INTEGRATED COVERAGE MEASUREMENT SAMPLE DESIGN FOR CENSUS 2000 DRESS REHEARSAL

David McGrath, Robert Sands, U.S. Bureau of the Census
David McGrath, Room 2121, Bldg 2, Bureau of the Census, Washington, DC 20233

Key Words: Differential coverage, Proportional allocation, Dual system estimate, Taylor series

INTRODUCTION

To plan for the 2000 Decennial Census, the Census Bureau is conducting the 1998 Census Dress Rehearsal in the following three sites: Sacramento, CA; Columbia, SC and surrounding counties; and the Menominee Indian Reservation, WI. For Census 2000, the Census Bureau plans to improve census coverage by combining results from the initial phase population count and the new Integrated Coverage Measurement survey to produce dual system estimates of the U.S. population. The Census Bureau will test the dual system estimation methodology by estimating the population of Sacramento and Menominee[1] during the 1998 Census Dress Rehearsal.

PURPOSE OF THE INTEGRATED COVERAGE MEASUREMENT SURVEY

Although the Census Bureau aspires to count every person living in the United States on census day, evaluations of previous censuses indicate the traditional census enumeration undercounts the true population. The Integrated Coverage Measurement (ICM) survey will improve census coverage by estimating for persons omitted in the initial phase, and estimating the number of erroneous or duplicate initial phase enumerations. This is accomplished by re-interviewing persons in a sample of block clusters, matching these data to the initial census enumeration, and subsequently estimating the population by producing dual system estimates (DSE) from the match of the initial enumeration and the reinteview (ICM) data.

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.

[1] The Census Bureau will produce a dual system estimate for the population of the South Carolina site, but this estimate focuses on evaluating census coverage rather than contributing to the census estimate.

Dual system estimation theory requires that the listing[2] and interviewing for the ICM survey be conducted independently of the initial enumeration (Wolter, 1986).

OVERVIEW OF INTEGRATED COVERAGE MEASUREMENT SAMPLE DESIGN

The sample design for the 1998 Census Dress Rehearsal (DR) is a stratified proportionate sample of block clusters. For geographic convenience and to satisfy cost constraints, we cluster ICM housing units into block clusters. ICM interviewers enumerate all persons in selected block clusters during the ICM survey[3].

Research has shown that not only does the Census undercount the total population, but that differential coverage by demographic groups also occurs. The probability of being enumerated in the census varies by race, ethnicity, tenure (owner / renter), and geographic area. For this reason, we stratified the ICM universe by these variables to ensure that each group was adequately represented in the sample. Sampling strata are further substratified by the housing unit size of the block cluster.

We selected the ICM sample in several stages. The first two stages were a systematic selection of block clusters within sampling strata and substrata. Next, small block clusters were subsampled to reduce field workloads. Finally, large block clusters were subsampled to reduce the homogeneity or the clustering of the sample.

SAMPLE SIZE

The Census Bureau calculated site-level person sample sizes to achieve a desired reliability level for each site level population estimate. Columbia and rural (surrounding counties) South Carolina were treated as separate sites for sampling purposes because we hypothesized that enumeration probabilities of persons in

[2] Listing refers to identifying all housing units in a selected area (block clusters in the census dress rehearsal).

[3] For blocks that continue through large block subsampling, all persons in selected groups of housing units are interviewed for ICM.

rural and urban areas would deviate. Reliability was measured by the coefficient of variation (CV) of the site level estimates. Desired reliability levels were as follows:

Table 1. Desired Reliability for Dual System Estimates of Population

Site	Coefficient of Variation (percent)
Sacramento, California	1.5
Combined South Carolina	1.5
Menominee, Wisconsin	5.5

Rough sample sizes for each site were calculated by solving the following equation for the sampling fraction (f):

$$CV = \sqrt{\frac{(1-f) * Z}{f * N}}$$

where: N is the site level population size in persons (from 1990 Census)
f is the sampling fraction
Z includes the design effect and population variance

These calculations yielded sample sizes of approximately 37,500 for Sacramento, 18,750 for rural SC, 18,750 for Columbia SC, and 2,000 for Menominee, WI.

