BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
Statistical Research Report Series
No. RR96/06

Preliminary Recommendations for Disclosure Limitation for the 2000 Census: Improving the 1990 Confidentiality Edit Procedure

Richard A. Moore, Jr.

U. S. Bureau of the Census
Statistical Research Division
Washington D.C. 20233

Report Issued: 5/7/96

Table of Contents

Executive Summary

Population censuses are essential to the well-being of any country. Data collected during these censuses are used to measure the rate and direction in which the country is progressing. The success of a census hinges on the ability of the statistical agency to obtain and disseminate accurate data. Yet agencies must guarantee respondents' confidentiality. This often involves invoking certain disclosure limitation techniques.

For the 1990 Decennial Census, the Bureau of the Census used a technique referred to as "the Confidentiality Edit" to mask the data. Although this technique preserved all population and housing unit counts mandated by law, it was not without its critics. Two of the main concerns were that (1) the technique offered no guarantee of confidentiality, and (2) only limited measures were put into place to quantify the extent of the distortion.

We believe that the 1990 Confidentiality Edit was a sound disclosure limitation technique. However, since this was the Census Bureau's first experience with a data-swapping technique, some details were overlooked. For the Year 2000 Census, we recommend that the Census Bureau continue to use a data-swapping approach.

This paper suggests ways in which the 1990 procedure can be modified to improve the amount of protection afforded to each individual respondent. We base these modifications on disclosure limitation concepts employed by other statistical agencies. In particular, we use the ARGUS approach (of Statistics Netherlands) to determine high risk key combinations. From the SAFE system (of the German Federal Statistics Agency), we learn that perturbing one carefully chosen key value of at least one respondent for each high risk combination can protect the data for all respondents with those combinations.

The paper suggests some additional ways to ensure confidentiality. These include (1) targeting cells with very few respondents as candidates for a swap, (2) limiting the key variables to those either legally mandated or those essential to our cross-tabulations, (3) limiting the size of the geographies for which tabulations are published, and (4) swapping geographic codes.

The 1990 procedure did limit the key variables to those mandated and swapped geographic codes. Both seemed to work well. However, there was only a limited attempt to target records which appeared to have a high risk of re-identification. In addition, there was no attempt to limit the size of the geography for which detailed tabulations were produced. Incorporating these additional features into the disclosure limitation methodology should make the procedure more secure.

Guaranteeing confidentiality was not the only concern expressed about the 1990 procedure. The D-statistic was the only criterion used to measure distortion to the file. This primarily measured the net changes to the frequency counts. No criterion was developed to measure the distortion to continuous auxiliary data. This paper proposes two such measures, BIAS and PERCENTAGE BIAS.

Not only do the BIAS and PERCENTAGE BIAS evaluate the amount of distortion, but they also can be used for assigning a sort order by which information on similar records are swapped. Using these measures, we establish an equivalence class hierarchy of key variables, then construct the hierarchy, so that the most variable classes occur early in the sequence. These classes require little inter-class swapping to preserve anonymity of their respondents. As more levels are added to the hierarchy, more inter-class swapping is necessary to preserve confidentiality. However, the latter classes have the least between-class variation. Therefore, this method should construct a sort order which will "nearly" minimize the amount of distortion to the data.

By incorporating these changes into the 1990 Confidentiality Edit, we believe that the US Bureau of the Census will develop a first-rate disclosure limitation technique. It is our hope that this technique will serve as a benchmark by which other techniques are judged for the year 2000 and beyond.

Preliminary Recommendations for Disclosure Limitation for the 2000 Census: Improving the 1990 Confidentiality Edit Procedure

Richard A. Moore, Jr.
Statistical Research Division
Bureau of the Census
Washington, DC 20233

Abstract

Population censuses are essential to the well-being of any country. Data collected during these censuses measure the rate at which the country is progressing. The success of a census hinges on the ability of the statistical agency to obtain accurate information for each individual (or household). This often requires the Bureau of the Census to release data in a manner which does not compromise the confidentiality of any particular individual's response. To achieve this, the Census Bureau uses certain disclosure limitation techniques. Although these techniques mask the data, they also distort the true statistics. An ideal masking procedure provides useful statistical data, while ensuring each respondent's anonymity.

This paper examines some of the limitation procedures available for the 2000 Census. It takes a close look at the Confidentiality Edit, the procedure adopted for the 1990 Census, and compares it to some of the newer software developed by other statistical agencies. The paper also suggests some additional techniques to measure the extent to which the data has been protected and distorted. By incorporating these into the 1990 Confidentiality Edit, the Census Bureau can develop a top-notch disclosure limitation technique for the initial decennial census of the next millennium.

I. Introduction

Any successful census relies on three principles: (1) the ability of the statistical agency to obtain accurate information, (2) the ability of the agency to process this information in a timely manner, and (3) the ability of the agency to disseminate accurate data to researchers, planners, and policy-making bodies. In order to achieve the first principle, statistical agencies often provide the respondents with a guarantee of confidentiality. In the U.S., for example, Title 13 ensures that the Bureau of the Census releases data in a manner which does not violate the confidentiality of any response. This conflicts with the third principle. The Bureau cannot provide the detailed data which users need without first distorting some individual responses. This induces inaccuracy into the final publicly disseminated products (e.g., microdata sample files and tabulations). Before processing the data, the Bureau must decide on a disclosure limitation strategy. The ideal strategy involves a sufficient amount of protection, while limiting the distortion to essential statistics.

For the 1990 Census, the Census Bureau studied several disclosure techniques. These included suppression (Porter, 1986) and controlled rounding (Greenberg, 1986). Both methods have their advantages and disadvantages. These are explained in greater detail in the referenced material. To mask the 1990 data, the Census Bureau used a method referred to as the "Confidentiality Edit" (Griffin, et al, 1989). This method uses data swapping principles (Dalenius and Reiss, 1982) to mask the file. It preserves the total population count and the total housing unit count for all geographic areas. Counts of particular sub-populations (e.g., sub-populations determined by sex, race, and/or age) are also not altered. This procedure also allows for the production of meaningful tabulations of auxiliary variables (e.g., Mean Income of Hispanic Women, age 60 and above, in Paramus, NJ).

The Confidentiality Edit is not without its critics (Fienberg, 1995). Among the major criticisms are (1) that the procedure offers no guarantee of confidentiality, and (2) that the measure used to quantify the amount of distortion is inadequate. This paper addresses these issues in detail. It suggests enhancements to the procedure that will assure an adequate amount of protection. In addition to the index of dissimilarity, which measures changes in frequency counts, the paper introduces a measure which reflects the bias induced by this procedure to statistics of the continuous variables. It also measures the protection afforded to each individual's responses.

Other international statistical agencies are rigorously attacking confidentiality problems. The German Federal Statistical Agency is presently designing SAFE, an end-load piece of software which ensures that publicly released tables of frequency counts do not violate the privacy of the respondents. Statistics Netherlands has developed ARGUS, a software package for transforming a microdata file into one which has been suitably masked for public release. This paper examines some of the underlying principles of each system and the manner in which they are applied. It suggests methods to incorporate these principles into the Confidentiality Edit system.

Sections II through IV discuss the 1990 Confidentiality Edit procedure, its advantages, and its disadvantages.

Sections V and VI discuss two approaches -- SAFE and ARGUS -- taken by other statistical agencies.

Sections VII through XI discuss improving the 1990 procedure. Section VII states some additional concepts to be incorporated into the procedure; VIII lists the enhancements; IX summarizes the recommendations; and X gives the recommendations in detail. Section XI discusses the use of the mean square bias as a measure of the induced distortion, while Appendix A gives the technical details of such a measure.

Sections XII and XIII discuss test results of the procedure. Tests were performed using the 1993 Annual Housing Survey Public Use Microdata Sample file.

Section XIV summarizes the work. Section XV lists the references.

II. The 1990 Confidentiality Edit Procedure, An Overview

The 1990 Confidentiality Edit is based on the principle, "If you create uncertainty in the mind of an intruder about whether a record has been altered, then you have protected the file." This is achieved by selecting a small sample of census households and swapping their responses with a household which has identical key characteristics, but which is located in a different geographic area in the same state. For the 1990 Census, different key variables were used for the long and short form of the Census. The keys were based on responses in the following categories.

