New Directions in Disclosure Limitation at the Census Bureau

*Laura Zayatz, Richard Moore, and B. Timothy Evans, U.S. Bureau of the Census**

The U.S. Bureau of the Census is in the business of collecting and disseminating data. The technological revolution of the 1980's and the accessibility of personal computing to the general public have fueled a rising demand for this data. This technology is allowing data users to handle increasingly large and detailed data sets and tabulations. Unfortunately, the more information the Bureau provides, the greater the possibility that a user can determine exact data values belonging to a particular respondent.

Title 13 requires that the Bureau ensure the confidentiality of data provided by all responding entities (individuals, households, and economic establishments). However, the Bureau's traditional disclosure limitation techniques cannot keep pace with the growing demand for an ever-wider variety of data products. As a result, we are unable to completely fill all requests. This paper gives an overview of some new disclosure limitation techniques under investigation at the Bureau, with respect to both microdata files and tabular data, that may facilitate broader dissemination of data.

■ Current Disclosure Limitation Techniques

In general, disclosure limitation can be done in two ways. One can limit the amount of data that is given out, or one can mask the data by adding noise, swapping values, etc. We are not going to go into detail on each method of disclosure limitation. For such detail, see Federal Committee on Statistical Methodology (1994). However, we would like to point out that the Census Bureau has traditionally opted for the approach of limiting the amount of information given out.

For demographic microdata, we limit the geographic detail. All identified geographic areas must contain at least 100,000 persons in the sampled area. We often categorize continuous variables or combine sparse categories in an attempt to prevent matching. And we topcode many variables. For example, rather than publishing a record showing an income of $1,000,000, the record may only show a representative value for the upper tail of the distribution, such as the mean for the tail (Cox and Zayatz, 1995).

For establishment tabular data, we use cell suppression. We remove from publication all cell values, which we feel pose too much risk of disclosure (if published, users could closely estimate a respondent's value). Because our tables are additive, we then have to remove sufficient other cells (complementary suppressions) to make sure the "risky" cells cannot be derived or estimated closely through addition and subtraction of cells that are published. By not publishing many cells, we are limiting the amount of information that we are giving out.

Census Bureau staff have traditionally chosen to limit the amount of information given out rather than mask the data because they do not want to alter the statistical qualities of the data by using noise or data swapping. An exception to this trend is the Confidentiality Edit used for tabular data from the 1990 Census of Housing and Population (Griffin, Navarro, and Flores-Baez, 1989). The Confidentiality Edit involved data swapping on the microdata file. This technique was chosen over cell suppression, which was used for the 1980 tabular data. Cell suppression must be performed on each table separately, and then suppression patterns must be coordinated among all tables (not an easy task). Also, cell suppression greatly limits a user's ability to aggregate data, and much information is lost in the form of complementary suppressions.

■ Why the New Directions?

The new directions stem from the technological

*This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

revolution and the huge demand for data. Specifically, for microdata, there is a lot of data available to the general public or to universities or other government agencies, which could possibly be matched to Census Bureau microdata files. There is a large amount of computing power and matching software available that could be used for this purpose. And there is an increasing interest in the idea of the Census Bureau using administrative data in many of our surveys and censuses. This would greatly increase the risk of our files because the agency supplying the administrative data would have great matching capability. Simply reducing the detail of variables on our files would not be enough to prevent such matching. The next section discusses our approach to this problem.

In terms of decennial census data, we would like to see some improvements on the Confidentiality Edit of 1990, which has been criticized for lacking proof that it actually protects the data. The fifth section discusses our work for the 2000 Census.

For establishment tabular data, we are looking at an alternative to cell suppression. As stated before, cell suppression must be performed on each table separately; suppression patterns must be coordinated among all tables; cell suppression greatly limits a user's ability to aggregate data; and much information is lost in the form of complementary suppressions. Among these disadvantages, the worst is probably the required coordination of cell suppression patterns among all tables. Given the many standard tables that we publish and the many special requests for tables that we receive, this is an enormous job. Also, the Census Bureau is working on a Data Access and Dissemination System (DADS) that seeks to allow users to define and create their own tables. Should this become a reality, suppression patterns would have to be coordinated among all tables that were created. That would be an impossible task. The sixth section discusses our approach to this problem.

