DISCLOSURE AVOIDANCE FOR THE 1990 CENSUS

Richard A. Griffin, Alfredo Navarro and Linda Flores-Baez
U.S. Bureau of the Census

I. INTRODUCTION

The design, development and implementation of disclosure avoidance methods are among the primary responsibilities of the Census Bureau. As part of our 1990 Census planning, we investigated several different disclosure avoidance techniques for use in the 1990 Census. The Census Bureau collects information from respondents under a guarantee of confidentiality. We are required by law (Title 13 of U.S. code) to release data in a way that does not identify an individual. Thus, our objective is to maximize the level of useful statistical information provided subject to the condition that confidentiality is not violated. This is the trade off for which we are trying to find a balance - maximize the availability of statistical data while providing adequate protection. This paper will briefly discuss the types of procedures which have been investigated for dealing with disclosure risk in the 1990 Census. The 1990 Census disclosure avoidance procedures for 100 percent and sample data will be presented, and the effects on data will be described. The planning and research aspects of developing these procedures will be described.

II. DISCLOSURE AVOIDANCE TECHNIQUES - 100 PERCENT DATA

Three types of procedures for dealing with disclosure risk for 100 percent data in the 1990 Census have been investigated. These are Suppression, Controlled Rounding, and Confidentiality Edit. Further details of these procedures are contained in [1].

A. Suppression

Suppression, as the name implies, is a disclosure avoidance technique which provides protection by not publishing data when there is an unacceptably high risk of disclosing confidential information. For the decennial census, the Census Bureau would suppress all information when tabulations fall below a selected threshold value. Complementary suppression is sometimes required so that suppressed data cannot be determined by subtraction using published data.

There were three suppression techniques considered for 1990. These are cell suppression, universe suppression and 1980 suppression. Cell suppression suppresses individual data cells whose value falls below a chosen threshold. Universe suppression is based on the universe of each matrix or tally and suppresses the entire distribution if the total for the universe is below the determined threshold. The 1980 suppression is an adaptation of universe suppression. In 1980, the following counts were never suppressed:

- Population counts by race and Hispanic Origin.
- Housing unit counts by vacancy status.
- Occupied housing unit counts by race and Hispanic Origin of the householder.

In general, the following rules were implemented:

- Detailed characteristics collected for total population or any suppression universe defined by race or Hispanic Origin were suppressed if there were 1 to 14 persons in the specified suppression universe.
- Detailed characteristics for families or households were suppressed for suppression universes defined by the race or Hispanic Origin of the householder if there were 1 to 4 occupied housing units in the specified group.
- Detailed housing characteristics were suppressed for suppression universes defined by vacancy status and tenure if there were 1 to 4 housing units in the relevant universe.
- Complementary suppression was applied between the major racial groups with pre-established rules for sequence of suppression, and between housing characteristics of owners and renters.
- There was no complementary suppression across geographic areas.

These suppression techniques are explained in detail in [2].

Advantages of the suppression methodologies are:

- Data are published as collected in the Census.
- Suppressed data cells make it evident the Census Bureau did something in order to preserve confidentiality of the data.

Disadvantages are:

- The methodology must be implemented separately for each data product.
- Often data users aggregate tracts or block groups forming user defined areas for analysis purposes. There are a lot of matrices with suppressed cell counts resulting in a limited ability to aggregate data.
- Loss of data due to complementary suppression.

B. Controlled Rounding

Controlled rounding is a rounding technique in which all data items are rounded to a suitably chosen base while preserving summation to table totals and subtotals. Through rounding, one

attempts to provide disclosure protection while preserving the usefulness of the data, providing unbiased estimates of derived statistics and maintaining the additive structure of data tables. The scheme considered for 1990 was controlled rounding base 3, which is explained in more detail in [3].

Advantages of controlled rounding are:

- All data cells are shown. Their values are each potentially altered slightly. Showing of all data cells is especially important since it allows for the aggregation of small areas into larger ones.
- It is evident to data users that disclosure avoidance steps have been taken (i.e. all published values are multiples of 3).

Disadvantages are:

- Inconsistencies will appear between tables in the released figures. Counts in different tables that should be identical by definition may be different due to independent rounding for each table.
- The methodology must be applied for every data product.
- Reviewing products for errors is made more difficult because of inconsistencies due to rounding.

