Working Paper No. 15
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (iv): Confidentiality issues for small areas

DISCLOSURE LIMITATION FOR CENSUS 2000 TABULAR DATA

Invited paper

Submitted by the Bureau of the Census, United States[1]

[1] Prepared by Laura Zayatz (laura.zayatz@census.gov).

Disclosure Limitation for Census 2000 Tabular Data[1]

Laura Zayatz
Bureau of the Census, Statistical Research Division, 3209-4, Washington, D.C. 20233

Abstract. This paper describes the statistical disclosure limitation techniques to be used for all U.S. Census 2000 tabular data products. Many of these tables are published for very small geographic areas. The paper includes procedures for short form tables, long form tables, special tabulations, and an online query system for tables. Procedures include data swapping, rounding, collapsing categories, and applying thresholds. Procedures for the short and long form tables are improvements on what was used for the 1990 decennial census. Several procedures we are using for the special tabulations are new and will result in less detail than was published from the 1990 decennial census. Because we did not previously have the online query system for tables, all of those procedures are newly developed.

1 Introduction

The Bureau of the Census is required by law (Title 13 of the U.S.Code) to protect the confidentiality of the respondents to our surveys and censuses. At the same time, we want to maximize the amount of useful statistical information that we provide to all types of data users. We are in the last stage of applying the disclosure limitation techniques for all tabular data products stemming from Census 2000. The techniques are designed to protect data confidentiality while preserving data quality.

This paper describes the Census 2000 disclosure limitation techniques for tables. In Section 2, we briefly describe the procedures that were used for the 1990 Census. In Section 3, we explain why some changes in those techniques were necessary. In Sections 4-7, we describe the procedures for Census 2000, including procedures for the 100% (short form) census tabular data, the sample (long form) tabular data, the special tabulations, and the Advanced Query System in American FactFinder. Section 8 offers a conclusion.

---

1 This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

2 Disclosure Limitation for 1990 Census Tabular Data

2.1 Procedure for the 100% (Short Form) Data

The Census Bureau attempts to get information on characteristics such as sex, age, Hispanic/NonHispanic, race, relationship to householder, and tenure (owner or renter) from 100% of the population through what we call "short form" questions. The 100% data are published in the form of tables. Many of the tables are published at the block level. The average block contains 34 people. Some of the more detailed tables are published at the block group level. The average block group contains 1,348 people. Thus these data are published for very small geographical units. For the 1990 census, the procedure used to protect the short form (100%) data was the Confidentiality Edit [1]. A small sample of census households from the internal census data files was selected. The data from these households were swapped with data from other households that had identical characteristics on a certain set of key variables but were from different geographic locations. Which households were swapped was not public information. The key variables were number of people in the household of each race by Hispanic/NonHispanic by age group (<18,18+), number of units in building, rent/value of home, and tenure. All tables were produced from this altered file. Thus census counts for total number of people, totals by race by Hispanic/NonHispanic by age 18 and above (Public Law 94-171 counts --- also known as Voting Rights counts) as well as housing counts by tenure were not affected. A higher percentage of records was swapped in small blocks because those records possess a higher disclosure risk. All data from the chosen households were swapped except for Indian Tribe. It was felt that it did not make sense to move a member of one tribe into a location inhabited by another tribe.

One advantage of the Confidentiality Edit is that it only needs to be implemented once on the internal microdata file in order to protect all tables produced from the file. A requirement for the Advanced Query System of American FactFinder, a table query system described in Section 3.3, is that the majority of disclosure limitation techniques be applied to the underlying data rather than to individual tables. We wanted to avoid techniques such as random rounding, cell suppression, and perturbation which are often applied on a table by table basis. An additional advantage of the confidentiality edit is that no data are suppressed, so aggregation of data is not a problem. The disadvantage is that there are no obvious changes in the tables that would make evident our disclosure limitation efforts.

2.2 Procedures for the Sample (Long Form) Data in Tabular Form

Approximately a one in six sample of the population receives the long census form which, in addition to the 100% information, collects information on characteristics such as marital status, school attendance and grade level, ancestry, language, place of birth, citizenship, military service, income, industry, and occupation. The sample data are also published in the form of tables. Some of the tables are published at the block group level. The average block group contains 1,348 people. Some of the more detailed tables are

published at the tract level. The average tract contains 4,300 people. Thus these data are also published for very small geographical units. In 1990, it was felt that the fact that it was a sample provided protection for all areas for which sample data were published except for small block groups. In small block groups, some values from one housing unit's record on the internal file were blanked and imputed using the 1990 census imputation methodology. This altered file was used to create all tables. Which values were altered was not public information.

2.3 Procedures for the Special Tabulations

The Census Bureau publishes hundreds of thousands of tables from the short form and the long form data. Still, users may not find the tables they want in the standard summary files. When this happens, they can request and pay for a special tabulation. In 1990, the altered 100% and sample files were used to create the special tabulations.

