Statistical Methodology for the Census 2000 Public Use Microdata Samples

Philip M. Gbur and Mary Frances Zelenak, U.S. Census Bureau

Abstract[1]
Public Use Microdata Samples (PUMS) files are data files that contain individual housing unit and person records with their associated characteristics. To protect confidentiality, identifying information and detailed geography is excluded. To allow data users to produce any tabulations of interest and to conduct detailed analyses, such as regression analysis and investigation of the relationships among variables, the U.S. Census Bureau creates PUMS files from decennial census results. Subject to the limitations on sample size and geographic identification, the user may generate tabulations interrelating any desired set of variables. This paper presents the statistical methodologies used in sampling, disclosure avoidance, and variance estimation for the United States, Puerto Rico, Guam, and U.S. Virgin Islands Census 2000 PUMS files.

I. Introduction

Every person and housing unit in the United States was asked basic demographic and housing questions (for example, race, age, and relationship to householder). A systematic sample of addresses received a long form (sample) questionnaire in Census 2000 which asked more detailed questions about items, such as income, occupation, and housing costs. The sampling unit for Census 2000 was the housing unit, including all occupants.

Estimates derived from the census sample files are expected to differ from the 100-percent figures because they are subject to sampling and nonsampling errors. Sampling error in data arises from the selection of people and housing units included in the sample. Nonsampling error affects both sample and 100-percent data and is introduced as a result of errors that may occur during the data collection and processing phases of the census.

Data tabulations based on a sample survey often provide most access to the data required by data users. However, some uses such as regression analyses, calculation of correlations, and detailed tabulations, require access to detailed microdata. As access to the raw data is often restricted, Public Use Microdata Sample (PUMS) files are often created for data users from sample survey results to fulfill this need. The PUMS files for Census 2000 were chosen from the universe of long form records.

In the PUMS, the basic unit is an individual housing unit and people living in occupied housing units or group quarters (GQ). However, microdata records in these samples do not contain names or addresses. The PUMS methodology for Census 2000 is similar to that used for the 1990 Census as described in [1].

The following sections provide an overview of the long form statistical methodologies and describe the statistical methodologies used for the U.S., Puerto Rico, Guam, and U.S. Virgin Islands PUMS files.

II. United States and Puerto Rico PUMS

A. Overview of Long Form Statistical Methodology

1. Sampling

The addresses that were to receive the long form questionnaire were chosen by taking a systematic, variable rate sample of addresses. The ultimate goal was to sample roughly 17 percent of all addresses nationwide. This was achieved through appropriate application of the selected sampling rates to each collection block based on the size of the governmental unit or census tract in which the block was located. Application of the rates for Census 2000 was based on the interim census tract delineation, as updated census tracts were not yet available. Governmental units were defined as states, counties, cities, incorporated places, school districts, American Indian Reservations, Tribal Jurisdiction Statistical Areas (now known as Oklahoma Tribal Statistical Areas), minor civil divisions in selected states, and census designated places in Hawaii.

The rates used were: 1-in-2, 1-in-4, 1-in-6 and 1-in-8, and were applied based on a governmental

[1]This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

unit's or tract's predetermined measure of size. The estimated number of occupied housing units was used as the measure of size.

The sampling rates were applied at the collection block level. For blocks that fell into more than one sampling stratum (such as a small incorporated place in a large county), we applied the higher sampling rate.

The sampling strata and their cutoff points were:
- 1-in-2 for governmental units < 800 housing units;
- 1-in-4 for governmental units between 800 and 1200 housing units;

if not 1-in-2 or 1-in-4 then
- 1-in-6 for census tracts < 2000 housing units; and
- 1-in-8 for census tracts $\geq$ 2000 housing units.

