A COMPARISON OF ALTERNATIVE ESTIMATION METHODOLOGIES FOR CENSUS 2000

James Farber and Richard Griffin, U.S. Bureau of the Census
James Farber, DSSD Room 2505-2, Bureau of the Census, Washington, DC 20233

Key Words: Weighting, Hot Deck, Imputation, Nonresponse Followup

Abstract

Recent decennial censuses have followed trends of decreasing mail return rates and accuracy, and increasing data collection expenses. In response, the Census Bureau plans a number of sampling operations for Census 2000, including sampling for nonresponse follow-up (NRFU) and sampling of undeliverable-as-addressed (UAA) vacants. UAA vacant addresses are those that are identified by the United States Postal Service as vacant. NRFU addresses are those that are not UAA vacant and that do not self-respond to the census. Although sampling of these addresses will save time and control costs in the census, it also means that a fraction of the population will not be physically enumerated. An estimation method is required to account for the population residing at nonrespondent and UAA vacant addresses not in either the NRFU or UAA vacant samples.

A number of methodologies have been proposed for NRFU and UAA vacant estimation. This paper outlines the underlying theory of these methods, and the advantages and disadvantages of each. In addition, this paper describes the results of empirical research conducted to compare the alternative estimation methods and to identify the method that can optimally be implemented in Census 2000.

I. Introduction

Census 2000 will occur in two major phases. The initial phase corresponds closely to the traditional census-taking process, in which an attempt is made to contact and count the entire population of the country. The second phase of Census 2000 includes Integrated Coverage Measurement (ICM), the program that provides an independent population roster at about 750,000 housing units. The ICM results can be compared to the initial phase results to determine the demographic and geographic distribution of undercoverage so that the final census estimates will have an undercount that is significantly lower than those of previous censuses.

To allow enough time for completion of the ICM phase and to control costs, the initial phase will include two operations that differ from those traditionally used: nonresponse followup (NRFU) sampling and undeliverable-as-addressed (UAA) vacant sampling. In the 1990 Census, enumerators visited all addresses that did not voluntarily respond to the census, an operation that required hundreds of thousands of temporary employees, lasted several weeks, and cost hundreds of millions of dollars. An exhaustive NRFU of this sort would be even more expensive and time-consuming in Census 2000 if the trend of decreasing response rates continues as expected. Thus sampling for NRFU will be used to save time and control costs. In particular, NRFU sampling will give ICM enough time to be completed so that its results may be combined with the initial phase results to produce a one-number census before the legally-mandated deadline of December 31, 2000.

Closely related to NRFU sampling is sampling of undeliverable-as-addressed (UAA) vacant addresses, which are addresses identified by the United States Postal Service as vacant. The 1990 Census included an additional followup to verify the status of vacant and nonexistent addresses, which, like exhaustive NRFU, consumed census resources. In Census 2000, the Postal Service will help to identify vacant addresses. However, previous research has demonstrated that a nontrivial number of UAA vacants are actually occupied (Green and Vazquez 1996). UAA vacant sampling will serve as a quality check on the identification of vacant addresses.

These two sampling operations mean that a portion of the population, the persons residing at nonsampled addresses, will not be physically counted in the initial phase of the census. Thus an estimation method is needed to account for this missed population. In the remainder of this paper, we will review the NRFU and UAA vacant sample designs, describe some alternative estimation methods, and present the results of empirical research that was undertaken to identify the optimal initial phase estimation method for Census 2000.

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.

II. Sample Design

As currently planned, NRFU sampling will follow the design that has been denoted as Direct Sampling to 90% Response. In each of the more than 60,000 census tracts, an initial completion rate, defined as $\frac{\text{\# of Respondents + \# of UAA Vacants}}{\text{Total \# of Addresses in the Census Tract}}$, will be computed about two weeks after Census Day. NRFU samples will be selected independently in each census tract, with sampling rates large enough to raise each completion rate to at least 90%. In census tracts with an initial completion rate of at least 85%, the NRFU sampling rate will be 1-in-3. For example, a census tract that has 60% initial completion will have a 3-in-4 NRFU sampling rate. The NRFU samples will be selected systematically with independent random starts in each census tract from the universe of nonrespondent addresses, the addresses that do not self-respond and are not identified as UAA vacant.

