ESTIMATION IN THE CENSUS 2000 DRESS REHEARSAL

Richard Griffin and Elizabeth Ann Vacca, US Bureau of the Census
Richard Griffin, Room 2025, Blding 2, DSSD, Bureau of the Census, Suitland MD 20233

Key Words: Census 2000; Dress Rehearsal, Estimation

The Census Bureau is paving the way into the future with its innovative plan for Census 2000. A major goal of Census 2000 is to reduce the differential undercount, in particular, the longstanding disparity in census coverage of racial and ethnic groups. Scientific methodology exists that allows us to correct the differential undercount, so the Census Bureau is no longer willing to provide a count that does not accurately represent the population. This methodology has been in the research and development stage for many years and the Bureau believes it is time for the census-taking environment to change with our ever-changing society. The plan for Census 2000 assures accurate representation of all U.S. residents at a lower cost than earlier methodology.

This paper provides an overview of the sampling methodology and details of the estimation methodology for the Census 2000 Dress Rehearsal. In the Dress Rehearsal, the Census Bureau is using traditional enumeration methods in Columbia, SC with a Post Enumeration Survey (PES) as a coverage measurement survey. The Census 2000 sampling and estimation plan is being used in Sacramento, CA; that is, sampling for nonresponse followup (NRFU), undeliverable as addressed(UAA) vacant followup and integrated coverage measurement (ICM). A modified Census 2000 sampling and estimation plan is being used in Menonimee, WI; that is, sampling for ICM only.

1. Sampling for Nonresponse and Undeliverable as Addressed Vacant Followup

The purpose of sampling for nonresponse followup and UAA vacant followup is to reduce cost and staffing requirements and save time. For the Dress Rehearsal in Sacramento CA, a sample of housing units not responding to the mail form will be selected to achieve a 90 percent or higher response rate in each census tract. All nonresponding housing units in ICM sample blocks will be included in this sample. Sampling replaces the traditional method of following up all of the nonresponding housing units. A systematic sample of housing units sorted by geography and form type (long vs. short form) will be selected to ensure we distribute the sample evenly across all nonresponding addresses in the census tract.

The sample will be selected on a prespecified date of the mail return period using a variable sampling rate that depends on the response rate in the tract at the time of sample selection. To obtain information from 90 percent of the housing units in each eligible census tract, tracts with lower initial response rates will have a larger proportion of housing units sampled. For example, we will sample the non-ICM sample blocks in a census tract with an 80 percent initial response rate at a 1-in-2 rate and the non-ICM blcoks in a census tract with an 70 percent initial response rate at a 2-in- 3 rate. To ensure equitable treatment of census tracts with high response rates, we use a 1-in-3 rate for all census tracts with an initial response rate of 85 percent or higher.

The bureau will sample postal service identified UAA vacants in non-ICM blocks from each eligible census tract at a 3-in-10 rate. All UAA vacants in ICM sample blocks will be included in sample. The UAA vacants will be sorted by geography and form type to ensure the sample is equitably distributed across the census tract. UAA vacant sampling and NRFU sampling will take place at the same time, but samples will be selected independently.

For the Census 2000 Dress Rehearsal, the Fiscal Year 1998 Budget Agreement instructed the Census Bureau not to use sampling in the South Carolina site, and the Menominee, WI site is not eligible for NRFU or UAA vacant sampling since it is an Indian Reservation. In ICM blocks, **all** nonrespondents and **all** UAA vacants will be in sample. All other census blocks are eligible for NRFU and UAA vacant sampling.

2. Sampling for Integrated Coverage Measurement

The ICM will be used to estimate the total population while accounting for the differential undercount by race and ethnicity. The ICM sample for the Census 2000 Dress Rehearsal is a stratified, systematic selection of clusters of geographically contiguous housing units designed to meet specific sample size requirements. The purpose of the ICM is to correct for the differential undercount in population estimates, particularly the differential undercount by race, ethnicity, and tenure. The goal of the ICM sampling methodology is to meet specific

Note: 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.

reliability criteria for total population estimates, while also allowing reliable estimates for population subgroups, particularly those that have historically been undercounted.

