Journal of Official Statistics
Vol. 9. No. 2, 1993, pp. 427–454
© Statistics Sweden

Statistical Disclosure Limitation Practices of United States Statistical Agencies[1]

Thomas B. Jabine[2]

Abstract: One of the topics examined by the Panel on Confidentiality and Data Access was the use of statistical disclosure limitation procedures to limit the risk of disclosure of individual information when data are released by U.S. 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 early in 1991. This paper is an updated version of that summary.

Key words: Statistical disclosure limitation; tabulations; microdata.

1. Introduction

This paper provides a description of the policies, practices and procedures used by United States federal statistical agencies for statistical disclosure limitation (SDL). SDL methods are applied by the agencies to limit the risk of disclosure of individual information when statistics are disseminated in tabular or microdata formats. The paper is based on a document prepared by the author for the Panel on Confidentiality and Data Access, Committee on National Statistics. The purpose of that document, which was submitted to the Panel in June 1991, was to furnish background information for use by the Panel in developing its findings and recommendations on this topic. The document presented to the Panel was based primarily on statistical agencies' responses to a request by the Panel for information about their confidentiality and data access policies and practices, with some follow-up by the author in a few instances. Some information was taken from an appendix to *Statistical Policy Working Paper 2, Report on Statistical Disclosure and Disclosure-Avoidance Techniques* (Office of Federal Statistical Policy and Standards 1978), which described the SDL practices used at that time by seven federal agencies. The initial draft was reviewed by agency representatives and revised to reflect their comments.

The next section of this paper presents the relevant information for each agency. If an agency is not included, either there was no submission by that agency or its submission did not include any information

[1] This paper was prepared for the Panel on Confidentiality and Data Access of the Committee on National Statistics. Material from this paper has been included in Chapter 6 of the panel report, *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.*

[2] Statistical Consultant, 3231 Worthington Street N.W., Washington D.C. 20015-2362, U.S.A.

relevant to SDL practices. The agency summaries are followed by a general discussion of the current status of SDL policies, practices and procedures, based on the agencies' information, and some concluding remarks. Copies of selected agency documents are included as attachments.

2. Agency Summaries

2.1. Department of Agriculture

2.1.1. Economic Research Service (ERS)
A statement of "ERS Policy on Dissemination of Statistical Information," dated September 28, 1989, 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, which is commonly applied to aggregate data in economic surveys, is often referred to as an (n, k) concentration rule. In this instance $(n, k) = (1, 0.6)$.

2.1.2. National Agricultural Statistics Service (NASS)
NASS has a formal "Standard for Suppressing Data Due to Confidentiality," Policy and Standards Memorandum No. 12–89, effective July 12, 1989 (shown in Exhibit 1). It covers only aggregate data, because NASS does not release public-use microdata (some external researchers are allowed on-site access to microdata under special arrangements). 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 represent more than 60% of the published value. There is a procedure for obtaining waivers which permits publication of values when these conditions apply. The standard also requires that appropriate steps be taken, when values are suppressed, to avoid complementary disclosure.

2.2. Department of Commerce

2.2.1. Bureau of Economic Analysis (BEA)
A three-page statement, "International Investment Division Primary Suppression Rules" (Landefeld 1990), covers the Division's SDL procedures for aggregate data from its surveys of companies. No cells can be published with fewer than three reporters and the top two reporters must account for less than an unspecified percentage of the total. The value of that percentage 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." The statement includes several "special rules" covering rounded estimates, county 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.
BEA's Regional Economic Measurement Division publishes estimates of local area personal income by major source. Quarterly data on wages and salaries paid by county are obtained from the Bureau of Labor

Statistics (BLS) and the BEA is obliged to follow SDL rules that satisfy BLS requirements. Exhibit 2, taken from the BEA publication *Local Area Personal Income, 1983–88*, Volume 1, describes the SDL procedures used: a combination of suppression, rolling up (e.g., combining data for two or more counties or industries) and disturbing data.

2.2.2 Bureau of the Census

The Census Bureau's past and current practices in the application of SDL 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, the Census Bureau had to develop suitable SDL 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 U.S. statistical agencies when these agencies have become more sensitized to the need to protect the confidentiality of individually identifiable information when releasing tabulations and microdata.

The Census Bureau's current and recent SDL practices and research are summarized in Greenberg (1990, 1991). Disclosure avoidance procedures for the 1990 Census of Population are described by Griffin, Navarro and Flores-Baez (1989). Earlier perspectives on the Census Bureau's SDL practices are provided by Cox, Johnson, McDonald, Nelson and Vazquez (1985) and Barabba and Kaplan (1975). Many other references will be found in these five papers.

Following are a few highlights of historical developments and current practices:

- In the 1960s and 1970s, the Census Bureau pioneered in the development of automated procedures for primary and complementary cell suppression applied to tabulations of aggregate data from the quinquennial economic censuses. Many refinements have been developed since then, such as the introduction of objective criteria for deciding which of several possible suppression patterns should be selected. Earlier efforts used minimization of the number of cells suppressed; more recent procedures minimize the total value suppressed. Identifying all possible complementary disclosures in a table of three or more dimensions continues to be a problem.

- For small-area count data based on the 100% items from the 1980 Census of Population and earlier censuses, a cell-suppression procedure was used to limit the risk of statistical disclosure. This method was found to have significant disadvantages, such as the need to suppress data in some large cells to avoid complementary disclosures and the difficulty of identifying all of the suppressions needed to avoid complementary disclosures. Consequently, a different procedure, involving data swapping, has been adopted for the 1990 Census. According to Greenberg (1991), "The Confidentiality Edit selects a small sample of census household records from the internal census data files and interchanges their data with other households which have identical characteristics on a set of selected key variables but are in different geographic locations." The procedure will not disturb census counts of total number of persons and totals by race, Hispanic origin and two age categories, under 18 and 18 and above. These counts provide information needed for implementation of the Voting Rights Act (U.S. Public Law 94–171). At a March 1991 Disclosure Limitation Conference organized by the Panel, Census Officials described the Census Bureau's consultations with representatives of the American Civil Liberties Union and other advocacy groups about the properties of the new procedures.

- In February 1981 the Census Bureau established a formal Microdata Review Board, being the first agency to do so. Approval of the Board is required for each release of a new microdata file. Exhibit 3 shows the instructions for submitting a proposed release for review by the Board. One criterion used by the Board is that geographic codes included in microdata sets should not identify areas with less than 100,000 population. This cutoff was adopted in 1981; previously a figure of 250,000 had been used. For the views of a representative of another agency concerning the Board's position on release of microdata from a survey conducted for that agency by the Census Bureau, see comments below under Department of Veterans Affairs (2.12).

2.3. Department of Defense

2.3.1. Defense Manpower Data Center

The Panel received no written material on SDL procedures, but was assured by its contact person that the Defense Manpower Data Center uses SDL procedures when data from its surveys are published.

2.4. Department of Education

2.4.1. National Center for Education Statistics (NCES)

In January 1991, at the request of the NCES, the Panel on Confidentiality and Data Access organized a Workshop on Confidentiality of and Access to NCES Data. Documentation on NCES's SDL policies and procedures was supplied in the briefing materials for the workshop.

A standard published in February 1987 (CES Standard 87-04-03, "Standard for Maintaining Confidentiality") said that "... care must be taken to ensure that individual respondents cannot be identified where confidentiality has been promised." The only specific requirement was that data cells based on fewer than three respondents be suppressed in ways such that they cannot be reconstructed from data in other cells.

Subsequent to the 1988 passage of legislation (the Hawkins-Stafford Amendment) with new requirements for protecting the confidentiality of NCES data, the need for systematic application of SDL procedures to NCES tabulations and microdata became more evident. A Disclosure Review Board was established in 1989 with responsibility for reviewing disclosure analyses and making recommendations to the Commissioner concerning release of all new public use data tapes. In assessing disclosure risks, the Board takes 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.