The following rates were used for certain data collections and special populations:
a. People living in GQ were sampled at a 1-in-6 rate.
b. Service Sites (shelters and soup kitchens) were sampled at a 1-in-6 rate.
c. Remote Alaska was sampled at a 1-in-2 rate.
d. The Telephone Questionnaire Assistance operation took incoming calls for requests for mailing questionnaires and for interviews. Interviews were done for short forms only and individuals providing an interview were not eligible for long form sampling. Individuals who telephoned to request a questionnaire received either their designated form type or were subject to a 1-in-6 sampling rate, depending upon whether they had their census identification number.
e. Addresses added to the mailout universe after the initial sampling were sampled according to the sampling rate of the stratum that the addresses' block was in.

Further details on the Census 2000 sampling process may be found in [2].

2. Weighting

As in every census since 1940, when we first asked questions in the census on a sample basis, the iterative proportional fitting methodology was used in Census 2000 to calculate weights for estimating characteristics of the entire country based on the long form sample. We carry out this methodology, also known as raking, within weighting areas.

Weighting areas, the geographic level at which we conduct the weighting, were formed within counties. Weighting areas were required to have a minimum of 400 sampled persons. As necessary, small counties with fewer than the prescribed number of cases were allowed to stand alone as weighting areas.

To ensure that we have a basic minimal sample within the weighting areas, there was augmentation (imputation of sample information for selected entire households or GQ people) of the long form sample using a set of predetermined rules, as needed. This was done to attain a minimum observed sampling rate within each area, reducing the associated variance. Long form data were imputed based on the reported short form data for sample augmentation. After augmentation, weighting proceeded separately for people, occupied housing units, and vacant housing units.

For each sample unit we set an initial weight equal to the inverse of the observed sampling rate (100 percent count divided by the number of sample cases received, including augmented cases). We then carried out the iterative proportional fitting methodology, also known as raking. Raking was performed in several stages.

For person weighting, for each weighting area, we formed a four-dimensional matrix using household type and size (such as family with own children with four people and family without own children with four people), sampling rate, whether the person is a householder, and Hispanic origin by race and age/sex. For occupied housing units, we used three dimensions: household type by size; race and Hispanic origin of the householder by tenure; and sampling rate. Vacant housing units were weighted based on a three cell array: “for sale;” “for rent;” and “other.” If a given classification/cell was not sufficiently large, then it was collapsed with another classification following a predefined pattern.

Raking is an iterative proportional adjustment of the cross-classified cell counts. The interior cell counts within a classification were multiplied by the ratio of the 100 percent count (for that classification) to the initially inflated sample total (for that classification). An iteration of the raking consists of one stage of adjustment for each dimension. Each stage adjusts all interior cell counts by the appropriate cell ratio. The

raking progressed until a predefined stopping criterion was reached.

The final step in the weighting process was to integerize the post-raking weights using a controlled rounding procedure.

Further details on the Census 2000 weighting process may be found in [3].

3. Direct Variance Estimation

For Census 2000, we used the Successive Difference Replication (SDR) methodology to calculate direct variances at the weighting area level. The SDR methodology was developed by Fay [4], based on the successive difference variance estimator for systematic samples. A successive difference estimator calculates the variance from the sum of squares of differences from overlapping pairs of sample units. This allows order of selection to be taken into account when the units' order of selection is maintained within the calculation.

Replicate weights were calculated using replicate factors and the integerized final weights from the weighting process. Fifty-two replicate weights were calculated for every housing unit and person in the observed long form sample.

Standard errors were calculated separately for characteristics of persons, families, and housing units/households. Replicate factors were multiplied by the final weights to produce replicate final weights. Once replicate final weights were produced, the SDR method estimated the standard error, $S_{SDR,t}$, of the estimate for the t^{th} data item through the formula:

$$S_{SDR,t} = \left(\frac{4}{52}\sum_{r=1}^{52} \left[x_{rt}\left(\frac{W_c}{W_{rc}}\right) - x_{ot}\right]^2\right)^{\frac{1}{2}}$$

Where:

x_{rt} is the weighted total of the r^{th} replicate for data item t where $r = 1, \ldots, 52$;

W_c is the weighted total of the sample for data type c where types are people, housing units, or families;

W_{rc} is the weighted total of the r^{th} replicate for data type c; and

x_{ot} is the weighted total of the full sample for data item t.