■ Microdata

For disclosure limitation of demographic microdata files, the Bureau of the Census is presently investigating the use of masking techniques, which involve additive noise and data swapping. Current research is directed in three areas. The first approach (Kim and Winkler, 1995), discussed below, has already been implemented. The second technique (Greenberg, 1987) is being given serious thought as an alternative to artificially generated data. The final method (Fienberg,1995) involves the use of a log-linear model to generate artificial responses. We are considering the feasibility of such an approach. All methods focus on the need to protect the confidentiality of the respondents, while ensuring the means and the variance-covariance structures of arbitrary subdomains are not significantly distorted.

The Kim-Winkler Approach

In 1995, the Department of Health and Human Services (HHS) contracted with the Bureau of the Census to produce a specially requested public use microdata file. This file supplemented information on the March 1991 Current Population Survey (CPS) public use file with Form 1040 information from each respondent's 1990 Internal Revenue Service (IRS) tax return. Because the IRS could use its data to re-identify individuals on the CPS file, great care had to be taken to mask the HHS file. In doing so, Jay Kim and Bill Winkler (1995) developed a very effective two-stage masking procedure. The first step involves the addition of randomly generated multivariate noise. In most cases, this distorts the values of sensitive data items (e.g., income, home value, mortgage, etc.) enough so that the record cannot be re-identified with the corresponding record on the original IRS file. In the second step, a portion of those distorted records that *can* be re-identified are subjected to a swap.

This procedure is very attractive since the user can control both (1) the amount of noise used to distort the file, and (2) the percentage of re-identifiable records to be swapped. In this way, not only are we assured that the file is adequately masked, but also that it has retained much of its analytical value.

The Bureau of the Census recognizes that the Winkler-Kim approach is a very powerful disclosure limitation tool. To employ it effectively, the user must exhibit some expertise in the setting of various parameters. Current research (Moore, 1996a) concentrates on (1) the development of a standard procedure for determining an ac-

ceptable amount of noise, and on (2) the development of a standard for the maximum percentage of re-identifiable "high risk" records in a publicly released file.

Rank-Based Proximity Swapping

For several years, data swapping has been an acceptable method of protecting microdata while preserving various frequency counts. This procedure has its limitations. Frequency counts (such as the number of black male doctors) may be retained, but the auxiliary statistics (such as the salaries of the black male doctors on the microdata file) may be seriously distorted. Rank-based proximity swapping (Greenberg, 1987) diminishes this problem. The procedure involves sorting (in ascending order) the values in each continuous field and swapping, so that the ranks of exchanged values differ by less than a prescribed amount.

Although Greenberg indicated that such a swap might retain analytic utility, he stopped short of guaranteeing anything. Recent research (Moore, 1996b) in this area has focused on deriving the "prescribed swapping difference" for each continuous variable subject to certain constraints. Moore has found that swapping will diminish the covariate relationships by a factor of R_0 ($0 < R_0 < 1$). The prescribed index differences (which yield the value R_0) are functions of (1) R_0, (2) the top and bottom codes for each continuous variable, and (3) the variability of the data between the top and bottom codes.

Synthetic Log-Linear Modeled Data

The Census Bureau is currently contracted with WESTAT and Stephen Fienberg (of Carnegie Mellon University) to investigate the feasibility of developing a model using true data and using it to synthetically generate artificial data. Fienberg suggests that Federal agencies combine various sources of error (e.g., sampling nonresponse, editing, imputation, and matching) with errors induced to ensure confidentiality. He then suggests that a new data set, with each response containing some unified error, be created from the original. A major concern at the Census Bureau is that such a method may destroy the interdependence between the categorical and continuous variables, resulting in a model that is inappropriate for the analyses of some data users.

■ The 2000 Census of Housing and Population

The U.S. Bureau of the Census has begun preliminary research for the development of a disclosure limitation strategy for the 2000 Census. For the 1990 decennial census, the major disclosure limitation technique was the Confidentiality Edit. This method involved subjecting a small sample of randomly-chosen respondents to a data swap. No effort was made to target specific records for swapping. Information between records was exchanged if and only if the two records agreed on six key fields. As a result, all census counts mandated by law (e.g., counts of total persons by race and Hispanic origin, and counts of housing units by various characteristics) were preserved.

