STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 31
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Methodological Issues Involving the Integration of Statistics and Geography
(Neuchâtel, Switzerland, 10-12 April 2000)

Topic (iv): Spatial analysis in a statistical context and disclosure control procedures

DISCLOSURE LIMITATION FOR CENSUS 2000

Submitted by U.S. Bureau of the Census [1]

Contributed paper

ABSTRACT

The decennial census of population and housing produces both 100% data (from a short form) and more detailed sample data (from a long form). Data from Census 2000 will be published in the form of conventional tables and public use microdata, as well as through an electronic query system (American FactFinder). A variety of disclosure limitation methods, including data swapping, will be employed on both the100% and sample data prior to any form of publication. We describe the 1990 procedures for disclosure limitation and the changes proposed for 2000, as well as limitations required for the electronic query system.

Keywords: confidentiality, microdata, swapping, census

I. INTRODUCTION

1. 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 investigating techniques that will be used for disclosure limitation (confidentiality protection) for all data products stemming from Census 2000.

2. This paper describes preliminary proposals for disclosure limitation techniques. In Section 2, we briefly describe the procedures that were used for the 1990 Census. In Section 3, we describe why some changes in those techniques may be called for. In Section 4, we give our initial proposals for procedures for Census 2000, including procedures for the 100% census tabular data, the sample tabular data, the microdata, and American FactFinder. In Section 5, we briefly describe methods of testing the resulting data in terms of retaining the statistical integrity of the data and giving adequate protection. Section 6 contains references.

[1] Prepared by Laura Zayatz, Philip Steel and Sandra Rowland.

II. DISCLOSURE LIMITATION FOR THE 1990 CENSUS

II.1 Procedure for the 100% Data

3. The 100% data are published in the form of tables. Most of the tables are published at the block level. The average block contains 36 people. Some of the more detailed tables are published at the block group level. The average block group contains 400 people. Thus these data are published for very small geographical units. 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, and tenure (own or rent). 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.

4. One advantage of the Confidentiality Edit is that it only needs to be implemented once on the internal mirodata file in order to protect all tables produced from the file. A requirement for the American FactFinder is that the majority of disclosure limitation techniques be applied to the underlying data rather than to individual tables. We wish 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.

II.2 Procedures for the Sample Data

II.2.1 Sample Data (long form) in Tabular Form

5. 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 400 people. Some of the more detailed tables are published at the tract level. The average tract contains 4060 people. Thus these data are also published for very small geographical units. 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.

II.2.2 Sample Data (long form) in Microdata Form

6. The microdata file contains records from 5% of all households in the nation. The microdata was created from the internal file after the blanking and imputation described in 2.2.1., so some protection was provided by that procedure. All identified geographic areas (PUMAs --- public use microdata areas) contained at least 100,000 people. Income values and some other continuous values such as age and rent were topcoded. Some very detailed categories from items such as Ethnicity and Indian tribe were collapsed into broader categories. And, of course, all identifying information such as name and address were stripped from the file.

III. WHY SHOULD THE 1990 PROCEDURES BE CHANGED?

III.1 Main Improvement: Targeting the Most "Risky" Records

7. 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 would like to extend the idea of targeting the most risky records for swapping. We would only swap records that were unique based on some set of key variables. Those are the records with the most risk. We would not swap households for which all data were imputed. They are not at risk. We would take into account the protection already provided by the rate of imputation. Records representing households containing members of a race category, which appears in no other household in that block, are easily identifiable and present a special risk. A very large percent of those records will be swapped. And finally, we would let the swapping rate differ among blocks and have an inverse relationship with block size (in terms of number of households). We believe it would be easier to identify a person or household in a small block than it would be in a large block.

III.2 Multiple Race Issues

8. 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 reponse to the race question. In 2000, people will be asked to check more than one box, if applicable. Thus we now have 63 possible answers to the race question. This leads to changes in disclosure risk as well as processing procedures because of the additional detail in the tables.