To empirically verify the adequacy of these sample sizes, DSE estimates and variances were simulated for the four areas. To calculate variances for the estimated number of erroneous enumerations and omissions, we used simple random sampling variances for estimated proportions with design effects. Design effects and proportions of erroneous enumerations and omissions for Sacramento were estimated from 1995 Census test data from Oakland, while these data for South Carolina and Menominee were estimated from the 1990 Post Enumeration Survey. The formulae below were used to calculate DSEs and variances by poststrata:

$$DSE = \frac{C * (1-P_e)}{1-P_o}$$

where: C is the Census estimate for the poststrata
P_e is the estimated erroneous enumeration rate for the poststrata
P_o is the estimated omission rate for the poststrata

$$Var(DSE) = \frac{C^2}{n*(1-P_o)^2} [Deff_e * P_e Q_e + R^2 * Deff_o * P_o Q_o + 2R\sqrt{Deff_e Deff_o} * P_e P_o]$$

where: $Deff_e$ is the design effect for erroneous enumerations in the poststrata
$Deff_o$ is the design effect for omissions in the poststrata

$$R = \frac{1-P_e}{1-P_o} = \frac{Q_e}{Q_o}$$

n is the person sample size for the poststrata
Q_e $1 - P_e$
Q_o $1 - P_o$

The expression for Var(DSE) is derived from a Taylor Linearization Approximation of DSE and the fact that $COV(P_e,P_o) = (-P_e P_o) / n$ since a person can't be both an omission and an erroneous enumeration (Griffin and Sands, 1997). DSE estimates and DSE variances were calculated using the rough sample size calculations. These simulations produced expected CVs of 1.6% for Sacramento, 1.2% for Columbia, SC, 1.2% for rural SC, and 5.4% for Menominee. These CVs are close to the desired levels of precision so the rough sample sizes were used for the DR.

SAMPLE ALLOCATION

Next, we allocated the sample size to sampling strata. We empirically tested several stratification methods including proportional allocation and several optimal allocation schemes to determine which technique produced the smallest DSE variances. We conducted simulations using 1990 Census population data to create population estimates and CVs. All of the sample allocation methods produced similar CVs because of the large ICM survey sample size. Proportional allocation was selected for the DR because this technique fairly allocates the sample across race/ ethnicity groups, it is easy to implement, and it is readily understood by the public.

FORMATION OF BLOCK CLUSTERS

The ICM primary sampling unit is the block cluster. Block clusters contain approximately 30 housing units and are comprised of one or more geographically contiguous blocks. Blocks are clustered to create field

work units of roughly equal sizes. Large and small blocks (defined later) were not clustered.

CREATION OF POSTSTRATA

The ICM sample was designed to provide sufficient precision for DSE estimates of total population for the ICM poststrata. ICM poststrata are the lowest level of detail for which DSE population estimates are produced. Poststrata are defined by characteristics of persons and contain relatively homogeneous groups of people. Variables used to define poststrata are race, ethnicity, and tenure. Poststrata are not consistent across sites because we created a poststratum if a demographic group accounted for more than two percent of a sites population. For example, American Indian with 'owner' tenure is a poststratum in Menominee but not in Sacramento.

Note: These poststrata are for sample design purposes. Estimation for the Dress Rehearsal will use more detailed poststrata.

CREATION OF SAMPLING STRATA AND ASSIGNMENT OF COLLECTION BLOCKS

Sampling strata were created that closely resemble the poststrata for which DSE estimates will be produced. Demographic variables race, ethnicity, and tenure were used for stratification to ensure that persons from these groups were adequately represented in the ICM sample. 1990 census block level data were used for this stratification.

Ideally, we would assign DR collection block clusters to strata based on the current demographic composition of the cluster. This was impossible because the source for demographic data was the 1990 Census which contains population data according to 1990 tabulation geography. Since tabulation geography often corresponds with government boundaries and not physical boundaries, 1990 tabulation blocks do not directly match 2000 DR collection blocks. We used the Census Bureau's equivalency file, which relates 1990 census tabulation blocks to Census 2000 DR collection blocks, to assign 1990 persons to DR collection blocks. We estimated the demographic composition of the DR collection block clusters by proportionally distributing 1990 persons from tabulation blocks to collection blocks based on the number of Master Address File (MAF) housing units associated with a tabulation block. This technique should provide a relatively accurate estimate of the composition of DR collection block clusters.

Census collection blocks were assigned to a stratum if they contained a desired concentration of intended persons. Block clusters were stratified into mutually exclusive and exhaustive strata based on the demographic composition of the block cluster. For instance, a block was assigned to the Black Renter sampling stratum in Sacramento if it contained more than 10% black renters. Simulations were conducted to empirically test which cutoffs (both percentage of persons and order of criteria) produced the lowest CVs for DSE population estimates. The selected criteria for Sacramento based on lowest CV are shown below:

Table 2. Algorithm for Assigning Clusters to Sampling Strata (Sacramento, CA)