1. State
2. Number of White Individuals in the Household
3. Number of Black Individuals in the Household
4. Number of American Indians or Aleutian Eskimos in the Household
5. Number of Asian Individuals or Pacific Islanders in the Household
6. Number of Individuals of Other Races in the Household
7. Number of Hispanics in the Household
8. Number of non-Hispanics in the Household
9. Number of Individuals over 18 Years of Age in the Household
10. Number of Family Units at the Address
11. Trailer/Mobile Home Designation (Yes / No)

Suppose you divide the census file for each state into two pieces. The first contains a small (e.g., 1 percent) random sample of the households in the Census; the second contains the remaining unsampled records. Choose a record from the random sample, search the remaining records in the sample for a household which agrees on the 11 keys. If one is found, swap the geographic codes; if not, search the records in the unsampled portion for a match. If a match is found, then swap the geographic codes of the in-sample record with those of the out-of-sample record.

How does such a procedure affect the Census counts? Published statistics for the 1990 Census include census counts for total persons, totals by race, by Hispanic origin, and by age 18 and above. The Census also provides housing counts by tenure and by rent and value. None of these statistics is affected by the Confidentiality Edit.

The Census also provides some aggregate statistics at the block level. During testing, it was discovered that only a small percentage of the blocks with less than a prescribed number of housing units were represented in the swap sample. To ensure that any respondent's confidentiality was not violated by the publication of block level statistics, the Census Bureau first identified these blocks and sampled them at a higher rate. (For tabulations based on the long-form sample, it used a blanking out and hot-deck imputation approach to prevent disclosure.)

Due to confidentiality concerns, the exact percentage of records in the sample is not able to be revealed. It was revealed that a match was found either elsewhere in the sample or in the unsampled portion of the universe for 99.7 percent of the sampled records.

Example 1 below demonstrates such a swap. Assume no other household in the sample had the same combination of responses for the 11-key categories, but a match was found in the non-sampled file. The Confidentiality Edit would swap the geographical codes (i.e., the Place) on the two records. The population and housing counts for all geographic areas would remain unaltered. The values of auxiliary statistics such as Income and Home Value statistics would be distorted within geographic areas. Distributions of categorical variables (such as occupation) within these geographies would also be slightly changed. It was hoped that the small sampling fraction would protect confidentiality, while not drastically distorting statistics derived from non-key variables.

Example 1. Data Interchange Methodology for the 1990 Confidentiality Edit

	Before Swap		After Swap	
Id No.	001	118	001	118
File	1% Sample	99% Non-Sample	1% Sample	99% Non-Sample
State	NJ	NJ	NJ	NJ
# Whites	9	9	9	9
# Blacks	0	0	0	0
# AI/ AE	0	0	0	0
# Asian/PI	0	0	0	0
# Others	0	0	0	0
# Hispanics	0	0	0	0
# non-Hispanics	9	9	9	9
# Over 18	4	4	4	4
# Units	2	2	2	2
Trailer/ Mobile Home	NO	NO	NO	NO
Place	Princeton	Trenton	Trenton	Princeton
Occupation	Lawyer	Clerk	Lawyer	Clerk
Income	95,000	25,000	95,000	25,000
Home Value	250,000	125,000	250,000	125,000

Refer again to Example 1. The procedure is equivalent to the "blank out and impute" technique for masking microdata files, especially if you consider it equivalent to the following procedure.

1. Divide the file into two disjoint segments: (1) a one percent sample and (2) a 99 percent non-sampled portion.

2. Duplicate the one percent sample.

3. "Blank out" some highly visible categorical field(s) (e.g., the geographic codes) on each record of the duplicated portion file.

4. Use the hot-deck imputation approach to impute for the blanked out field(s) on the duplicated file. Never use the original version of the record as its own donor. Do not use the same record as a donor for more than one record.

5. "Blank out" the donated fields on the donor records. Hot deck impute, using the original one-percent sample as a donor file.

6. Reassemble the file, replacing the original records with their imputed counterparts.

III. Attractive Features of The 1990 Confidentiality Edit

The 1990 Confidentiality Edit has several attractive features. They are described in more detail below.

1. For a select number of critical statistics, the population and housing unit counts within geographic areas were not altered. All statistics mandated by law were preserved.

2. As mentioned above, the 1990 Confidentiality Edit is similar to a blanking out and imputing procedure, an acceptable method of data masking (Subcommittee on Disclosure Limitation Avoidance of the Federal Committee on Statistical Methodology, 1994).

3. The method was easy to program and implement. It involved no more than matching a small number of records on 11 keys and interchanging a few geographic codes.

4. For the entire universe, all frequency distributions were preserved. In addition, the multivariate relationships between continuous variables were not altered.

5. Since the percentage of swapped records was relatively small, analysis on sub-domains (e.g., the residents of Princeton, NJ) with a relatively small number of respondents (e.g., about 20) should not be adversely affected.

To measure the analytical effect of the Confidentiality Edit, the Census Bureau calculated a number of distributions for demographic data at small geographic levels. These calculations reflect the characteristics of the geographic levels before and after the swapping operation. The distributions were compared using the index of dissimilarity, or D-statistic. It is described below.

Let X_i = the count before the data interchange operation in the i-th row.
Let Y_i = the count after the data interchange operation in the i-th row.
Let r = the number of rows in the table.
Let $X = \sum X_i$ and $Y = \sum Y_i$.

Then D, the index of dissimilarity, caused by the Confidentiality Edit is

$$D = \frac{1}{2} * \sum_{i=1}^{r} \left| \frac{X_i}{X} - \frac{Y_i}{Y} \right|.$$

This index can be interpreted as the net proportion of the total which has been changed by the swapping operation. As one can see, the calculation of this statistic is a straight-forward procedure. Example 2 illustrates this.

Example 2. The D-Statistic for Age Group Change in a Given Block

Age Range	Before the Edit	After the Edit
Under 5	3	2
5 to 17	4	5
18 to 39	4	3
40 to 64	5	7
Over 65	4	3
Total	20	20

D = 0.15

IV. Unattractive Features of the Confidentiality Edit

The Confidentiality Edit also has some unappealing features. Below are some of the major concerns of the procedure.

1. The Census Bureau swapped only a small percentage of records. It sampled smaller population groups more heavily, but took no precautions to guarantee that the unique individuals in these (or larger) groups were more likely to be in the sample.

2. The Census Bureau never really quantifies how reliable the statistics are. This is particularly true of continuous data. Although the D-statistic gives one a measure of how certain frequency counts have been altered, it does not handle the problem of auxiliary statistics on such sensitive variables as household income.

3. The Census Bureau never really quantifies how safe the data are. The Confidentiality Edit is based on a data-swapping technique of Dalenius and Reiss (1982), yet the former method differs significantly from the latter. Hence, all the confidentiality claims made by Dalenius and Reiss do not necessarily apply to the Confidentiality Edit procedure. The Census Bureau must prove that their method provides a sufficient amount of protection.

Other international statistical agencies have also addressed similar concerns. Sections V and VI illustrate the approaches of the German Federal Statistical Agency and Statistics Netherlands. Section VII suggests enhancements to the present procedure which will somewhat alleviate these problems.

V. The SAFE Approach to Confidentiality (German Federal Statistical Agency)

SAFE (Standardierte Anymisierungs-Functionen bei Einzelangenaben, or Standardized Anonymization Functions for Microdata) is the name given to the approach taken by the German Federal Statistical Agency (Appel,et al date unknown). The project began in 1989, when the demands for special tabulations became a burdensome problem for the centralized agency. As a result, a cooperative effort of more than 50 cities and municipalities began to develop a personal computer-aided system which satisfied basic disclosure limitation requirements as well as produced tabulations suitable for analysis.

Following the 1987 population census in Germany, each city and municipality was allowed to receive census microdata records for its inhabitants. Each local area required this data for the research, analysis, and evaluation of past and present policy on the infrastructure and quality of life of its inhabitants. Results of such studies would allow elected officials to draft plans and policies which would allow the municipality to progress in a positive direction.