A new standard (NCES 1992, Standard IV-01-92) states that "In reporting on surveys and preparing public use data tapes, the goal is to have an acceptably low probability of identifying individual respondents." For tabulations, the standard requires that "... each publication cell must have at least three (unweighted) observations in it and subsequent tabulations (e.g., cross–tabulation) must not provide additional information which would disclose individual entities." Supporting material for the new standard provides detailed guidelines for disclosure analyses of proposed public-use data tapes.

2.5. Department of Energy

2.5.1. Energy Information Administration (EIA)

The *Energy Information Administration Standards Manual*, issued in April 1989, includes a one-page standard (88-05-06) "Nondisclosure of Company Identifiable Data in Aggregate Cells" and a set of "Guidelines for Implementation of a Disclosure Avoidance Rule" (Exhibit 4). Data cells must be based on three or more respondents, and a cell dominance rule called the "p/q rule" is applied, where p/q is 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. Values of p/q selected for specific surveys are not published and are considered confidential.

In the EIA standard, p denotes the prior accuracy of one company's estimate of another's reported value and q denotes its posterior accuracy, that is, subsequent to publication of the data. For example, if the prior estimate were accurate within 30% of the actual value and the posterior estimate were accurate within 10%, we would have $p/q = 3$. In the literature on this subject, it is more usual to reverse the meanings of p and q, so that one would be dealing with a q/p rule.

The Standards Manual does not cover SDL techniques for microdata files. EIA has issued several public-use files based on its periodic Residential Energy Consumption Surveys. In these files, various standard SDL procedures are used to protect the confidentiality of individual households and persons. No geographic identifiers below the Census Division level are included. Data on local area weather conditions (such as the number of heating and cooling degree-days during the reference period) and electricity prices have the potential for identifying specific areas. Error inoculation procedures are therefore used to disguise the weather and electricity cost data.

2.6. Department of Health and Human Services

2.6.1. National Center for Health Statistics (NCHS)

The *NCHS Staff Manual on Confidentiality*, published in 1984, includes section 10, "Avoiding Inadvertent Disclosures in Published Data" and 11, "Avoiding Inadvertent Disclosures Through Release of Microdata Tapes" (Exhibit 5). A variety of straightforward standard SDL techniques are presented. No quantity figures should be based on fewer than three cases and a (1, 0.6) concentration rule is prescribed. 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. For microdata sets, geographic places with fewer than 100,000 people are not to be identified. All new microdata sets must be approved for release by the Director or Deputy Director.

2.6.2. Social Security Administration (SSA)

No current information on SDL procedures was included in SSA's submission to the Panel. However, Office of Statistical Policy and Standards *Statistical Policy Working Paper 2* (1978) included a fairly detailed SSA statement on "Policies and Procedures for Avoiding Disclosure in the Release of Statistical Tabulations and Microdata" and the Panel was informed that these policies and procedures are still in effect. Part of this statement consists of a set of guidelines issued by the Assistant Commissioner, Office of Research and Statistics in 1977. The guidelines include basic rules for count data and dollar amounts and for the release of tabulations based on earnings and benefit data merged with data for the same persons supplied by outside researchers. No dollar amounts are to be published for cells containing fewer than three members.

2.7. Department of Interior

2.7.1. Bureau of Mines (BOM)

The Bureau of Mines has a "Standard for Handling Proprietary Survey Data", No. 1–85, dated August 22, 1985. Section VI, "Disclosure Analysis" (Exhibit 6), describes the SDL procedures that are to be used. In general, publishable cells must have at least three companies and must comply with a multiple concentration rule: (1, 0.75) and (2, 0.90).

The Panel was informed, in a letter (Absalom 1991) that the BOM standard would be revised in the near future. The letter noted that:

> This will probably entail revising the Standard to allow publication of data cells with 3 or more respondents where, e.g., two respondents account for 90 percent of the cell total if both dominant companies waive the confidentiality of their survey data.

2.8. Department of Justice

2.8.1. Bureau of Justice Statistics (BJS)

No information on SDL procedures was included in BJS's submission. The National Crime Survey is conducted for BJS by the Census Bureau, so presumably the latter determines the SDL procedures to be applied to the public-use microdata sets from that survey.

2.9. Department of Labor

2.9.1. Bureau of Labor Statistics (BLS)

Commissioner's Order No. 2-80, "Confidential Nature of Bureau Records," dated July 3, 1980, requires that:

> 7e. Publications shall be prepared in such a way that the figures will not reveal the identity of any specific respondent or will not knowingly allow the data of any specific respondent to be imputed from the published figures.

A subsequent provision allows for exceptions under conditions of informed consent. Only the Commissioner can authorize such exceptions.

A statement submitted to the Panel by the BLS (1990) observed that SDL techniques vary by program. The most commonly used is the (n, k) concentration rule. Specific values are cited for two programs: (1, 0.8) for total employment in the BLS's monthly sample surveys of establishments and (1, 0.5) for cell weights in the Producer Price Index program. Generally, values of k range from 0.3 to 0.8. In addition, most establishment surveys have requirements for a minimum number of respondents or observations per cell.

For microdata files from surveys like the Current Population Survey (a monthly labor force survey), the SDL procedures are determined by the Census Bureau, which conducts these surveys for BLS under an interagency agreement.

2.10. Department of Transportation

2.10.1. National Highway Traffic Safety Administration (NHTSA)

NHTSA operates a National Accident Sampling System in which a national sample of traffic accidents is selected for investigation in sufficient detail to support NHTSA's standards development and evaluation programs. The Panel received an NHTSA (1981) document "Agency Procedures for Release and Security of Research Data Collected Under the National Accident Sampling System." Much of the document is about procedures applied by agency quality control contractors for designated geographic regions to remove explicit identifiers from accident documentation, which includes police reports, driver records, vehicle records, medical records, death certificates, etc., prior to sending data to NHTSA for inclusion in its automated files. One potential identifier, the Vehicle Identification Number (VIN), is retained by NHTSA because it is needed for analyses by vehicle make and model, which can be determined from the VIN. However, the serial portion of the VIN is deleted for any releases of data outside the agency.

The same procedures concerning deletion of personal identifiers and release of partial VINs are used in the agency's Fatal Accident Reporting System.

2.11. *Department of the Treasury*

2.11.1. Internal Revenue Service, Statistics of Income Division (SOI)

The Panel received Chapter VI of the *SOI Division Operating Manual*, issued January 7, 1985. The only specific SDL rule mentioned in Chapter VI is 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, e.g., counties, require at least ten observations. The Statistics of Income Division has sponsored research on SDL techniques, notably work by Spruill (1982, 1983) in the early 1980s, which was directed at the evaluation of masking procedures for business microdata. On the basis of Spruill's findings, the SOI released some microdata files for unincorporated businesses. Except for this one instance, U.S. statistical agencies have not issued public-use microdata sets of establishment or company data, presumably because they judge that application of the SDL 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.

2.12. *Department of Veterans Affairs*

The Panel's contact person informed the Panel that there were no department-wide confidentiality provisions. His letter to the Panel (Dientsfry 1991) described his experiences negotiating with the Census Bureau's Microdata Review Board for access to microdata files from the 1987 Survey of Veterans, which the Census Bureau conducted for the Veterans Administration. He felt that the Review Board required excessive restrictions in order to eliminate the possibility that the survey data might be matched to information in his Department's files.

2.13. *Independent agencies*

2.13.1. General Accounting Office (GAO)

The excerpts from the GAO's *Project Manual* (1989, 1990) that were sent to the Panel provided extensive information on how the GAO gains access to information for its studies. They discuss the circumstances under which pledges of confidentiality may be given to persons providing information and give some examples of appropriate language, e.g., "Although individual answers may be disclosed in our report, they will not include any information that could be used to identify individual respondents."