3 Why Were the 1990 Procedures Changed?

3.1 Main Improvement: Targeting the Most "Risky" Records

As we stated in Section 2, small blocks and small block groups had higher rates of swapping and blanking and imputation because the records from small geographic areas possess a high disclosure risk. We wanted to extend the idea of targeting the most risky records for the disclosure limitation techniques. We swapped records that were unique based on some set of key variables. Those were the records with the most risk. We did not swap households for which all data were imputed. They were not at risk. We took into account the protection already provided by the rate of imputation. Records representing households containing members of a race category, which appeared in no other household in that block, were easily identifiable and presented a special risk. A very large percent of those records were swapped. And finally, we let the swapping rate differ among blocks and have an inverse relationship with block size (in terms of number of households). We believed it would be easier to identify a person or household in a small block than it would be in a large block.

3.2 Multiple Race Issues

In 1990, a person could only be identified by a single race. That is, people were only supposed to check one box on the questionnaire in response to the race question. In 2000, people were asked to check more than one box, if applicable. Thus we now have 63 possible answers to the race question. This led to changes in disclosure risk as well as processing procedures because of the additional detail in the tables.

3.3 The Advanced Query System (AQS) of American FactFinder (AFF)

AFF [2] has been developed to allow for broader and easier access to the standard

summary files and to allow users to create their own data products. One part of AFF is the AQS. The goal of the AQS is to allow users to submit requests for user-defined tabular data electronically. A request would pass through a firewall to an internal Census Bureau server with a previously swapped, recoded, and topcoded microdata file. The table would be created and electronically reviewed for disclosure problems. If it was judged to have none, the table would be sent back electronically. This is a new way of publishing tabular data, so we needed to develop new disclosure limitation procedures for the AQS. Note that some users want very large sets of tables, and it would not be practical to request them all via AQS. Those users must request special tabulations.

3.4 The Disclosure Review Board (DRB)

In 1990, the Census Bureau had a Microdata Review Panel. This group of people reviewed all microdata files for potential disclosure problems prior to their release. Files were often modified as a result of the review process. There was no formal review process for tabular data. Decisions on which tables to release and which should be withheld for confidentiality reasons were made on an adhoc basis by different people. In 1995, the Disclosure Review Board was formed. This group of people must review all microdata files and all tabular data products prior to their release. The Board had to approve the disclosure limitation techniques used for the tabulations and develop other requirements that the tabulations must meet. Changes are often made as a result of the review process.

4 The Procedure for the 100% Data

As we did in 1990, we swapped a set of selected records. Unlike 1990, the selection process was highly targeted to affect the records with the most disclosure risk. There was a threshold value for not swapping in blocks with a high imputation rate. Only records which were unique in their block based on a set of key demographic variables were swapped. A unique record was selected for swapping with a probability of:

$$C_1 * \frac{1}{block\ size} \quad .$$

That is, the probability of being swapped had an inverse relationship with block size. In addition, records representing households containing members of a race category which appeared in no other household in that block had an additional P_1 probability of selection. Pairs of households that were swapped matched on a minimal set of demographic variables. All data products are created from the swapped file. For any tables that were iterated by race, there had to be at least 100 people of a given race in a given geographic area for that table to be released.

5 The Procedure for the Sample Data in Tabular Form

Swapping (rather than blanking and imputation) was performed to protect the data. This increased the amount of distortion (giving us more protection). Swapping had the nice quality of removing any 100% assurance that a given record belonged to a given household. It was consistent with the 100% procedure. And, it retained relationships among the variables for each household.

As with the 100% data, we used 2 different sets of key variables --- one to identify the unique records and one to find the swapping partners. We held several variables fixed (unswapped). For example, travel time to work and place of work for a household would not make sense if swapped with a household geographically far away.

The procedure for producing the masked file was very similar to the procedure for the 100% data. Blockgroup replaced block because blockgroup is the lowest level of geography for publishing sample data. The threshold value for not swapping in blockgroups with a high imputation rate differed, and the probability that unique record was swapped was:

$$C_2 * \frac{sampling\ rate\ for\ that\ blockgroup}{blockgroup\ size}$$

We gave the chance of being swapped an inverse relationship with blockgroup size. We also gave the chance of being swapped a direct relationship with blockgroup sampling rate. The lower the sampling rate, the more likely that the sample unique was not unique in the entire blockgroup population. So a smaller sampling rate led to a lower chance of being swapped. For any tables that were iterated by race, there had to be at least 50 unweighted people of a given race in a given geographic area for that table to be released.

6 The Procedure for the Special Tabulations

All special tabulations are generated from the swapped data files. All cell values are rounded according to the following scheme:

0 rounds to 0
1-7 rounds to 4
8 or greater rounds to the nearest multiple of 5

Totals are constructed before rounding, thus universes remain the same from table to table, but the tables are no longer additive. Qunatiles (percentiles) may be calculated in

1 of 2 ways. If they are calculated as an interpolation from a frequency distribution of unrounded data, no additional rounding is required. This is the technique used in the standard summary files. If they are point quantiles generated using SAS and Proc Univariate, they are rounded to 2 significant digits, and there must be 5 nonoverlapping cases on either side of each quantile point. Thresholds on universes are often applied to avoid showing data for small geographic areas or small population groups. We often require 100 cases for 100% data and 50 unweighted cases for sample data. Occasionally we require 3 unweighted cases for sample data for very small tables. Percents and rates are calculated after rounding. We allow some exceptions when the numerator and/or denominator is not shown. Usually tables have no more than 3 or 4 dimensions, and the DRB does consider mean cell size.