In order to receive this microdata, each municipality had to commit itself to the same disclosure limitation standards as the central statistical office. In addition, the central office had the responsibility of coordinating the effort and designing a general procedure which would satisfy the needs of each local government. The following is the approach taken.

1. Each municipal office receives a complete set of microdata records for its inhabitants. These records will not be masked in any way.

2. Any individual who uses the records for research must be sworn to privacy.

3. Using the original data, the sworn individuals determine the statistics necessary to analyze and evaluate past and present policy and to suggest a future course of action.

4. For documentation, researchers may want to release some of these tabulations to the public. The principal publicly-released product will be tabulations of n-dimensional frequency counts. Tables must be published so that each non-zero valued cell appears to have at least a minimum number of respondents. This number is referred to as the tolerance.

5. SAFE is "end-load" software, which examines and masks these tabulations. Tabular values are disturbed so that all interior cells below the tolerance are set to either zero or a value above the tolerance. Values in other cells are adjusted so that all marginal totals are retained. In Example 3 below, a municipality has 134 households. A researcher cross-tabulates the regions (North, East, South, and West) of the municipality and by highest level of education of head of household. Assume a tolerance of 5 units per cell is used. Below are listed the original and the "SAFE-adjusted" tabulations.

Example 3. The SAFE Approach to Protect Frequency Counts
Before SAFE

	Education Level -- Head of Household				
	HS or Below	**2-Year Degree**	**4-Year Degree**	**Adv. Degree**	**Total**
North	19	10	10	15	54
East	4	10	10	1	25
South	11	14	8	2	35
West	15	1	2	2	20
Total	49	35	30	20	134

After SAFE

	Education Level -- Head of Household				
	HS or Below	**2-Year Degree**	**4-Year Degree**	**Adv. Degree**	**Total**
North	18	11	10	15	54
East	5	10	10	0	25
South	11	14	10	0	35
West	15	0	0	5	20
Total	49	35	30	20	134

The SAFE approach is equivalent to masking re-identifiable records (i.e., those in cells with counts below tolerance) in the following way. Think of each dimension of an n-dimensional table as a key variable.

1. Find a similar record that matches on n-1 of the key variables.

2. Blank out the value of the n-th variable on one of the two records.

3. Use the other record to hot-deck impute.

4. Retabulate the file, holding the marginals fixed.

5. Find two marginal totals that are now unbalanced with the sum of their interior cells. Choose one record from each of these interior cells. Repeat Steps 1, 2, and 3 to rebalance the sums.

6. Repeat Steps 4 and 5 until all marginals are balanced.

7. Choose another record from a cell whose total is below tolerance. Repeat Steps 1-6.

8. Repeat Step 7, until no cells with below tolerance counts exist.

SAFE was designed solely to adjust tabular frequencies. It was not designed to adjust individual microdata records. Suppose one wants to use the above procedure to adjust the microdata. He has an n -dimensional table with "t" cells which are below tolerance. The frequency counts for these cells are $k_1, k_2, ..., k_t$. Even the most efficient program could take as many as $(k_1 + ... + k_t) * 2^n$ imputations.

Although one can adjust the frequencies on a table very quickly, it is a much more laborious job to adjust the records so that they tabulate to a given table. Such a microdata swapping procedure would not be feasible for the 100 million household records in the census file. Let us now turn our attention to the ARGUS approach.

VI. The ARGUS Approach to Confidentiality (Statistics Netherlands)

The ARGUS system is a disclosure limitation software package developed by Statistics Netherlands (Pieters and de Waal, 1995). It was developed for the purpose of protecting a microdata file in a systematic and efficient way. It uses the following approach.

1. Assume any intruder is able to obtain accurate information about certain targeted individuals for only a limited number of variables. One can quickly determine and list all key combinations for which an intruder might be able to obtain complete information.

2. Disclosure risk is inversely related to the number of observations in each key combination. Set a level for the amount of disclosure risk which you can tolerate. This amount corresponds to the minimum number of observations per key combination cell. Hence, one can define the tolerance level as the minimum number of respondents per combination in order for unperturbed data and statistics to be released.

3. Identify all combinations which fail tolerance.

4. Identify the records of all respondents for these combinations.

5. Perturb each record which falls in a below-tolerance level combination. This can be done in a variety of ways (e.g., collapsing codes, truncating values of key variables to a certain number of significant digits, suppressing information in cells). After perturbation, all modified key combinations satisfy tolerance and are allowed to be published.

Before invoking the ARGUS system, the user must produce three input files. These include (1) the original microdata file, (2) a parameter file containing the record layout and the tolerance level, and (3) a file of all sensitive key combinations. When invoked, the system queries the microdata set for key combinations which fail tolerance. It then produces (on screen) a list of these combinations and the number of respondents for each.

Based on his on-sight analysis, a masking agent "suggests" ways in which the microdata file can be modified. He may decide to recode variables (e.g., instead of many different classifications for RACE, he may use only WHITE and NON-WHITE), he may truncate values (e.g., the last digit of AGE or of INDUSTRIAL/OCCUPATION CODE may be truncated), or he may choose to suppress as many values as necessary to achieve tolerance. The system then stores the latest battery of suggestions.

The ARGUS system is extremely attractive because the masking agent interacts with the system. After each series of suggestions, the system modifies the original file and reproduces a list of combinations which fail tolerance. A skilled individual can immediately recognize the effects of his latest disclosure limitation procedures. If not satisfied with a particular modification (e.g., a 4-digit OCCUPATION CODE should be truncated to 2 digits rather than 3, and/or AGE should be collapsed into ranges instead of truncated), he can change the existing modifications.

Let's assume that the agent has set certain collapsing and truncation patterns, but the modified file still contains key code combinations which are below tolerance. The masker can invoke the "automatic suppression" routine. The system will then use linear programming techniques to suppress the minimal number of values necessary to meet tolerance. When the system can suppress one of several variables to mask a respondent, it refers to a hierarchy (pre-defined by the agent) on the importance of each variable. It always suppresses the value of the least important variable, so the user has some control over the type of information that gets suppressed.

VII. Concepts for Enhancing the 1990 Confidentiality Edit

Both the SAFE and the ARGUS systems identify records in cells which fail tolerance. While SAFE tabulates to identify risky key combinations, ARGUS assumes the subject matter area can identify them beforehand. ARGUS has several major advantages over SAFE.

1. Assume one can quickly sort the file on the key variables. He can use sequential processing to readily identify risky key combinations.

2. Assume a record falls in a high risk combination. ARGUS allows the user to prioritize the key variables. If more than one of the key combination values can be suppressed to protect the record, ARGUS chooses the field with the lowest priority.

Combine these two concepts with the following SAFE concept.

3. Assume a record falls in a high risk cell in an n-dimensional table. Think of each dimension as a key variable. By perturbing the value of one carefully chosen key variable, one can protect the identity of the respondent.

We hope to use these concepts to enhance the masking ability of the 1990 Confidentiality Edit. The Census Bureau provides two types of public-use products. The first type is the microdata sample file which contains information from the long form sample. Each file contains 5 (or less) percent of the population. The second is tabulations based on the entire population. Some of these tabulations contain such fine detail that some of the interior cells may contain only one or two respondents. The Census Bureau is interested not only in disclosing that the frequency is small, but in prohibiting intruders from deriving information from auxiliary statistics about some particular individual in one of these cells (such as Average Income or Average Home Value for a block of three households). ARGUS techniques quickly identify high risk disclosures, while SAFE techniques are better suited to mask them. Can we form some hybrid technique that addresses both simultaneously?

VIII. Possible Enhancements to the 1990 Confidentiality Edit

Consider the following approach. First, suppose the Census Bureau determines that statistics for cells with less than Q respondents are too risky to be publicly released. However, if some perturbation is done to one or more of its respondent's data, then statistics for the cell can be released. This was one of the basic principles of the 1990 Confidentiality Edit. Second, assume a cell has less than Q respondents. A 5 percent microdata sample is randomly chosen. The probability that any given individual in that cell is contained in the sample is less than Q/20. Assume one individual in the block is chosen and that his data has been distorted, then the probability that the one of the cell's contributors is in the sample and the information accurately reflects one of its respondents is less than or equal to (Q-1)/20. For low values of Q (e.g., $Q < 8$), this means the microdata is very likely to be incorrect or not in the microdata sample file.