The GAO does not provide specific guidelines on SDL procedures that should or might be used to ensure that such pledges would be adhered to. According to a letter (Grosshans 1991):

> Because of the diversity of data collection instruments GAO uses, we do not specify a step-by-step process for breaking the link between the respondent and the response, but believe our guidance establishes the requirement to break that link. We provide our managers with the flexibility to implement our requirements in the manner that best serves the need of the particular assignment.

2.13.2. National Archives and Records Administration (NARA)

The Panel's initial response from NARA stated that confidentiality restrictions are determined by the agencies from which they receive records. A subsequent letter (Thibodeau 1991) provided clarification:

> Subpart I [of 44 U.S.C. 2107] outlines the terms under which the Archivist of the U.S. administers the transfer of records to the National Archives. It provides that agency heads may identify restrictions on the use or examination of records transferred to the National Archives, provided they identify the statutory basis or Freedom of Information Act exemption that pertains. The Archivist of the U.S. will not remove restrictions so identified, without the written concurrence of the head of the agency from which the records were transferred. However, such restrictions remain in force only until the records have been in existence for 30 years, 'unless the Archivist determines for specific bodies of records that the restrictions shall remain in force for a longer period.'

2.13.3. National Science Foundation (NSF)

In November 1988, at the request of the NSF, the Committee on National Statistics convened a Workshop on Confidentiality of and Access to Doctorate Records. One of the topics discussed at that workshop was the nature of SDL procedures that would have to be applied in order to release public-use microdata files based on NSF's Doctorate Records File (based on an annual census of doctorate recipients) and Survey of Doctorate Recipients (a longitudinal sample survey). A report prepared for NSF (Boruch and Kehr 1983) had shown conclusively that if records from the Doctorate Records File were released with names deleted, it would still be possible to identify most records with only a few items of collateral information readily available from other sources. Disclosure risks were especially high for women and minorities. Participants in the Workshop agreed with this assessment and agreed that the content of any public-use file would have to be very limited. It was suggested that such files be issued as a means of introducing potential users to the two data bases, but that restricted access arrangements would need to be developed for researchers desiring more detailed information.

Although the NSF subsequently considered the possibility of releasing a public-use microdata file, it has not done so. Some researchers have gained access to microdata files under written agreements which require them to work at sites where their access and use can be supervised by NSF employees or contractors.

3. Discussion

3.1. Variation among agencies in their general approaches to SDL

Many U.S. statistical agencies have standards, guidelines or formal review mechanisms that are designed to ensure that adequate disclosure analyses are performed and appropriate SDL techniques applied prior to release of tabulations and microdata. The Census Bureau has a formal review mechanism for microdata releases, but did not provide any formal agency standard or guidelines for tabular data. It is possible that such materials do exist but are not considered releasable because they contain specific parameter values associated with disclosure limitation rules. Standards and guidelines exhibit a wide range of specificity: Some contain only one or two simple rules while others are much more detailed. Examples of more detailed formal documentation or procedures include those of the National Center for Health Statistics, the Energy Information Administration (for tabulations only), the Census Bureau (for microdata) and the 1977 Social Security Administration guidelines.

3.2. SDL procedures for tabulations

Most standards or guidelines provide for minimum cell sizes and some type of concentration rule. Some agencies (for example, the Economic Research Service, the National Agricultural Statistics Service, the National Center for Health Statistics, the Bureau of Mines and the Bureau of Labor Statistic publish the values of the parameters they use in (n, k) concentration rules, whereas others do not, on the grounds that outside knowledge of these values increases disclosure risks.

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 0.3 to 0.8. The Bureau of Labor Statistics uses different values of k in different programs. The Bureau of Mines uses a two-stage rule with parameter values (1, 0.75) and (2, 0.90). The most elaborate rule included in standards or guidelines was EIA's "p/q rule" (see Exhibit 4). This rule has the property of subadditivity, i.e., if two cells are nonsensitive their sum is also nonsensitive, and it allows for flexibility in specifying how much gain in information about its competitors by an individual company is acceptable. Also, it provides a somewhat more satisfying rationale for what is being done than does the arbitrary selection of parameters for an (n, k) concentration rule. As noted in connection with the Bureau of Mines response to the Panel, 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 they would be suppressed or masked under the agency's normal SDL procedures. Other agencies, including the National Agricultural Statistics Service, the Census Bureau and some of the state agencies that cooperate with the Bureau of Labor Statistics in its federal-state statistical programs, also use this type of procedure for some surveys. None of the agency documentation indicated that small respondents included in such data cells were also being asked for their permission to publish the data.

Except for the 1977 Social Security Administration guidelines and the NCHS guidelines, there is little discussion in the various standards and guidelines of the effects of zero cells on the disclosure risks associated with publication of multidimensional frequency count tables. Several of the standards and guidelines make no mention of the need to do something to prevent complementary disclosures when one or more cells are suppressed to prevent primary disclosures.

3.3. *SDL procedures for microdata*

Only about half of the agencies included in this review had something to say about SDL procedures for microdata, and some that did merely indicated 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 by the Census Bureau. Major releasers of public-use microdata – the Census Bureau, the National Center for Health Statistics and more recently the National Center for Education Statistics – have all established formal procedures for review and approval of new microdata sets. In general these procedures, unlike those used for tabulations, do not rely on parameter-driven rules. Instead, they require judgements by reviewers who 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 and the expected amount of error in the data. Geography is an important factor. The Census Bureau and the National Center for Health Statistics specify that no geographic codes for areas with a population of less than 100,000 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 percentage 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 is provided by the Energy Information Administration'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.

Topcoding 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 whose values exceed the cutoffs. There are at least two rather complex issues associated with topcoding. One is how to deal with variables that represent the sum of several components, for example, total income and income in several different categories. Another is the effect of topcoding on time series and longitudinal analyses. Because of inflation, cutoff values are likely to change over time, causing breaks in time series data. Analysts who wish to track changes in real income may have their work complicated not only by topcoding, but by the use of class intervals instead of exact values below the cutoffs.

Some of these problems can be ameliorated if agency disclosure analysts consult with knowledgeable users. The analysts can choose from a wide range of masking procedures, some of which may cause less trouble for users than others. Greenberg closed his 1991 paper by saying "In the design of a data release strategy many options are available...'. It is important that data users contribute to the planning process by contributing to the discussion of options and choices by indicating both needs and preferences."

3.4. *Hired hackers*

Only one statistical agency informed the Panel of any agency-sponsored efforts to crack its system, that is to contract with an outsider to attempt to identify individuals in its public data products. The agency representative, who notified the Panel orally of their use of this procedure, said that it was useful, implying that it helped them to spot weaknesses in their SDL procedures and to eliminate them. However, the Panel was requested not to identify the agency in its report.

4. Concluding Remarks

The purpose of the document on which this paper was based was to give the Panel on Confidentiality and Data Access information on current (at the time) statistical disclosure limitation policies and practices of federal statistical agencies. Some of the information has been updated, but the author cannot guarantee that all relevant changes up to the time of final submission for publication have been identified.

Starting late in 1991, the Statistical Policy Office of the U.S. Office of Management and Budget took the lead in organizing an Ad Hoc Committee on Disclosure Risk Analysis, with representatives from several statistical agencies. The goals of the Committee were to describe current SDL practices and policies of the statistical agencies, identify common elements and develop an agenda for research that would serve the needs of several agencies. Each participating agency was asked to provide a description of its current SDL practices and policies.

After its first meeting, the Committee formed two subcommittees, one to review SDL methodology and one to review perceptions of disclosure risks associated with the dissemination of statistical tabulations and microdata. The Subcommittee on Methodology has produced various documents, including an annotated bibliography on SDL methods and papers on SDL procedures for tabulations and microdata sets. These and other products of the Committee will probably be included in a Statistical Policy Working Paper which will be, at least in part, an update of Statistical Policy Working Paper 2 (Office of Federal Statistical Policy and Standards 1978).

The Panel on Confidentiality and Data Access has expressed its support for many of the recommendations that appeared in Statistical Policy Working Paper 2, and has developed additional recommendations on SDL policies, procedures and research. Readers who are interested in the Panel's recommendations on these topics will find them in Chapter 6 of the Panel's report.

