

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint ECE/Eurostat Work Session on
Statistical Data Confidentiality**
(Thessaloniki, Greece, 8-10 March 1999)

Working Paper No. 6
English only

Topic (i): new applications of disclosure control methods

MAKING DATA MORE ACCESSIBLE IN A CLIMATE WHERE PERCEPTION MATTERS

Submitted by the U.S. Bureau of the Census¹

Contributed paper

I. Overview

1. The U.S. Census Bureau is developing a comprehensive series of products based upon a mix of products for print, CD-ROM and the Internet with varying degrees of subject and geographic detail. The results of the next U.S. population census in 2000 will be provided to the public through the new Internet data access tool, Amercian FactFinder, which will permit access to pre-defined and custom tabulations. To preserve confidentiality, tabular products will make use of the data swapping technique used in 1990, which has proved extremely popular with data users. As a new feature on the Internet, users will be able to generate custom tabulations through controlled access to the internal microdata which have been swapped and passed through confidentiality filters that prevent the generation of sparse tables.

2. There are confidentiality issues arising from these and other activities that are being addressed by security and disclosure limitation methods. Nevertheless, issues remain about unknown threats from intruders who have new tools to break disclosure protections and concerns that the public may perceive that the agency has not done enough. The concerns are heightened by the current attention to privacy threats created by the use of new technologies such as the Internet. This paper discusses the real and perceived threats to confidentiality and privacy from new dissemination activities and how the Census Bureau and other statistical agencies can address them.

II. The Reality of Confidentiality

3. Although the law² permits no fudging when it comes to confidentiality and insists that no data be released that could identify an individual respondent, there are no absolute guarantees that data released to the public are disclosure-proof. Decisions to release data involve tradeoffs between use and confidentiality protections. At the Census Bureau, a Disclosure Review Board (DRB) reviews all proposed data products prior to release to help ensure the proper balance. This board consists of

¹ Prepared by Gerald Gates.

² Title 13, United States Code, Section 9 (a) (2) states "Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, may...make any publication whereby the data furnished by any particular establishment or individual under this title can be identified."

Census Bureau staff from many of the directorates involved in the production and analysis of the data as well as the research division and policy office. All requests to release data (as tables or microdata) are submitted with a checklist identifying potential disclosure concerns and describing how they have been addressed. Once the board decides on whether to release the data, the sponsoring division must comply or appeal the decision to a senior level group. In practice, the DRB has been successful in achieving its objectives without senior staff intervention.

4. In deciding to release microdata products, the DRB recognizes that the fact that the file represents a sample of the population is a very powerful source of protection in itself. If someone knows that a person or business is represented on a file, the likelihood of identifying them is much greater than if there is a low probability that the person/business is present. Conversely, if it is likely that a person/business is on a file, the risk is magnified. The fact that large businesses are sometimes sampled such that they have 100 percent chance of selection is the primary reason why public use microdata on businesses is so rare. For tabular survey data, sampling creates the uncertainty needed to ensure that cells with small values (ones or twos) do not reveal individual persons or businesses. Traditionally, tabular data from censuses or business surveys use suppression techniques to protect against disclosure. In the 1990 Census, the Census Bureau used a “confidentiality edit” comprised of data swapping as well as forced imputation to mask the data while permitting the publication of small cells.³

5. Another important criterion in deciding on the acceptability of public use microdata is the proposed level of geography. The smaller the geographic area attributed to a record, the greater the likelihood that the individual associated with that record will be identified. The DRB establishes geographic minimums for public use microdata that usually require populations of 100,000 or more persons. For one particularly detailed longitudinal survey, the geographic minimum was raised to 250,000 persons. In addition to geography, the existence of matching databases is the second greatest threat to disclosure. Specific populations represented on public use files may also be represented on data files available to other government agencies, private organizations, or the general public. Similar variables that are contained in these databases can be linked and the identity of the survey respondent may be revealed. If record linkage is expected to be a problem, data masking techniques can be applied to the variables in question to thwart matching efforts. These and other measures to protect microdata are described in detail in Statistical Policy Working Paper 22.⁴