The goal for the Census 2000 Dress Rehearsal is to obtain a 1.5 percent coefficient of variation (CV) on the population estimate for two of the sites (Sacramento, CA and Columbia, SC and the surrounding area) and a 5.5 percent CV for Menominee, WI. The ICM sample is being used in the South Carolina site to evaluate the undercount of the population count.

To simplify ICM data collection and reduce nonsampling error (e.g. matching error) and variability in primary sampling unit size, we form block clusters. Geographically we group adjacent blocks together to form a block cluster. The block cluster is the primary sampling unit. We classify all block clusters for each Dress Rehearsal site into homogenous groups known as sampling strata. These strata are based on the demographic characteristics of each block cluster, such as the racial and ethnic composition and the proportion of renters. The strata definitions correspond to major demographic groups that are historically undercounted and depend on the concentration of these groups within each site. We design the strata definitions to permit production of reliable dual system estimates of select subpopulation groups. We sort the block clusters geographically within each stratum and a systematic sample of block clusters is selected with equal probability, resulting in proportional allocation to sampling strata.

3. NRFU AND UAA Vacant Estimation

A hot deck imputation will be used to obtain population estimates under NRFU and UAA vacant sampling for the Census 2000 Dress Rehearsal in Sacramento and for Census 2000. The hot deck method is called the nearest-neighbor or systematic hot deck. This method, which we have applied extensively for missing data purposes in previous decennial censuses, involves the substitution of data for a nonsampled unit from the nearest neighbor sampled donor unit, with several constraints:

- the donor address must be in the same census tract as the nonsampled address;
- a NRFU sampled address can substitute only for NRFU nonsampled addresses, and likewise for UAA vacant addresses;
- a sampled address cannot be represented in the initial phase estimates more than the inverse of its probability of selection (i.e., it's weight);
- a sampled address in an ICM block cannot serve as a donor;
- and when possible, a nonsampled address in a multi-unit structure will receive donor data from a sampled address in the same structure.

The first and second constraints preserve the independence of NRFU and UAA vacant census tract population estimates. The third and fourth constraints ensure that the hot deck estimate agrees in expectation with the estimate from the standard weighting methodology. The last constraint on the hot deck methodology uses the serial correlation among census items to improve estimation. That is, housing units and their occupants in the same multi-unit structure tend to be more alike on census items such as tenure(whether the housing unit is owned or rented), race, and household size than housing units and persons outside of the structure. Since these items will help to define the poststrata for Integrated Coverage Measurement, preserving their geographic distribution as much as possible is important, and the nearest-neighbor hot deck accomplishes this.

It is the serial correlation among housing units that provides the primary support for the nearest-neighbor hot deck for NRFU and UAA vacant estimation. In particular, the high NRFU sampling rates means that in many census tracts donors will not be used more than once, and therefore we will be able to impute from donors that are very close to the nonsampled donees. The ICM estimation methodology requires a number of data items for poststratification. Since all characteristics for nonsampled NRFU and UAA vacant units are imputed, such units may be placed in poststrata without an additional imputation step.

For each tract, the NRFU weight which is needed for the third constraint is

$$W_{NRFU} = \frac{\#\ Units\ in\ NRFU\ Universe}{\#\ Units\ in\ NRFU\ Sample} \ ,$$

while the UAA vacant weight is

$$W_{UAA} = \frac{\#\ Units\ in\ UAA\ Universe}{\#\ Units\ in\ UAA\ Sample} \ .$$

The plan for Census 2000 includes many response opportunities available to the population, such as replacement forms, Be Counted Forms, and Telephone Questionnaire Assistance. It is highly likely that this multitude of response options will increase the frequency of census forms returned after NRFU and UAA vacant sample selection. When a late return is received from a housing unit that is in either the NRFU or UAA vacant sample, we will have two census forms from that unit: the late return and the enumerator return. In this case the Primary Selection Algorithm (PSA) will determine which form or combination of forms will represent the unit.