5. References

Absalom, S.T. (1991). Bureau of Mines. Letter to V.A. de Wolf, June 26, 1991.

Barabba, V.P. and Kaplan, D.P. (1975). U.S. Census Bureau Statistical Techniques to Prevent Disclosure – The Right to Privacy vs. the Need to Know, presented at the 40th Session of the International Statistical Institute, Warsaw, 1975.

Boruch, R.F. and Kehr, W. (1983). On Use of the Doctorate Records File and the Survey of Doctorate Recipients: Privacy and Research Utility. Northwestern University, Report No. A-159-4, prepared for the National Science Foundation.

Bureau of Labor Statistics (1980). Confidential Nature of Bureau Records, Commissioner's Order No. 2-80. Washington, D.C.: Department of Labor.

Bureau of Labor Statistics (1990). Information Requested by Panel on Confidentiality and Data Access. Washington, D.C.: Department of Labor.

Center for Education Statistics (1987). Standards and Policies. Washington, D.C.: Department of Education.

Cox, L., Johnson, B., McDonald, S., Nelson, D., and Vazquez, V.(1985). Confidentiality Issues at the Census Bureau. Proceedings, Annual Research Conference Bureau of the Census, 199–218.

Dientsfry, S. (1990). Department of Veterans Affairs. Letter to V.A. de Wolf, December 4, 1990.

Economic Research Service (1989). ERS Policy on Dissemination of Statistical Information. Washington D.C.: Department of Agriculture.

Energy Information Administration (1989). Energy Information Administration Standards Manual. Publication No. DOE/EIA-0521, Washington, D.C.: Department of Energy.

General Accounting Office (1989). Obtaining Access to Information. Chapter 7.1, Project Manual (pages are dated September 1989 and January 1990).

Greenberg, B. (1990). Disclosure Avoidance Research at the Census Bureau. Proceedings, Annual Research Conference, Bureau of the Census, 144–166.

Greenberg, B. (1991). Disclosure Avoidance Practices at the Census Bureau. Statistical Policy Working Paper 20: Seminar on Quality of Federal Data, Part 3. Washington D.C.: Office of Management and Budget, 3 67–376.

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, 516–521.

Grosshans, W. (1991). General Accounting Office. Letter to V.A. de Wolf, June 21, 1991.

Landefeld, J.S. (1990). Bureau of Economic Analysis. Letter to V.A. de Wolf, September 27, 1990. Attachment D, International Investment Division Primary Suppression Rules.

National Center for Education Statistics (NCES) (1992). NCES Statistical Standards. Publication No. 92–021. Washington D.C.: Department of Education.

National Center for Health Statistics (1984). NCHS Staff Manual on Confidentiality. DHHS Publication No. (PHS) 84-1244. Washington D.C.: Department of Health and Human Services.

National Highway Traffic Safety Administration (1981). Agency Procedures for Release and Security of Research Data Collected Under the National Accident Sampling System, revised December 1981. Washington D.C.: Department of Transportation.

Office of Federal Statistical Policy and Standards (1978). Report on Statistical Disclosure and Disclosure-Avoidance Techniques. Statistical Policy Working Paper 2. Washington D.C.: Department of Commerce.

Spruill, N. (1982). Measures of Confidentiality. Proceedings of the Section on Survey Research Methods, American Statistical Association, 260–265.

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

Statistics of Income Division, Internal Revenue Service (1985). Operating Manual. Manual Issuance VI-1, revised release, January 7, 1985.

Thibodeau, K. (1991). National Archives and Records Administration. Letter to V.A. de Wolf, June 18, 1991.

Exhibit 1. National Agricultural Statistics Service, Policy and Standards Memorandum No. 12-89, Standard for Suppressing Data Due to Confidentiality, July 12, 1989

FOR ACTION BY: State Statistical Offices and Headquarters Units
REFERENCE: Policy and Standards Memorandum 7-88

Approved by: ______________________
Deputy Administrator for Programs

I. PURPOSE: This Policy and Standards Memorandum (PSM) outlines NASS policy for suppressing estimates and summary data to preserve confidentiality. This policy applies to all estimates (official and unofficial) and summary data published in releases (such as grain storage capacity, plant population, row widths, etc.) and Research and Staff reports. This policy applies even when unpublished summary data and estimates can be released as outlined in PSM 7-88.

II. SUPPRESSION DUE TO CONFIDENTIALITY: To avoid disclosure of individual operations for a given item of interest, summary data and estimates must not be published or released if either: (1) the nonzero value for the item of interest is based on information from fewer than three respondents; or (2) the data for one respondent represents more than 60 percent of the value to be published.

Exceptions to this rule are granted only when written and signed permission is given by an authorized officer of each firm (or respondent) concerned. If such permission is obtained, the written authorization must be retained in permanent files, and must be updated every five years. When the estimates or summary data are published or released by Headquarters, copies of the signed statement must be forwarded to the Branch Chief for the commodity in question through the Chairperson of the Agricultural Statistics Board. Each State is responsible for preparing the permission statement so that it corresponds to the individual respondent's situation. However, the following example may be used as a guide:

> We greatly appreciate your continued response to our broiler surveys and statistics program. As we discussed over the telephone, your dominant position as a major broiler producer in this State requires us to obtain your permission in order to publish State estimates. Even though you have given this permission over the phone, we would like for you to sign the following statement:
>
> I agree to the publication of the annual broilers production and value data in *State*
>
> Signed ______________________
>
> Title ______________________

> This does not mean we will publish your name with the estimates. This will not change our publication of monthly broiler hatchery or placement data which is combined with other States.

When data must be suppressed due to confidentiality, State offices must indicate "not for publication" and circle the item on the estimate worksheet form. When electronic submission of recommendations is used, the data entry layout (or menu driven screen) will include a data field (or prompt) for confidentiality, when necessary, as commodities are included in the Agricultural Statistics Board Data Base. States must key this field when data are to be suppressed, and they should also explain the reason for suppression in their comments.

Suppressed data may be aggregated to a higher level. However, care must be taken to ensure that the suppressed data can not 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. The suppressed data must be combined with at least one other State (Agricultural Statistics District, county, etc), so that in effect, the data are suppressed for two or more States (Agricultural Statistics Districts, counties, etc.).

Exhibit 2. Bureau of Economic Analysis, section on "Disclosure Avoidance" from Local Area Personal Income, 1983-88, Volume 1. U.S. Department of Commerce, July 1990

Disclosure avoidance

BEA's heavy reliance on the administrative record files of other government agencies makes it particularly important that BEA be aware of, and observe, the legal requirements established to safeguard the privacy of persons and firms by avoiding the disclosure of confidential information. BEA, like other statistical agencies, must balance its responsibility to avoid disclosure with its responsibility to release as much useful information as possible. This balancing has led to a policy of limiting release of estimates to the two-digit SIC level for regions, States, and local areas, although more detailed source data are available to BEA.

As described in the section on wage and salary disbursements, BEA receives county ES-202 data files from the State employment security agencies (ESA's), through the Bureau of Labor Statistics, at the four-digit SIC level.[30] These aggregations by county contain information covering one or more firms in an industry classification; the disclosure of information about any particular firm is prohibited by law. Three basic techniques for disclosure avoidance are available: Suppression, rolling up, and disturbing.[31] BEA uses a combination of all three techniques.

After completing its two-digit SIC estimates of wage and salary disbursements—based mainly on the ES-202 data—BEA examines the files to identify potential disclosures.

[30] Other examples of administrative record files used for State and local area income estimation that contain information about individuals are those from the Social Security Administration and from the Department of Veterans Affairs. These files are summarized to aggregate totals by program and county, and each county record or cell contains enough individuals to preclude the identification of any single person.

