes for demographic public use microdata, and (d) design of data release and masking strategies for economic microdata. The author discusses these projects focusing on objectives, progress to date, current activities, and future work.

Greenberg, B. (1990b), "Disclosure Avoidance Practices at the Census Bureau," presented at the Seminar on Quality of Federal Statistics sponsored by the Council of Professional Associations on Federal Statistics, Washington, DC.

A data collection agency has the obligation to release as much information to the public as possible while adhering to pledges of confidentiality given to respondents. The author discusses the trade-offs between the completeness and the accuracy of microdata and tabular data. For microdata this reduces to releasing fewer variables and collapsing categories versus adding noise to the data or to trading completeness for one data attribute at the expense of completeness for another. For tabular data one either suppresses information and collapses categories or introduces noise. Both these actions can be thought of as data masking. The first option reduces completeness, while the second option reduces accuracy.

Greenberg, B. and Voshell, L. (1990a), "Relating Risk of Disclosure for Microdata and Geographic Area Size," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 450-455.

This paper examines the percent of records on a microdata file that represent individuals or households with a unique combination of characteristics variables. In particular, the authors describe the relationship between the percent of population uniques on a file from a specific geographic region and the size of that region.

Greenberg, B. and Voshell, L. (1990b), "The Geographic Component of Disclosure Risk for Microdata," Statistical Research Division Report Series, Census/SRD/RR-90/12, Bureau of the Census, Statistical Research Division, Washington, DC.

The relationship between the percent of population uniques on a microdata file from a specific geographic region and the size of that region is described. The authors also introduce the idea of using random subsets of microdata records to simulate geographic subsets of microdata records.

Greenberg, B. and Zayatz, L. (1992), "Strategies for Measuring Risk in Public Use Microdata Files," Statistica Neerlandica, Vol. 46, No. 1, pp. 33-48.

Methods of reducing the risk of disclosure for microdata files and factors which 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, pp. 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.

*Jabine, T. B. (1993a), "Procedures for Restricted Data Access," Journal of Official Statistics, Vol. 9, No. 2, pp. 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 which illustrate both successful modes and procedures for providing access, and failures to gain the desired access.

Jabine, T. B. (1993b), "Statistical Disclosure Limitation Practices of United States Statistical Agencies," Journal of Official Statistics, Vol 9., No. 2, pp. 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.

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.

Jones, D. H. and Adam, N. R. (1989), "Disclosure Avoidance Using the Bootstrap and Other Re-sampling Schemes, "Proceedings of the Bureau of the Census Fifth Annual Research Conference," Bureau of the Census, Washington, DC., pp. 446-455.

Methods to protect confidentiality from cleverly designed complex sequences of queries are classified under four general approaches: conceptual modeling, query restriction, data perturbation, and output perturbation. The authors present data coding as the basis of a new perturbation method, and also propose an output perturbation approach based on the bootstrap.

Keller, W. J. and Bethlehem, J. G. (1992), "Disclosure Protection of Microdata: Problems and Solutions," Statistica Neerlandica, Vol. 46, No. 1, pp. 5-19.

Identification and disclosure problems related to the release of microdata in the Netherlands are discussed. The authors discuss both population and sample uniqueness. An argument is presented that disclosure avoidance should be achieved by legal provisions and not by reducing the amount or quality of data releases.

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 little-known 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.

Keller-McNulty, S. and Unger, E., (1993), "Database Systems: Inferential Security," Journal of Official Statistics, Vol. 9, No. 2, pp. 475-499.

The problems of data security and confidentiality have been studied by computer scientists and statisticians. The areas of emphasis within these disciplines on data security are different but not disjoint. One of the main differences is how one views data release. Statisticians have focused on aggregate data release and on single static files of microdata

records. Computer scientists have focused on data release through sequential queries to a database. An initial integrating factor of the two fields is the concept of information stored as a federated database. This paper synthesizes the research done in both of these disciplines and provides an extensive review of the literature. Some basic definitions integrating the two fields are given and data security and confidentiality methodologies studied in both disciplines is discussed.

Kelly, J. P. (1990), "Confidentiality Protection in Two- and Three-Dimensional Tables," Ph.D. Dissertation, University of Maryland, College Park, MD.