In addition, standard errors were calculated assuming a 1-in-6 simple random sample (S_{SRS}) for each, t^{th}, data item as follows:

If $x_{ot} \leq 0.98\ W_c$, then

$$S_{SRS,t} = (5x_{ot}[1-(x_{ot}/W_c)])^{1/2};$$

else if $x_{ot} > 0.98\ W_c$, then

$$S_{SRS,t} = (5x_{ot}[1 - 0.98])^{1/2}.$$

A design factor (DF) was calculated from these standard errors for the t^{th} data item at the weighting area level as: $DF = S_{SDR,t} / S_{SRS,t}$.

The DFs reflect the effects of the sample design and part of the complex ratio estimation procedure used for the Census 2000 sample data.

Further details on the Census 2000 variance estimation process may be found in [5].

4. Generalized Variance Estimation

The long form sample can be the basis of a myriad of estimates calculated at many geographic levels. The Census Bureau has a commitment to provide estimates of sampling error for all estimates and to minimize burden on data users by not overwhelming them with volumes of error estimates. Thus, we provided a set of 59 generalized DFs to approximate sampling errors.

DFs were calculated for selected data items within each weighting area. Due to space limitations, generalized DFs were made available across four percent-in-sample categories or intervals: <15, ≥15-<25, ≥25-<35, 35+. The percent-in-sample was defined at the weighting area level to be the observed unweighted sample count divided by the 100 percent count, which was equal to the final weighting area observed sampling rate multiplied by 100. The count was of persons for population characteristics and of housing units for housing characteristics.

Data items were arranged into groups and subgroups based on characteristic. For each state, the District of Columbia, and Puerto Rico, generalized DFs for each group and subgroup were calculated over each of the percent-in-sample intervals as a weighted average DF. They were also calculated at the national level.

Data item groups were examined for homogeneity of variance. Specific data item DFs which were determined to be outliers were down-weighted.

Further details on the Census 2000 generalized variance estimation process may be found in [6].

B. Sampling

Two PUMS files were created, a 1-percent and a 5-percent sample. A stratified systematic selection procedure with equal probability was used to select each of the PUMS. The sampling universes were defined as all occupied housing units including all occupants, vacant housing units, and GQ people in the census sample. The sample units were stratified during the selection process. The stratification was intended to improve the reliability of estimates derived from the PUMS by defining strata within which there is a high degree of homogeneity among the census sample households with respect to characteristics of major interest.

The occupied housing unit stratification was performed using a matrix containing 34,080 cells made by combining 71 race groups, 5 Hispanic origin groups, 3 family types, 2 tenure groups, 4 groups based on maximum age of household members, and the 4 long form sampling rates. In the case of occupied housing units the primary sampling unit selected by the systematic selection process was housing units and all person records were extracted after the housing units were chosen. Therefore, the race and Hispanic origin correspond to the householder. The maximum age variable, in contrast, could come from any household member.

The vacant housing unit stratification was performed within a matrix consisting of 12 cells made by combining the four long form sampling rates with three vacancy statuses.

The GQ stratification used a matrix of 2,840 cells made by combining 71 race groups, 5 Hispanic origin groups, 4 age groups, and 2 types of GQ. For GQ people, the race, Hispanic origin, and age were those of the individual GQ person.

The sample selection procedures were performed separately for each of the three subsampling universes: occupied housing units (including all people in them), vacant housing units, and GQ persons, as follows. The number of 1-percent public use microdata samples for a given state was determined by the full census sample size for that state. For instance, if the full census sample for a state was 20 percent, then the census sample was divided into 20 subsamples of approximately equal size. The 1-percent public use microdata sample was designated at random from the 20 subsamples. From the remaining 19 subsamples, five 1-percent subsamples were designated at random and merged to produce the 5-percent public use microdata sample.