This disclosure strategy has not been without its critics (Fienberg et al., 1995). In particular, there are questions as to how well the method protected the data and how much the auxiliary statistics were distorted. These concerns do not appear to have been appropriately addressed during the processing of the last decennial census. The Bureau of the Census is meticulously examining the 1990 procedure (Moore, 1996c). For the 2000 disclosure limitation procedure, the following questions will be addressed:

- ❒ How unique does a record have to be in order to be considered risky?
- ❒ What set of categorical variables readily identifies high-risk individuals?

Assuming a respondent is at risk, the following additional concerns must be addressed:

- ❒ In order to preserve anonymity, what information needs to be swapped?
- ❒ Is there a procedure to swap records, so that the auxiliary statistics are not significantly distorted for relatively large (e.g., 20 or more respondent) sub-domains?

Current research not only focuses on developing algorithms, which provide solutions to the questions, but also

on developing the measure(s) necessary to quantify the extent to which the file is protected and/or distorted.

■ Establishment Tabular Data

The Census Bureau's traditional approach to disclosure limitation with establishment tabular data has been cell suppression. (The Bureau does not release establishment microdata.) Each cell in a table is subjected to a sensitivity criterion to determine which cells are disclosure risks. All cells that fail the criterion are suppressed in the publication, along with sufficient other cells, to prevent data users from recovering the values of the sensitive cells by manipulating additive relationships among cell values and row/column totals. By suppressing cells, we effectively limit the amount of information available to data users. In fact, users often complain that we suppress too much.

In addition to limiting the amount of information we can provide, using cell suppression also necessitates coordinating suppression patterns between interrelated tables, which can be a very complicated and difficult process. In particular, it makes the production of special tabulations extremely tedious. With work proceeding on DADS, as already mentioned, special tabulations are poised to become the rule rather than the exception, which will make coordinating suppression patterns among all requested tables virtually impossible.

As an alternative to cell suppression, we are investigating the addition of noise to establishment microdata as a disclosure limitation technique (Evans, Zayatz, and Slanta, 1996). Specifically, we would perturb each respondent establishment's data by a small amount, say, 10 percent. Then, if a single establishment dominates a cell, the value in the cell will not be a close approximation of the dominant establishment's value because that value has had noise added to it. (What constitutes a "close" approximation is open to debate.) By adding noise, we would avoid disclosing the dominant establishment's true value.

Noise would be added to an establishment's data by means of a multiplier. For our 10-percent example, the multiplier would be near either 0.9 or 1.1 and would be applied to all of the establishment's data items prior to tabulation. Because the same multiplier would be used with an establishment wherever that establishment was tabulated, values would be consistent from one table to another. That is, if the same cell appeared on more than one table, it would have the same value on all tables.

We could use a variety of distributions to generate the multipliers, provided that they were centered at or near 0.9 and 1.1. It is a key requirement, however, that the overall distribution be symmetric about 1. That is, the distribution around 0.9 and the one around 1.1 should be "mirror images" of each other. In this case, the expected value of any multiplier will be 1, even though, in practice, no multiplier will ever actually equal 1. Hence, the expected value of the *amount* of noise in any establishment will be zero. This should prevent the noise from introducing any bias into our level estimates.

In the degenerate case, however, where a cell contains only a single establishment, the cell value would contain about 10-percent noise. Other sensitive cells, in which one large establishment dominates the cell value, would also contain large amounts of noise because the amount of noise in the cell total would resemble the amount of noise in the dominant establishment (roughly 10 percent). The more dominant the large establishment is, the more closely the cell resembles the single-contributor case. These are precisely the cells that are at risk for disclosure and need to be protected.

On the other hand, we would like to assign the multipliers so that we minimize the effect of the noise on important aggregate estimates that are not disclosure risks. We can do this by strategically sorting the establishments before assigning the multipliers. Suppose that, for the survey in question, the most important estimates are produced by SIC × geography. We would sort establishments by SIC × geography × measure of size and then assign the multipliers in a pairwise-alternating fashion. The direction of perturbation (multiplier greater than 1 vs. less than 1) for the first establishment would be chosen randomly, and, henceforth, each successive pair of establishments would be perturbed in the opposite direction from the pair immediately preceding it.

