SAMPLE DESIGN FOR THE CENSUS 2000 ACCURACY AND COVERAGE EVALUATION

Randal ZuWallack, Matthew Salganik, and Vincent Thomas Mule, Jr., U.S. Census Bureau
Randal ZuWallack, U.S. Census Bureau, Rm 2501, Bldg 2, Washington DC 20233

Keywords: Accuracy and Coverage Evaluation, Census 2000, Stratified Systematic Sampling

Introduction

Every ten years the Census Bureau attempts to enumerate every person living in the United States. Although a complete count is desired, past experience indicates it is virtually unattainable. According to past census evaluations using demographic analysis, the undercount has ranged from 2.8 million in 1980 to 7.5 million in 1940 (Bureau of the Census, 1997). Beginning with the 1950 census, the Census Bureau began conducting post-enumeration evaluations to estimate census coverage. These evaluations took a case by case matching approach to identify people who were missed and those who were counted. More recent evaluations of this type include the 1980 Post-Enumeration Program (PEP) and the 1990 Post-Enumeration Survey (PES). For the PEP, information based primarily on the Current Population Survey was used to estimate people not counted in the census enumeration (Fay, 1988). A second part of the PEP involved selecting a sample of census records to estimate the number of erroneous census enumerations. Improvements were introduced for the 1990 PES. Rather than using information that was not specifically designed for measuring census omissions, a survey was designed with this sole purpose in mind. As was done in 1980, a sample was also selected for estimating erroneous census enumerations.

In the tradition of improving census evaluations, the Census Bureau is conducting the Accuracy and Coverage Evaluation (A.C.E.) following the Census 2000 enumeration. Similar to the PES, the A.C.E. checks the quality of the census in two ways. One is by comparing data from the census to data collected from an independent sample of housing units to estimate the number of people missed. The other is by selecting a sample of census records to estimate the number of erroneous census enumerations. This information is combined to determine dual system estimates of the total population and many demographic groups, which is then compared to the census results to estimate coverage rates. This paper discusses all phases of the A.C.E. sample design, how the design was effected by the recent Supreme Court decision on sampling for the Census (Department of Commerce v. United States House of Representatives, 1997), and changes made to the design based on an evaluation of the Census 2000 Dress Rehearsal design.

The authors are mathematical statisticians in the Decennial Statistical Studies Division of the U.S. Census Bureau. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

P Sample and E Sample

Because there are two types of coverage errors, missed people and erroneous inclusions, two samples are selected to evaluate census coverage --the population sample (P Sample) and the enumeration sample (E Sample). The P Sample consists of the people living in the housing units designated for A.C.E. interviews. These units are randomly selected from an address list which is compiled independently of the census list for a sample of geographic areas. The list is referred to as the Independent List. The P-sample people are matched back to the census to determine if they were counted or missed. The E Sample consists of people living in a sample of housing units enumerated in the census. The E-sample people are checked to determine whether they were correctly counted in the census, or whether they were erroneously included. Erroneous enumerations include duplicates, fictitious names, people who were born after census day or people who died prior to census day.

Table 1. P Sample and E Sample Comparison

	P Sample	**E Sample**
Estimates	Omissions	Erroneous Inclusions
Universe	All housing units in US[1]	Census housing units
PSUs	Block Clusters	Block Clusters

[1]All housing units in the United States are eligible to be selected except housing units in Remote Alaska.

Block Cluster

The primary sampling units are block clusters, which are one or more geographically contiguous census blocks grouped together. Census blocks are formed by streets, roads, railroads, streams, etc. Forming block clusters involves a complicated hierarchical algorithm involving many rules and constraints. In general, the goal of block clustering is to produce sampling units that average about 30 housing units.