Two types of direct disclosures must be identified. The first, termed "reporting-unit disclosure," oocurs when a given cell contains fewer than a prescribed number of firms. The second, termed "dominant-firm disclosure," occurs when—regardless of the number of firms contained in the cell—a single firm accounts for some predetermined, significant percentage of the total, thus dominating the cell. For the first type, the ES-202 files that BEA receives contain reporting-unit information that permits determination of the number of firms in each cell. For the second type, cells at the four-digit SIC level containing dominant-firm disclosures are identified by the State ESA's using the individual employers' records. The ESA's also provide ES-202 tabulations stratified by size of firm for the first quarter of each year. From this information, BEA identifies the two-digit SIC cells for which a single firm might account for more than the allowable percentage of the cell wage total.

The items identified as disclosures, either by reporting-unit or dominant-firm criteria, are referred to as "primary wage and salary disclosures." To prevent direct release of this confidential information, BEA's disclosure-avoidance procedures for regions, States, and counties utilize a combination of two techniques: Systematic rolling up and dominant-cell suppression. The first is systematically to "roll up" wages and salaries, other labor income, and proprietors' income to the sum of the three—total earnings by industry and county. The second is to test the primary wage disclosure file against the total earnings file by county to see whether wages account for a predetermined significant portion of earnings such that the primary wage disclosure results in an earnings disclosure. Where earnings are not sufficiently large to mask or cover the primary wage disclosure, a suppression indicator appears on the earnings file. This combination of techniques—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—yields the final primary earnings disclosure file, which indicates the cell suppressions necessary to prevent direct disclosure of two-digit SIC information for counties.

Two additional types of cell suppressions—secondary and complementary suppressions—are necessary to prevent the derivation (indirect disclosure) or primary disclosure cells. Secondary suppressions are additional industry cells that are suppressed to prevent indirect disclosure of the primary (two-digit SIC) disclosure cells through subtraction from higher level industry totals. Complementary suppressions are 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 by testing a multidimensional matrix consisting of industry and county cells for each State and region. Computer programs impose a set of rules and priorities in order to select additional cells for suppressions until the entire multidimensional matrix of suppressions is balanced so that indirect disclosure is impossible from any direction in the matrix.

The selection process maximizes the amount of information that can be released at higher industrial and geographic levels at the expense of the more detailed industrial and

[31] Suppression is the deletion of a value and its replacement with a symbol—usually (D)—to indicate that it is being withheld. Rolling up is the combination of the cell containing sensitive information with another cell. This may be done systematically through the combination of entire sets of estimates to create a single set in which each cell contains the sum of the corresponding cells in the input sets. Disturbing is the alteration of a number enough to prevent exact disclosure but not enough to impair the usefulness of the information.

geographic information. For example, if possible, the secondary selection process will suppress additional two-digit industries rather than the higher level industry division total. Likewise, if possible, additional counties will be suppressed rather than the State totals. In some cases, discretionary decisions are superimposed on the outcome of this process to preserve regional or national totals.

A variant of the "disturbing" technique, along with dominant-cell suppression, is used to prevent disclosures stemming from the aggregations of counties to metropolitan areas. Under this approach, the metropolitan area total for each industry represents one of three situations. (1) If there are no county suppressions, the actual metropolitan area total is shown. (2) If the dollar amount of county suppressions is small relative to the sum of all the metropolitan area's counties, a partial metropolitan area total is shown, marked with an "*" indicator flag; in these situations, the amount shown constitutes the major portion of the actual total. In effect, the metropolitan area total is "disturbed" through the omission of the suppressed county amounts(s). (3) If the dollar amount of the county suppressions is large enough to impair the usefulness of the partial total, the entire amount is suppressed, and a "(D)" is shown.

Exhibit 3. Bureau of the Census, Instructions for Submitting a Proposal Requesting Approval to Release a Demographic Microdata File. Attachment to memorandum, "Public-use Microdata Review Process" from Microdata Review Panel, Jan. 14, 1988

The project manager should submit *nine* copies of each of the materials listed below to the Chairperson of the Microdata Review Panel at least 1 month prior to the time approval is needed. The chairperson will arrange a Panel meeting to discuss the proposal. The project manager usually will be requested to attend the meeting or send a representative who is knowledgeable about the proposal and the corresponding survey. The Panel's decision to approve or reject release of the file will be documented in a memorandum to the appropriate Division Chief.

Required Materials (Nine copies of each)

1. Cover memorandum from the Division Chief that includes a *brief* description of the purpose and design of the survey and any other relevant information; e.g., the date by which approval is needed.
2. Specifications that show the categories proposed for each variable or item that will be on the file. If time or costs prevent advance preparation of tape specifications for the initial *MRP* review, submit the survey questionnaire marked-up to show the items proposed for the tape with a description of how write-in entries will be coded and a list of any other information that will be on the tape (e.g., sample weights and geographic information).
3. A completed "Checklist" providing information needed by the Panel to evaluate the disclosure potential of the file. The Checklist asks about items that you are proposing to delete or change for confidentiality reasons *and* about items where you are not sure whether such treatment is necessary. One copy of the Checklist is attached; additional copies should be reproduced, as needed, by using this copy.
4. A table documenting that the population of every geographic area identified on the file

proposed for release has at least 100,000 inhabitants. This population minimum is required by the Criteria for Disclosing Public Use Microdata issued by the Census Bureau in February 1981. These criteria do not allow selective exemptions from the minimum population requirement; however, the Panel may determine that a higher population threshold is necessary to compensate for file content with greater-than-normal disclosure potential.

The Panel considers the minimum population requirement met if each area to be identified has 100,000 inhabitants in the areas subject to sampling (e.g., a PSU) as of the most recent census. Population estimates nearer the survey date may be used, if desired, unless the intended geography includes urban/rural, size of place, or other categories available only from census data. Use of population data from another source, for example a prior census, must be approved in advance by the Panel.

The table should show the total population in sampled areas (PSUs) cross-tabulated by every geographic identifier to be shown on the file (see the example in Attachment A). Geographic information, for this purpose, does not generally include information provided by the respondent; for example, farm status. The source of the population figures used in the table must be specified. If this file was previously released (or will be released again) with different geographic identifiers, the table must show both sets of geography. *Note that subsequent releases, with different geography, must take into account that if the files were to be combined, areas of less than 100,000 persons may be identified. The proposed release must eliminate this possibility.*

Each cell in the table for an identified area or the remainder should show a population of 100,000 or more; if any cell falls below 100,000, the geographic specifications must be revised to meet the requirement and reflected in the table submitted to the Panel. If the sample was a PSU-based design, the total of all cells in the table should be the total population in sampled PSUs. If the sample was not selected from within designated sample areas (e.g., a sample from a list of addresses not in PSUs), the total of all cells should be the total population of all areas subject to sampling (often the entire U.S.). Please contact a Panel representative for more information on how to prepare the table when the sample is not a PSU-based design.

Exhibit 4. Energy Information Administration, excerpts from Energy Information Administration Standards Manual, U.S. Department of Energy, April 1989

ENERGY INFORMATION ADMINISTRATION STANDARD 88-05-06

SUBJECT: NONDISCLOSURE OF COMPANY IDENTIFIABLE DATA IN AGGREGATE CELLS

Superseded Version: 84-3-04, effective 10/85

PURPOSE: To ensure the nondisclosure of confidential company identifiable data.

APPLICABILITY: To tables based on positive valued survey data where EIA has determined (1) that company specific responses may be proprietary and prohibited from public disclosure by 18 U.S.C. 1905, or (2) that responses by individuals, families, or households are entitled to protection under the Privacy Act of 1974 (Public Law 93-579). This applies to all data released to persons outside EIA, including published and unpublished data.

EXCEPTION: In publications where data historically have been published without suppression of sensitive cells, a waiver to this Standard may be obtained if a Federal Register Notice announcing that such a publication will continue does not meet with negative response. This change must also be explained on the collection form and in the forms clearance package sent to OMB.