6. When data are made available under a restricted access arrangement, the data are protected by ensuring that the researcher is liable if he/she discloses the information or uses it improperly. This process works if the agency has the authority to ensure researcher compliance and if adequate security can be maintained. Agencies that routinely process confidential data have the appropriate security in place and have staff that know the rules and understand the implications of breaches. When data are used offsite or at satellite facilities, the security issues become more important. Often, it is suggested that others don’t have the same “culture of confidentiality” needed to ensure the data are not misused. This “culture of confidentiality” is not something one derives overnight and requires exposure to a long period of handling sensitive information. As a requirement for becoming a Census Bureau Research Data Center (RDC), a university must show a high level of commitment to confidentiality and agree to employing a full-time Census Bureau employee who has been exposed to the Census Bureau’s “culture of confidentiality.”⁵

³ Griffith, R.A., Navarro, A., and Flores-Baez, L, “Disclosure Avoidance for the 1990 Census,” Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp 516-521.

⁴ “Report on Statistical Disclosure Limitation Methodology,” Statistical Policy Working Paper 22, Statistical Policy Office, Office of Management and Budget, Washington D.C., May 1994.

⁵ Cooper, Joyce M.R., Mcerrell, David R., Nucci, Alfred R., and Reznick, Arnold P., “Protecting Confidential Data at Restricted Access Sites: Lessons Learned from Census Bureau Research Data Centers,” 1998 Proceedings of the Government Statistics and Social Statistics Sections, American Statistical Association,

III. The Perception of Confidentiality

7. Protecting confidentiality is not always so straightforward. Those involved know that there is as much art as there is science to deciding how much protection is enough. Consequently, agencies may be nervous about data releases and worry that they may have underestimated the risk or that the public will perceive that the risk is greater than they can tolerate. In the end, it is a matter of continuing to earn the public's trust that you have done your best to protect their interests. Once that trust is lost, it becomes a huge obstacle to gaining future cooperation in surveys. The perception that the agency cannot be trusted can be as damaging to future response rates as an actual breach. In today's climate, a cynical public can lose trust even by association. For instance, lack of trust in government can translate into lack of trust in a specific agency. Also, lack of trust in private sector survey research organizations can spill over to government statistical agencies. Recent opinion surveys show that many Americans don't trust government and many believe that all government computers are connected.⁶

8. A modern issue in the role of perception is the explosion of new technologies to process and disseminate information worldwide. The proliferation of computers and networks leading to the World Wide Web has greatly improved the lives of many and made information readily available. The same technology has led to privacy abuses that are hard to detect but are none the less real. The U.S. newspapers routinely contain stories of privacy abuses created on the Internet. Users of the Web don't understand how it works but they like the advantages it provides. They also fear that they are losing control over their information and how it is used.⁷ Web sites are just beginning to post privacy policies and many do so in a very general way. This sense of the unknown contributes to the perception that it may not be as safe to provide personal information over the Internet as it was through the mails, for instance.

9. Another factor in the perception of confidentiality is the degree of knowledge about procedures used to process and protect information. Messages that are too complicated or too simple do not reassure or inform. Sometimes messages are not received because they are not delivered effectively. Statistical agencies devote considerable energies to protection against disclosure but at the same time provide few resources to preparing and delivering messages that can effectively deal with perceptions.

IV. Perception Scenarios

10. As mentioned previously, several possible perceptions have been identified for planned Census Bureau activities. Most of these activities are directly related to efforts to make the most effective use of new technologies to disseminate data. The following possible perceptions have been targeted as the most serious and requiring immediate attention.

- a) Perception: The Census Bureau releases tabulations that contain cells that show characteristics for a single census respondent.

Data that are swapped in Census 2000 form the basis for all predefined tables similar to the printed tables released after the 1990 Census. Tables will contain data for blocks, block-groups, census tracts and other geographic units. In some of the finer breakdowns, the tables will contain cells with one or two households, seemingly in violation of confidentiality. However, since a percentage of the data will have been exchanged (swapped) with data from households in nearby blocks, one

forthcoming..