Contains proof that the integer programming problem of finding an optimal set of complementary suppressions is "NP-hard"; that is, the number of computations increases (roughly) exponentially with the number of primary suppressions.

Kelly, J. P., Assad, A. A. and Golden, B. L. (1990), "The controlled Rounding Problem: Relaxations and Complexity Issues," OR Spektrum, Springer-Verlag, 12, pp. 129-138.

The three-dimensional controlled rounding problem is described and proved to be NP-complete. For tables where a solution does not exist, a series of relaxations of the zero-restricted problem is described that can lead to solutions. Examples of tables that need various orders of relaxation are given.

Kelly, J. P., Golden, B. L., and Assad, A. A. (1990a), "Cell Suppression: Disclosure Protection for Sensitive Tabular Data," Working Paper MS/S 90-001, College of Business and Management, University of Maryland, College Park, MD.

This paper formulates and develops solution techniques for the problem of selecting cells for complementary suppression in two dimensional tables. (Sensitive cells must not be able to be estimated to within a specified tolerance interval). The authors present a network flow-based heuristic procedure for the complementary suppression problem. The objective function is the minimization of the total of the suppressed values. The authors use the network flow based heuristic procedure currently used by the Census bureau (a sequence of network flow models) to find a feasible solution, then implement a "clean-up" procedure to improve the solution. The paper also develops a lower bounding procedure, which can be used to estimate the quality of the heuristic solution. It can also generate a starting point for the heuristic procedure. Extensive computational results based on real-world and randomly generated tables demonstrate the effectiveness of the heuristic.

Kelly, J. P., Golden, B. L., and Assad, A. A. (1990b), "Cell Suppression Using Sliding Protection Ranges," Working Paper Series MS/S 90-007, College of Business and Management, University of Maryland, College Park, MD.

Complementary suppression cells must be selected to protect those cells identified as primary suppressions. Traditionally, the protection required is defined by an interval centered around the value of each primary suppression. This paper formulates and

develops solution techniques for the problem where sliding protection ranges are allowed; this represents a relaxation of the traditional problem. In this problem, the protection ranges have fixed widths but are free to slide; the only restriction is that they must contain the values of the primary suppressions.

The authors present a network flow-based heuristic for this modified cell suppression problem and use a lower-bounding procedure to evaluate the performance of the heuristic. Extensive computational results based on real-world and randomly generated tables demonstrate that sliding protection ranges can significantly reduce the total amount of suppressed data, as compared to the traditional suppression scheme.

Kelly, J. P., Golden, B. L., and Assad, A. A. (1990c), "A Review of the Controlled Rounding Problem," Proceedings of the 22nd Symposium on the Interface, Interface Foundation of North America, Springer-Verlag, pp. 387-391.

A review of the state of the art in controlled rounding. Notes that three dimensional controlled rounding does not lend itself to a network representation on which to base an effective solution. They quote previous work which demonstrates that the zero restricted controlled rounding problem is NP-Complete. It has also been shown that not every three-way table has a zero-restricted controlled rounding solution. They relax the zero-restricted requirement and discuss their linear programming solution to the relaxed problem (discussed in detail in one of their 1990 papers.) Their procedure (ROUND and BACK) has an advantage in that it either finds a solution or proves that none exists. They also discuss pre-processor heuristics to speed convergence. These are called "Round-Round and Back,""Quick-Round and Back" and "Anneal-Round and Back". The latter appears to be quickest and most successful to date.

Kelly, J. P., Golden, B. L., Assad, A. A. and Baker, E. K. (1988), "Controlled Rounding of Tabular Data," Working Paper MS/S 88-013, College of Business and Management, University of Maryland, College park, MD. (also published in Operations Research, Vol. 38, No. 5, pp. 760-772.)

The authors describe the use of a binary tree search algorithm based on linear programming techniques for solving three-dimensional controlled rounding problems. The algorithm determines whether or not a solution exists, and effectively finds a solution if one does exist. Computational results are presented. A technique for decreasing the running time of the algorithm is also described.

Kim, J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 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. Kim

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. (1990a), "Masking Microdata for National Opinion Research Center," Final Project Report, Bureau of the Census.

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.

Kim, J. (1990b), "Subpopulation Estimation for the Masked Data," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 456-461.