REQUIRED ACTIONS:

1. Sensitive cells are identified in one of two ways:
   a. If the nonzero value for a cell is the reported value from one respondent or is the sum of the reported values from two respondents, the cell is considered sensitive.
   b. For all other applicable cells use the p/q rule alone or in conjunction with some other subadditive rule. (See the reference for a definition of the *p*/*q* rule and a discussion of its use).
2. If a cell is sensitive, suppress publication of that cell and apply complementary suppression to other cells to assure that the sensitive value cannot be reconstructed from published data.
3. Use the symbol "W" to denote data that have been suppressed, along with a footnote, explaining that "W" represents "Withheld".
4. Do not reveal disclosure avoidance rules that are used to protect confidentiality. However, the rules must be documented, and the documentation available within EIA.
5. Do not reveal "weights" from a sample survey.

REFERENCE:

1. Guidelines for Implementation of a Disclosure Avoidance Rule, p. 83.

Guidelines for Implementation of a Disclosure Avoidance Rule

(Standard 88-05-06)

These guidelines assist in understanding and implementing disclosure avoidance procedures using the *p*/*q* rule, as specified in EIA Standard 88-05-06. Section I presents a background introduction to disclosure avoidance, Section II describes the *p*/*q* rule, Section III presents a discussion of the impact of parameter selection and describes using a combination rule for use in special situations, Section IV provides a brief discussion of complementary suppression, Section V provides guidelines for applying disclosure when imputation for nonresponse is performed, Section VI describes disclosure avoidance when reported data are negative, Section VII describes procedures to use when an item is the difference between two positive reported items, and Section VIII describes procedures to use when an item is the weighted average of positive reported quantities. If program offices identify other situations which require special treatment, please call Nancy Kirkendall (EI-70) on 586-2276.

I. Introduction

Statistical disclosure is the release of confidential information by the cross-tabulation of microdata. This information is in the form of an attribute which may be uniquely identified with a particular responding unit. Release may be exact or approximate. For example, if a cell was published based on only two companies, each would be able to use the published total to identify the other's value exactly. Generally, when more companies contribute to a cell, the total merely allows a company to form a better estimate for the other companies' values. There are two types of disclosure possible, internal and residual. Internal disclosure occurs when members of a cell use their own data and the cell value to obtain confidential information about others in the same cell. Residual disclosure involves the mathematical manipulation of the data from other cells to obtain the confidential information in the sensitive cell.

Sensitive cells are cells which, if released, would disclose confidential information. Such cells are usually identified by the use of a sensitivity rule. Sensitivity rules determine whether or not internal disclosure would occur, and the extent of the disclosure. In EIA, the procedure for preventing disclosure is to withhold sensitive cells from release. Cells which have been identified as being sensitive by the use of a sensitivity rule and are suppressed for that reason are called primary suppressions. If totals are to be released, and cells requiring primary suppression exist, then certain other cells must also be suppressed to prevent residual disclosure. This type of cell suppression is called complementary suppression.

The appropriate application of disclosure avoidance requires identification of cells which require primary suppression, and additionally, identification of a set of appropriate cells for complementary suppression.

II. The Recommended Primary Rule (the p/q rule)

The primary suppression rule which is recommended for use in EIA is the p/q rule, one of the so called "priori-posteriori ambiguity rules." It was selected because it satisfies the required property of subadditivity, and provides protection among regions. Subadditivity guarantees that if two cells are both nonsensitive, their sum is also nonsensitive. This sensitivity measure is applicable only when publishing totals of nonnegative reported values, and is necessary when the reported values are volumetric data such as production, stocks, or sales volume. Volumetric data have the greatest potential for disclosure because they can show the greatest diversity among the reported values of a single company, or set of companies.

The sensitivity measure for the p/q rule is given by

$$S(x) = x_1 - (p/q) * (T - x_1 - x_2)$$

where the cell is sensitive if $S(x)$ is nonnegative, and

x_1 is the largest reported value in the cell

x_2 is the second largest reported value in a cell (x_1 can be equal to x_2)

T is the total to be released. In a census survey it is the sum of the reported values, in a sample survey it is the weighted sum of the reported values.

p/q is the input sensitivity parameter. It represents the maximum permissible gain in information when one company uses the published total and its own value to create better estimates for its competitors' values. p/q must be greater than one, but need not be an integer. The particular choice of p/q is to be made by the program office. (See Guidelines in Section III.) The particular value selected and applied should be considered as confidential.

III. Parameter Selection Criteria and Uses of a Combination Rule

Although the p/q rule does not exactly correspond to any of the rules based on the contribution of the largest company or the sum of the largest two companies, it is possible to calculate the percent contributions of x_1 and $x_1 + x_2$ for which suppression is guaranteed by different values of p/q. We use the following expression for the region where $S(x)$ is positive (the region of sensitivity).

$S(x)$ will be positive and disclosure will occur for any x_1 and x_2 such that:

$$(x_1 + x_2) * p/q + x_1 > T * p/q.$$

Since $(x_1 + x_2) > x_1$ for all x_2, we know that if

$$x_1 * p/q + x_1 > T * p/q$$

then $(x_1 + x_2) * p/q + x_1 > T * p/q$, and the cell is sensitive. Thus, if the largest cell contributes a larger proportion to the cell total than $(p/q)/(1 + p/q)$, the cell is sensitive.

Similarly, the smallest value of $(x_1 + x_2)$ which guarantees disclosure, regardless of the contribution of the other companies, is found at the intersection of the disclosure boundary with the line $x_1 = x_2$. Thus, substituting $x_2 = x_1$ into the inequality and solving for x_1 yields $x_1/T > (p/q)/(1 + 2 * p/q)$. Rewriting this in terms of $(x_1 + x_2)$ gives:

$$(x_1 + x_2)/T = 2 * x_1/T > 2 * (p/q)/(1 + 2 * (p/q)).$$

Thus, if $(x_1 + x_2)$ contributes a larger fraction to the total than $2 * (p/q)/(1 + 2 * p/q)$, then the p/q rule will determine that the cell is sensitive.

The following table summarizes these results for selected values of p/q.

Cell suppressed by p/q rule if x_1, or $(x_1 + x_2)$ exceeds the following percentage of the total:

p/q	x_1 exceeds:	$x_1 + x_2$ exceeds:
2	66.7%	80.0%
3	75.0%	85.7%
6	85.7%	92.0%
9	90.0%	94.7%

Note that the larger values of p/q permit release of cells where x_1 or $x_1 + x_2$ represent larger percentages of the total. That is, larger values of p/q permit more information to be gained by releasing the total, T. As an example, if x_1 accounted for 85 percent of the total, x_2 accounted for 5 percent of the total, and $p/q = 9$, then $S(x) = -.05 * T$ and the cell would not be classified as being sensitive.

If program offices prefer to use a large value of p/q, the p/q rule alone may permit too much disclosure when the cell is dominated by one large company, as in the above example. However, a combination rule may be drafted to overcome this problem. A combination rule is defined to be the maximum of the sensitivity measure defined for the p/q rule and the sensitivity measure defined for some other subadditive rule. An example of another subadditive rule is the $n - k$ rule. In the $n - k$ rule, the cell is sensitive if the largest n companies account for k percent or more of the total. If $n = 1$ and $k = 80$, for example, the sensitivity measure would be:

$$S_1(x) = x_1 - .8 * T$$

and, as before, the cell is sensitive if $S_1(x)$ is nonnegative.

The combination rule is subadditive since the maximum of two subadditive sensitive measures is subadditive. In this example the sensitivity measure would be:

$$S* \ (x) = \max \ (S(x), S_1(x)).$$

In the above numerical example, $S(x) = -.05 * T$, and $S_1(x) = .05 * T$. Thus, the maximum is $.05 * T$, and the cell is sensitive.

IV. Complementary Suppression

Determining the optimal pattern of cells for complementary suppression is a complicated procedure because it potentially requires a search over all possible cells to select the fewest number of cells, with the smallest possible total value, which adequately protect the cells requiring primary suppression. The Office of Statistical Standards has available a computer program obtained from Statistics Canada which performs complementary suppression for both two and three way tables. The program will be made available to the program offices.