To illustrate, suppose the first establishment was assigned a multiplier close to 1.1. Then, the second and

third establishments would have multipliers close to 0.9; the fourth and fifth establishments, close to 1.1; the sixth and seventh, close to 0.9, etc. This procedure assures that, for any establishment in a given SIC and geographic area, there will be on average another establishment in the same SIC and geographic area that is about the same size but that has been perturbed in the opposite direction. Thus, when the aggregate estimates are computed, the noise present in these two establishments should have a tendency to cancel out. This pairwise cancelling of noise in the summation should result in the SIC × large geographical area estimates containing very little noise. This is desirable since these aggregate estimates are generally not sensitive and do not need to be protected.

In general, the amount of protection provided to an estimate by the noise would depend on the amount needed. This property, combined with the fact that noise would only have to be added once, would greatly simplify the production of special tabulations. We could produce as many tabulations as necessary, and, for each one, the noise would naturally end up being greater in the sensitive cells. We would also no longer have to keep track of suppressed cells between tables.

The percentage of noise in a cell would be defined as the percent by which the noise-added value differed from the true, noise-free value. We would have to calculate both values for each cell in order to quantify the amount of noise each cell contained. All cells exceeding a certain noise threshold, say, 7 percent, would be flagged. In this way, users would be alerted to cells whose values contained a lot of noise and hence may be unreliable. Also, the description of the flag would explain how and why the noise was added, thus assuring users (who may be surprised not to see any suppressed cells) that disclosure limitation had indeed been performed.

We have tested the noise technique with the Research and Development Survey, and the results are encouraging (Evans, Zayatz, and Slanta, 1996). Sensitive cells that would normally be suppressed typically exceeded the noise threshold (7 percent in our example), while nonsensitive cells received less noise. Also, the noise did not appear to introduce any bias into the estimates. In light of our initial results, we plan to continue investigating the effect of noise on such things as trend estimates and regression analysis to see if noise can be used to protect sensitive data without disrupting the kinds of analysis that data users typically perform with Census Bureau data.

■ Summary

Because of the growing demand for data and the increase in risk due to technological advances, the Census Bureau may need to change its approach to disclosure limitation. Our current techniques, which limit the amount of information provided, are at odds with our desire to incorporate administrative data and our desire to allow users to define their own data products. The Confidentiality Staff is developing and testing disclosure limitation methods involving the addition of noise and data swapping. These techniques would allow for the release of an ever-wider variety of data products.

■ References

Cox, L. H. and Zayatz, L. (1995), "An Agenda for Research in Statistical Disclosure Limitation," *Journal of Official Statistics*, Vol. 11, No. 2, pp. 205-220.

Evans, B.T.; Zayatz, L.; and Slanta, J. (1996), "Using Noise for Disclosure Limitation of Establishment Tabular Data," *Proceedings of the 1996 Annual Research Conference*, U.S. Bureau of the Census.

Federal Committee on Statistical Methodology (1994), *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, Washington, D.C.: U.S. Office of Management and Budget.

Fienberg, S. E. (1995), "Taking Uncertainty and Error in Censuses and Surveys Seriously," *Proceedings of the Statistics Canada Symposium 95*, Ottawa, Ontario, to appear.

Fienberg, S. E.; Steele, R. J.; and Makov, U. (1996), "Statistical Notions of Data Disclosure Avoidance and Their Relationship to Traditional Statistical

Methodology: Data Swapping and Loglinear Models," *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C., to appear.

Greenberg, B. (1987), "Rank Swapping for Masking Ordinal Microdata," U.S. Bureau of the Census, unpublished manuscript.

Griffin, R.; Navarro, F.; and Flores-Baez, L. (1989), "Disclosure Avoidance for the 1990 Census," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 516-521.

Kim, J. J. and Winkler, W. E. (1995), "Masking Microdata Files," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, to appear.

Moore, R. A. (1996a), "Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets," Statistical Research Division Report Series, **RR 96-04**, U.S. Bureau of the Census, Washington, D.C.

Moore, R. A. (1996b), "Analysis of the Kim-WinklerAlgorithm for Masking Microdata Files--How Much Masking Is Necessary and Sufficient? Conjectures for the Development of a Controllable Algorithm," Statistical Research Division Report Series, **RR 95-05**, U.S. Bureau of the Census, Washington, D.C.

Moore, R. A. (1996c), "Preliminary Recommendations for Disclosure Limitation for the 2000 Census: Improving the 1990 Confidentiality Edit Procedure," Statistical Research Division Report Series, **RR 96-06**, U.S. Bureau of the Census, Washington, D.C.