During the sample selection operation, consecutive two-digit subsample numbers from 00 to 99 were assigned to each sample case as it was selected for the 5-percent or 1-percent sample to allow for the designation of various size subsamples and, as discussed in section II.D, to allow for the calculation of standard errors. As an example, for a 1-percent public use microdata sample, the choice of records having subsample numbers with the same "units" digit (e.g., the two "units" digit includes subsample numbers 02, 12, 22, ..., 92) will provide a 1-in-1000 subsample.

Samples of any size between 1/20 and 1/10000 may be selected in a similar manner by using appropriate two-digit subsample numbers assigned to either of the microdata samples. Care must be exercised when selecting such samples. If only one "units" digit is required, the units digit should be randomly selected. If multiple “units” digits are required, then the first should be randomly selected and the others should be appropriately spaced to maximize the spread. For example, if two "units" digits are required, the first should be randomly selected and the second should be either 5 more or 5 less than the first. Failure to use this procedure, e.g., selection of records with the same "tens" digit instead of records with the same "units" digit, would provide a 1-in-10 subsample but one that would be somewhat more clustered and as a result subject to larger sampling error.

Further details on the Census 2000 PUMS sampling procedures may be found in [7].

C. Weighting

The weights that appear on the PUMS files are the product of the long form weight and the PUMS sampling weight. The long form weights were obtained from the procedure described above in Section II.A.2 resulting in the assignment of a weight to each sample person and housing unit record. The PUMS sampling weight was the inverse of the sampling rate used for a given state equivalent area. These weights were then integerized using a controlled rounding procedure.

D. Variance Estimation

Variances are calculated for PUMS estimates by calculating an unadjusted standard error and then multiplying the result by the appropriate generalized DF calculated as described, above, in Section II.A.4. For estimated totals, the formula for the unadjusted standard error for the 1-percent PUMS is:

$$SE(\hat{Y}) = \sqrt{99(\hat{Y})(1 - \frac{\hat{Y}}{N})}$$

Where:

$\hat{Y}$ = estimate of characteristic total and

N = size of geographic area.

For estimated percentages, the formula is:

$$SE(\hat{p}) = \sqrt{\frac{99}{B}\,\hat{p}\,(100 - \hat{p})}$$

Where:

$\hat{p}$ = estimate of percentage and

B = size of base.

For estimates from the 5-percent file, the "99" would be replaced with "19." As an aid to users, a table of unadjusted standard errors is provided for selected estimated totals (percentages) and geographic area (base) sizes.

Use of tables or formulas to derive approximate standard errors is simple and does not complicate processing. Nonetheless, a more accurate estimate of the standard error can be obtained from the samples themselves, using the random group method [8]. Using this method it is also possible to compute standard errors for means, ratios, indexes, correlation coefficients, or other statistics for which the tables or formulas do not apply.

The random group method does increase processing time since it requires that the statistic of interest, for example a total, be computed separately for each of up to 100 random groups. The variability of that statistic for the sample as a whole is estimated from the variability of the statistic among the various random groups within the sample.

To obtain the random group standard error for an estimated total, $\hat{X}$, the formula is:

$$var(\hat{X}) = (\frac{t}{t-1}) \sum_{g=1}^{t} [x_g - \frac{1}{t} (\sum_{g=1}^{t} x_g)]^2$$

or the computational formula:

$$var(\hat{X}) = (\frac{t}{t-1}) \sum_{g=1}^{t} x_g^2 - t\,\bar{x}_g^2$$

It is suggested that t = 100 for estimating the standard error of a total since, as discussed in section II.B, each of the sample records was assigned a two-digit subsample number sequentially from 00 to 99. The two-digit number can be used to form 100 random groups.

For example, a sample case with 01 as the two-digit number will be in random group 1. All sample cases with 02 as the two digit number will be in random group 2, etc., up to 00 as the one-hundredth random group. The reliability of the random group variance estimator is a function of both the kurtosis of the estimator and the number of groups, t. If t is small, the coefficient of variation will be large, and therefore, the variance estimator will be of low precision. In general, the variance estimator will be more reliable with a larger t.