Kim derives estimators for the mean and variance/covariance parameters of a subgroup under different uncorrelated and correlated data conditions when the data set as a whole is masked by additive noise or by additive noise plus transformation. He illustrates the good results using a mixed population data base. He concludes that it is safe to mask the whole data set once and to let the users estimate the mean and variance/covariance of subpopulations using his formulas.

Kumar, F., Golden, B. L. and Assad, A. A. (1992), "Cell Suppression Strategies for Three-Dimensional Tabular Data," Proceedings of the Bureau of the Census 1992 Annual Research Conference, Bureau of the Census, Washington, D. C.

The authors present a cell suppression strategy for three-dimensional tabular data which involves linear programming techniques and heuristic search. The linear programming techniques find a sufficient set of complementary suppressions and a lower bound on the total value that must be suppressed to obtain a sufficient set. If the set obtained has a much higher total value than the lower bound, a heuristic search attempts to find a better solution. An analysis of results is given.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm,," Journal of Official Statistics, Vol. 9, No. 2, pp. 313-331.

The definition of disclosure depends on the context. Sometimes it is enough to violate anonymity. Sometimes sensitive information has to be revealed. Sometimes a disclosure is said to occur even though the information revealed is incorrect. This paper tires 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.

Little, R. J. A. (1993), "Statistical Analysis of Masked Data," Journal of Official Statistics, Vol. 9, No. 2, pp. 407-426.

A model-based likelihood theory is presented for the analysis of data masked for confidentiality purposes. The theory builds on frameworks for missing data and treatment assignment, and a theory for coarsened data. It distinguishes a model for the masking selection mechanism, which determines which data values are masked, and the masking treatment mechanism, which specifies how the masking is carried out. The framework is applied.

Lougee-Heimer, R. (1989), "Guarantying Confidentiality: The Protection of Tabular Data," Master's Degree Thesis, Department of Mathematical Sciences, Clemson University.

Looks at selection of complimentary cells for three-way tables of magnitude data; describes the Census Bureau's current two-way procedure and demonstrates a three-way procedure based on finding linear dependent sets of vectors in a system of linear equations. (Procedure sequentially fixes complementary suppression cells in stages by solving linear programming subproblems.) The suppression method quoted is to protect each primary cell to within a "tolerance". That is upper and lower bounds are specified (in an undisclosed way) for each primary cell. The problem is to select complimentary cells so that it is impossible to estimate the value of the primary cells more accurately than their tolerance regions.

Lunt, T. F. (1990), "Using Statistics to Track Intruders," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC.

The author presents a detailed account of a real-time intrusion-detection expert system that identifies computer users who exhibit unexpected or suspicious behavior. Security systems of this type, while certainly applicable to disclosure avoidance in microdata files, do not fall into the general class of analytic approaches to confidentiality protection.

McGuckin, R. H., (1992), "Analytic Use of Economic Microdata: A Model for Researcher Access With Confidentiality Protection,", Center for Economic Studies Discussion paper CES 92-8, Bureau of the Census, Washington D. C.

This paper describes the benefits of analytic research with economic microdata and describes the administrative arrangements that have been developed by the Census Bureau to provide access to microdata files by selected researchers who are appointed as special sworn employees of the Census Bureau and work on site at the Center for Economic Studies. The author proposes expansion of facilities for user access, including provision of access at regional centers located in universities or Census Bureau regional offices. He also recommends that the Census Bureau use somewhat broader criteria to decide which research projects are relevant to Census Bureau program needs and therefore meet statutory requirements for this mode of access to the Census Bureau's economic microdata.

McGuckin, R. H. and Nguyen, S. V. (1988), "Use of 'Surrogate Files' to Conduct Economic Studies with Longitudinal Microdata," Proceedings of the Fourth Annual Research Conference, Bureau of the Census, Washington, DC., pp. 193-209.

Essentially same paper as one below.

McGuckin, R. H. and Nguyen, S. V. (1990), "Public Use Microdata: Disclosure and Usefulness," Journal of Economic and Social Measurement, Vol. 16, pp. 19-39.

The authors discuss and compare methods for masking economic microdata for public use data files, given the economic data characteristics of uniqueness of particular information and skewed size distribution of business units. They examine summary statistics methods such as data grouping under the (n,k) rule, and providing the variances, covariances and means of the original data. They also discuss using surrogate files involving stochastic and deterministic data transformations.