C. Confidentiality Edit

Confidentiality Edit is based on selecting a small sample of census households from the internal census data files and interchanging their data with other households which have identical characteristics on a set of selected key variables, but are in different geographic locations. The matching and interchanging operations are controlled on the key variables of number of persons in household, population characteristics of race, Hispanic origin and age, and on housing characteristics of units at building, rent/value and tenure.

The result of the controls described above is that census counts for total persons, totals by race, Hispanic origin and age 18 and above (Public Law 94-171 counts) as well as housing counts by tenure are not affected by the confidentiality edit.

Advantages of Confidentiality Edit are:

- Need to implement only once on internal files to obtain protection for all data products.
- All data cells are shown so there is no interference with data aggregation.
- More data are made available than in 1980.

The major disadvantage is:

- There are no obvious changes in published tables so that our efforts for disclosure protection are not evident.

III. INITIAL PLANNING AND RESEARCH - 100 PERCENT DATA

A disclosure avoidance methodology was sought which would:

- provide adequate protection;
- not alter the Public Law 94-171 census counts; in particular the total population count and the total housing unit count for all geographic areas;
- yield meaningful and useful data for the user;
- be cost efficient; and
- be usable on all forms of media used for dissemination of data from the 1990 Census;

It was decided that the Confidentiality Edit methodology should be investigated in detail as a possible disclosure avoidance mechanism for the 1990 Census. This methodology was simulated using 1980 Census 100 percent data from the state of New Jersey. The findings of this research are described in detail in [4].

In our research study, the following issues were addressed:

- The existence of a matching housing unit for all cases in sample.
- Disclosure avoidance provided by the procedure, particularly for small areas.
- The effect on census statistics after interchanging records geographically.

Several sampling fractions for selecting the household records to be interchanged were chosen for study. In using disclosure avoidance methodologies there is a trade off between the level of protection and the distortion of the derived statistics. For the case of Confidentiality Edit, the higher the sampling fraction (or the more household records interchanged) the more the protection, but the statistics are distorted to a greater degree. The rationale for a procedure such as this is that we introduce sufficient uncertainty into the published data so that no one can say with certainty that displayed data is for a given household or individual due to the introduction of noise from the data interchange.

A. Availability of Matching Records

The expected matching rate is a function of the proportion of "unique" households in the population (households that match to no other households in the state). For all of the sampling rates simulated, the matching rate was at least 99.7 percent.

B. Disclosure Protection

The level of disclosure protection provided by the Confidentiality Edit methodology was discussed in [5] and is summarized below. The main question was the level of protection provided for small areas.

For small groups of housing units, the Confidentiality Edit methodology will enable the data user to determine

that the "true" value (for count data) is likely one of 2 or 3 possible values. Without knowing the sampling rate for data interchange or the sample design, the user is not able to determine a probability distribution associated with this set of possible "true" values.

To contrast Confidentiality Edit with the 1980 suppression, if the value was from a suppression universe smaller than the threshold, the value was suppressed. Otherwise, the "true" value was published.

For the controlled rounding base 3 approach, suppose the "true" value is 2. This value could be obtainable in one source as 0 and in another source as 3. If this happens, then the data user knows the "true" value is 1 or 2 and each of these values is equally likely.

The protection provided by Confidentiality Edit for small groups of persons is greater than for small groups of housing units due, in part, to the possibility of more than one person from the group being from the same household. If k is the maximum number of persons of the group from the same household, the user will be able to determine that the "true" value is one of 2k+1 values (k is usually unknown to the user).

Even with this uncertainty, there was concern about the protection for small areas since small areas are less likely to be represented in the Confidentiality Edit sample. The smallest area for which 100 percent data is published is the block. The problem of providing adequate protection for small areas under a disclosure avoidance program based primarily on the Confidentiality Edit was investigated further. A decision was made to simulate a data interchange program with a higher sampling rate for small blocks. More details of this simulation are given in section IV.

C. Effect on Census Statistics due to Confidentiality Edit

The effect on census statistics of Confidentiality Edit was studied and reported in [4]. A number of distributions were calculated for demographic data at the tract and block level. These distributions were prepared to reflect the data for tracts and blocks before and after the data interchange operations.