III.3 American FactFinder (AFF)

9. AFF [3] is being developed to allow for broader and easier access to the data and to allow users to create their own data products. The goal is to allow users to submit requests for 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 will not affect the disclosure limitation procedures for the public use microdata. Those disclosure limitation techniques will have been applied to this data before it is made available on AFF or any other Census Bureau web site. However, this is a new way of publishing tabular data, so we need to develop new disclosure limitation practices for AFF.

IV. INITIAL PROPOSALS

IV.1 Initial Proposal for the 100% Data

10. As we did in 1990, we will swap a set of selected records. Unlike 1990, the selection process will be targeted. There will be a threshold value for not swapping in blocks with a high imputation rate. Only records which are unique in their block based on the set of key variables will be swapped. The key variables are still under discussion but will be based on general demographics. A unique record will be selected for swapping with a probability of:

11. That is, the probability of being swapped will have an inverse relationship with block size. In addition, records representing households containing members of a race category which appears in no other household in that block will be have an additional P1 probability of selection. All data products will be created from the swapped file. We will test and evaluate values for the various parameters using data from the 1995 and 1996 Census tests and the 1998 Dress Rehearsal (see Section 5). The current plan is to hold Indian Tribe fixed (unswapped) as was done in 1990 (see 2.1).

12. In testing these procedures on the Dress Rehearsal data, we have found that there remain a very small number of records which we consider to have a high disclosure risk and for which we can find no matching households. For those records, we may drop one or more key variables and allow those values for those few records to be changed (swapped with non-matching values).

13. We must stress that this is the current proposal, but it remains under discussion and is certainly subject to change.

IV.2 Initial Proposal for the Sample Data

IV.2.1 Sample Data in Tabular Form

14. We propose that swapping (rather than blanking and imputation) be performed to protect the data. This will increase the amount of distortion (giving us more protection). Swapping has the nice quality of removing any 100% assurance that a given record belongs to a given household. It is consistent with the 100% procedure. And, it retains relationships among the variables for each household.

15. Note that for the 100% data, we used the same set of key variables to locate the unique (risky) records and to find matching households (swapping partners) for those records. For the sample data, we may use 2 different sets of key variables --- one to identify the uniques and one to find the swapping partners. We also may hold a few more variables fixed (unswapped). For example, travel time to work and place of work for a household may not make sense if swapped with a household geographically far away.

16. The procedure for producing the masked file then is very similar to the procedure for the 100% data. Blockgroup replaces 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 may differ, and the probability of a unique record being swapped is:

17. We have given the chance of being swapped an inverse relationship with blockgroup size. We have also given the chance of being swapped a direct relationship with blockgroup sampling rate. The lower the sampling rate, the more likely that the sample unique is not unique in the entire blockgroup population. So a smaller sampling rate should lead to a lower chance of being swapped.

IV.2.2 Sample Data in Microdata Form

18. The disclosure limitation procedures for the microdata files will be similar to the procedures used for the 1990 data. Because of additional concerns this decade about advances in technology and the abundance of databases in the private sector, the disclosure limitation techniques may be a bit more conservative. For example, some additional variables such as different income types and travel time to work may be rounded. Other variables such as month of birth may be dropped. The microdata will be created from the internal file after the swapping described in 4.2.1. All PUMAs will contain at least 100,000 people, and the individual states will help the Census Bureau to determine how to define those standard geographic areas.

19. Income values and some other continuous values such as age and rent will be topcoded. Topcodes for variables that apply to the total universe will include at least 2 of 1 percent of all cases. Topcodes for variables that apply to subpopulations will include either 3 percent of the appropriate cases or 2 of 1 percent of all cases, whichever is the higher topcode. Some very detailed categories from items such as Ethnicity and Indian tribe will be collapsed into broader categories. And, of course, all identifying information such as name and address will be stripped from the file.

IV.3 Initial Proposal for American FactFinder

20. American FactFinder 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 AFF 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 will be weighted. The query and the resulting table must each pass through a filter.