The authors address theoretical aspects of various transformations that might be applied to longitudinal datasets in order to protect confidentiality. The focus is on the ability of the transformed dataset to yield unbiased estimates of parameters for economic models. Both stochastic and deterministic transformations are considered, all are found to be flawed in one way or another. The authors conclude that it may be more useful to release variance-covariance matrices than to develop transformed microdata files.

*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 micro-data 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.

Mokken, R. J., Kooiman, P., Pannekoek, J. and Willenborg, L. C. R. J. (1992), "Assessing Disclosure Risks for Microdata," Statistica Neerlandica, Vol. 46, No. 1, pp. 49-67.

The authors provided methods to estimate the probability that the release of a microdata set allows users to obtain confidential data for population unique individuals that are known by the user to exist. This paper covers the situation were there are multiple users of a geographically stratified micro data release that contain multiple identification variables.

Mokken, R. J., Pannekoek, J., and Willenborg, L. C. R. J. (1992), "Microdata and Disclosure Risks," Statistica Neerlandica, Vol 46, No 1.

The authors provide a method to calculate the risk that an investigator is able to re-identify at least one individual in an microdata set. This risk is shown to depend on some variables that are readily controlled by the releasing agency such as the coarseness of the key variables and the size of the subsample that is released.

Mugge, R. H. (1983a), "Issues in Protecting Confidentiality in National Health Statistics," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 592-594.

Statistical programs must be administered in such a way as to bring the best scientific results but also to protect any participating subjects. This paper discusses four particular policy issues as they relate to NCHS.

Public use microdata files are protected by, first, suppressing all direct identifiers, and then, suppressing or categorizing other variables that might lead to identification. The author notes that most NCHS releases cannot be linked to known comparison files, so that it is impossible to test the risk of disclosure. He also notes the small sampling fractions that are used in most NCHS surveys, which he believes to be an additional safeguard against disclosure.

Mugge, R. H. (1983b), "The Microdata Release Program of the National Center for Health Statistics," Statistical Policy Working paper 20, pp. 367-376.

The author presents an overview of the NCHS microdata release procedures, which include suppression of some data elements, but not transformation. He notes that NCHS data typically include enough noise so that purposeful addition of more noise would probably be redundant. No testing program exists at the agency, so there are no measures of the level at which confidentiality is protected. However, as far as is known, there has never been a breech of confidentiality.

*Navarro, A., Flores-Baez, L., and Thompson, J. (1988), "Results of Data Switching Simulation," presented at the Spring meeting of the American Statistical Association and Population Statistics Census Advisory Committees.

This paper documents some of the findings from a simulation of a data switching procedure. This procedure is one part of a disclosure limitation technique termed the "Confidentiality Edit" that will be used on data from the 1990 Decennial Census prior to forming demographic tables of frequency counts. From this simulation, it was discovered that the data from small blocks needed additional protection. The success of the procedure and its effect on the statistical properties of the data are described.

Paass, G. (1988), "Disclosure Risk and Disclosure Avoidance for Microdata," Journal of Business and Economic Statistics, Vol. 6, pp. 487-500.

Paass 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. He also 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.

Paass, G. (1989), "Stochastic Generation of a Synthetic Sample from Marginal Information," Proceedings of the Bureau of the Census Fifth Annual Research Conference, Bureau of the Census, Washington, DC, pp. 431-445.

In this paper the author describes a stochastic modification algorithm (SMA) used to construct a synthetic sample X from different input sources, the sources being independent samples or summary statistics from an underlying population. The first step in the process is to construct an X as a best fit to the data by a maximum likelihood or minimum cost criterion, and the second step is to generate a sample with a cost value near the minimum which also has maximum entropy. Paass tests his method on income tax data for the German Treasury.

*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.

*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.

Rowe, E. (1991), "Some Considerations in the Use of Linear Networks to Suppress Tabular Data," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 357-362.

The paper discusses network flow theory and its application to finding optimal complementary suppressions in two dimensional tables. The network flow methodology uses closed paths in the table. This analysis considers defining costs to try to assure that the selected path both includes all primary suppressions and minimizes the sum of the suppressed cells (total cost). The paper points out the problems associated with treating one primary cell at a time in terms of finding the "least cost" path.

*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, pp. 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.

