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 Estimates will not be published from sample surveys unless: (1) sufficient nonzero reports are received for the items in a given class or data cell to provide statistically valid results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) 3+	Threshold Rule 3+	No	Yes
NASS	(n,k), (1,.6) 3+	Threshold Rule 3+	No	Yes
BEA	p-percent c=1	1+ Not Sensitive for Est. Surveys	No	No
BOC	(n,k), p-percent (Ag Census), Parameters Confidential	1+ (Economic Census), Confidentiality Edit (Demographic Census), Accuracy Requirements (Demographic Surveys)	Yes -- Microdata Review Panel	Yes
NCES	3+ Accuracy Standards	3+ Accuracy Standards	Yes -- Disclosure Review Board "Protected" Data File	No
EIA	pq, Parameters Confidential	Accuracy Requirements	Yes -- Agency Review	Yes
NCHS	(n,k), (1,.6)	3+	Yes -- Review by Director or Deputy	No
SSA	3+	Threshold Rule 5+ Marginals 3+ Cells	Yes -- Agency Review	No
BJS	N/A	10+, Accuracy Requirements	Yes -- Legislatively Controlled, Agency Review	No
BLS	(n,k) Parameters Vary by Survey	Minimum Number Varies by Data Collection	BOC Collects Title 13	Yes
IRS	3+	3+	Yes -- Legislatively Controlled	No
EPA	Minimum Number Varies by Data Collection	Minimum Number Varies by Data Collection	Yes -- 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)