The data distributions were compared using a dissimilarity index or D-Statistic [6]. A description of this D-Statistic is given below for an arbitrary column in a two-way table for a particular publication area.

Let X_i = the count before Confidentially Edit for the cell in the i^{th} row.

Y_i = the count after Confidentially Edit for the cell in the i^{th} row.

r = the number of rows in the given table.

$$X = \sum_{i=1}^{r} X_i$$

$$Y = \sum_{i=1}^{r} Y_i$$

The index of dissimilarity, D, between the two columns is given by

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

This index can be interpreted as the minimum proportion of a column that would have to be moved or redistributed so that the proportion of the total in a row will be the same as the proportion of the total in the row prior to the data interchange operation.

The following hypothetical table is used to illustrate our use of this statistic.

Age	Before	After
Under 5 Years	3	2
5 to 17 Years	4	5
18 to 64 Years	10	10
65 Years or Over	3	3
Total	20	20

D = .05.

One household has 1 person under 5 years and 2 persons age 18 to 64. This household is involved in a data interchange with a household with one person age 5 to 17 and two persons age 18 to 64. The value of the D-Statistic is .05. The total population is 20 so that only 1 person would have to be redistributed within the age categories in order to get the same distribution as before the data interchange. Specifically, by moving one person from the 5 to 17 category in the "after" distribution to the under 5 category, the new distribution will be exactly equal to the one prior to data interchange. This illustrates that it is important to examine the D-Statistic in conjunction with the number of people who are moved by the Confidentiality Edit.

We viewed the D-Statistic as a measure of the data distortion introduced by the interchange operation of the Confidentiality Edit. The D-Statistic was calculated for each census tract in the study state of New Jersey for a representative set of 100 percent data distributions described in [4]. A number of tracts with high D-Statistic values were examined by Census Bureau subject matter experts. No levels of distortion were found which would adversely affect the use of the data. As part of this process, the D-Statistic tract values

were cross-tabulated by the number of individuals moved for a number of data distributions. We observed that large values of the D-Statistic occurred when only a small number of people were involved in the data interchange operation in small areas. It was also evident that when a large number of people were moved the D-Statistic was small due to a large base (the number of persons in a column). Thus, even though a large number of people had been moved, the distortion to derived statistics was seen to be small.

The distortion induced by Confidentiality Edit on the block level was also studied. The level of distortion at the block level was not severe.

D. Conclusions

The findings of this initial planning and research indicated that the use of small samples for data interchange do not introduce unacceptable levels of distortion into the 100 percent data. Confidentiality Edit was determined to be a means of providing sufficient uncertainty into the data to allow for adequate protection against disclosure while at the same time providing reliable data. The problem of providing adequate protection for small areas remained a concern and further research and results are discussed in Section IV.

IV. PLANNING AND RESEARCH - SMALL AREA 100 PERCENT DATA

The research described in Section III. indicated that Confidentiality Edit did not provide sufficient protection for small blocks. Applying one sampling rate for data interchange to the internal detail file resulted in a large number of small blocks which did not contain any records selected for the data interchange sample. A decision was made to simulate a data interchange program with a higher sampling rate for small blocks. This was done using the 1980 Census 100 percent edited detail file for the state of New Jersey.

A. Availability of Matching Records

The expected matching rate is a function of the proportion of "unique" households in the population (households that match to no other households in the state). Since it was possible that "unique" households were clustered in small blocks, there was a concern that a higher sampling rate in small blocks would result in a lower matching rate. However, the result of this simulation was a matching rate of about 99.7 percent, the same level as in the simulation described in Section III.

B. Disclosure Protection

The level of disclosure protection provided by Confidentiality Edit with a higher sampling rate for small blocks was determined to be sufficient. For small blocks, the higher sampling rate resulted in a sufficient proportion of small blocks having a household selected for the data interchange sample.

C. Effect on Census Statistics due to Confidentiality Edit

Since the overall sampling rate was higher for this simulation of Confidentiality Edit due to the higher sampling rate in small blocks, there were more records involved in data interchange. Thus, it was necessary to re-examine the distortion in Census statistics.

An evaluation very similar to that described in Section III.C. using the D-Statistic was performed. The results were similar, no levels of distortion were found which would adversely affect the use of the data.