Sande, G. (1984), "Automated Cell Suppression to Preserve Confidentiality of Business Statistics, "Statistical Journal of the United Nations, ECE 2, pp. 33-41.

Discusses in general terms the application of linear programming to complementary suppression. Also outlines the CONFID program developed by Sande at Statistics Canada.

*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, pp. 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.

Skinner, C. J. (1992), "On Identification Disclosure and Prediction Disclosure for Microdata," Statistica Neerlandica, Vol 46, No. 1, pp. 21-32.

Skinner discusses how to estimate the probability of disclosure for two types of disclosure (identification and prediction.) In particular he demonstrates how a Poisson-gamma model can be used to estimate the number of population unique records.

Skinner, C. J. and Holmes, D. J. (1992), "Modelling Population Uniqueness," Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, pp. 175-199.

Authors present various statistical models to be used to estimate the number of population unique records using data collected from a sample from the population. In particular there are examples that demonstrate the effectiveness of a Poisson-lognormal model.

Skinner, C. J., Marsh, C., Openshaw, S., and Wymer, C. (1990), "Disclosure Avoidance for Census Microdata in Great Britain," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC, pp.131-143.

The authors lay out in detail the structural logic of estimating the risk of disclosure and its dependence on factors which can be controlled by those masking methods that preserve the integrity of the data (not by contamination methods). They characterize the type of investigator, the degree of identification, the scenario by which the investigator attempts identification, and the two steps that must be achieved for identification, i.e. i) locate a record that matches the key individual on all the variables common to the microdata and to the additional information file, and ii) infer with some degree of confidence that this record does belong to the target individual.

The authors then summarize the four conditions under which identification is possible as 1) the target individual does appear in the microdata; 2) the common variable values of the target individual are recorded identically in the additional information and the microdata; 3) the combination of common variable values for the target individual is unique in the population; and 4) the investigator infers with some degree of confidence that the combination of common variable values is unique in the population.

The assessment of the probability of each of the four conditions in the context of census microdata is then set forth in some detail, and their product is proposed as the estimate of the risk of disclosure.

Spruill, N. L. (1982), "Measure of Confidentiality," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 260-265.

Spruill uses a monte-carlo procedure to measure the effectiveness of five disclosure avoidance procedures: adding random noise, multiplying by random noise, aggregation, random rounding and data swapping.

Spruill, N. L. (1983), "The Confidentiality and Analytic Usefulness of Masked Business Microdata,"Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 602-607.

This paper presents empirical results on the trade-off between confidentiality protection and data usefulness. Both simulated and real data are used, and several masking procedures are used with each dataset: additive random error, multiplicative random error, grouping, rounding, and data swapping. Confidentiality protection is measured in terms of the proportion of entities that can be correctly linked to a public file.

The results indicate that, when the number of matching variables is small (4 to 6) all masking procedures can be used successfully. When this number is high (20 to 32), masking is much more problematic, although grouping becomes a more attractive procedure. It is noted that the proportion of zeroes in the data set can be an important consideration.

Strudler, M., Oh, H. L. and Scheuren, F. (1986), "Protection of Taxpayer Confidentiality with Respect to the Tax Model," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 375-381.

The Tax Model, a microdata file of a sample of individual taxpayer' returns, is made available to the public. This paper describes the methods used to minimize disclosure risk from the Tax model, and the results of empirical testing of the resulting file. Disclosure risk is reduced by rounding, by "blurring" independently across sensitive variables, and by lowering subsampling rates in the high-income strata. Testing was based on the "Spruill method" of finding best matches by minimizing the sum of absolute deviations,

taken across variables for which it is believed that true data may be known with certainty. Testing indicated that a useful file could be released to the public.

Sullivan, C. M. (1992a), "An Overview of Disclosure Practices," Statistical Research Division Research Report Series, Census/SRD/RR-92/09, Bureau of the Census, Washington, DC.

This paper gives a general definition of the problem of disclosure avoidance for tabular data. The author describes sensitive data, cell suppression, primary suppression rules, cost functions, and feasible ranges for suppressed values.

Sullivan, C. M. (1992), "The Fundamental Principles of a Network Flow Disclosure Avoidance System," Statistical Research Division Research Report Series, Census/SRD/RR-92/10, Bureau of the Census, Washington, DC.