When a late return is received from a housing unit that is not in either sample, we will accept the late return to represent the unit and will not impute data for that unit. In forming the imputation donor file, we will remove the unit from the nonsampled portion of the sampling universe, and adjust the sample proportionally to prevent bias in the initial phase estimates. The proportional removal occurs by computing the ratio

$$r_{NRFU} = \frac{A}{B},$$

where A = the number of NRFU sampled addresses in non-ICM blocks in the tract and B = the number of nonsampled addresses in non-ICM blocks in the tract. This ratio is the number of sampled cases that each nonsampled case represents. Then

$$X_{OCC} = r_{NRFU} \; x \; C,$$

where C = the number of occupied nonsampled late returns, is the number of occupied sampled case to be removed. If possible, all of the removed occupied sampled cases will come from occupied late return sampled addresses. But if there are not enough of these sampled cases, then as many occupied non-late returns will be removed from the sample as necessary to reach $X_{OCC.}$ We will use a systematic sample to identify the specific cases to remove from sample. The removal procedure occurs separately for the NRFU sampling universe and the UAA vacant sampling universe. That is, nonsampled late returns in the NRFU universe are accounted for only by the removal of cases from the NRFU sample, while nonsampled UAA vacant returns are accounted for only by UAA vacant sampled addresses.

4. Integrated Coverage Measurement Estimation

For the Census 2000 Dress Rehearsal, dual system estimation, similar to that used for the 1990 PES, will be used to produce direct estimates in Sacramento and Menominee, and used to produce coverage measures for Columbia, SC. The plan is to use a combination of poststratification, raking methods and dual system estimation (DSE) to produce these estimates.

4.1 Poststratification and Estimation

We will use DSE methodology for the Census 2000 Dress Rehearsal. An assumption underlying DSE for census coverage error is that the capture probabilities for the initial phase of the census be equal or that the capture probabilities for the independent survey (ICM phase) be equal. The capture probability for the initial phase does not have to be equal to the capture probability for the ICM phase. Since capture probabilities are not uniform for all members of the population, the goal is to form poststrata based on variables shown to be correlated with coverage error, such as tenure, race and Hispanic ethnicity, age, sex, and urbanization. Then the assumption assumes that the capture probabilities are uniform within these poststrata. Violation of the equal capture probability assumption for the initial phase and ICM is known as heterogeneity and results in biased estimates. Certainly the poststrata improve the estimation over what we would achieve without it. Poststratification is the level at which dual system estimates are calculated within dress rehearsal sites.

For Census 2000, within each state, forming poststrata to group people that have similar coverage properties will be necessary, such as by race/origin groups by age/sex by tenure. If we define six race/origin groups, seven age/sex groups, two tenure categories and three geographic groups in a typical state, a total of (6×7×2×3) = 252 poststrata would be required, more than the sample will support. Iterative proportional fitting, or raking, is a well-known method to weight survey data simultaneously to multiple dimensions of control variables. Raking was first used in the decennial census in 1940 to estimate the weights for the census long form data. In most uses, raking controls the weights of a small survey so that the weighted population equals the results of a recent census or much larger survey. Raking, as defined for the Census 2000 Dress Rehearsal, reverses this situation because the results of the ICM phase are considered superior to the results of the larger initial phase due to the enhanced address list, lower enumerator workload, more intensive interviewing and dual system estimation. Raking in the Census 2000 Dress Rehearsal ICM will allow more poststrata, addressing the heterogeneity concern, while controlling the high variances in the DSE estimates. For example, for Black male renters age 18-29, we would define a different coverage factor for owners and renters. We know that owners and renters have different coverage properties.

The rake will use two marginal sets of poststrata. Before any necessary collapsing, one set will be 42 race/origin by age/sex poststrata and the other set will be two poststrata for tenure (owner/renter). We want to avoid the heterogeneity that would result from calculating DSEs for each marginal control (for example, a DSE for Black males age 18-29 including both owners and renters). Thus, we will first calculate direct Dual System Estimates (DSEs) for each of the eighty four cells of the cross-classification of these two sets of poststrata. We will sum these DSEs to obtain the two marginal sets of poststrata. We will then rake the post NRFU/UAA estimates to these two sets of marginal constraints. Research on the best characteristics to use to define the marginal constraints and on the number of dimensions to use for the raking matrix is continuing. The decision to rake in the Dress Rehearsal does not automatically mean we will use raking in Census 2000. The research will compare possible raking matrices with one dimensional poststratification in terms of bias and variance so that a decision on raking vs. poststratification and the best set of characteristics to define raking controls or poststrata can be made for Census 2000.