Integrated Coverage Measurement Survey

Until January 25, 1999, when the Supreme Court ruled that statistical sampling could not be used for the House of Representatives reapportionment, the Census Bureau had planned to conduct an Integrated Coverage Measurement (ICM) Survey. The primary goal of the ICM was to produce accurate and reliable direct state estimates, which would then be used for the reapportionment. Preliminary calculations indicated that the ICM allocation may result in coefficients of variation for the Dual System Estimate of approximately 0.5% in all states and standard errors of about 60,000 in the larger states (Schindler, 1998).

The Supreme Court ruling produced a change in the requirements. Direct state estimates were no longer needed for the reapportionment process, and consequently neither was a 750,000 housing unit sample. In contrast to the ICM, which incorporates the information into the population estimates, the A.C.E. results in a second set of estimates which will be used to evaluate the census and potentially for other purposes.

Because the Supreme Court ruling came too late to entirely redesign the sample, we will select an initial sample of block clusters using the ICM design. The independent list will be comprised of the housing units in these selected clusters, called the A.C.E. listing sample. The sample will be reduced during a later process called the A.C.E. Block Cluster Reduction. This has some limitations. The ICM was designed for efficient direct estimates for state total population. The primary goal for A.C.E., however, is to generate reliable demographic group estimates for the purpose of measuring differential coverage. The ICM sample is being selected using proportional allocation within a state. While this might be efficient for total population estimates, it is not efficient for estimating the population of smaller demographic groups. Overall, due to an increased sample size, we expect the reliability to be better for most of the poststrata estimates than the 1990 PES. Also, we expect the state total population estimates to be more reliable than for the 1990 PES.

Stratification and Sort Variables

Historically, coverage rates in the census have varied for many different groups in the population. In 1990, coverage rates were calculated for 357 poststrata identified by region, geographic area, race, Hispanic origin, age, sex, and tenure (own/rent). Although the estimated undercount for the total population was 1.6%, the estimated undercounts for the 357 groups ranged from -8.29% to 21.27% (Thompson, 1992). The poststrata definitions for Census 2000 are currently being researched and thus are not known. However, we are assuming they will be based on similar variables as in 1990 to account for the differential undercount. In order to estimate the coverage rates for several different poststrata with acceptable precision, there must be an adequate amount of sample selected for each of these poststrata. Since the characteristics of people within a block cluster vary, exact sample sizes for these groups are unattainable. However, the variation in the sample sizes for these groups can be improved by grouping similar block clusters together and selecting a systematic sample across these groups. In an attempt to better control the sample sizes from these different groups, block clusters will be classified into categories based on their estimated size, demographic composition, and level of urbanization.

Block clusters will initially be stratified into four mutually exclusive groups within each state: small block clusters (0-2 housing units), medium block clusters (3-79 housing units), large block clusters (80 or more housing units), and American Indian Reservation (AIR) block clusters. These groups will be sampled at different rates during the selection of the A.C.E. listing sample.

Although there will be no differential sampling within these four sampling strata, the clusters will be sorted by several variables in an attempt to sample a diverse set of block clusters. The first sort variable is the American Indian indicator, which has three categories:

- AIR or trustland
- tribal jurisdiction statistical area, Alaska Native Village statistical area or tribal designated statistical area
- all other areas

The second sort variable is the demographic group. Block clusters will be grouped with other block clusters containing similar demographic proportions based on 1990 census data. Assigning this variable to block clusters is described in more detail in the following paragraph. A third variable used for sorting the clusters is the level of urbanization. Each block cluster will be categorized as an urbanized area with 250,000 or more

people, an urbanized area with less than 250,000 people, or a non-urban area. Finally, the clusters will be sorted geographically using county and cluster number.

To aid in selecting a sample that is well represented by the 6 major race/origin groups as well as owners and renters, block clusters will be classified into 12 demographic groups. Although many block clusters tend to have a large proportion of one demographic group, rarely are they entirely composed of only one, thus many clusters may fit well in two or more categories. To ensure that each cluster is assigned to only one group, a hierarchical assignment rule was developed so that when a cluster exceeds the group threshold, it is assigned to that group. These group thresholds were developed by grouping similar 1990 blocks together using a multivariate clustering method[2]. Table 2 lists these threshold values. The order of the hierarchy gives the smaller demographic groups priority over the larger ones and renters priority over owners.

