Should We Continue To Release Public-Use Microdata?[1]

Cynthia Z. F. Clark, Bureau of the Census

In its information processes, the Bureau of the Census seeks to meet two sometimes contradictory objectives:

(1) to facilitate the broadest use of data that it collects and publishes--as a public good--and

(2) to protect the confidentiality of collected household and establishment data and to ensure the privacy of individual responses (as required by Title 13).

The Census Bureau assures its respondents that the business and personal information it collects will be restricted to statistical uses and analyses, i.e., to provide information for government and societal decision making. This is the sole justification for the agency's request for information. Maintaining the Census Bureau's "culture of confidentiality" is a responsibility felt critically at all levels of the Bureau. Tension arises in fulfilling the two objectives of the Census Bureau. A critical balance is required between protection and release strategies.

■ The Historical Census Bureau Strategy for Public Data Release

The Census Bureau initiated the production of public use microdata files[2] in the mid-1960s. The Bureau provides public-use files for demographic and decennial data but not for economic data. To minimize the risk of disclosure in these public-use files, the Bureau employs a variety of strategies such as: limiting the amount of geographic information, including only a sample of the records from the data collection; top coding of sensitive variables such as income; changing continuous variables to categorical ones; introducing additive noise; and swapping of data items.

The Bureau also releases tabular products. The methods used to protect the confidentiality of data released in tabular format are linked to those used in the production of public-use microdata. This linkage will become more important as we move to the future. Examples of disclosure limitation methods follow. For more information about these techniques, see Federal Committee on Statistical Methodology (1994).

(1) In tabulation from economic censuses, the Bureau uses an automated disclosure system that identifies and suppresses sensitive cells and selects cells for secondary suppression.

(2) Prior to releasing tables from the decennial census, the Bureau has utilized the technique of 'data swapping' for household records and the 'blank and impute' disclosure technique for some values in order to increase the protection accorded to unique records.

(3) In demographic surveys, where a sample of individuals are surveyed, the weighting of the sample usually provides adequate protection.

The Bureau has established several vehicles permitting qualified researchers to use confidential data under controlled conditions that restrict access. Researchers are granted Special Sworn Status to come to the Census Bureau to carry out specific projects that result in benefits both to the researchers' fields and to the Census Bureau's data programs. This status requires researchers to work under the same restrictions as Census Bureau employees. In 1978, the ASA/NSF/Census Bureau Research Fellows program was begun. For more information on this program, see the Census Bureau's web site, http://www.census.gov; under "Subjects A to Z," go to "Fellowship Opportunities."

In the early 1980s, some of the Fellows were given access to economic (establishment and firm) data at the newly established Center for Economic Studies (CES). Demand for access to these data had always existed but could not previously be satisfied in a systematic way (McGuckin and Pascoe, 1988). For these data, it has

been almost impossible to create public-use files without either risking disclosure of identifiable data or destroying the analytic usefulness of the data. This flows from the extremely skewed distributions of these data, coupled with publicly accessible information about the largest companies. Access to the data by ASA/NSF/Census Bureau Research Fellows and other researchers contributed greatly to the value of the information for public policy use. Today, researchers at CES working on approved research projects may be given restricted access to certain demographic microdata sets at secure facilities--with appropriate safeguards to protect data confidentiality.

In 1994, the CES restricted access program expanded outside the Washington, D.C. area with the opening of the Boston Research Data Center (RDC) in partnership with the National Bureau of Economic Research and the NSF. In 1997, the Carnegie Mellon Census Research Data Center was opened in Pittsburgh in partnership with the Heinz School of Public Policy and Management. This year, in partnership with the NSF, the Census Bureau began a limited expansion of the program. With the University of California (UC), the Bureau is establishing a new California Census Research Data Center. This data center will have two offices--set to open in late 1998 or early 1999--at UC Berkeley and at UC Los Angeles. In addition, non-public-use versions of demographic data files have been made available at CES and the RDC's. For more information on CES and the RDC programs, see Reznek, Cooper, and Jensen (1997); Cooper, Reznek, Merrell, and Nucci (1998); also the CES web pages at http://www.census.gov/ces/ces.html.

■ Impact of the Current Societal Environment on Disclosure Risks

Today, the situation is much more complex than it was twenty years ago. It has become harder to protect the confidentiality of data. The current societal environment has characteristics that increase the opportunity for re-identification of confidential survey data. Some of the factors that increase the risk of re-identification include:

(1) easy access to public data bases extant in electronic form, thus increasing one's ability to link various sources of data. For example, data that previously existed only in paper format for a limited array of uses are now being used for a multitude of things that were not previously considered;

(2) increased use of the Social Security Number in society;

(3) increased sensitivity to the availability of data;

(4) enhanced computer capabilities that increase the probability and possibility of linking files;

(5) more sophisticated record linkage software--another factor that increases the probability of linking files;

(6) establishment of data warehouses;

(7) use of data mining to profile records and discover patterns in data bases;

(8) existence of advanced, automated look-up services and search capabilities;

(9) advances in the use of geographic information systems in conjunction with other data bases, facilitating re-identification;

(10) expansion of the Internet with its ever-broadening availability of data bases; and

(11) changes in the characteristics of those who are potential users of public-use microdata files. In the past, academic or government researchers were the users; now, a broader spectrum of Internet users capable of performing data analysis can access census information.