This paper provides a very clear explanation of how to translate the problem of cell suppression in a table into a problem that can be solved with network flow methodology. In particular it explains how to use a network to describe tables which are related additively in one dimension and additive with a hierarchial structure in the other dimension.

Sullivan, C. M. (1993a), "A Comparison of Cell Suppression Methods," ESMD-9301, Economic Statistical Methods Division, Bureau of the Census, Washington, DC.

Two techniques for removing superfluous suppressions when network flow methodology is used to apply complementary suppression are described. Also discussed are problems encountered when applying the techniques to actual economic census data.

Sullivan, C. M. (1993b), "Adjustment Techniques to Supplement a Network Flow Disclosure Avoidance System," Proceedings of the International Conference on Establishment Surveys.

This is an extended version of Sullivan (1993a).

Sullivan, C. and Rowe, E. (1992), "A Data Structure Technique to Facilitate Cell Suppression Strategies," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 685-690.

The authors describe the network flow methodology currently used at the Census Bureau for identifying complementary suppressions for a primary suppression. They then introduce an integer programming technique which may be used after the network flow technique has identified complementary suppressions for all primary suppressions to release superfluous complementary suppressions.

Sullivan, C. and Zayatz, L. (1991), "A Network Flow Disclosure Avoidance System Applied to the Census of Agriculture," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 363-368.

Since rounding and perturbation are unsatisfactory for aggregate magnitude data, the Economic and Agriculture Divisions have always chosen a cell suppression technique to protect aggregate data. The objective in applying complementary suppressions is to ensure the protection of the sensitive data value at minimum cost. Commonly, the original data value that would have appeared in the publication is assigned as the cost. Minimizing the cost incurred through complementary suppression produces a publishable table with maximum data utility; that is, the greatest amount of usable data is provided. (Note: The solutions obtained are not optimal.)

For the 1992 Census of Agriculture, research was conducted on the cell suppression technique using the network flow system of applying complementary suppressions. However, the existing network flow system was not optimal for agricultural data because of the complexity of the data structure and its use of systems of three dimensional tables. Therefore the network flow methodology required customizing. This paper discusses the formulation of the customized network methodology and the limitation encountered with the customized version when applied to agricultural data.

The network-flow methodology for agricultural data was successfully adapted to some extent. However, in its present form, the authors feel it is still unsuitable.

Sullivan, G. R. (1989), "The Use of Added Error to Avoid Disclosure in Microdata Releases," Unpublished Ph.D. Dissertation, Iowa State University, Ames, Iowa.

This paper discusses methods of adding error to observations by means of a masking algorithm that creates a data set that is statistically representative of the original data records in three ways. First, the masked data set should have the same first and second moments as the original data set. Second, the correlation structure of the original and masked data sets should be nearly identical. Third, the univariate distribution functions of the original and masked data should also be the same. The paper also investigates the statistical usefulness of the masked data set by comparing statistical analyses performed on the original and masked data and it evaluates the effectiveness of the mask with regard to disclosure avoidance.

Sullivan, G. R. and Fuller, W. A. (1989), "The Use of Measurement Error to Avoid Disclosure," Proceedings of the Section on Survey Research, American Statistical Association, Alexandria, VA, pp. 802-807.

The authors describe the general technique of adding a random error vector to mask each data vector. On the basis of an approach an intruder would use to construct predictions of confidential variables from conditional probabilities derived from data already known, they illustrate how to select an error covariance matrix for masking normally distributed

data. To find a balance between an error variance large enough to sufficiently lower the probability of matching a record but not to severely distort the data, they illustrate the efficacy of adding vectors of error that have a covariance matrix equal to a multiple of the covariance matrix of the original unmasked data vectors.

Sullivan, G. R. and Fuller, W. A. (1990), "Construction of Masking Error for Categorical Variables," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 435-439.

The authors present a method for masking categorical variables to reduce the risk of attribute disclosure, in which each classification variable is transformed into a Bernoulli variable, and then further transformed into a standard normal variate using the sample univariate distribution functions. The masking is completed by adding a normally distributed error vector to each transformed vector of normalized data. They illustrate how to back-transform the data to the original scale, and then to convert the Bernoulli variables back to their categorical values, and provide an example in terms of its correlation structure.

Tendrick, P. and Matloff, N. S. (1987), "Recent Results on the Noise Addition Method for Database Security," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 406-409.