Table 2. Assignment Rule for Census 2000 A.C.E.

Order	Proportion	Threshold
1	Hawaiian and Pacific Islander Renters	0.10
2	Hawaiian and Pacific Islander Owners	0.10
3	American Indian and Alaska Native Renters	0.10
4	American Indian and Alaska Native Owners	0.10
5	Asian Renters	0.20
6	Asian Owners	0.20
7	Hispanic Renters	0.20
8	Hispanic Owners	0.20
9	Black Renters	0.25
10	Black Owners	0.25
11	White and other Renters	0.30
12	White and other Owners	all others

A.C.E. Listing Sample Selection

For each state, a systematic sample is selected for each of the four strata listed in the previous section. In the following paragraphs, the sampling for the medium and large clusters is discussed, followed by the small block clusters and finally the AIR clusters.

As stated earlier, the Census Bureau was preparing to conduct an ICM during the early stages of the sample design. Thus the 25,000 block clusters were allocated to the states to approximately meet the ICM sample requirements, while maintaining a minimum of 300 block clusters per state. Selecting a sample of block clusters within each state results in approximately 2 million housing units to list. The sampling is done in two steps to guard against a listing workload that would be too formidable to complete in time. If the first systematic sample of block clusters results in a workload that is 10% more than the number of housing units allowed for listing, a second systematic sample is drawn from the first to approximately meet the listing constraint. Large block clusters are selected at a higher rate than medium clusters during the A.C.E. listing sample selection. These higher rates coupled with large block subsampling will result in more clusters represented in sample while keeping the total number of designated interviews within budget.

Small block clusters are generally sampled at a lower rate than both medium and large clusters. This is due to cost considerations which are further explained in a later section. These lower sampling rates cause some small cluster to have high weights, which may disproportionately affect the dual system estimates. In an attempt to avoid the problems associated with the high weights we will initially sample 5,000 small block clusters. Using information about these 5,000 clusters we will attempt to target potential problem clusters in the subsampling operation which will reduce the number of small clusters in sample. These initial 5,000 small clusters were allocated to states proportionately to their projected number of housing units in small blocks. This allocation was bounded by two constraints -- a 20 block cluster minimum and a minimum expected sampling rate of 1 in 1000.

To ensure sufficient sample for calculating accurate undercount rates for American Indians on reservations, 355 block clusters will be selected from the block clusters on AIR nationwide. Small block clusters on AIR will not be included in this 355 block clusters. These clusters will be eligible for selection in the small cluster stratum. These 355 clusters were allocated to 26 states proportional to the 1990 population of American

[2]PROC FASTCLUS in SAS uses a multivariate clustering technique called nearest centroid sorting . For details, refer to pages 824-850 of the SAS/STAT User's Guide, Volume 1, Version 6, Fourth Edition.

Indians on reservations. Ten states contained AIR clusters with little or no American Indian population. These clusters are not be included in an AIR stratum, but instead are eligible for selection in the other strata. The remaining 14 states and the District of Columbia contain no block clusters on AIR.

A.C.E. Block Cluster Reduction

As previously stated, the ICM sample will be reduced via the A.C.E. Block Cluster Reduction. This process is the first of three operations that will reduce the 2 million housing units listed down to approximately 300,000 housing units, which is nearly twice the sample size of the 1990 Post-Enumeration Survey (PES). The other two operations are described in the sections that follow. The sample was allocated to the states and the District of Columbia proportional to state population, with a minimum of 1,800 housing units designated for interview per state. The reduction will possibly have variable sampling rates within each state based on race, ethnicity and tenure classification of the block clusters. This differential sampling will help to provide sufficient sample sizes for providing estimates for several different poststrata. In order to provide sample for reliable AIR estimates, the AIR block clusters will not be reduced.