Sampling Stratum	Criterion
1. Asian / Pacific Islander Renter	Block Clusters with more than 10% API Renter persons
2. Hispanic Renter	Block Clusters not in 1 with more than 10% Hispanic Renter persons
3. Black Renter	Block Clusters not in 1 or 2 with more than 10% Black Renter persons
4. Minority Owners	Block Clusters not in 1, 2, or 3 with more than 30% Black+API+Hispanic Owner persons
5. Other Renters	Block Clusters not in 1, 2, 3, or 4 with more than 50% Renter persons
6. All Other	Block Clusters not in 1, 2, 3, 4, or 5

CREATION OF SUBSTRATA

Block clusters are further stratified into small, medium, and large clusters because these substrata were sampled at different rates. We assigned block clusters to substrata based on the number of housing units contained in the cluster according to the MAF or Address Control File (ACF)

Table 3. Substrata Definitions

Substrata	Number of MAF/ACF Housing Units
Small	0-2
Medium	3-79
Large	≥ 80

CALCULATION OF SAMPLING RATES

Earlier, we created sample sizes for desired numbers of persons; however, the ICM is a sample of block clusters, not persons. To convert the sample size from persons to block clusters, we first converted persons to housing units by using 1990 Census data to calculate the average number of persons per household for each sampling stratum. Next, we converted the housing unit sample size to the block cluster sample size by using the average number of MAF/ACF housing units per block cluster for each sampling stratum/substratum combination.

These block cluster sample sizes apply only to the medium and large substrata. Approximately half the block clusters in the U.S. are small, but they cumulatively still contain very few persons. For this reason, the sample size for small clusters was developed independently from the medium and large sample size. Field division estimated that resources permit the listing and interviewing of an additional 20 percent of the medium and large cluster sample size for small clusters. The independent calculation of the small cluster sample size introduces weight variation because small clusters are now sampled at a different rate. The impact of the weight variation on variances is expected to be small because of the low volume of persons in these blocks.

Strata sample sizes are now allocated to the medium and large substrata. For large blocks, higher first stage sampling rates, in conjunction with the large block subsampling operations (discussed later), result in approximately a self-weighting design in the medium and large strata. The increased rate used was a factor of the average number of housing units in large clusters divided by the average number of housing units in medium clusters for that stratum. For example, if the Black Renter stratum in Sacramento contained an average of 120 housing units in large clusters and 30 housing units in medium clusters, large clusters would be sampled at a four times higher rate in the first stage. The design remains approximately self-weighting because roughly one in four housing units would be selected from this large cluster during large block subsampling.

SAMPLE SELECTION - FIRST STAGE

A systematic sample of block clusters was selected within each sampling stratum/substratum combination where the TakeEvery equaled the inverse of the sampling rate. To stratify the clusters implicitly according to geography, each sampling strata/substrata file is sorted by geographic block location before sample selection.

Implementation of this sample design required the listing of more housing units than the budget allotted. The following table shows the number of housing units budgeted for listing and the number of housing units selected after first stage of sampling:

Table 4. First Stage Sampling Results

Site	Listing Budget (Number of Housing Units)	Number of Housing Units Selected in First Stage
Sacramento, CA	21,000	48,118
Rural SC	10,500	21,719
Columbia, SC	10,500	20,716
Menominee, WI	780	1,822

Selecting large clusters at a higher sampling rate in the first stage caused the violation of the listing budget. Although large block subsampling allows us to achieve the correct person sample sizes, resources do not permit the complete listing of this quantity of housing units.

SAMPLE SELECTION - SECOND STAGE

Solving the listing problem required either spending additional money to list all selected housing units, or drawing a subsample of block clusters selected in the first stage. For the Dress Rehearsal, we decided to attempt to adhere to the budget and select a second stage sample of block clusters. Large clusters were targeted for subsampling because they contain most of the housing units. We subsampled block clusters at a rate that approximately accommodates the listing constraints stated above. A formula was developed to maximize the number of clusters selected without exceeding the listing budget:

$$Sampling\ Rate = \frac{HUs\ budgeted\ for\ listing\ in\ large\ clusters}{HUs\ in\ large\ clusters\ selected\ in\ 1st\ stage}$$

where: The number of housing units budgeted for listing in large clusters (numerator) equals the total housing unit listing budget less the number of housing units selected in medium clusters in the first stage.

For Sacramento and both South Carolina sites, it was only necessary to subsample within large clusters. For Menominee, the housing units in the medium clusters alone exceeded the listing budget. Therefore, medium and large block clusters were sampled at a rate of 1 in 2 for Menominee.

The second stage sampling causes the selection of a larger proportion of housing units from fewer large block clusters. This causes an increase in the variance of estimates because of the homogeneity of persons in these clusters.