⁶ Gates, GW, and Bolton, D, "Privacy Research Involving Expanded Statistical Use of Administrative Records," 1998 Proceedings of the Government Statistics and Social Statistics Sections, American Statistical Association, Forthcoming.

⁷ Gates/Bolton, forthcoming.

cannot assume that the household has been identified. In 1990, few users raised concerns and those that did were reassured that this was not a violation of confidentiality. In 2000, many new, unsophisticated users will have access to these tabulations through American FactFinder and the potential for confusion will be greatly magnified.

b) Perception: Data provided to researchers at the new Research Data Centers (RDCs) cannot be protected as well as when it is restricted to Census's Headquarters.

The Census Bureau has opened two RDCs, one in Boston Massachusetts and one in Pittsburgh, Pennsylvania. Two new centers will open soon in California. These centers are a useful alternative for researchers who want to use confidential data for special research studies but do not wish to relocate to the Census Bureau's headquarters in Suitland, Maryland. One condition for designation as a RDC is that the institution provide a secure environment for the data to ensure that data are not purposely or unintentionally misused. A full time Census Bureau employee staffs the center when it is in use. In addition, only summary statistics, such as regressions, medians, means, etc., are permitted to be removed from the center and, then, only after they have been reviewed for disclosures.

Despite this tight security, and because of the relative newness of the RDC approach, the Census Bureau has determined that some of its particularly sensitive data cannot reside at an RDC. For instance, data linkages that combine data from sources outside the Census Bureau with survey data are of particular concern. The Census Bureau has decided that even though the data can be protected at any of the RDCs, it will take a wait-and-see approach to providing direct access to researchers at its RDCs. Instead, data from these linkages will first be masked to give them an added level of protection.

c) Perception: You can get to confidential data through the Census Bureau's American Factfinder data retrieval system on the Internet.

Through the American Factfinder, one will be able to submit requests for special tabulations to be run in real time against the internal data file, which has been through the data swapping procedure but is not safe for public release. For these user-defined tables, the Census Bureau is prepared to implement filters and firewalls to screen requests for tabulations. Tabulations that would reveal, directly or indirectly, characteristics that have not been swapped will be denied. Final plans for the American Factfinder disclosure methods are to be completed by the end of 1998. Among the disclosure issues to be considered are the broad availability of data to unsophisticated users, the inexpensive powerful new tools to link and mine data files, the potential for collaboration among users to defeat the protections, and the public's perception that the security controls may not be adequate.

d) Perception: Proposed data sharing legislation will open the doors to weakening the confidentiality protections of the Census Bureau.

In the U.S., legislation has been proposed to permit the sharing of confidential information among a limited group of key statistical agencies. This legislation is designed to improve data and reduce overall costs and burden on respondents. Specifically, it would permit agencies to evaluate and improve their survey frames, evaluate the quality of data collected by the Federal statistical system, and conduct joint research using similar data. Under this legislation, each member of the sharing enclave would have strong legal protections to ensure confidentiality and to restrict uses to statistical purposes only. Each agency would be permitted, but not required, to share its data with the others. The Office of Management and Budget's Statistical Policy Office would arbitrate disputes among agencies. Under the proposed legislation, agencies would be required to provide

informed consent about the sharing. This legislation was not passed in the last Congress, but it is expected to be reintroduced soon after the new Congress convenes in January 1999.

In public opinion surveys conducted by the Census Bureau about confidentiality, the agency found that the public does not know or believe that their information will be protected or that the law requires it to be so. On the other hand, when focus groups were asked about the Census Bureau's planned use of administrative records, some participants were skeptical about the idea of data sharing and some believed that the Census Bureau already uses other agencies' lists. A number of participants viewed the sharing of data by other government agencies with the Census Bureau as the first step toward total erosion of confidentiality in government data collection. The idea of reversing the flow of information such that census data might be released to other government agencies met with the most opposition.⁸ It should be noted that other surveys have revealed support for the Census Bureau's access to administrative data for authorized statistical programs.⁹ Nevertheless, growing concerns over loss of control over information may generate concern about data sharing, even among statistical agencies when strict legal protections are in place. As our focus groups revealed, the public may perceive that once sharing starts there is no stopping it.