Adding random noise to sensitive variables maintains expected mean values, but adds bias to other common estimators. The authors investigate bias in percentile ranks and in regression coefficients, and, for the case of regression coefficients, develop methods for eliminating this bias. They also discuss the advantages of adding multivariate noise (when several sensitive variables are involved) that has the same covariance structure as the original data.

Vemulapalli, K. C. and Unger, E. A. (1991), "Output Perturbation Techniques for the Security of Statistical Databases," Proceedings of the 14th National Computer Security Conference, Washington, DC.

This paper addresses the technique of adding "noise" or perturbations to query answers in order to prevent disclosure. The study analyses not only the amount of protection, but also the amount of bias introduced. In particular, the analysis is done for **sum** queries. Ensuring that identical query sets return identical answers is proposed as a solution to compromise by averaging. The favored solutions offer high security while requiring relatively little extra processing time, and so are suitable for on-line systems.

Vogel, F. A. (1992), "Data Sharing: Can We Afford It?," Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, Dublin.

The author discusses several issues that have been raised concerning disclosure, confidentiality, and the sharing of data collected for the National Agricultural Statistics

Service estimating program. There is general agreement within NASS that public benefits can be derived from further analysis of survey data. However, under no circumstances will the preservation of the confidentiality pledge be violated. Special tabulations will be done and on-site analysis can occur only if all confidentiality provisions have been met. Full data sharing does not exist because of NASS supported confidentiality standards.

Voshell, L. (1990), "Constrained Noise for Masking Microdata Records," Statistical Research Division Report Series, Census/SRD/RR-90/04, Statistical Research Division, Bureau of the Census, Washington, DC.

The author presents two algorithms which transform data generated by a random number generator into data satisfying certain constraints on means and variance-covariance structure. Data sets such as these may be beneficial when used for introducing noise in order to mask microdata as a disclosure avoidance technique.

Wang, Q., Sun, X., and Golden, B. L. (1993), "Neural Networks as Optimizers: A Success Story," College of Business and Management, University of Maryland, College Park, MD.

The authors apply a modified, continuous Hopfield neural network to attack the problem of cell suppression. They design an energy function and a learning algorithm to solve two-dimensional suppression problems. The approach is shown to perform well.

Wester, W. C. and Hemmig, R. (1984), "Disclosure Analysis for the Economic Censuses," Proceedings of the Business and Economic Statistics Section, American Statistical Association, Alexandria, VA, pp. 406-409.

Discusses the practical implementation of a complementary disclosure scheme for the complex set of tables published in the Economic Censuses.

Willenborg, L. C. R. J. (1992a), "Disclosure Risk for Microdata Sets: Stratified Populations and Multiple Investigators," Statistica Neerlandica, Vol 46, No 1.

Willenborg discusses the estimation of the risk of disclosure from the release of a geographically stratified microdata set to multiple investigators. The risk of disclosure is high when data is released for small geographic areas because an investigator is very likely to be aware of population uniques for small geographic areas .

Willenborg, L. C. R. J. (1992b), "Remarks on Disclosure Control of Microdata," Statistica Neerlandica, Vol. 46, No. 1.

Willenborg discusses what conditions need to hold so that a computationally easy function can be used to estimate disclosure risk. He also discusses use of subsampling and redefinition of key variables to reduce the risk of disclosure. In addition it is shown how contamination of the original data by adding noise to the key values can reduce the disclosure risk.

Willenborg, L. C. R. J., Mokken, R. J., and Pannekoek, J. (1990), "Microdata and Disclosure Risks," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, DC, pp. 167-180.

The authors introduce a measure of the disclosure risk of a microdata file. The measure involves the probabilities that a respondent is in the sample, that the intruder knows information about the respondent, that the information known by the intruder identifies the respondent to be unique in the population, and that the intruder knows that the respondent is unique and finds and recognizes the respondent in the microdata file. A Poisson-Gamma model which can be used to predict uniqueness in the population is described.

Wolf, M. K. (1988), "Microaggregation and Disclosure Avoidance for Economic Establishment Data," Proceedings of the Business and Economic Statistics Section, American Statistical Association, Alexandria, VA.

This paper describes the development of a microaggregate data file. The author evaluates the degree to which microaggregation preserves information contained in the original data. The effect of microaggregation on disclosure risk of a data file is also discussed.