Percentage estimates of zero and estimated totals of zero are subject to both sampling and nonsampling error. While the magnitude of the error is difficult to quantify, users should be aware that such estimates are nevertheless subject to both sampling and nonsampling error even though in the case of zero estimates the corresponding random groups estimate of variance will be zero.

Also, the standard error estimates obtained using the random groups method do not include all components of the variability due to nonsampling error that may be present in the data. Therefore, the standard errors calculated using the methods described in this section represent a lower bound for the total error. Data users should be aware that, in general, confidence intervals formed using these estimated standard errors do not meet the stated levels of confidence. Data users are advised to be conservative when making inferences from the data provided in this data product.

E. Disclosure Avoidance

Disclosure avoidance is the process for protecting the confidentiality of data. A disclosure of data occurs when someone can use published or released statistical information to identify an individual who provided information under a pledge of confidentiality. Title 13, United States Code, Section 9, prohibits the Census Bureau from publishing results in which an individual can be identified. Since microdata records are the actual housing unit and person records, the Census Bureau takes steps to prevent the identification of specific individuals, households, or housing units by modifying or suppressing some data on the PUMS files.

The Census Bureau's internal Disclosure Review Board sets the confidentiality rules for all data releases. Using disclosure avoidance procedures, the Census Bureau modifies or removes the characteristics that put confidential information at risk for disclosure. A checklist is used to ensure that all potential risks to the confidentiality of the data are considered and addressed. Although it may appear that the PUMS files show information about a specific individual, the Census Bureau has taken steps to disguise the original data while making sure the results are still useful. The techniques used by the Census Bureau to protect confidentiality in tabulations vary, depending on the type of data.

Data swapping is a method of disclosure avoidance designed to protect confidentiality in data (the number or percentage of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases. A sample of households is selected and matched on a set of selected key variables with households in other geographic areas that have similar characteristics. Data swapping procedures were first used in the 1990 census and also were used for Census 2000.

A major disclosure avoidance method used is to limit the geographic detail shown in the files. A minimum threshold of 10,000 for the national population was set for identification of groups within categorical variables in the state- level PUMS files. A geographic area must have a minimum of 100,000 population to be fully identified in the 5-percent file, and 400,000 for the 1-percent file. Furthermore, certain variables are topcoded where the actual values of the characteristics are replaced by a descriptive statistic, such as the mean of all values above a set value.

Further details on the disclosure avoidance methodologies used for Census 2000 and how they differed from 1990 may be found in [9].

III. Guam and U.S. Virgin Islands PUMS

A. Sampling

For the 2000 Guam and U.S. Virgin Islands Census, every person and housing unit received the same questionnaire with detailed content questions. There were no separate short and long form questionnaires. Thus, there were no sampling, weighting, or variance estimation methods required analogous to the U.S. and Puerto Rico long form methodologies.

A stratified 1-in-10 systematic selection procedure with equal probability was used to select the Guam and U.S. Virgin Island PUMS to create a 10-percent PUMS file. The sampling universe was defined as all occupied housing units including all occupants, vacant housing units, and GQ persons in the census. The sample units were stratified during the selection process. The stratification was intended to improve the reliability of estimates derived from the 10-percent sample by defining strata within which there is a high degree of homogeneity among the census households with respect to characteristics of major interest.

First, the units were divided into three major groups: occupied housing units, vacant housing units, and GQ population. For Guam, a total of 99 strata were defined: 72 strata for occupied housing units, 24 strata for GQ people, and 3 strata for vacant housing units. For the U.S. Virgin Islands the comparable number of strata were: 195, 144, 48, and 3.

The occupied housing unit universe was stratified by: family type, race or ethnic origin of the householder, tenure, and maximum age in the household for Guam. For the U.S. Virgin Islands, "race or ethnic origin of the householder" was replaced by "race and Hispanic origin of the householder." The vacant housing unit universe was stratified by vacancy status for both areas. Finally, the GQ population was stratified by GQ type (institutional, noninstitutional), race or ethnic origin, and age for Guam. For the U.S. Virgin Islands, "race or ethnic origin" was replaced by "race, Hispanic origin."