5. Treatment of Movers

For the Census 2000 Dress Rehearsal in Sacramento and Menominee and for Census 2000, we are using a procedure known as Procedure C. Procedure C identifies all current residents living or staying at the sample address at the time of the ICM interview plus all other persons who lived at the sample address on census day and have moved since census day. However, we match only the census day residents (nonmovers and outmovers) with the census questionnaire(s) at the sample address. The gross undercoverage rate for movers will be determined by matching the outmovers, i.e., Census Day residents who have left. This ratio is weighted by the number of inmovers. Estimates of the number of nonmovers, outmovers, inmovers, and the percent matched for nonmovers and outmovers, will then be made.

In 1990 and 1980, we used Procedure B. Procedure B identifies all current residents living or staying at the sample address at the time of the ICM interview. We asked that the respondent provide the address(es) where these persons were living or staying on census day. We then matched these persons against names on corresponding census questionnaire(s) at the nonmovers or inmovers census address. Estimates of the number and percent matched for nonmovers and inmovers can be made. The unresolved match rate for inmovers was around 13%. With sampling for nonresponse followup in Census 2000, inmover matching would have an even higher level of difficulty. Procedure B will not be used for Census 2000.

In the 1995 and 1996 Census tests, Procedure A was used. Procedure A reconstructs the households as they existed at the time of the census. We asked that a respondent identify all persons who were living or staying in the sample household on census day. We then matched these persons against names on the census questionnaire for the sample address (and surrounding area). From this information, estimates of the number and percent matched for nonmovers and outmovers can be made. An outmover match rate should be more accurate than an inmover match rate particularly with sampling for nonresponse. For outmovers, the interviewer attempts a proxy interview to obtain their name and new addresses and data that can be used for matching. Then we can attempt to trace the people to obtain an interview with a household member. We match the best available data for outmovers to their census day address in the same manner as used for the nonmovers.

The potential advantage of Procedure C is that the estimate of the number of movers uses inmover data that is more reliable since it is collected from the inmovers themselves. The match rate of the movers is estimated using the outmover match rate so that we avoid the difficulties of inmover matching. Outmover tracing is a problem, however, and often using proxy data for matching is necessary. But since we may not be able to get a good measure of the inmover match rate, Procedure C will be used for the Dress Rehearsal.

6. Calculating Direct DSEs

The dual system model classifies each person as being either included or not in the initial phase, as well as being either included or nor in the ICM:

	Initial Phase		
ICM	In	Out	Total
In	N_{11}	N_{12}	N_{1+}
Out	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N_{++}

In theory, all cells are observable except for N_{22} and any of the totals that include N_{22}. The model assumes independence between inclusion in the two phases. This means that the probability of being in the ijth cell, P_{ij}, is the product of the marginal probabilities, $P_{i+}P_{+j}$. With this assumption, the estimate of the total population, N_{++} is

$$N_{++}=(N_{+1})(N_{1+})/N_{11}$$

This is called the dual system estimator (DSE), see Wolter(1986).

We will calculate DSEs for each interior cell of the raking matrix. Sums of these DSEs will produce the marginal controls to which the initial phase estimates for each interior cell are raked. The initial phase estimates come from the Post NRFU/UAA file that includes the results of the hot deck NRFU/UAA estimation.