Assuming the above is a reasonable approach, we can use the concepts of Section VII to devise a procedure which will identify below-tolerance combinations and simultaneously alter at least one record in each. This should provide adequate protection for both the detailed tabulations and the microdata samples. The procedure, documented below, should accomplish this.

1. Suppose the microdata file has "k" key variables (KV_1, KV_2, ..., KV_k), all of which can be readily and accurately obtained by the most determined intruder. Sort the file by the key variables: KV_1 by KV_2 by ... by KV_k.

2. Read the file sequentially. Start by initializing the record counter at 0.

3. As each record is read,

   a. if the key combination agrees with that of previous record, increment the counter by 1;

   b. if the key combination differs, check the counter.
      (1) If the counter is Q, the tolerance, or more, reset the counter to 1.
      (2) If it is less than Q, swap some information with the previous record. You have now altered both the previous and current key combination statistics. Since there is no need to alter further records in the current combination, "re-initialize" the count to Q.

IX. Concerns and Recommendations for the New Confidentiality Edit Procedure

The above procedure requires that the Census Bureau address the following 5 concerns.

1. What is an appropriate value for the tolerance, Q?

2. What is the maximal set of key variables readily available to the most determined intruder?

3. What information should be swapped, if a combination fails tolerance?

4. Suppose a record must be swapped with one in a different key combination. The resulting file will be most analytically useful, if the continuous fields on the two records should be approximately the same. Is there any order in which the key variables can be sorted to ensure this?

5. How badly can the statistics be distorted in order to still be considered useful?

We make the following recommendations.

1. **Tolerance.** We recommend that a tolerance of Q=3 will sufficiently protect the data. This means that cells with one or two respondents will be perturbed, thus making it impossible for any individual to derive another's data correctly. For cells with n=3 or more respondents, an intruder who knows the data of one respondent will only be able to derive the aggregate statistics FOR the other n-1.

2. **Key Variables.** The set of key variables must contain all combinations which preserve counts mandated by law. Variables by which data is cross-tabulated must also be included. Geography need only be as fine as areas by which tabulations are produced (e.g., one-digit block group).

3. **Field Swapped.** We recommend that geographic codes be swapped. Not only did this work well for the 1990 Census, but we feel that swapping this variable provides the most protection. Its only short-coming is that it distorts SOME sub-state level statistics of auxillary data, but most swapping can be done within the tract.

4. **Sort Order for Key Variables.** To ensure the sub-state level statistics are distorted as little as possible, use an iterative method to determine the "near-optimal" order.

5. **Data Distortion.** Suppose we have an arbitrary data set with mean, $\bar{x}$, and standard deviation, s. Suppose that we want the swap so that the following two conditions are met.

   (1) For protection, we want to swap so that the average swapped value changes by about p percent.

   (2) The standard deviation is the amount of dispersion in the unswapped set. To preserve data utility, we want to increase the dispersion by no more than a factor of (1 + f). For some important continuous variable, calculate the mean square error of each observation from its value after the swap.

The data will be protected and not overly distorted if

$$(\frac{p}{100} * \bar{x})^2 < \frac{1}{N} * \sum (x_i - x_i')^2 < (2f+f^2) * s^2.$$

The next section will supply detailed explanations for each of the above recommendations.

X. A Detailed Explanation of the Recommendations

The previous section addressed concerns about a general Confidentiality Edit procedure. The reccomendations were brief. This section seeks to clarify these concepts.

An Appropriate Value for Q. The Census Bureau must distort the data so that no household's responses can be derived from the published statistics. Any individual can identify the cells in which his household falls. By subtracting its responses from the statistics, he can obtain statistics for the other Q-1 households. If Q = 2, each household in that cell can obtain statistics on the other. This number will be accurate unless the data of at least one of them has been perturbed. Hence Q must be 3 or more. Is it necessary that Q be greater than 3? We doubt it. Larger values of Q are required only if the data is tremendously skewed. Household data is rarely skewed enough to require more than 3 responses for adequate protection.

RECOMMENDATION: Use for a tolerance, Q=3.

Maximal Set of Key Variables. Swapping will only be necessary when tolerance is violated for tabular publications. The dimensions of the table can be thought of as key combinations. These identify households with unusual characteristics within a given area (e.g., The number of female Asian doctors in North Dakota is 1.) Low frequency counts for small areas do not necessarily pose a problem. The problem occurs when a frequency count for cell is low and the table divulges accurate aggregate information.

RECOMMENDATION: Re-examine the type of cross-tabulations published. Limit them, if possible. Do not try to get too fine with the geographic areas published. Otherwise expect a large number of swaps for rare characteristic households.

Information to be Swapped. There are several alternatives for swapping data. These include: (1) swapping only the sub-state geographic codes, (2) swapping responses of all continuous data items, or (3) swapping the first categorical response which causes uniqueness. Example 4 illustrates each of the 3 swaps. The swapped values are bolded.

Example 4. Various Swapping Algorithms

	Original		Method 1		Method 2		Method 3	
	Rec 1	Rec 2	Rec 1	Rec 2	Rec 1	Rec 2	Rec 1	Rec 2
State	NJ	NJ	NJ	NJ	NJ	NJ	NJ	NJ
Whites	9	10	9	10	9	10	**10**	**9**
Blacks	10	0	10	0	10	0	10	0
...								
Place	Trent	Prince	**Prince**	**Trent**	Trent	Prince	Trent	Prince
Income	100	250	100	250	**250**	**100**	100	250
Home Value	150	400	150	400	**400**	**150**	150	400

Method (1) slightly distorts both the categorical frequencies and the statistics of continuous data for the sub-state geographic areas. It does retain the statistics of all subsets. Method (2) retains the categorical frequencies at the sub-state level. It distorts sub-state level statistics and statistics of subsets defined by categorical variables. Method (3) retains sub-state level statistics, but distorts the categorical frequencies at this level.

All methods retain the means and variance-covariance relationships at the state level. Methods (1) and (2) provide reasonable protection to the data. Method (1) gives a false location, thus discouraging an intruder from obtaining a match, while Method (2) gives false information if a match is obtained. Method (3) provides the least amount of protection. Suppose an intruder was uncertain about the number of Whites in a housing unit in Trenton, NJ, but he knew that it was the only one with 10 Blacks. He could correctly identify it as Record 1 and obtain the correct household income and home value. For this reason, we would discourage the use of Method (3).

Is either of the other two methods more favorable? Method (1) retains the state-level statistics but not the frequencies for all categorical subsets, while Method (2) distorts these statistics but retains the frequencies. The most desireable trait is strictly up to the discretion of the subject matter analyst.

Method (1) may be favorable to Method (2) in that the former actually masks the identity of the high risk respondents. Method (2) does not really attempt to conceal the identity of the high risk respondents, it just distorts the continuous data. By cleverly assigning the sort order of the key variables, one can alleviate some of the differences between the two methods. If the two distort data in a similar way, Method (1) is superior over Method (2), since it actually masks the respondent's identity.

RECOMMENDATION: Use Method (1). It masks the file better than Method (2). When a record is marked for a swap, find a record with similar key variables and swap the sub-state level geographic codes.

Assigning the Sort Order of Key Combination Variables. Although the major objective of the swapping operation is to mask respondent's identity, it is necessary that it not diminish the analytic utility of the file. More specifically, if the file is to retain its analytic validity, the values of the continuous variables on the records swapped should differ as little as possible. Records are listed in key combination order. A swap occurs, only if a key combination is below tolerance. In this case, the final listed record of the below-tolerance class is swapped with the initial record of the succeeding class. Suppose one attempts to list the key combination classes so that each class has approximately the same means as its predecessor and successor class. If one can accomplish this, then sub-domain statistics should not be significantly distorted. How should one proceed?