During the sample selection operation, subsamples were identified as described for the U.S. and Puerto Rico PUMS files. As a result, samples of any size between 1/10 and 1/1000 may be selected by using appropriate two-digit subsample numbers assigned to the microdata samples.

Further details on the sampling procedures for the Guam and U.S. Virgin Islands PUMS files are in [10] and [11].

B. Weighting

The 2000 Guam and U.S. Virgin Islands PUMS were self-weighting. All persons or housing units in the PUMS have a weight of 10.

C. Variance Estimation

Since detailed data was collected for all addresses in Guam and the U.S. Virgin Islands, design factors had to be calculated and generalized for use in calculating standard errors on PUMS estimates. Design factors were calculated for each data item in a data grouping where the data groupings were individual measures and the data items were the possible categorical responses or grouped responses. For example, three of the data groupings were: age, household income in 1999, and tenure. The corresponding data items in these groupings were: age groupings, selected income intervals, and owner occupied / renter occupied. For Guam there were 55 groups and 650 items while for the U.S. Virgin Islands there were 57 groups and 694 items.

The design factors were calculated by first calculating estimates of the j^{th} item total from the i^{th} potential PUM sample for all i and j using the following formula:

$$\hat{Y}_{ij} = 10 \sum_{k=1}^{n_i} y_{ijk} \qquad (i=1,...,10;\ j=1,...,X)$$

X=650 for Guam and 694 for the U.S. Virgin Islands. Where:

n_i = size of the i^{th} sample (either total number of people, housing units, or families depending on the item) and

y_{ijk} = 1 if the k^{th} unit in the i^{th} sample has the attribute associated with the j^{th} item;
0 otherwise.

The complex standard error, $SE_c(\hat{Y}_j)$, of the j^{th} item total estimate from the designated PUM sample, $\hat{Y}_j$, is then calculated as:

$$SE_c(\hat{Y}_j) = \sqrt{10 \left(\sum_{i=1}^{10} (y_{ij})^2 - \frac{1}{10} Y_j^2 \right)}$$

Where:

$y_{ij} = \sum_{k=1}^{n_i} y_{ijk}$ = j^{th} item total (j = 1,...,X; X=650 for Guam and 694 for the U.S. Virgin Islands) within the i^{th} sample (i=1,...,10), and

$Y_j = \sum_{i=1}^{10} y_{ij}$ = j^{th} item total across all i samples (j = 1,...,X; X=650 for Guam and 694 for the U.S. Virgin Islands).

An estimate of the standard error for each data item assuming SRS based on the selected PUMS was then calculated as:

$$SE_{srs}(\hat{Y}_j) = \sqrt{90 y_{ij} \left(1 - (\hat{Y}_{ij} / N_x)\right)} \quad (j=1,...,X)$$

X=650 for Guam and 694 for the U.S. Virgin Islands. Where:

i is fixed and represents the designated PUM sample, and

N_x is the appropriate population total (x = p for person total, h for housing unit total, and f for family totals).

An item design effect (referred to as a design factor) for the j^{th} item is calculated as the ratio of the standard error given the PUMS design to the standard error assuming SRS using the formula:

$$DF_j = \frac{SE_c(\hat{Y}_j)}{SE_{srs}(\hat{Y}_j)} \qquad (j=1,...,X)$$

X=650 for Guam and 694 for the U.S. Virgin Islands.

A group design factor (DF_G) is calculated for each data grouping as a weighted average of the item design factors within the data group.

$$DF_G = \sum_{j \in G} (W_j \times DF_j) \quad (G=1,...,Z)$$

Z=55 for Guam and 57 for the U.S. Virgin Islands.