V. 1990 CENSUS 100 PERCENT DISCLOSURE AVOIDANCE MENTHODOLGY

Confidentiality Edit will be used as the 1990 Census 100 percent data disclosure avoidance methodology. This methodology is briefly described as follows:

A. Select a small sample of households from the internal census data files. The sampling rate is higher for small blocks.

B. Match the sample records, according to a set of well defined matching rules, to other records on the file in a different geographic location. The matching is controlled so that agreement on household size, race, Hispanic origin, age (18+), tenure and rent/value is achieved between each sample household and its matching household.

C. Interchange the matched household records according to a well defined "data interchange" operation. The "interchanged" file becomes the official version of the internal detail file and is used to prepare all subsequent census data products. The edit rules result in a controlled procedure such that:
   (1) population counts by total, race, Hispanic origin and persons aged 18 and above are not changed; and
   (2) housing unit counts by total, tenure and rent/value categories are not changed.

VI. DISCLOSURE AVOIDANCE - SAMPLE DATA

A. Background

The uncertainty of what is in sample provides adequate protection for most areas for which sample data will be published; the exception to this rule is areas such as small block groups. The smallest geographic unit for which sample data are published is block group and it was felt that a small block group may have an unacceptable disclosure risk even when only considering sample data.

In 1990, an imputation based methodology is proposed to reduce the risk of disclosure for sample data in small block groups. This methodology involves blanking of a sample of the data fields

(population and housing items) on the sample edited detail file for one of the sample housing units in each small block group and imputing using the 1990 Census imputation methodology. By a small block group, we mean block groups with fewer than a set number of housing units at least one of which is in sample.

B. Data Distortion

The Census Bureau's sequential hot deck imputation method is dependent upon the geographic ordering of households on the Sample Edited Detail File (SEDF). As the SEDF is sequentially processed, each household (person) record with a missing value on a specified variable is assigned that variable's value from the "last","similar" household (person). "Last" means geographically most proximate in terms of the geographic ordering of the file. "Similar" means that the donor household (person) has the same recorded values on a set of auxiliary variables known to be highly associated with the imputed variable in the population. That is, the donor household (person) falls into the same "classification group," defined on the basis of a unique combination of values of the auxiliary variables [7] and [8]. Using the "last" household (person) is advantageous enabling the imputation procedure to exploit the correlations between nearby records [9].

In the 1980 Content Reinterview Study [10], measures of response bias were reported for unedited data (before imputations and computer edits) and for edited data. The report concluded that "the level of bias seems to be about the same in the edited and unedited distributions for a particular characteristic."

The 1990 Census sample data disclosure avoidance methodology of blanking and imputing for a sample of the data fields for one of the sample units in a small block group results in a slight increase in the imputation rate for a tabulation area containing small block groups. An approximation for the variance of the sample mean taking into account the imputation rate is given in [11]. The theory from [11] used in this application is for an imputation procedure in which the immediately preceding observed value in the sequence is imputed for each missing value. These imputation procedures usually impute for missing observations for a variate by dividing the sample into adjustment cells based on ancillary variables known for all sample cases, and then substituting for each missing observation the preceding observed value within the adjustment cell in the sequence (in our case geographic) the sample file is passed for imputation. Since the variance of the imputation mean for the entire sample is dependent on the variances of the imputation means for the individual adjustment cells, the theory is developed assuming the entire sample consists of a single adjustment cell.

Assume the sample is a random sample without replacement of size n from a population of size N. Let m denote the number of missing values in a given sample, and $q = E(m)/n$.

For $i = 1,\ldots, n$ let x_i denote the variate value for the i-th sample unit, whether observed or not, and let $w_i = 1$ if x_i is observed, $w_i = 0$ if x_i is missing. It is assumed that x_i and w_i are independent for all i and that for a fixed m all possible arrangements of the missing values are equally likely. It is also assumed that $\mathrm{cov}(x_i,x_j) = 0$ for $i \neq j$. Note that for simple random sampling without replacement $\mathrm{cov}(x_i,x_j) = -S^2/N$ for $i \neq j$. A large absolute value for this covariance is for N = 250 persons and $S^2 = .25$ (for a 50 percent variable). In this case $\mathrm{cov}(x_i,x_j) = -.001$.