Recall the key variables are denoted KV_1 , KV_2, ..., KV_n and the key combination classes are expressed by the n-dimensional vectors $(a_1, a_2, ..., a_n)$. The value of a_1 alone should define very few below tolerance classes. If it does, maybe the acceptable responses for data item KV_1 should

be collapsed. The (a_1, a_2) combination segregates each of the a_1 classes into smaller groups. Again, there should be very few below tolerance level classes at this second level. If there are, one should consider collapsing the acceptable responses for variable KV_2. The same occurs at the third level with (a_1, a_2, a_3), the fourth level, etc. At each level we expect to see additional combinations with counts below tolerance. The number of extra tolerance failures should increase as we add levels (e.g., the first level may have 0 failures, the second level 1 failure, the third level 3 failures, the fourth level 10 failures, ...). Most swaps will be caused at the n-th level. It is at this level that we must exert the most control over the variation between classes. In a step-discrimination-like procedure, the following iterative procedures constructs a key combination order which attempts to minimize this variation.

Step 1. Choose a continuous variable, X, with which all highly sensitive continuous variables will be reasonably well correlated. We will attempt to minimize variation on this variable and hope that the variation for all other continuous variables follow.

Step 2. (a) Calculate the within-class variance of X for each value of a_1 for field KV_1.

(b) Call the maximum of these variances $MVAR_1$.

(c) Perform (a) and (b) for each of the fields $KV_2, KV_3, \ldots$.

(d) Suppose $MVAR_{j1} = \min \{ MVAR_1, MVAR_2, \ldots, MVAR_n \}$. KV_{j1} will be the last key variable in the swap sequence.

What is the motivation behind this choice of variables? At worst, the ordering of the first n-1 key variables sorts the file so that the rank of the x_i are random. By forcing the within-class variance of the a_i to be bounded by as small a number as possible, you have limited the maximum amount of distortion caused by a swap at the final stage.

Step 3. Perform the following operations for every key variable except KV_{j1}.

(a) Calculate the within-class variance of X for each combination of (a_1, a_j) for field KV_1 by KV_{j1}.

(b) Call the maximum of these variances $MVAR_1$.

(c) Perform (a) and (b) for each of the n-1 fields KV_2 by KV_{j1}, KV_3 by KV_{j1}, ...

(d) Suppose $MVAR_{j2} = \min \{ MVAR_1, MVAR_2, \ldots MVAR_{j-1}, MVAR_{j+1}, MVAR_n \}$. KV_{j2} will be the next to last key variable in the swap sequence.

The motivation here is similar to that of the previous. At worst, the first n-2 variables in the sort list the x_i in random order. We want to minimize the variation at the last 2 steps.

Succeeding Steps. Modify Step 3 by adding 1 level at a time. Continue until the order of all n levels are determined.

Although this algorithm does not guarantee a "patented perfect" solution, it does determine a procedure for defining a sort order which logically may limit distortion of the final statistics. Keep in mind that there are some deficiencies in the process. These are listed below.

(1) Only one continuous variable can be used to determine the order. It assumes that there exists one which is relatively highly correlated with most continuous key variables. If none exists, one may have to "calibrate" by the most sensitive variable (e.g., income).

(2) This approach assumes that the values of the variable are randomly ranked by each of the previous sorts. Any correlations between the key variables chosen in the first k steps and those still remaining to be assigned in the order are lost.

(3) We have to develop a measure to determine whether the sort order is, in fact, optimal or close to optimal.

(4) By attempting to minimize the distortion to the resulting file, the method is only interested in preserving the file's analytic utility. Continuous variables can also be used to re-identify respondents. Should the swap be designed to guarantee a certain amount of distortion?

RECOMMENDATION: Swapping can seriously distort statistics for individual records. A great deal of care must be taken to ensure the perturbed file retains its analytic utility. Invoking the procedure should reduce the amount of distortion to individual records.

An Acceptable Amount of Distortion. The issue of an acceptable amount of distortion is actually a two-part problem. First, one must induce a sufficient amount of noise to protect the data in senstitive cells. Second, too much distortion renders the file analytically useless. Some easy to implement criteria are needed to place bounds on the distortion.

Criterion 1 (Lower Bound). Assume that we want to swap so that the swap distorts the average individual record's ~~data~~ by about p %. P_{BIAS} is the average percentage change over the entire sample; but only a small number, call it n_s, are swapped from the universe total, N. Thus, we actually want to restrict the calculation of P_L to those records involved in the swap. So,

$$P_L = p * \frac{n_s}{N}.$$

For example, if p =10 percent, then

$$P_L = 10 * (n_s / N).$$

Criterion 2 (Upper Bound). Let s and $\bar{x}$ be the standard deviation and mean of some continuous variable. Suppose we want the swapping to induce an error of at most (1+f) times the standard deviation. The swapping is nothing more than an induced error. For the data user, this bias compounds errors caused by sampling. For the typical random sample from an unmasked universe, the only quantifiable error is sampling error, s. Now suppose we swap data in such a way that the average piece of data is disturbed by the bias, BIAS. The new net error, z, is related to the previous two by the equation

$$z^2 > s^2 + BIAS^2,$$

where

$$BIAS^2 = \frac{1}{N} * \sum (x_i - x_i')^2.$$

We want to induce bias, so that the resulting net error is no more than (1+f) times the original standard deviation. The above equation becomes

$$(1+f)^2 * s^2 > s^2 + BIAS^2.$$

This means

$$BIAS^2 < (2f + f^2) * s^2.$$

Hence the result.

XI. Using the Mean Square Bias as a Measure of Distortion

By the work in Section X, we have determined the bounds on the mean square bias, BIAS. By construction of the bounds, we have seen that the BIAS is directly related to the amount of distortion. The more distortion induced, the higher the value of the BIAS. However, the magnitude of the BIAS is also a function of the magnitude of the continuous variables, $\{x_i\}$. A more readily intuitive measure is the PERCENTAGE BIAS, defined as 100 percent times the BIAS divided by the mean of x. This measure is equivalent to the coefficient for variation. It can be used as a measure in two ways.

The PERCENTAGE BIAS as a Measure of Distortion. The bounds for a suitable amount of distortion can also be expressed in terms of the PERCENT BIAS, P_{BIAS}, as follows.

$$\therefore p * \frac{n_s}{N} < P_{BIAS} < 100 * (2f + f^2)^{1/2} * \frac{s}{\bar{x}}.$$

Thus, if one knows the mean, standard deviation, and a suitable value for f (the maximum proportion of additional distortion to be added), he can determine whether the swap masks the file sufficiently while not distorting its analytical utility.

The PERCENTAGE BIAS as a Measure for the Optimal Sort. The PERCENTAGE BIAS can also be used to ensure that the sort on the key variables is near optimal. Recall that the ideal optimal sort order minimizes the expected square of the pointwise differences. This is nothing more than minimizing the BIAS and the PERCENTAGE BIAS. If one suspects the sort is not optimal, he need only sort the file in a different order and recalculate the PERCENTAGE BIAS.

We have already noted that the algorithm which determines sort order is not fail-safe. However, it is relatively quick and easy to implement. It should give near-optimal results. Therefore, it may be possible to find a different sort order which gives a slightly lower bias. However, the algorithm which identifies this order may be much more involved and require considerably more time and computer resources to implement.

XII. Testing the Enhancements

In Section VIII, we recommended a procedure which may enhance the masking ability of the 1990 Confidentiality Edit. In this section, we will describe the procedure used to test the performance of this technique. We tested the procedure on the 64,998-record Annual Housing Survey Public Use Microdata Sample. The following fields were extracted from each record:

(1) IDNUM - a unique 12-digit identifying number,
(2) REGION - region in which the housing unit was located,
(3) BEDROOMS - the number of bedrooms in the unit,
(4) BATHS - the number of bathrooms in the unit,
(5) YR_BLT - the year in which the unit was erected,
(6) INCOME - annual household income,
(7) HOME_VAL - the market value of the housing unit,
(8) MORTGAGE - monthly mortgage payment,
(9) MAINTAIN - annual maintenance cost, and
(10) TAXES - monthly property taxes.