■ Risk Mitigating Characteristics of Microdata Files

The long list above is tempered by a set of mitigating factors that lessen the risk of re-identification. First, the non-Census Bureau data bases often do not have the same time frame as the data items from the Census Bu-

reau surveys. The data ages fairly quickly--people move relatively frequently, and time causes changes in variables. Second, most data bases and survey data contain noise. This makes matching difficult. Third, if you really wanted to know about an individual, there are easier ways to obtain such information than through exploration of a microdata file released by a statistical agency.

■ New Approaches to Data Access and Release

The Census Bureau recognizes the impact that the factors described above have had on attempts to protect the privacy of individual data. In addition, the Census Bureau acknowledges that different users have different needs. It is developing a multifaceted approach to accommodate such differences. The Bureau is seeking to expand its portfolio to include the following options:

Data Access and Dissemination System

The Data Access and Dissemination System (DADS) will release tables and summary statistics. It will be tailored for various levels of users. Tier 1 will release predefined tables for decennial, demographic, and economic census data. Tier 2 will allow users to customize tables from the decennial census and American Community Survey public-use files. These tables will be subject to current disclosure limitation techniques for predefined geographic areas having a population of 100,000 or more. In Tier 3, DADS will allow users to request customized tables from the entire decennial data base. Tier 3 is still being developed (Hoy and LaPlant, 1998).

Luxembourg Income Study (LIS) Model

The LIS project allows users to submit, remotely, a request for statistical analysis. The Census Bureau is considering such an approach for the future whereunder data users would submit a statistical computer program after de-bugging it by using a "dummy" data file. The analysis is done behind a firewall. A Census Bureau staff member would review the output before it is returned to the requester. Presently, research is being conducted to see how such a model could work in an automated environment independent of human review.

Other approaches include the use of noise in microdata files or creating synthetic data with the characteristics of survey data. The success of either approach hinges on the degree of customer satisfaction with these types of data product. As mentioned earlier, an expansion of the Research Data Centers is planned that will provide more opportunities for users accorded protected access. These procedures are illustrative of the two approaches used to protect data confidentiality: (1) restricted data options and (2) restricted access options (Jabine, 1993). By looking at a broad array of options, we can better serve the public and our growing set of data users.

■ Continuing Research Program Assessing Confidentiality Risks Associated with (1) Disclosure Protection Procedures and (2) Data Access Procedures

The Bureau is investigating these new approaches in the context of how well they protect the confidentiality of the data. The re-identification risks associated with each alternative have not been quantified. The Census Bureau is conducting research in re-identification, following the lead of researchers in Europe and at the National Center for Education Statistics. The Census Bureau plans to examine new tools, such as the idea of "local suppression" used by Statistics Netherlands, whereunder unique microdata records are eliminated by suppressing one or more values in each of them (Willenborg and de Waal, 1996, p. 77). In addition, the Census Bureau will continue to examine different approaches in an ongoing research program directed toward evaluation of current and emerging statistical methods to limit disclosure.

Data collected by statistical agencies are used in a multitude of ways. Statistical data are a public good. The keepers of the data must respond to public needs, ensuring that data are available for societal decisions. Data users can let the Bureau know what defects they see in reported information--flowing either from disclosure avoidance procedures or data quality issues. This dialogue will lead to overall improvements in the quality of the Census Bureau data series.

■ Footnotes

[1] This paper reports the results and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

[2] Public-use microdata files contain nonidentifiable person or household records.

■ References

Federal Committee on Statistical Methodology (FCSM) (1994), "Report on Statistical Disclosure Limitation Methodology," *Statistical Policy Working Paper 22*, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

"Establishing New Research Data Centers," *Federal Register*, 63(14), pp. 3309-3310, January 22, 1998.

Cooper, J.; Merrell, D.; Nucci, A.; and Reznek, A. (1998), "Protecting Confidential Data at Restricted-Access Sites: Lessons Learned from Census Bureau Research Data Centers," *American Statistical Association 1998 Proceedings of the Sections on Government Statistics and Social Statistics,* Alexandria, VA.

Hoy, C.E. and LaPlant, Jr., William (1998), "Protecting Demographic Census and Survey Data for the Year 2000 and Beyond," *American Statistical Association 1998 Proceedings of the Survey Research Section*, Alexandria, VA.

Jabine, Thomas B. (1993), "Procedures for Restricted Data Access," *Journal of Official Statistics*, 9(2), pp. 537-589.

McGuckin, R. and Pascoe, G. (1988), "The Longitudinal Research Database: Status and Research Possibilities," *Survey of Current Business,* November, pp. 20-26.

Reznek, A.P.; Cooper, J.M.R.; and Jensen, J.B. (1997), "Increasing Access to Longitudinal Survey Microdata: The Census Bureau's Research Data Center Program," *American Statistical Association 1998 Proceedings of the Sections on Government Statistics and Social Statistics*, pp. 243-248.

Willenborg, L. and de Waal, T. (1996), "Statistical Disclosure Control in Practice," *Lecture Notes in Statistics # 111,* New York: Springer-Verlag.