Define the following for each person j in poststrata i:

E-Sample is the persons in ICM sample block clusters enumerated in the initial phase

P-Sample is the persons in ICM block clusters enumerated during ICM

W_{ICM} is the ICM sample weight

PR_{CE} is the probability of a correct enumeration

Mover Status:

NM - Nonmover

IM - Inmover

OM - Outmover

W_{NR} ICM nonresponse weight

PR_R Residence probability

PR_M Match probability

N_C to be the estimated number of initial phase persons (post-NRFU estimate)

II_C to be the estimated number of initial phase persons with incomplete information (non-data defined)

N_E to be the estimated number of E-Sample persons

$$N_E = \sum_{\substack{j\in E\text{-}Sample\\ j\in i}} W_{ICM,j}$$

N_{CE} to be the estimated number of correctly enumerated persons

$$N_{CE} = \sum_{\substack{j\in E\text{-}Sample\\ j\in i}} W_{ICM,j} \times PR_{CE,j}$$

$N_{P,NM}$ be the estimated number of P-Sample nonmovers

$$N_{P,NM} = \sum_{\substack{j\in P\text{-}Sample\\ j\ nonmover\\ j\in i}} W_{ICM,j} \times W_{NR,j} \times PR_{R,j}$$

M_{NM} to be the estimated number of P-Sample nonmover matches

$$M_{NM} = \sum_{\substack{j\in P\text{-}Sample\\ j\ nonmover\\ j\in i}} W_{ICM,j}{\times}W_{NR,j}{\times}PR_{R,j}{\times}PR_{M,j}$$

Similarly define $N_{P,IM}$, $N_{P,OM}$, and M_{OM} to be the estimated number of P-sample inmover and outmover persons and the number of P-sample outmover matches.

$ADJR_{OM}$ to be an adjustment factor for resident status for the outmovers

$$ADJR_{OM} = \frac{\displaystyle\sum_{\substack{j\in P\text{-}Sample\\ j\ outmover\\ j\in i}} W_{ICM,j} \times W_{NR,j} \times PR_{R,j}}{\displaystyle\sum_{\substack{j\in P\text{-}Sample\\ j\ outmover\\ j\in i}} W_{ICM,j} \times W_{NR,j}}$$

Then the DSE coverage factor for poststrata i is given by:

$$\hat{CF}_{DSEi} = \frac{N_c - II_c}{N_c} \times \frac{N_{CE}}{N_E} \times \frac{N_{P,NM} + N_{P,IM} \times ADJR_{OM}}{M_N + \dfrac{M_{OM}}{N_{P,OM}} \times N_{P,IM} \times ADJR_{OM}}$$

and $DSE_i = (CF_{DSEi})(N_{Ci})$.

7. Raking

The marginal dual system estimates for the raking estimates are obtained by adding the cell DSEs for the corresponding rows (p) or columns (h). Since the raking is done for a two dimensional matrix, the notation from section 6 will now be changed. DSE_i which is the estimate for poststratum i will now be denoted $DSE_{p,h}$ For example for the poststratum for black male renters age 18-29, p is for black males age 18-29 and h is for renters:

$$DSE_p = \sum_h DSE_{p,h}$$

and $DSE_h = \sum_p DSE_{p,h}$

Define a matrix M of dimension 100xpxh.

Let $M(1,p,h) = I_{p,h}$ (the initial phase count in cell p,h)

```
Set: k = 2
    for p=1,.., if $\sum_h M(1,p,h) > 0$, set
        $RATIO_p = \dfrac{DSE_p}{\sum_h M(1,p,h)}$.
    Else set $RATIO_p$=1
    for p=1,..., and h=1,..., set
    $M(2,p,h) = RATIO_p \times M(1,p,h)$
An iterative process is now begun (*):
    let: k = k + 1
        for h=1,..., if $\sum_p M(k-1,p,h) > 0$, set
            $RATIO_h = \dfrac{DSE_h}{\sum_p M(k-1,p,h)}$
        else set $RATIO_h = 1$
        for p=1,..., and h=1,..., set
        $M(k,p,h) = RATIO_h{\times}M(k-1,p,h)$
    let: k = k + 1
        for p=1,..,35, if $\sum_h M(k-1,p,h) > 0$, set
            $RATIO_p = \dfrac{DSE_p}{\sum_h M(k-1,p,h)}$
        else set $RATIO_p = 1$
        for p=1,..., and h=1,..., set
        $M(k,p,h) = RATIO_p{\times}M(k-1,p,h)$
            Set $MAX{=}\underset{h}{MAX}(|DSE_h - \sum_p M(k,p,h)|)$
        If MAX ≥ 1 and k<100, return to the top of the
        iterative process (*). Otherwise, stop.
```