We use Field (1) strictly for identification purposes, Fields (2) through (5) to determine relatively unique key combinations, and Fields (6) through (10) to measure the effect of the enhancements and the sort order on the statistics of the resulting file. The following testing procedure was implemented.

(1) Determine the "near-optimal" sort order for the four key combination variables (REGION, YR_BLT, BEDROOMS, and BATHS) and the continuous variable (INCOME).

(2) Extract IDNUM and the key variables to a separate file.

(3) Sort the key variables in "optimal" order.

(4) Process the sorted file in sequential order. Each different key combination is considered a separate equivalence class. For every equivalence class whose count fails tolerance, swap the REGION of the last record in that class with the REGION of the first record in the next class.

Notes: (a) Because the Annual Housing Survey is a 1 in 1500 sample, we set the tolerance to Q= 2 (i.e., only unique key combinations were targetted for swap). For the 2000 Census, we may want to set the tolerance to 3 or more. See Section VIII, "An Appropriate Value for Q".

(b) In Section X, "Information to be Swapped," we stated three possible masking swaps. (See also Example 4.) We have tested only Method (1), swapping of the geographic codes. This was chosen for two reasons. First, this was the procedure used in the 1990 Confidentiality Edit. Second, as stated in Section X, we feel this alternative best masks the data.

(5) After swapping the necessary geographic information, update those codes on the public use file.

(6) Calculate the resulting BIAS and PERCENTAGE BIAS. Use the equivalence classes determined by the key combinations. Ensure the file is not severely distorted.

XIII. Performance Testing Results

The following results were obtained for some of the various stages of the testing procedure outlined in Section XII.

Determination of the "Near-Optimal" Sort Order. First, let's illustrate the technique to determine the "near-optimal" sort order. In Section X, we outlined an iterative procedure. This calculated within-class variances at various levels and took the minimum of the maximum at each level. One would then add the variable corresponding to this value to the hierarchy. Afterwards, one would proceed with analysis on the "unused" variables restricted to the hierarchy, already in place. The following "walks" the reader through the four steps. The chosen variable at each step is marked by a double asterisk (**). All maximum within-class variances are listed in Table 5.

Table 5. Maximum Within-Class Standard Deviations

	Maximum Standard Deviations				
CLASS	INCOME	HOME VAL	MORTGAGE	MAINTAIN	TAXES
REGION	17,873	95,039	405	703	14.8
YR_BLT	22,430	92,397	469	867	17.2
BEDS	41,088	109,488	1,078	1,651	19.0
BATHS	36,133	148,327	796	1,427	22.2
REG-YR	26,920	104,148	507	1,618	19.9
REG-BEDS	44,378	147,507	1,051	3,536	30.3
REG-BATHS	47,061	176,777	822	1,915	30.0
YR-BEDS	57,983	208,597	1,078	3,536	38.2
YR-BATHS	57,426	187,383	886	4,738	31.1
BEDS-BATHS	63,640	141,421	1,144	3,536	30.4
REG-YR-BEDS	67,104	219,203	1,074	2,858	38.2
REG-YR-BATHS	70,658	223,446	1,124	4,738	43.1
REG-BEDS-BATHS	70,427	187,144	1,051	3,536	42.4
YR-BEDS-BATHS	70,711	194,454	877	3889	38.9

Step1. Calculate the maximum within class standard deviations for the variable INCOME. Choose the minimum of these maximum.

REGION 17,873 **
YR_BLT 22,430
BATHS 36,133
BEDROOMS 41,088

Step 2. Calculate the maximum within-class standard deviations for the classes REGION by _____. Choose the minimum of the maximum of these.

REGION by YR_BLT 26,920 **
REGION by BEDROOMS 44,378
REGION by BATHS 47,061

Step 3. Calculate the maximum within-class standard deviations for the classes REGION by YR_BLT by _____. Choose the minimum of the maximum.

REGION by YR_BLT by BEDROOMS	67,104 **
REGION by YR_BLT by BATHS	70,658

STEP 4. By default BATHS is the final variable chosen. The "near-optimal" sort order is then the reverse of the order selected or BATHS by BEDROOMS by YR_BLT by REGION.

For this data set, the "near-optimal" sort order which minimizes the PERCENTAGE BIAS of INCOME was determined to be

(1) BATHS by BEDROOMS by YR_BLT by REGION.

Refer to Table 5. This gave a PERCENTAGE BIAS (average pointwise bias divided by the mean of the variable) of 17.95 percent. Several other sort orders were tested. These include

(2) BATHS by BEDROOMS by REGION by YR_BLT,
(3) BEDS by BATHS by YR_BLT by REGION, and
(4) REGION by YR_BLT by BEDROOMS by BATHS.

Order (1) was determined to be "near optimal" for income. Orders (2) and (3) test the effect of juxtapositioning one adjacent pair of fields from the "near optimal" sort. Order (4) tests the effect of using the reverse of the "near optimal" order. The resulting PERCENTAGE BIAS's are listed in Table 6 below. It also contains PERCENTAGE BIAS's for the other continuous variables.

Although Order (1) appears to be "near optimal", the bias of Order (2) is not much larger. Of particular interest is the performance of Order (4). The reverse of the "near-optimal" order actually performed better than the "near-optimal" order. Upon further inspection, one would find that the means for the REGION-BATHS, REGION-BEDROOMS, and REGION-BATHS-BEDROOMS combinations are much more distorted by Order (4) than by Order (1). However, suppose equivalence classes are defined by the values of YR_BLT. Orders (1) and (4) give similar biases.

Introducing the YR_BLT into the equivalence class definition has some equalizing effect on the bias. It is not immediately obvious why this occurs. However, the key variable, at which the uniques are identified, determines where the swap occurs. For Order (4), the REGION-YR_BLT combination does not define any unique households in the sample. When BEDROOMS are added to the definition, about 30 percent of the uniques appear. The maximum within-class standard deviation of income for these classes at this level is 67,104. For Order (1), most uniques (about 35 percent) are identified at the BATHS-BEDROOM-YR_BLT combination. The maximum within-class standard deviation here is 70, 711. Consequently, Order (4) has a lower bias.

Table 6. Percentage Bias for Various Sort Orders and Various Continuous Variables For Equivalence Classes defined by BATHS...BEDROOMS...YR_BLT...REGION

	Sort Order			
	BATH..BED.. YR_BLT. REG	BATH..BED.. REG..YR_BLT	BED..BATH.. YR_BLT..REG	REG..YR_BLT ..BED..BATH
Variable				
INCOME	17.95	18.15	23.40	15.90
HOME VAL	6.05	7.10	8.40	7.50
MORTGAGE	3.95	3.60	4.90	4.10
MAINTAIN	23.15	23.75	23.60	17.30
TAXES	6.45	6.65	7.70	5.80

Perhaps a better measure of "optimality" is the average of the PERCENTAGE BIAS's listed in Table 7. This would eliminate situations such as the one mentioned above. Indeed, sub-domains will probably not be defined at our "building block" level, but rather as combinations of these blocks or even based on some other "arbitrary" criteria. Consequently, as the Census Bureau did with the index of dissimilarity, base the amount of distortion induced by a swap on average of PERCENTAGE BIAS's over several levels and not just the reading at the finest level.

Table 7. Percentage Bias for Income For Various Sort Orders and Various Equivalence Classes

	Sort Order			
	BATH .BED. YR_BLT.REG	BATH.BED. REG.YR_BLT	BED.BATH YR_BLT.REG	REG.YR_BLT .BED.BATH
Equivalence Class Definition				
REG-BATH	2.95	4.30	3.70	43.70
REG-BED	2.85	4.50	1.60	41.50
REG-BATH-BED	7.30	9.95	6.20	13.40
REG-YR-BATH-BED	17.95	18.15	23.40	15.90
REG-YR	2.65	2.60	3.20	0.00
REG-YR-BED	12.20	12.90	12.90	11.90
REG-YR-BATH	13.35	13.35	16.70	14.00
AVERAGE	8.46	9.39	9.67	20.06

Calculation of a Range for the PERCENTAGE BIAS. In addition to determining the "near optimal" sort order, we have to ensure that the swap has accomplished its purposes : (1) adequately masking the data, and (2) not severely distorting the analytic utility of the file. We ensure this by confirming that the value of P_{BIAS} falls within certain prescribed limits. From Section XI, we have calculated bounds on the acceptable range for PERCENTAGE BIAS, namely

$$p * \frac{n_s}{N} < P_{BIAS} < 100 * (2f + f^2)^{1/2} * \frac{s}{\bar{x}}.$$

Here

- n_s = the number of records on which data has been swapped;
- N = the number of observations in the universe;
- f = the maximum percentage by which we are willing to increase an arbitrary subset's variation;
- s = the variance of the universe;
- $\bar{x}$ = mean of the universe; and
- p = average percentage distortion to each record.