Program offices may choose to perform complementary suppression manually, using their industry knowledge, and maintaining the same or similar patterns of suppression from one release to the next.

Implementation of a disclosure avoidance procedure must be augmented by an audit feature which permits an evaluation of proposed patterns of suppression to assure that there is no residual disclosure.

V. Treatment of Imputation for Nonresponse to Disclosure Analysis

There are two general causes for the application of disclosure analysis when imputation procedures are used to adjust for nonresponse:

1. imputed values are based on the other respondents' data, as in adjusting weights or "hot decking.";
2. imputed values are based on data submitted by the nonresponding company in a previous time period.

In the first case, there can be no disclosure of values associated with the nonresponding companies. Only the reporting companies are at risk. To assess this risk, the imputed values are included in the total, T, but the imputed values are not counted as reported values for identification of the largest two companies. (Recall that in applying the sensitivity measure to sample surveys, T is calculated based on the weighted total, and x_1 and x_2 are defined based on actual unweighted survey responses.)

In the second case, the theoretical justification for the imputation procedure is that there is a high correlation between values reported by the same company in different time periods. Hence this type of imputed value contains sensitive data for that nonresponse. Thus, the imputed data should be treated as reported data for purposes of disclosure analysis. That is, the imputed values should be included in the total T, and should be considered as reported values when selecting the largest two values in the cell.

VI. Negative Reported Values

If all reported values are negative, apply the p/q rule to the absolute values of the reported data. Complementary suppression must be applied if marginal totals are released.

VII. Differences of Positive Reported Values

If the published item is the difference between two positive reported quantities (e.g., net production equals gross production minus inputs), then apply the p/q rule as follows:

> If the resultant difference is generally positive, as is the case for net production of distillate fuel oil, apply the p/q rule to the first item (gross production in the above example). Suppress cells for which the first item is sensitive.
>
> If the resultant difference is generally negative, as is the case for net production of an item which is used primarily as inputs, apply the p/q rule to the second item (inputs in this example). Suppress cells for which the second item is sensitive.
>
> If the difference can be positive or negative, and is not dominated by either, simply make sure that there are at least three nonzero respondents contributing to each cell.

Complementary suppression must be applied if marginal totals are released.

VIII. Weighted Averages

If a released item is the weighted average of two positive reported quantities (such as volume weighted price), apply the p/q rule to the weighting variable (volume in this example).

Suppress the average cell if the weighting variable is sensitive. It is not necessary to use complementary suppression on the average variable.

Both primary and complementary suppression must be applied to the weighting variable, if it is also a released item.

Exhibit 5. National Center for Health Statistics, excerpt from NCHS Staff Manual on Confidentiality, U.S. Department of Health and Human Services, Sept. 1984

10. Avoiding Inadvertent Disclosures in Published Data

10.1 *Problem.* In their zeal to make available to the public a full set of information on a given subject, statisticians may—and sometimes do—present so much detail in published tabulations that they accidentally reveal confidential information about particular study subjects. This may happen in several ways. For example,

A. One line y_i of a cross-tabulation contains a total of two individuals. On reading the table an individual with the y_i characteristic now knows the x characteristic of the other individual in the population having the y_i characteristic.

B. All cases in line y_i of a statistical table fall in the cell in column x_i. We then know that any individual in the population with characteristic y_i also has characteristic x_i.

C. Cell x_iy_i gives the total income of all individuals with characteristics x_i and y_i. If there are only two individuals, *a* and *b*, in the population with that combination of characteristics, then *a*, knowing his own income, will be able to determine *b*'s income by simple subtraction, and *b* will also be able to determine *a*'s income.

D. A table gives the total annual receipts for all five nursing homes in county *m*. However, nursing home *a* is much larger than all the rest combined; it accounts, in fact, for three-fourths of all nursing home receipts in the county. Knowing the county total, the manager of nursing home *a* is able to calculate the incomes of the other four homes, at least within some fairly narrow limits.

E. A Standard Metropolitan Statistical Area (SMSA) contains two counties, *a* and *b*. Four hospitals are located in county *a* and only one in county *b*. A statistical report is published, giving confidential hospital data totaled for each SMSA. Another report is published with confidential data on hospitals by county, but only for counties with three or more hospitals. Using the two reports one can subtract the data for county *a* from the SMSA data, deriving the confidential data for the lone hospital in county *b*.

F. The maximum Social Security benefit for an individual retired person is, say $235 per month. A published table shows that white males aged 70 to 74 in county *a* receive an average benefit of $235 per month. It is now known that *every* white male aged 70 to 74 in county *a* who receives a Social Security payment receives $235 a month.

These examples imply the existence of several general types of situations in which statistical disclosure may occur. An additional possibility may be found in a group of three or more tables of subsets of a given population from which disclosures are possible through the solution of simultaneous equations. Center guidelines as set forth in Section 10.3 take into account the several possible disclosure situations.

10.2. *Types of Disclosure.* Center policy recognizes and attempts to deal with several classes of disclosure:

A. *Exact versus approximate disclosures.* Exact disclosure is the disclosure of a specific characteristic, such as race, sex, or a particular pathological condition. Approximate disclosure is the disclosure that a subject has a characteristic that falls within a certain range of possibilities, such as being between 45 and 55 years of age or having an income between $15,000 and $25,000. An approximate disclosure *may* in a given situation be considered harmless because of its indefinite nature.

B. *Probability-based versus certainty disclosures.* Data in a table may indicate that members of a given population segment have an 80-percent chance of having a certain characteristic; this would be a probability-based disclosure as opposed to a certainty disclosure of information on given individuals. In a sense, every published table containing data or estimates of descriptors of a specific population group provides probability-based disclosures on members of that group, and only in unusual circumstances could any such disclosure be considered unacceptable. It is possible that a situation could arise in which data intended for publication would reveal that a highly specific group had an extremely high probability of having a given sensitive characteristic; in such a case the probability-based disclosure perhaps should not be published.

C. *Internal versus external disclosures.* Internal disclosures are those that result completely from data published from one particular study. External disclosures occur when outside information is brought to bear upon the study data to create disclosures. This possibility must be recognized in any disclosure analysis.

10.3. *Special Guidelines for Avoiding Disclosure.* Except where otherwise indicated, the following guidelines apply to all Center publications of statistics:

A. In no table should all cases of any line or column be found in a single cell.

B. In no case should the total figure for a line or column of a cross-tabulation be less than 3.

C. In no case should a quantity figure be based upon fewer than three cases.

D. In no case should a quantity figure be published if one case contributes more than 60 percent of the amount.

E. In no case should data on an identifiable case, nor any of the kinds of data listed in preceding items A-D, be derivable through subtraction or other calculation from the combination of tables published on a given study.

F. Data published by NCHS should never permit disclosure when used in combination with other known data.

Report writers and editors in the Center are to follow these guidelines. If a guideline appears unreasonable in a given situation, approval for a special exception to the guideline should be requested from the Director or the Deputy Director. The following types of cases represent exceptions to the above guidelines which do not require special approval from the Director or Deputy Director:

A. It has been a longstanding tradition in the field of vital statistics not to suppress small frequency cells in the tabulation and presentation of data. For example, it has been considered important to know that there were two deaths from rabies in Rio Arriba County, N. Mex., in a given year, or that there were only one infant death and two fetal deaths in Aitkin County, Minn. These types of exceptions to general NCHS practices in other programs are followed because they have been accepted traditionally and because they rarely, if ever, reveal any information about individuals that is not known socially.

B. Tables may show simple *counts* of number of persons, even though the number in a cell is only "1" or "2," provided the classifying data are not judged to be sensitive in the context of the table. For example, publication of counts of health manpower personnel by occupation by area is considered acceptable, if not accompanied by other distinguishing characteristics, or other cross-classifications that have the effect of adding descriptive information about the same persons. However, publication of counts of personnel for a specified occupation by area by income is not acceptable for cells of less than three persons because that would reveal sensitive income data.