Using this notation and these assumptions, we have:

$$\mathrm{Var}(\bar{x}) = (1/n-1/N)S^2\left[\frac{1+q}{1-q} + \frac{-2q+2q^{n+1}}{n(1-q)^2}\right]$$

For n larger than 125 (N = 250 persons sampled at a rate of .5) and the q values (imputation rates) of our application the second term of the summation can be ignored. Thus, we have:

$$\mathrm{Var}(\bar{x}) = (1/n-1/N)S^2\left[\frac{1+q}{1-q}\right] \quad (1).$$

The increase in imputation rate for a tabulation area due to this sample data disclosure avoidance methodology is directly proportional to the proportion of the tabulation area population that is in small block groups. This proportion will be very small (less than 5 percent) for most tabulation areas. Using equation (1) and a conservative (i.e. high) estimate of the increase in imputation rate due to disclosure avoidance, relative increases in coefficients of variation (C.V.) were computed for several sample data items. These relative increases are functions of the "before" and "after" disclosure avoidance imputation rates. The "after" imputation rate is directly proportional to the proportion of the population of the tabulation area that is in small block groups. The higher this proportion, for a given sample data item, the higher the "after" imputation rate. Three sample data items which are expected to have high imputation rates in 1990 on the basis of 1980 Census results were examined. These were language spoken at home (8.2% of persons 5 years or older were allocated in 1980), income in the year preceding the census (11.5%

of incomes of persons 15 and over were allocated), and unemployment in the year preceding the census (15.9% of persons 16 and over were allocated). A sample item expected to have a low imputation rate was also examined. This item was work disability (4.4% of persons 16 to 64 were allocated).

Results showed that as long as less than 5 percent of the population of a tabulation area are in small block groups, the relative increase in CV will be less than 2.5 percent of the original CV (i.e., a 30.0% CV will not increase to more than 30.8%).

C. Conclusions

The proposed 1990 Census disclosure avoidance methodology for sample data is basically as follows:

1. The sample itself provides adequate protection for all areas for which sample data are published except for small block groups.
2. An imputation methodology will be used to provide disclosure avoidance for sample data in small block groups. This methodology involves blanking of a sample of the data fields (population and housing items) for one of the sample housing units in each small block group and imputing using the 1990 Census imputation methodology.
3. Once sample data imputation is completed the resulting sample data file (for which disclosure avoidance has been applied) is used to prepare all subsequent census sample data products.

This data imputation methodology for providing disclosure avoidance for sample data will add very little to the level of error of the estimates. A major reason for this is that the relative increase in imputation rates is expected to be very small.

REFERENCES

1. Griffin, R. and Thompson, J., Confidentiality Techniques for the 1990 Census, presented at a Concurrent session of the Joint Census Advisory Committee, October, 1987.
2. Porter, G., Suppression Methodology and Decennial Census Data, presented at the 1990 Census Data Products Fall Conference, November 1986.
3. Greenberg, B., Designing a Disclosure Avoidance Methodology for the 1990 Decennial Census, presented at the 1990 Census Data Products Fall Conference, November 1986.
4. Navarro, A., Flores-Baez, L. and Thompson, J., Results of Data Switching Simulation, presented to the American Statistical Association and Population Statistics Census Advisory Committees, April 1988.
5. Thompson, J., Level of Disclosure Protection of Proposed Disclosure Avoidance Technique for the 1990 Census, internal Census Bureau memorandum, July 1987.
6. Shryock and Siegel, The Methods and Materials of Demography, Vol. 1, 3rd. Ed., 1975.
7. Johnson, R., 1990 Census Research and Evaluation Program Proposal for Research on Imputation Error, internal Census Bureau document, 1988.
8. Ford, B.L., "An Overview of Hot-Deck Procedures," pp. 185-206 in Incomplete Data in Sample Surveys, Vol. 2, Theory and Bibliographies, 1983.
9. Sande, I.G., "Imputation in Surveys: Coping with Reality," The American Statistician, August 1982.
10. Bureau of the Census, 1980 Census of Population and Housing. Evaluation and Research Reports. Content Reinterview Study: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview, PHC80-E2, 1986.
11. Ernst, L.R., "Variance of the Estimated Mean for Several Imputation Procedures," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1980.