Small Block Cluster Subsampling

Small block clusters, those with between 0 and 2 housing units, get special attention in the A.C.E. These clusters have only a few housing units and are not a cost-effective workload for interviewing and follow-up operations. In order to wisely use our fixed resources we will sample small clusters at a lower rate than both medium and large clusters. Because of these uneven sampling rates the people in small clusters will have high weights. These high weights can disproportionately affect the dual system estimates. In 1990 only about 2.4% of the P sample people and 1.7% of the E sample people lived in small clusters. Yet these clusters contributed almost 10% to the net undercount and 15% to the estimated variance (Fay, 1998). In an attempt to improve our estimates we have developed a special design component to deal with small clusters.

Initially we will select 5,000 small clusters that will be a part of the A.C.E. address listing operation. Then through the small cluster subsampling operation we will reduce the number of small clusters in sample while at the same time attempting to achieve two other goals. First, we would like to prevent any small clusters from having weights that are extremely high compared to other clusters in the sample. Second we would like to limit the weights on the few clusters which we expected to be small, but turned out to be larger. Both of these goals would help to reduce the variance of the Dual System Estimator.

To achieve these goals we will use differential subsampling where the subsampling rates are based on the number of housing units on the Independent Listing and the number of housing units on the Census List. We are in the process of determining the methodology for attaining both goals.

Large Block Cluster Subsampling

Large block cluster subsampling is the final stage in selecting the housing units that are designated for an A.C.E. interview. The underlying concept of large block subsampling is to select a wide range of clusters, while still remaining within the budgeted number of housing units for interview. Assuming that people within a cluster are similar, interviewing all of them is not the most efficient use of resources. Instead, interviewing a smaller piece of several different clusters should provide a more geographically diverse sample.

This stage involves selecting a portion of each block cluster containing 80 or more housing units[3]. Housing units are selected by dividing each large cluster into segments of adjacent housing units, that differ by no more than one housing unit. Then, a sample of segments is selected by taking one systematic sample across all large clusters in a state. All housing units in the selected segments are designated for A.C.E. interview. The sampling rate is determined so that the number of units selected for interview in large clusters added to the number selected in non-large clusters is approximately equal to the interviewing budget. In other words, since all housing units in non-large clusters are designated for interview, the difference between this number and the budgeted number of interviews is the target number of designated interviews from the large clusters.

E Sample Identification

Once the housing units have been selected for A.C.E. interview the next operation is to select the housing units that are in the E Sample. The information gathered from these housing units will be used to estimate the number of erroneous inclusions in the census. Although an overlapping P Sample and E Sample is not necessary, it is more cost efficient. If the E Sample includes many of the same people we can use the

[3]Clusters on American Indian Reservation are not subject to Large Block Cluster Subsampling.

information from the P-sample interview to determine whether they were correctly enumerated and thus do not require a follow-up visit.

In an attempt to create overlapping samples, and thus save money, we will map the block clusters and segments of block clusters that are used to select the P Sample onto the census address list. If this step yields any cluster which will require more than 80 follow-up interviews, the E-sample housing units in these clusters will be subsampled.

Changes from Census 2000 Dress Rehearsal

In 1998, the Census Bureau conducted a Dress Rehearsal to refine the Census 2000 operations. The Dress Rehearsal revealed a few areas in the sample design that needed improvement. Many of the changes were minor operational details, but there are a few enhancements worth noting, two of which involve the treatment of small blocks.

The first change involves the formation of block clusters. Small blocks were not clustered with their neighbors for the Dress Rehearsal. Under certain conditions in 2000, small blocks are clustered with their neighbors. This reduces the total number of small clusters and thus reduces their weights. Overall, this change reduced the number of small clusters by about 65%, from 2,968,956 to 1,029,185. Under the new clustering procedure the initial weights for housing units in small clusters vary from 25 to 632 with an average of 221. Had improvements not been made, they would have ranged from 56 to 1,010 with an average of 588. Figure 1 shows the weight distributions of the 50 states, the District of Columbia, and Puerto Rico using both methods.