V. Strategies to address perception concerns:

11. If any or all of these scenarios prove to be real concerns for the public, the Census Bureau will have a major public relations problem to deal with. The visibility of the data products available on the Internet certainly raises the stakes that misunderstandings will occur and that these will be broadcast widely. To alleviate potential concerns that confidentiality has been breached, the Census Bureau is exploring weaknesses in data protection methods. To deal with the perception issues, the Bureau is preparing an outreach effort to explain its procedures and elicit public reaction. This campaign will attempt to provide reassurance and limit misunderstandings. The process that is envisioned includes:

- a) Evaluating current security and disclosure limitation activities to ensure that they are adequate in light of new technologies and potential intruders. Computer experts have been commissioned to try to break into microdata products to determine whether they are safe and where vulnerabilities may exist. State-of-the-art "firewall" technology in being employed to protect computer systems from would-be hackers.
- b) Creating a culture of confidentiality wherever confidential data are used. Since people are the key ingredient in data use and protection, their role in protecting confidentiality cannot be overstated. Awareness of the great importance placed on data protection is essential to avoiding intentional or inadvertent breaches. Organizations that process sensitive data have built up an institutional climate that values security and impresses on employees how important it is to the overall work. To ensure this culture is maintained in an environment where change is constant, the Census Bureau will need to strengthen security training for staff--especially newly hired staff. At Research Data Centers, training of staff and users will be all the more critical, since the remoteness from Headquarters will prevent staff integration into the secure environment established at Headquarters.
- c) Developing and testing messages that reassure both users and the general public that confidentiality is being maintained. If the messages are too complicated they will not be understood and if they are too simple they may seem condescending or superficial. This should begin through cognitive testing of messages to ensure that words have the same meaning with different audiences. Messages should also be tested with focus groups comprised of key stakeholders, such as academic researchers, advocacy groups, and Internet

⁸ Gates/Bolton, forthcoming.

⁹ Gates/Bolton, forthcoming.

- users. Final message wording should resonate with these groups such that individual interests are addressed. For instance, messages should highlight how apparent weaknesses are not real and how protections have been designed to ensure the validity of results using the data.
- d) Focusing confidentiality messages on anyone who happens upon the data, not just serious users. For data from the 1990 census, a confidentiality message on the Census Bureau web site explaining the “confidentiality edit” is directed to data users under the heading “accuracy of the data.” This focus is understandable in an environment when only sophisticated users search for data to analyze and their primary concern is with data quality. In our new paradigm, messages will need to explain techniques such as data swapping up front to all web surfers. In addition, all surfers will need to be informed that public use microdata files have been modified to protect the confidentiality of individual respondents and that they can’t extract individual records through the American Factfinder.
 - e) Communicating messages effectively. Messages should be developed and tested with a communications strategy in mind. Various avenues should be explored including official publication of agency decisions (such as the Federal Register in the U.S.), presentations at professional conferences and symposia, notices at agency Web sites, and press releases. The choice will depend on proximity to the data release. Federal Register notices and presentations at professional meetings will identify potential problems early on, while Web site messages and press releases will inform users what to expect from the data they are about to view. The World Wide Web offers a great mechanism for communicating messages to all users. Links can ensure that all users pass through confidentiality notices prior to requesting data. Of course, there is no assurance that the message will be read, but well placed, effective messages should increase the chances and will provide a reference to those who do protest to the agency, media, or their elected officials.

VI. Conclusion

12. This paper has attempted to focus on an important and little understood aspect of confidentiality. Public perception is illusive and often hard to predict or change. Nevertheless, when we have reason to believe that our activities are likely to raise concerns we must be willing to find out the extent of that concern. If the concern is based on misunderstanding, we must try to explain ourselves to those involved and attempt to dissuade our detractors. As we make greater use of new technologies to expand the utility of our data, it is probable that without an effective communication strategy we will lose public support for our efforts, not because we did something wrong, but because we did not make ourselves clear.