Define the raked coverage factor and the raked dual system estimate for each cell (p,h) by

$$DS\tilde{E}CF_{p,h} = \frac{M(k,p,h)}{I_{p,h}} \quad \text{and} \quad D\tilde{S}E_{p,h} = M(k,p,h),$$

respectively.

8. Small Area Estimation

Simple synthetic estimation will be used to calculate estimates for blocks within poststrata (interior cells of the raking matrix) for all data products, including Public Law #94-171, redistricting data. In each block, for each poststratum with a estimated undercount (coverage factor greater than 1) person records will be created by replicating initial phase person records (with controlled rounding) in the same block and poststratum. As a result of tabulations from the resulting file, for a particular poststratum, the block estimate is obtained by multiplying the initial phase estimate by the poststratum coverage factor. For example, suppose the coverage factor for black male renters age 18 - 29 in a given site is 1.05. Then the initial phase estimates of black male renters age 18-29 in all blocks in the site will be multiplied by 1.05 (with controlled rounding) to produce the block census estimates.

For poststrata with a estimated overcount (coverage factor less than 1) we need to reduce counts from the initial phase. We do not want to eliminate from the count any person who responded to the initial phase. Thus we will only allow whole person imputations to be replicated and placed in a special overcount category (tabulated with an effective weight of -1). We will distribute the overcount to the blocks in proportion to the number of whole person imputations in the block in the poststratum with controlled rounding.

9. Service-Based Enumeration (SBE)Estimation

To estimate the population of persons without a usual home in the Census 2000 Dress Rehearsal in Sacramento and Menominee, the Census Bureau will use multiplicity estimation and replication imputation to create a file of all persons in this population. The multiplicity question asked of all persons enumerated on a "random" day of the week at a soup kitchen, mobile food van, or shelter is of the following form: How many days during the past week, including today, have you visited shelters, soup kitchens, or mobile food vans? After unduplicating enumerated persons, the multiplicity estimator is as follows:

$$N = 7\sum_{k=1}^{M} \frac{1}{A_k} = \sum_{k=1}^{M} w_k$$

where N is the estimated SBE population, M is the number of unduplicated persons enumerated on the selected day, and A_k is the response of person k to the multiplicity question. The weight, w_k, for person k may not be an integer. For example, if $A_k = 4$, then $w_k = 1.75$. If the weight for a person is an integer, the person's record will be replicated a number of times equal to the weight minus 1. If the weight is not an integer, the number of replications will not exceed the greatest integer component of the weight (i.e., if the weight is 1.75 the number of replications will be 0 or 1). Controlled rounding will be used.

All Be Counted Form (BCF) persons with no usual residence who do not match to a SBE person will be included in SBE counts with effective weight equal to 1. This is an attempt to count persons with no usual residence who do not use services.

If a BCF person with no usual residence matches to a SBE person, the BCF will be discarded and the SBE person will be flagged. The SBE person will not be replicated regardless of their multiplicity weight (i.e., only counted once). Assuming perfect matching and that all persons with no usual residence who use services fill out a BCF, this will correct weighted counts from some of the double counting of persons who use services but not on the day of interview. A false non-match for a BCF with no usual residence will mean an overcount since (1) we will not discard the BCF and (2) we will replicate the SBE person who should of matched as indicated by their multiplicity weight. A person with no usual residence who uses services but not on the day of interview who does not fill out a BCF will result in an undercount since (1) there is no BCF form and (2) we will not replicate matches assuming BCFs have been filled. Even if there is perfect matching and all persons who use services fill out a BCF there will still be double counting. This will occur when a person with no usual residence who uses services but not on the interview day fills out a BCF. He/she will not match to a SBE person so the BCF will be retained while another person in SBE will be replicated to account for this BCF person.