Also different in the Dress Rehearsal is the allocation of small clusters to states. In the Dress Rehearsal small clusters were allocated proportionately to the number of medium and large sample clusters in each site. This methodology is inefficient since many states have a large population but very little of it is contributed by small blocks whereas other states have a higher percentage of their population in small blocks. To account for this, in 2000 the small clusters were allocated proportional to the number of housing units projected in small clusters. This generally benefits states with larger proportions of the population residing in small clusters. The two allocations are listed in Table 3 for the states with the five highest and five lowest proportions of the population residing in small blocks.

Much of the A.C.E. operational planning was based on 1990 census data. For instance, the estimated number of housing units for creating the Independent List for each state was estimated based on 1990 information. Since these numbers were then used for renting office space and hiring staff in different areas of the country, exceeding these numbers may poise workload problems. Thus, these estimates became the listing constraints. To help keep the listing close to the listing constraints, two adjustments were built into the design. The first involves an adjustment prior to selecting a sample which is based on expected values. If it appears the listing would be too much based on the preliminary sampling rate, then the sampling rate was decreased. The second adjustment comes in the form of a two step sample. If the clusters selected during the first step surpass the listing constraint, a second sample from the first sample is selected. Without these two procedures, the listing would have surpassed the constraints by over 7.5 percent.

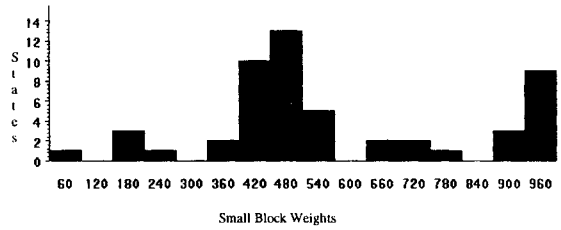
As can be seen by the sampling of changes listed in the above paragraphs, the A.C.E. sample design is continuously being updated and improved. Although there are still details to develop, such as the sampling rates for the small block subsampling and the possible strata for A.C.E. reduction, the framework is in place to provide reliable estimates of census coverage.

Table 3. Initial Small Block Cluster Weights for Selected States

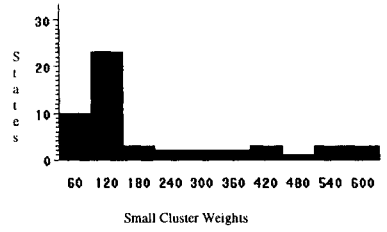
State	Percent 1990 Hus in Small Blocks	Dress Rehearsal Method Weight	Census 2000 Method Weight
North Dakota	11.67%	299	148
South Dakota	9.14%	246	139
Nebraska	5.47%	222	94
Kansas	4.64%	365	113
Wyoming	3.46%	529	617
Rhode Island	0.37%	11	41
New Jersey	0.32%	92	218
California	0.29%	156	467
Hawaii	0.24%	102	306
DC	0.06%	6	25

Figure 1. Frequency of Small Cluster Weights

(a) Dress Rehearsal Clustering



(b) Census 2000 Clustering



References

Bureau of the Census. (1997). Report to Congress: Plan for Census 2000. Washington, D.C.: Bureau of the Census.

Department of Commerce v. United States House of Representatives, No. 98-404 (U.S. filed Jan. 25, 1999).

Fay, R.E. (1988), "Evaluation of Census Coverage from the 1980 Post Enumeration Program (PEP): Census Omissions as Measured by the P Sample", Census Bureau Memorandum, March 10, 1988.

Fay, R. E. (1998), "Small Blocks in the 1990 PES", Census Bureau Memorandum, August 1998 (DRAFT).

SAS Institute Inc., *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 1*, Cary, NC: SS Institute Inc., 1989. 943 pp.

Schindler, E. (1998), "Allocation of the ICM Sample to the States for Census 2000," *Proceedings of Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, to appear.

Thompson, J. (1992), "CAPE Processing Results", Census Bureau Memorandum, March 20, 1992.