Example. Suppose we let p =10 and we swap information on uniques from the 1993 Annual Housing Survey Public Use Microdata Sample file. We use the variable INCOME to determine the optimal swapping order (BATHS, BEDROOMS, YR_BLT, REGION). It has a mean and variance of 11,369 and 16,572 respectively. From this we obtain the "near-optimal" sort. This sort requires that geographic codes on n_s = 582 of the N = 64,998 be swapped. From this we can calculate the following acceptable range for P_{BIAS}.

$$0.1 < P_{BIAS} < 20$$

From Table 7, we can see that the value of P_{BIAS} (at the finest equivalence class level) is 17.95. This is within the prescribed range and the swap is acceptable.

Reducing the value of the PERCENTAGE BIAS. Suppose we have determined our swapping order and tolerance level. We then calculate the PERCENTAGE BIAS and find that it falls outside the prescribed range. To reduce the value, we have to swap less records. There are two ways in which this can be obtained.

(1) We can lower the tolerance. Suppose that we have calculated the frequencies $\{f_1, f_2, ..., f_I\}$ for each of the I equivalence classes. Assume that we have set a tolerance at $Q = q_0$, which yields U equivalence classes which fail tolerance. If $q_1 < q_o$, then a tolerance of $Q = q_1$ will cause less equivalence classes to fail tolerance. Hence, less records will be swapped.

(2) We may only desire to swap a percentage of the records. This will not guarantee that we have protected the identity of every readily identifiable record. Instead, an intruder would only be able to ascertain that we have possibly protected the identity of any given respondent.

Is there any way to determine how we should set the tolerance, $Q = q_1$, or p', the percentage of records to be swapped, based on the ratio $p = P_U / P_{BIAS}$? Appendix A provides the answers to these questions. In particular, the results of Theorem A.2 and A.3 give reasonable estimates for p'. This value can also be used to detemine the value of $Q = q_1$, where the number of classes which fail tolerance is p'*U.

Theorem A.2. Suppose we originally set the tolerance at q_o, causing U classes to fail tolerance, which causes $P_U = p * P_{BIAS}$. We can lower the P'_{BIAS} under the prescribed limit P_U by choosing a difference tolerance q_1, which will cause only $p^2 * T$ equivalence classes to fail tolerance.

Theorem A.3. Let

T = Total Number of Records with Changed Values (when $Q = q_0$);
N = Number of Records from Above Tolerance Classes Used in the Swap; and
$p = P_U / P_{BIAS}$.

Suppose we want to swap $p'^2 * T$ records, to force $P_{BIAS}' = P_U$. A good approximation for p' is

$$p' = \frac{(T-N)-\sqrt{(T-N)^2-[p^2*T*(T-2N)]}}{(T-2N)}.$$

Example. Table 8 illustrates that P_{BIAS} can be effectively reduced by using p'. The illustration was tested using the 1993 Annual Housing Survey Public Use Microdata file. It was sorted in BATHS...BEDS... YR_BLT...REGION order. A tolerance of Q = 2 was used. This yielded U = 473 uniques (below tolerance classes). To execute the swap algorithm, data on T = 582 records were changed. This meant N= 109 non-uniques were required. The PERCENTAGE BIAS's for the 5 contiuous variables are listed in Column 2 of Table 8. Suppose we were required to reduce the PERCENTAGE BIAS by a factor of 2/3. These expected values are shown in Column 3. Since we could not lower Q, we selected a random sample of $p^2 = 4/9$ (p = 2/3 or 0.67) to be swapped. As one can see in Column 4, the observed bias values were too high. Finally, we adjusted using the result of Theorem 3. The values corresponding to p' = 0.56 are found in Column 5. As one can see, they are extremely close to the expected values.

We must address one other point. We have a swapping algorithm. We can determine a "near-optimal" sort order and control the amount of bias induced by such a swap. For a large file, can we do this in a reasonable amount of time? The results of the following sub-section, lead us to believe that we can.

Table 8. PERCENTAGE BIAS
As a Function of the Number of Records on Which Data Is Exchanged

	Original	Expected (p=0.67)	Observed (p=0.67)	Observed (p'=0.56)
VARIABLE				
TOTAL RECORDS SWAPPED	582	259	331	253
% BIAS INCOME	18.6	12.4	14.7	12.7
%BIAS HOME VAL	6.1	4.1	4.7	4.1
%BIAS MORTGAGE	4.1	2.7	2.7	2.5
%BIAS MAINTAIN	22.5	15.0	14.9	13.4
%BIAS TAXES	6.2	4.1	4.5	4.1

Processing Time Requirements. The processing would require approximateley 6 steps: (1) extract the key variables and the identification number, (2) sort the extracted file on the key variables, (3) process the sorted file to identify tolerance failures and swap the geographic codes, (4) sort the output by the identification number, (5) sort the original file by the identification number, and (6) update the geographic codes on the original file. Processing the 1993 Annual Housing Survey Public Use Microdata Sample test file required only 41 seconds of Computer Processing Unit (CPU) time to execute. This is only a 65,000 record file, sorted on 4 key variables. Larger files with more key variables will require more time to process. The sorts required about half of the processing time (20 CPU sec.) on the VAX9000 (MCVX09). There is no reason to believe that this ratio will change for larger files. If the time required to sort the Year 2000 Census files can be accurately estimated, the total time should be able to be accurately estimated also. Below are the CPU times for each step of the procedure. These times do not include the time to determine the "near optimal sort." This would be determined by a sample of the universe.

Step 1 ... Extract the key variables and IDNUM into File 1	3.97 CPU sec.
Step 2 ... Sort File 1 on the 4 key variables	7.18
Step 3 ... Identify Below Tolerance Classes and Swap	9.49
Step 4 ... Sort File 1 by IDNUM	6.22
Step 5 ... Sort File 2 by IDNUM	6.69
Step 6 ... Update geographical codes	7.29
Total	40.64 CPU sec.

XIV. Conclusions

To mask the 1990 Census data, the U.S. Bureau of the Census developed a swapping technique, which it refers to as the "Confidentiality Edit". This technique has several very attractive features. Namely (1) it preserves certain housing unit and population counts (some of which are mandated by law); (2) it is equivalent to the "blank out and impute" procedure, which is a widely accepted procedure for masking microdata files; (3) it is easy to program and implement; (4) it preserves the multivariate relationships for the universe; and (5) it does not drastically distort analysis on relatively sparse (e.g., approximately 20 respondents) sub-domains.

The 1990 Confidentiality Edit procedure is not without its critics. Criticisms include (1) the file may not have been adequately masked, and (2) the Census Bureau never really quantified how accurate the statistics (particularly for continuous data) really are.

We believe that the 1990 Confidentiality Edit was a sound disclosure limitation technique. However, since this was the first occasion on which the Census Bureau used a data-swapping technique, some details were overlooked. For the Year 2000 Census, we recommend that the Census Bureau again use a data-swapping approach. We recommend that the 1990 approach be modified slightly to improve the amount of protection afforded to the identity of each individual respondent.

To improve the next Decennial Census, we recommend that the Bureau target records which have a "high risk" of re-identification. Records are considered "too risky" if only a very few contain similar traits within the same geographic region as delineated on the Public Use Microdata Sample files. In this case, at least one randomly chosen record from each "high risk" equivalence class will be swapped with a record from a different equivalence class with very similar traits (and, hopefully, very similar continuous data). A procedure has been developed to group the equivalence classes by similarity.