10.4. *Evaluating a Disclosure Problem.* There may be mitigating circumstances in a given situation which may make it acceptable to publish data that, strictly speaking, could result in "disclosures." Such circumstances could provide grounds for requesting the "special exception" to the previously noted rules:

A. When data in a study are based upon a small-fraction sample, for example, less than 10 percent of the universe, it might generally be assumed that disclosure will not occur through published tabulations. However, there could be exceptions. So much detail may be presented that an individual unique in the population is identified through the tables, or a member of the sample may find himself and others in the data. The usual rules precluding publication of sample estimates that do not have a reasonably small relative sampling error should prevent any disclosures from occurring in tabulations from sample data.

B. The existence of errors or imputations in the data brings some small reduction in the likelihood of disclosure through table publication.

C. Incompleteness of reporting, which often occurs even where studies are supposed to include 100 percent of a given group in the population, also reduces the certainty of any disclosure taking place through publication of data.

D. In some instances the danger of disclosure might be mitigated by the fact that the data in question have no sensitivity. They may already have appeared in a published directory, or they may involve entirely obvious characteristics, or they may relate to an earlier time. Since that time, many changes have occurred, so that the data have become completely innocuous.

10.5. *Measures To Avoid Disclosure.* Two methods customarily used in the Center to prevent disclosures from taking place through tabulations:

A. The table is reduced in size when rows or columns are combined into larger categories, eliminating the particular cells that would otherwise produce disclosures.

B. Unacceptable data in cells are suppressed. When this is done, it is necessary also to suppress other cells in the table to prevent determination of the unacceptable cell figure through subtraction. It is usually necessary to suppress four cells in a cross-tabulation in order to avoid disclosure through one cell—the offending cell (x_iy_i), another cell in the same row (x_jy_i), another cell in the same column as the offending cell (x_iy_j), and also the cell (x_jy_j) at the intersect of the additional row and column involved in the newly suppressed cells.

11. Avoiding Inadvertent Disclosures Through Release of Microdata Tapes

It is Center policy to make its files on individual elementary data units available at cost to the scientific community so that additional analyses can be made of these data for the country's benefit. The scientific community has shown great interest in such tapes, and many requests for the Center's tapes are received each year. Except when the file contains no confidential information, these "public use microdata tapes" are released without any identifiers, such as name or address, of the reporting units.

11.1. *Problem.* Even though all personal identifiers are removed from cases in a microdata file, a few items of information which appear as variables on the tape may identify data subjects to any person who has access to the information from another source. Thus, for example, if tape descriptors indicate that the data subject is an attorney who graduated from the University of Maryland but show nothing else about his personal characteristics, the subject is not identified. If the tape goes on to indicate that the attorney graduated in 1964, has a wife who graduated from Radcliffe in 1966, has three children, and lives in Fairfax County, Va., the subject is now probably identified uniquely, and all information in the file about the subject would be disclosed to anyone with access to the file who could then identify the person from the given set of characteristics. The place of residence, especially when it is not a heavily populated area, is particularly useful in the identification process.

The low-ratio sample that the Center uses in its surveys would usually frustrate a person who is trying to locate a known individual in the Center's survey files. Thus, if a survey involves a 1 : 1,000 sample, the investigator would have only one chance in a thousand of finding in the file a particular individual whose data he is searching. However, if the investigator goes on a "fishing expedition" to find "anyone" in a file who might be identified, chances are much better.

11.2. *Mitigating Circumstances.* The only absolutely sure way to avoid disclosure through microdata tapes is to refrain completely from releasing any microdata tapes, but this would deprive the Nation of a great deal of very important health research. Therefore the Center must make a determination as to when the public's need is sufficiently great to justify the risk of disclosure. It is the Center's policy to release microdata tapes for purposes of statistical research only when the risk of disclosure is judged to be extremely low. Some factors bearing upon the acceptability of this risk are the following:

A. As noted, when a survey involves only a proportionately very small sample of establishments or individuals, there is small chance that it will identify a given case of interest. (This assumes, of course, that an investigator would not have a means of determining what cases fell into the sample.) Identification of an individual case would require a great deal of outside information not likely to be found outside the survey itself. However, if the sample is stratified and cases in certain strata are selected with high probability, there is little or no advantage in reducing risk of disclosure through sampling as far as cases in those strata are concerned.

B. Identifying individuals in the microdata file would usually be an expensive undertaking, hardly justified by the kind of information in the file. Public-use microdata files, in fact, should not contain any information that would likely harm or embarrass the individual or establishment if it should happen to be disclosed.

11.3. *Rules*. The following rules apply to all microdata tapes released by NCHS which contain any information about individual persons or establishments, except where the supplier of information was told, prior to his giving the information, that the information would be made public:

A. The tape must not contain any detailed information about the subject that could facilitate identification and that is not essential for research purposes (e.g., exact date of the subject's birth).

B. Geographic places that have fewer than 100,000 people are not to be identified on the tape.

C. Characteristics of an area are not to appear on the tape if they would uniquely identify an area of less than 100,000 people.

D. Information on the drawing of the sample which might assist in identifying a data subject must not be released outside the Center. Thus, the identities of primary sampling units are not to be made available outside the Center.

E. Before any new or revised microdata tapes are published, they, together with their full documentation, must be approved for publication by the Director or Deputy Director.

F. A microdata tape containing confidential data on unidentified individuals or facilities may not be released to any person or organization outside the NCHS until that person, or a responsible representative of that organization, has first signed the statement on the Order Form, whereby he gives assurance that the data provided will be used only for statistical reporting or research purposes.

If it appears in any particular instance that the strict application of one or more of these rules is inappropriate, a request should be submitted to the Director or Deputy Director to allow an exception.

Some organizations have the policy of introducing random errors into their public-use microdata tapes in order to reduce the probability of disclosure. This practice is undesirable from the standpoint that it inevitably lessens the value of the tape for making sensitive statistical analyses. Center staff are encouraged to study the feasibility and advisability of taking such steps in order to reduce further the risk of disclosures through use of the Center's public-use tapes.

Exhibit 6. Bureau of Mines, excerpt from Standard for Handling Proprietary Survey Data, No. 1-85, U.S. Department of the Interior, August 22, 1985

VI. Disclosure Analysis

A. Mandatory Steps

1. There are two mandatory steps required in performing disclosure analysis. These steps will ensure protection of company proprietary data, in most situations. In practice, these steps are to be strictly followed and the results reviewed by the responsible commodity specialists to ensure proper protection of data which must be concealed.

2. Step One – *"Rule of Three"* – In order to publish *totals*, the following rules must be applied:
   a. If there are *one* or *two* companies in the frame, a "*yes*" response to the proprietary data question must be given by each reporting company for the surveys employing this question. The *total* of one or two companies cannot be published for other surveys not having the question on the form.
   b. If there are at least three companies surveyed contributing to the total under consideration (e.g., national, State, county) and their data (1) are not released by the company and (2) are *not available elsewhere*, the total meets the "Rule of Three" for publication.
3. Step Two – *Dominant Company Principle*
   a. In addition to meeting the Rule of Three, (Step One) the total *must* meet the "Dominant Company Principle" which requires that no *one* company contribute more than 75% of the total and that no *two* companies contribute more than 90% of the total in order to publish the total.
4. Judgement Review
   a. There are situations, however, in which the rules stated above may not be appropriate. It is not practical to attempt to delineate all possible cases and prescribe methods for handling each. For example, when there are less than three companies and the mines or plants of a multiplant company give mixed answers to the confidentiality (proprietary data) question, judgement beyond these rules is required to determine if the total can be published. Judgement *will* be necessary in these cases. This judgement can best be employed by the responsible commodity specialist.
   b. A "judgement review" *in writing* should be requested of the commodity specialist responsible for the commodity(ies) by the Proprietary Information Custodian when there are questions regarding the release of proprietary information.

Received September 1992
Revised March 1993