SMALL BLOCK SUBSAMPLING

Small block subsampling is the process of selecting a subsample of previously selected small blocks for ICM interviews. Prior to subsampling, all small blocks selected in the first stage were listed during the ICM Independent Listing operations. Small blocks were subsampled because the effort required to conduct interviews is resource intensive compared with the effect these blocks have on population estimates. Sampling small blocks at a different first stage rate combined with small block subsampling creates significant sample weight variation and may have a large effect on variance estimation.

One scenario where a small block can dominate an estimate or variance component is when the block actually contains substantially more housing units than were present on the frame (MAF/ACF). Block clusters were initially classified as small, medium, or large prior to ICM block cluster sampling. After the ICM listing, a more accurate housing unit count is available for small blocks. To protect against extreme growth in a small block, we do not subsample small blocks if 10 or more housing units are discovered during ICM listing. If the small block does not encounter this growth, we select a systematic subsample of small block clusters at a rate of 1-in-10.

LARGE BLOCK SUBSAMPLING

The final stage of ICM sampling is the subsampling of large block clusters. Large block clusters are subsampled because we want to maximize sampling efficiency by reducing the clustering effect caused by the homogeneity natural to persons within a geographic area. To achieve the smallest possible variance, the goal is to spread the sample across more block clusters and select fewer persons within each cluster. A second purpose of large block subsampling is to maintain consistent field interviewer workloads across block clusters.

Block clusters are eligible for large block subsampling if they contain 80 or more ICM listed housing units. The subsampling cutoff of 80 housing units was chosen as a reasonable interview workload. The subsample is drawn by forming segments of adjacent housing units within large block clusters and selecting a subsample of segments.

All large clusters in a sampling stratum will contain a fixed number of segments. The number of segments is based on the TakeEvery for that stratum to ensure the selection of at least one segment from each cluster. Determining the number of segments per large cluster balances sampling and nonsampling error. The smallest variance on sample size is achieved when a large number of segments containing few housing units are created. However, creating a large number of segments increases non-sampling error because interviewers must identify more segments boundaries during field work. This process increases the likelihood that enumerators will select an incorrect unit or make similar non-sampling errors. Standardizing the number of segments per large cluster at the stratum level causes the number of housing units per segment to vary across clusters within the same stratum.

The actual number of housing units in selected medium clusters, according to the ICM listing, will not directly correspond with the numbers from the MAF/ACF. To compensate for this variation, we incorporated both the desired housing unit sample size and the more current housing unit counts from the ICM listing into the large block subsampling TakeEvery calculation. This will result in an achieved sample size very close to the desired sample size.

$$TE = \frac{\text{Number of Listed (ICM) HUs in Large Clusters}}{\text{Desired HU Sample Size in Large Clusters}}$$

For the dress rehearsal, this method partially compensates for loss in reliability due to the unanticipated second stage selection of block clusters.

After calculating the TakeEverys for each stratum, we sample large block clusters on a flow basis after the clusters proceed through the housing unit matching phase. A systematic sample of segments will be selected in each sampling stratum from a frame of all large block clusters. Selecting one systematic sample from a sampling stratum, rather than a separate sample from each large cluster, reduces sample size variability. This allows us to achieve an actual sample size very close to the original desired sample size.

DRESS REHEARSAL SAMPLE RESULTS

The final block cluster and housing unit sample sizes for the Census 2000 Dress Rehearsal are shown in Table 5 on the next page.

Table 5. Summary of ICM Sample Results

Site	Block Cluster Size (substrata)	Number of Block Clusters	Number of Housing Units
Sacramento, California	Small	7	97
	Medium	63	10,181
	Large	320	6,141
	Total	**390**	**16,419**
Rural South Carolina	Small	10	42
	Medium	283	6,450
	Large	25	1,961
	Total	**318**	**8,453**
Columbia, South Carolina	Small	7	60
	Medium	333	6,942
	Large	15	2,222
	Total	**355**	**9,224**
Menominee, Wisconsin	Small	1	2
	Medium	16	327
	Large	4	465
	Total	**21**	**794**
Total Dress Rehearsal	Small	25	201
	Medium	695	23,900
	Large	364	10,789
	Total	**1,084**	**34,890**

Note: Block cluster and housing unit data shown in this table are post small and large block subsampling. Block clusters are displayed in small, medium, and large size categories according to their original classification using housing unit data from the MAF/ACF (i.e., The ICM independent listing provides more accurate housing unit counts that were unavailable during the original assignment of block clusters to substrata.)

REFERENCES

Griffin, Richard and Robert Sands. "Results of ICM Sample Size Determination for the 1998 Dress Rehearsal", Internal Census Memorandum A-2, June 1997.

Wolter, Kirk M. "Some Coverage Error Models for Census Data" (1986). *Journal of the American Statistical Association*, Vol 81, June 1986.