In addition, we have developed a measure based on the mean square bias, induced to the means of mutually exclusive yet exhaustive equivalence classes. All swapping will take place between these classes. We believe that this mean square bias can be used in two ways: (1) to ensure that the file has been adequately masked, and (2) to ensure that the analytic validity of the file has been retained.

We believe that by incorporating these changes and additions into the 1990 Confidentiality Edit procedure, the US Bureau of the Census will develop a first-rate disclosure limitation technique, befitting the importance and significance of the initial Census for the next millennium.

XV. References

1. Appel, G., Kinzel, S. And Noite, D. (Date unknown). "SAFE - A Generally Usable Program System for the Anonymization of Individual Data in Official Statistics".

2. Dalenius, T. And Reiss, S. P. (1982). "Data-swapping: A Technique for Disclosure Control," Journal of Statistical Planning and Inference, **6**, 73-85.

3. Fienberg, S. E. (1995). "Review of Current Plans for Disclosure Limitation in 2000 Decennial Census Data," Bureau of the Census Contract 50-YABC-2-66205, Task Order 8, US Bureau of the Census, Washington, DC.

4. Greenberg, B. (1986). "Designing a Disclosure Avoidance Methodology for the 1990 Decennial Censuses," presented at the 1990 Census Data Products Fall Conference, Nov. 17-18, Arlington, VA.

5. Griffin, R. A., Navarro, A. And Flores-Baez, L. (1989). "Disclosure Avoidance for the 1990 Census," Proceedings of the Section on Survey Research Methods, American Staistical Society.

6. Pieters, A. J., and de Waal, T. (1995). "A Demonstration of ARGUS," (user documentation).

7. Porter, G. (1986). "Suppression Methodology and decennial Census Data," presented at the 1990 Census Data Products Fall Conference, Nov. 17-18, Arlington, VA.

8. Subcommittee on Statistical Disclosure Limitation Methodology of the Federal Committee on Statistical Methodology (1994). Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper 22, Office of Management and Budget, Washington, DC.

Appendix A. A Technical Approach to the Analysis of the PERCENTAGE BIAS

Theorem A.1. Let s and $\bar{x}$ be the standard deviation and mean of some continuous variable. Suppose we want the swapping to induce an error of at most (1+f) times the standard deviation. Then the upper bound, P_U, on the BIAS is

$$P_U = 100 * (2f + f^2)^{1/2} * (s / \bar{x}).$$

Proof. The PERCENTAGE BIAS, P_{BIAS}, is nothing more than a measure for the amount of induced error. For the data user, this bias compounds errors caused by sampling. For the typical random sample from an unmasked universe, the only error is sampling error, s. Now suppose we swap data in such a way that the average piece of data is disturbed by the bias, BIAS. The new net error, z, is related to the previous two by the equation

$$z^2 > s^2 + BIAS^2.$$

We want to induce bias, so that the resulting net error is no more than (1+f) times the original standard deviation. The above equation becomes

$$(1+f)^2 * s^2 > s^2 + BIAS^2.$$

This means

$$BIAS^2 < (2f + f^2) * s^2.$$

However, the BIAS = $P_{BIAS} * \bar{x} / 100$, so

$$P_{BIAS} < 100 * (2f + f^2)^{1/2} * (s / \bar{x}).$$

Hence the result.

Example. Suppose we use the continuous variable INCOME, with a mean of 11,369 and a standard deviation of 16,572; and assume we want to disturb the standard deviation by at most 1 percent (f = .01). This yields a value for P_U of 20.

Reducing the value of the PERCENTAGE BIAS. Suppose we have calculated the frequencies $\{f_1, f_2, ...f_I\}$ for each of the I equivalence classes. Further, assume that we have set a tolerance of $Q = q_0$. This yields U equivalence classes which fall below tolerance, hence one respondent from each of these U classes must be below tolerance. The data are swapped, and P_{BIAS} is calculated. Suppose $P_{BIAS} > P_U$ (i.e., $P_U = p * P_{BIAS}$, with $p < 1$). We can lower the value of P_{BIAS} by lowering the value of U. This requires that we lower the tolerance q_0. Let's attack the following question, "How low should we set the tolerance $Q = q_1$ so that $P_{BIAS} < P_U$?"

Theorem A.2. Suppose we originally set the tolerance at q_0, causing U classes to fail tolerance. The swap causes $P_U = p * P_{BIAS}$. We can lower the PERCENTAGE BIAS under the prescribed limit P_U by choosing a difference tolerance q_1, which will cause only $p^2 * U$ equivalence classes to fail tolerance.

Proof. Recall that

$$BIAS^2 = \frac{1}{N} * \sum (x_i - y_i)^2,$$

where

x_i = the value of the i-th respondent (before the swap) and
y_i = the value of the i-th respondent (after the swap).

In this sum, approximately U of the observations are non-zero. This causes the BIAS to be too large. Suppose the perturbed values are distorted by approximately the same amount. If we can manipulate the tolerance so that only $p^2 * U$ equivalence classes fail tolerance, then only $p^2 * U$ of the non-zero terms will appear in the summand. Hence the new bias, BIAS', will be related to the original bias by the formula,

$$BIAS'^2 = p^2 * BIAS^2 \quad \text{or} \quad BIAS' = p * BIAS.$$

If one attempts to use Theorem A.2, it will generally over-estimate the value of q_1. More than the desired percentage of records change value. Why is this? Often times data on a record in a below tolerance class is exchanged with data on another below tolerant class record. By lowering the value of q_1, one changes the status of some of these classes to above tolerance. Hence, more records exchange values than one desires. Theorem 3 below compensates for this.

Theorem A.3. Let

T = Total Number of Records with Changed Values (when $Q = q_0$);
N = Number of Records from Above Tolerance Classes Used in the Swap; and
$p = \widehat{P_U} / P_{BIAS}$.

Suppose we want to swap $p'^2 * T$ records, to force $P_{BIAS}' = P_U$. Then a good approximation for p' is

$$p' = \frac{(T-N) - \sqrt{(T-N)^2 - [p^2 * T * (T-2N)]}}{(T-2N)}.$$

Proof. Let $\{u_1, u_2, ...\}$ denote the below tolerance records swapped and $\{n_1, n_2, ...\}$ denote the above tolerance records used in the swap. Two types of situations occur.

Type 1: Data on below tolerance records are swapped with data on other below tolernce records (i.e., $u_1 <-----> u_2$).

Type 2: Data on below tolerance records are swapped with data on above tolerance records (i.e., $u_3 <-----> n_1$).

Note that above tolerant records are only used in Type 2 swaps. Since N above tolerant records are used, Type 2 contains N pairs (or 2N total reocrds). Therefore Type 1 are the (T-2N) other below tolerance records.

Suppose the tolerance is lowered so that only p'^2 of the below tolerance equivalence classes remain. Further, suppose that these are randomly distributed in the order. Suppose we want the resulting distribution of below tolerance equivalence classes to require $p^2 * T$ records to exchange values. Then

(1) p'^2 of the original Type 1 pairs are retained (Count 1 = $p'^2 * (T - 2N)$);

(2) $p' * (1- p')$ of the original Type 1 pairs lose only the first above tolerant equivalence class when tolerance q_1 replaces q_0 (Count 2 = $p' * (1- p') * (T - 2N)$);

(3) $(1- p') * p'$ of the original Type 1 pairs lose only the second above tolerant equivalence class when tolerance q_1 replaces q_0 (Count 3 = $(1 - p') * p' * (T - 2N)$); and

(4) p' of the original Type 2 pairs are retained (Count 4 = $p' * 2N$).

Therefore,

$$p^2 * T = \text{Count 1} + \text{Count 2} + \text{Count 3} + \text{Count 4}.$$

Substituting for these counts and simplifying, we find

$$[p'^2 * T] - [p' * \{ 2 * (T - 2N) \}] + [p^2 * T] = 0.$$

To obtain the desired result, use the quadratic formula to solve for p'.

Corollary A.4. The results for Theorem A.2. and A.3. can be used, if one desires to swap a percentage, p or p', of all below tolerance records (as opposed to actually lowering the tolerance).