IV.3.1 The Query Filter

21. 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. Guidelines for requesting tabulations will include:

22. Levels of geography: The external user is advised in the user interface that the block 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 blocks or split tracts are not permitted. A minimum population requirement is also imposed.

23. Maximum number of table dimensions: The user interface permits no more than 3 dimensions (page, column, and row) not including geography.

24. Total population per geographic unit: Population size criteria will be determined from data in the summary files that indicate whether the population is large enough to pass the results filter. The user will be informed if the population size is too small.

25. The query filter also delimits the use of sensitive 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.

26. The system determines if the query includes race, Hispanic origin, group quarters and other sample data variables that by their nature could disclose confidential data when cross-tabulated with each other or with any variable except geography. Then the system determines if the query requests small areas - blocks, block-groups or user-defined geography with a population size that is less than average tract size (4060 in 1990), medium areas (population size 4060-99,999) or large areas (population size 100,000 or more). According to the population size of the area or areas requested, the system permits the use of appropriate combinations of short, medium or long lists of predefined categories of race, Hispanic origin, group quarters and other sample variables in the cross-tabulation. Only topcoded values of sensitive variables may be accessed.

27. 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 and modified sensitive variables.

IV.3.2 The Results Filter

28. Each resulting tabulation selected from the full microdata files obtained through American FactFinder must meet certain criteria or American FactFinder 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 rules and their parameters and population threshold values will be tested in 1999 and finalized for Census 2000.

29. 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 not excluded. If either the mean or median are less than n the system does not permit the tabulation.

30. 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. To address this issue, we have added a rule to the results filter to limit the proportion of cells with values of one. The rule counts the total number of nonzero cells in the cross-tabulation and the number of cells in the cross-tabulation with a value of 1 and then ensures that the ratio of the count of cells with a value of 1 to the count of total nonzero cells in the cross-tabulation is less than some preset parameter.

V. TESTING FOR DATA QUALITY AND ADEQUATE PROTECTION

31. Using data from the 1990 census and the Census 2000 Dress Rehearsal, we will examine the relationship between the rate of swapping and the amount of noise added to the data. For the 100% data, the distribution of age and age by race will be calculated before and after swapping. The effect of swapping disappears as the level of geographic aggregation increases. At the tract level, a pseudo-variance, based on the differences can be calculated. Based on research already conducted, county level tabulations differ considerably less than the imputation rate, for any reasonable rate of swapping. There should be little or no bias because the basic characteristics of the swapped households are fixed. Similar studies are planned for the sample data.

32. We will keep track of what percentage of uniques are swapped, what percentage of records are swapped, and the geographic levels at which swaps occur. We will also examine the proportion of uniques and swaps for different block sizes, since the selection favors small blocks and small blocks have a greater proportion of uniques. We will calculate the index of dissimilarity, the "D statistic" [2], as was used in 1990, for a number of distributions at various size geographic levels to determine the net proportion of the total which has been changed by the swapping operation.

33. We will test the filter rules using data from the Census 2000 Dress Rehearsal one hundred percent and sample microdata files in the American FactFinder system in 1999. We want to see, using these rules and real data, which tables would be released and which would be denied. The results of these tests may lead to changes in the rules. The proposed confidentiality testing would have three distinct and complementary components:

- best practice assessment
- determination of protection levels for complementary disclosure
- examination of the potential for linkage of records with unique characteristics

34. For the microdata, where appropriate, we will calculate the percentage bias [2] introduced by the swapping into some of the more important continuous variables, such as income, for different combinations of the key variables. [2] even suggests a range of acceptable values for the percentage bias that could be used as a guideline for ensuring that we have introduced enough noise to protect the data but not too much so as to distort it greatly. We will also perform a given set of regressions on the raw data and on the data after swapping to analyze the affect of swapping on the statistical properties of the data.

VI. 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] Moore, R. A. (1996), "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, DC.

[3] Zayatz, L. and Rowland, S. (1999), "Disclosure Limitation for American FactFinder," Proceedings of the Section on Government Statistics, American Statistical Association, to appear.