CHAPTER VI

Recommended Practices

A. Introduction

Based on its review of current agency practices and relevant research, the Subcommittee developed a set of recommendations for disclosure limitation practices. The Subcommittee believes that implementation of these practices by federal agencies will result in an overall increase in disclosure protection and will improve the understanding and ease of use of federal disclosure-limited data products.

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

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

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

Lastly, Recommendation 12 in Section B.4 pertains to microdata. Many decisions concerning the disclosure limitation methods used for microdata are based solely on precedents and judgment calls. The only recommendation presented here is to remove all directly identifying information, such as name, and limit the amount of information that is reported from other identifying variables. There is clearly a need for further research. Hence, Chapter VII "Research Agenda," focuses on microdata.

B. Recommendations

B.1. General Recommendations for Tables and Microdata

Recommendation 1: Seek Advice from Respondents and Data Users. In order to plan and evaluate disclosure limitation policies and procedures, agencies should consult with both respondents and data users. Agencies should seek a better understanding how respondents feel about disclosure and related issues. For example, whether they consider some data items more sensitive than others. The research done by Eleanor Singer for the Census Bureau provides one model (see Singer and Miller, 1993).

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

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

Because microdata represent a relatively greater disclosure risk, the Subcommittee recommends that agencies that release microdata and tables first focus their attention on the review of microdata releases. In agencies with small or single programs for microdata release, this may be assigned to a single individual knowledgeable in statistical disclosure limitation methods and agency confidentiality policy. In agencies with multiple or large programs, a review panel should be formed with responsibility to review each microdata file proposed for release and determine whether it is suitable for release from a statistical disclosure limitation point of view. Review panels should be as broadly representative of agency programs as is practicable, should be knowledgeable about disclosure limitation methods for microdata, should be prepared to recommend and facilitate the use of disclosure limitation strategies by program managers, and should be empowered to verify that disclosure limitation techniques have been properly applied.

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

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

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

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

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

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

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

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

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

Recommendation 5: Use Consistent Practices. Agencies should strive to limit the number of disclosure limitation practices they use, and to employ disclosure limitation methods in standard ways. Insofar as possible, agencies should be consistent in defining categories in different data products and over time. Such practices will make disclosure-limited tables and microdata more user-friendly. Examples include using consistent schemes for combining categories, establishing standardized practices for similar data such as categorizing or top-coding items like age or income, and moving towards standardized application of geographic size limitations. Software should be developed, made broadly available and used to implement these methods to assure both consistency and correct implementation.

B.2. Tables of Frequency Count Data

Recommendation 6: Research is Needed to Compare Methods. There has been considerable research into disclosure limitation methods for tables of frequency data. The most commonly used method at present is suppression. Besides suppression, other well-developed methods include controlled rounding and the confidentiality edit. The Subcommittee was unable to recommend one preferred method. Instead, we recommend that a research project be undertaken to compare these three methods in terms of data protection and usefulness of the data product. (Further discussion of this recommendation can be found in Chapter VII.)

If suppression is used, the guidelines listed in Recommendations 9 and 10 also apply to tables of frequency data.

B.3. Tables of Magnitude Data

Recommendation 7: Use Only Subadditive Disclosure Rules. Disclosure occurs in statistical tables when published cell values divulge or permit narrow estimation of confidential data. For example, a count of 1 or 2 in a frequency count table of race by income may divulge the income category of one respondent to another respondent or data user. Or, a cell representing total sales within an industry for a particular county may allow narrow estimation of the sales of a single company. Such cells are called **primary disclosure cells** and must be subjected to disclosure limitation.

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

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

Recommendation 9: Do Not Reveal Suppression Parameters. To facilitate releasing as much information as possible at acceptable levels of disclosure risk, agencies are encouraged to make public the kind of rule they are using (e.g. a p-percent rule) but they should not make public the specific value(s) of the disclosure limitation rule (e.g., the precise value of "p" in the p-percent rule) since such knowledge can reduce disclosure protection. (See Section IV.B.3 for an illustration of how knowledge of both the rule and the parameter value can enable the user to infer the value of the suppressed cell.) The value of the parameters used for statistical disclosure limitation can depend on programmatic considerations such as the sensitivity of the data to be released.

Recommendation 10: Use Cell Suppression or Combine Rows and/or Columns. There are two methods of limiting disclosure in tables of magnitude data. For single tables or sets of tables that are not related hierarchically, agencies may limit disclosure by combining rows and/or columns. Agencies should verify that the cells in the resulting table do not fail the primary suppression rule. For more complicated tables, cell suppression should be used to limit disclosure. Cell suppression removes from publication (suppresses) all cells that represent disclosure, together with other, nondisclosure cells that could be used to recalculate or narrowly estimate the primary, sensitive disclosure cells. Zero cells are often easily identified and should not be used as complementary suppressions. Suppression methods should provide protection with minimum data loss as measured by an appropriate criterion, for example minimum number of suppressed cells or minimum total value suppressed. These recommended practices also apply if suppression is used for tables of frequency count data.

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

B.4. Microdata Files

Recommendation 12: Remove Direct Identifiers and Limit Other Identifying Information. The challenge of applying statistical disclosure methods to microdata is to thwart identification of a respondent from data appearing on a record while allowing release of the maximum amount of data. The first step to protect the respondent's confidentiality is to remove from the microdata all **directly identifying information** such as name, social security number, exact address, or date of birth. Certain univariate information such as occupation or precise geographic location can also be identifying. Other univariate information such as a very high income or presence of a rare disease can serve both to identify a respondent and disclose confidential data. Circumstances can vary widely between agencies or between microdata files.

Agencies should identify univariate data that tend to facilitate identification or represent disclosure, and set limits on how this information is reported. For example, the Census Bureau presents geographic information only for areas of 100,000 or more persons. Income and other information may be top-coded to a predetermined value such as the 99th percentile of the distribution. Lastly, appropriate distributions and cross tabulations should be examined to ensure that individuals are not directly identified.

Sometimes the methods used to reduce the risk of disclosure make the data unsuitable for statistical analysis (for example, as mentioned in Chapter V, recoding can cause problems for users of time series data when top-codes are changed from one period to the next). In deciding what statistical procedures to use, agencies also need to consider the usefulness of the resulting data product for data users.

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