7 The Procedures for the Advanced Query System

The AQS does not provide an open-ended or unconstrained opportunity to construct any or all possible tabulations from the full microdata files. As stated previously a query for a table through the AQS would pass through a firewall to an internal Census Bureau server with a previously swapped, recoded, and topcoded microdata file. All tables generated from the sample data are weighted. The query and the resulting table must each pass through a filter.

7.1 The Query Filter

If a user requests a tabulation for more than one area or for a combination of areas, each area must individually pass the query filter.

The external user is advised in the user interface that the blockgroup is the lowest level of geography permitted for 100% data and the tract is the lowest level of geography permitted for sample data for an external user. Requests for split blockgroups or split tracts are not permitted. A minimum population requirement is also imposed for each area. The user interface permits no more than 3 dimensions (page, column, and row) and 1 universe not including geography.

The query filter also delimits the use of variables such as race, Hispanic origin, group quarters, cost of electricity, gas, water, fuel, property taxes, property insurance cost, mortgage payments, condo fees/mobile home costs, gross rent, selected monthly owner cost, household/family income and individual income types. External users may obtain only predefined categories or recoded values of these variables. Most variables have several sets of recodes that the user can choose from. So if the user is requesting a table from a large geographic area, he can choose a very detailed list of recodes. If a user is requesting a table from a small geographic area, he can choose a short list of recodes to try to ensure that the table will pass the results filter.

If the query passes the query filter rules, the query is sent from the external server outside the firewall to the internal server inside the firewall to the full microdata files. The full microdata files contain all of the predefined categories for race, Hispanic origin, group quarters, etc.

7.2 The Results Filter

Each resulting tabulation selected from the full microdata files obtained through the Advance Query System must meet certain criteria or the AQS will not provide the user with the tabulation. If a user requests a tabulation for more than one area or for a combination of areas, each area must individually pass the results filter. The criteria are designed to prevent the release of sparse tabulations which can lead to disclosure. If a tabulation does not meet the criteria, the user will receive a message stating that the tabulation cannot be released for confidentiality reasons.

The system computes the total mean and median population cell sizes of the tabulation. For both mean and median calculations, only the internal cell counts are used (not the marginal totals). For both the mean and median calculations, cells with zero are included. If either the mean or median is less than some number n the system does not permit the tabulation.

Our disclosure limitation rules are designed to prevent the release of sparse tables. They do not guarantee that there will be no cell values of 1. In fact, many of our predefined tables contain cell values of 1, and for those we rely on the data swapping procedure to protect the data. The Advanced Query System uses the swapped file in generating tables, but again we wish to avoid releasing very sparse tables. The third rule in the results filter limits the proportion of cells with values of one. The ratio of the number of unweigthed cells counts of one to the number of non-zero cells must be less than some given parameter.

In our testing, we found that the mean rule is unnecessary. Whenever it failed, either the median or the ratio of ones rule also failed. It may be taken out of the system.

8 Conclusion

Prior to production, we tested and evaluated the data swapping procedures with various parameters using data from the 1995 and 1996 Census tests and the 1998 Dress Rehearsal. Following Census 2000, we evaluated the procedure's performance on the census data in terms of reducing disclosure risk and maintaining data quality. The evaluation report is Census Confidential because it contains confidential details about the procedure. We were very pleased with the results of the procedure. We were able to swap the records with the highest risk without much change to the tables. The cells most effected were the cells with 1 or 2 households or people in them.

Requests for special tabulations continue to pour in. They are keeping the Disclosure Review Board very busy. The formal review process has led to a decrease in the amount of detail in special tabulations.

The Advanced Query System rules and their parameters and population threshold values were tested by Census Bureau staff and disclosure limitation experts at Carnegie Mellon University. They were judged to be in accordance with best practices [3]. Data users, led by an outside contractor, evaluated the usefulness of the AQS, and most were very pleased [4]. Their one complaint was that they often could not get the detail they wanted for very small geographic areas because of the confidentiality filters.

References

1. 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.
2. Rowland, S. and Zayatz, L. (2001), "Automating Access with Confidentiality Protection: The American FactFinder," Proceedings of the Section on Government Statistics, American Statistical Association, to appear.
3. Duncan, G., Roehrig, S., and Kannan, K. (2000), "Final Report on the American FactFinder Disclosure Audit Project for the U.S. Census Bureau," prepared under contract to the U.S. Census Bureau.
4. Schneider, P. J, (2002), "American FactFinder Advanced Query System - Assessment Report on Stage Two (Sample File) Beta Testing," prepared under contract to the U.S. Census Bureau.