The weight applied to each item design factor is the ratio of the unweighted item total to the sum of the unweighted item totals within the group

and is calculated as:

$$W_j = \frac{y_{ij}}{\sum_{j \in G} y_{ij}} \qquad (j=1,...,X)$$

X=650 for Guam and 694 for the U.S. Virgin Islands. Where:

i is fixed and represents the designated PUM sample and

G represents the individual groups (G = 1,...,Z; Z=55 for Guam and 57 for the U.S. Virgin Islands).

Variance estimation for estimates from the Guam and U.S. Virgin Islands PUMS files is conducted using the design factors as described for the U.S. and Puerto Rico in section II.D with the integer of "99" (used for the 1 percent sample) replaced with "9."

Further details on variance estimation for the Guam and U.S. Virgin Islands PUMS files are in [12] and [13].

D. Disclosure Avoidance

The methodologies used for disclosure avoidance in the Guam and U.S. Virgin Islands PUMS are the same as described above for the U.S. and Puerto Rico. Since a geographic area must have a minimum population of 100,000 to be fully identified, the only geography indicated on the PUMS are Guam and the U.S. Virgin Islands themselves.

Acknowledgments

The authors wish to thank all of the many people whose efforts supported the production of the PUMS files for Census 2000. The authors also thank Suzanne Dorinski, Steven Hefter, Donna Kostanich, Arnold Reznek, and Philip Steel for their helpful comments on drafts of this paper.

References

[1] Griffin, Richard A. and Alfredo Navarro (1991), "1990 Census Public Use Microdata Sample Design Issues," Proceedings of the American Statistical Association, Survey Research Methods Section, Alexandria, VA: American Statistical Association, pp. 437-442.

[2] Hefter, Steven P. (1999), "Long Form Sampling Specifications for Census 2000," DSSD Census 2000 Procedures and Operations Memorandum Series #LL-5.

[3] Hefter, Steven P. and Philip M. Gbur (2002), "Overview of the U.S. Census 2000 Long Form Weighting," 2002 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association.

[4] Fay, Robert E. and George F. Train (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," Proceedings of the Government Statistics Section of the American Statistical Association, pp. 154-159.

[5] Gbur, Philip M.and Lisa D. Fairchild (2002), "Overview of the U.S. Census 2000 Long Form Direct Variance Estimation," 2002 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association.

[6] Davis, Peter P. and Roger Shores (2004), "Long Form Generalized Variance Specifications for Census 2000," DSSD Census 2000 Procedures and Operations Memorandum Series #LL-19.

[7] Sissel, Dennis D. (2004), "Census 2000 United States and Puerto Rico Public Use Microdata Samples (PUMS) - Computer Sampling Specifications," DSSD Census 2000 Procedures and Operations Memorandum Series #GG-20.

[8] Wolter, Kirk M. (1985), *Introduction to Variance Estimation*, Springer-Verlag New York Inc., New York, NY.

[9] Zayatz, Laura (2002), "SDC in the 2000 U.S. Decennial Census", *Inference Control in Statistical Databases*, Josep Domingo-Ferrer editor, Springer, p.193-202.

[10] Zelenak, Mary Frances (2003), "Computer Specifications for Selecting the 2000 Public Use Microdata Sample (PUMS) for Guam," DSSD Census 2000 Procedures and Operations Memorandum Series #GG-11.

[11] Nguyen, Nganha T. (2003), "Computer Specifications for Selecting the 2000 Public Use Microdata Sample (PUMS) for the U.S. Virgin Islands ," DSSD Census 2000 Procedures and Operations Memorandum Series #GG-15.

[12] Hefter, Steven P. (2003), "Computer Specifications for Census 2000 Guam Public Use Microdata Sample Variance Calculations," DSSD Census 2000 Procedures and Operations Memorandum Series #GG-18.

[13] Hefter, Steven P. (2004), "Computer Specifications for the Census 2000 U.S. Virgin Islands Public Use Microdata Sample Variance Calculations," DSSD Census 2000 Procedures and Operations Memorandum Series #GG-19.