Wright, D. and Ahmed, S. (1990), "Implementing NCES' New Confidentiality Protections," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 446-449.

Public Law 100-297 (Stafford Hawkins Act) imposed new and more stringent confidentiality requirements on NCES. The authors provide an overview of the methods that have been developed and implemented to meet the new requirements, and of the process that led to these methods. With regard to public use tapes, the paper discusses the data masking and testing procedures that were used for various surveys, focusing on the identification of publicly available reference files, on the use of a Euclidean distance measure for matching sample schools to reference schools, and on the problems that arise when school coordinators know the identities of teachers who were sampled in their schools.

Zayatz, L. (1991a), "Estimation of the Number of Unique Population Elements Using a Sample," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 369-373.

The author introduces two methods of estimating the percent of unique population elements in a sample microdata file. Examples of performance are given.

Zayatz, L. (1991b), "Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample," Statistical Research Division Report Series, Census/SRD/RR-91/08, Bureau of the Census, Statistical Research Division, Washington, DC.

The author introduces two methods of estimating the percent of unique population elements in a sample microdata file. A third method is evaluated. Examples of performance of all three methods are given.

Zayatz, L. (1991c), "Estimation of the Percent of Records on a Death Information Microdata File that Could be Correctly Matched to CPS Microdata," Statistical Research Division Technical Note Series, No. RR-91/02, Bureau of the Census, Statistical Research Division, Washington, DC.

This paper describes the results of using two techniques to estimate the percent of records on one version of a microdata file (the Death Information file) that could be linked to a Current Population Survey microdata file. The Microdata Review Panel considered these results, made some changes to the file, and then approved the release of the file.

Zayatz, L. (1992a), "The Effect of Geographic Detail on the Disclosure Risk of Microdata from the Survey of Income and Program Participation," Statistical Research Division Report Series, Bureau of the Census, Statistical Research Division, Washington, DC, No. CCRR-92/03.

The author describes the relationship between the percent of population uniques and the geographic detail on a Survey of Income and Program Participation microdata file. The relationship between the percent of sample uniques and the geographic detail on such a file is also examined. The objective is to relate the consequences in terms of disclosure risk of lowering the required minimum number of persons in the sampled population per identified geographic region on SIPP microdata files.

Zayatz, L. (1992b), "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 which 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. (1992c), "Linear Programming Methodology for Disclosure Avoidance Purposes at the Census Bureau," Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 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 subtables. 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. V. (1992d), "Using Linear Programming Methodology for Disclosure Avoidance Purposes," Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute, pp. 341-351.

This paper is based on Zayatz (1992b). It describes the implementation of linear-programming to find complementary suppressions for a single primary suppression. The method identifies the complete set of complementary suppressions by considering the primary suppressions sequentially. The procedure is applicable to two or three dimensional tables. The three ways of improving results, listed above under Zayatz (1992b), are discussed.

**Reports Available in the Statistical Policy Working Paper Series**

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document Sales, PB86-211539/AS)
3. An Error Profile: Employment as Measured by the Current Population Survey (NTIS Document Sales, PB86-214269/AS)
4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)
5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)
6. Report on Statistical Uses of Administrative Records (NTIS Document Sales, PB86-214285/AS)
7. An Interagency Review of Time-Series Revision Policies (NTIS Document Sales, PB86-232451/AS)
8. Statistical Interagency Agreements (NTIS Document Sales, PB86-230570/AS)
9. Contracting for Surveys (NTIS Document Sales, PB83-233148)
10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)
11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276/AS)
12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)
13. Federal Longitudinal Surveys (NTIS Document Sales, PB86-139730)
14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)
15. Quality in Establishment Surveys (NTIS Document Sales, PB88-232921)
16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB90-205238)
17. Survey Coverage (NTIS Document Sales, PB90-205246)
18. Data Editing in Federal Statistical Agencies (NTIS Document Sales, PB90-205253)
19. Computer Assisted Survey Information Collection (NTIS Document Sales, PB90-205261)
20. Seminar on the Quality of Federal Data (NTIS Document Sales, PB91-142414)
21. Indirect Estimators in Federal Programs (NTIS Document Sales, PB93-209294)
22. Report on Statistical Disclosure Limitation Methodology (NTIS Document Sales, PB94-165305)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703)487-4650.