UAA vacant samples are also selected independently in each census tract from the universe of addresses that are designated as UAA vacant. The UAA vacant sampling rate in all census tracts is 3-in-10. Other types of UAA units may occur, such as an occupant who refuses to accept a mailed census form. These cases are considered as nonrespondents and are placed in the NRFU sampling universe.

Under both sample designs, there is 100% followup of nonrespondent and UAA vacant addresses in census blocks selected for the ICM sample to facilitate the requirements of ICM estimation. And for variance estimation purposes, the minimum sample size is two units for any census tract, unless only one unit is in the sampling universe, in which case that unit will be in the sample with certainty.

III. Alternative Estimation Plans

After the samples are selected, enumerators will visit only the sampled addresses. Following data collection, the initial phase population estimate for any geographic area can be decomposed as

$$\hat{Y}_{TOT} = Y_R + Y_{inNRFU} + \hat{Y}_{outNRFU} + Y_{inUAA} + \hat{Y}_{outUAA}$$

where

$\hat{Y}_{TOT}$ is the total population estimate for the census tract;

Y_R is the known population in respondent housing units;

Y_{inNRFU} is the known population in sampled nonrespondent housing units;

$\hat{Y}_{outNRFU}$ is the estimated population in nonsampled nonrespondent housing units;

Y_{inUAA} is the known population in sampled UAA vacant housing units;

$\hat{Y}_{outUAA}$ is the estimated population in nonsampled UAA vacant housing units.

This decomposition shows the two estimation steps that are needed: estimation of the nonsampled NRFU population and estimation of the nonsampled UAA vacant population. A number of alternative methods are available for these two types of estimation, three of which are discussed below.

A. Simple Weighting Method

Simple weighting is the standard survey estimation methodology (Cochran 1977). Under simple weighting, the population estimate for the sampling universe in any geographic area is

$$Y_{in} + \hat{Y}_{out} = \Sigma_{i=1}^{n} W_i Y_{in,i} \text{ , where}$$

$Y_{in,i}$ is the household size of the *i*th sampled unit from the sample of size *n*, and

W_i = 1 if unit *i* is in an ICM block since these units are in the sample with probability one, or

$$= \frac{\text{Sampling Universe Size in non-ICM Blocks}}{\text{Sample Size in non-ICM Blocks}} .$$

Simple weighting yields nearly unbiased results, even with the adjustment due to complete followup in ICM blocks. This method is also easy to explain, and has an extensive history of use in numerous surveys conducted by the Census Bureau. However, simple weighting may not be the optimal estimation method, particulary for small areas.

Two potentially damaging situations arise regarding estimates for small census blocks. The first is a possible absence of sample. For example, the lowest NRFU sampling rate in Census 2000 will be 1-in-3, which means that a census block with one or two nonrespondent housing units could potentially receive no NRFU sample, and thus the simple weighting block estimate $\Sigma_{i=1}^{n_{blk}} W_i X_i$ would be zero. Given that the address list shows that at least one nonsampled nonrespondent housing unit exists in the census block, we would have difficulty believing and explaining our estimate of zero population. And while this underestimate would be offset by an overestimate in some other census block since the census tract estimate is nearly unbiased, it is nonetheless advantageous to have "realistic" census block estimates for redistricting purposes. The second disadvantage of simple weighting for small area estimates is the possibility of undue influence for large households. A household that contains an exceptionally large number of persons may artificially boost the census block population estimate. For example, if a household of 12 persons makes the sample, then the estimate of the block containing that household will likely be highly influenced by that one household. This is of particular concern for small census blocks, where there may not be enough respondents or sampled nonrespondents to offset the large household.

B. Hot Deck Imputation Method

The hot deck is an imputation procedure that has a long history of use as a missing data procedure in the decennial census. Since nonsampled addresses have completely missing information, the hot deck can be used to fill in that missing data from, for example, the nearest sampled address. Since household size is one of the imputed characteristics, the hot deck can also provide population estimates for the nonsampled addresses. Each sampled unit receives a weight W_i using the simple weighting definition above. But instead of weighting up each sampled unit, the hot deck estimate is produced by replicating each sampled unit a maximum of W_i - 1 times into nonsampled units in the census tract, which produces a household size, either known or imputed, for every housing unit in the census tract. The hot deck population estimate for any geographic area can then be computed simply by summing the household sizes of the units in the area. The census tract estimates produced by this method are identical in expectation to those produced by simple weighting and hence are nearly unbiased. Each sampled unit can be replicated a maximum of W_i - 1 times since the sampled unit represents itself in its weight.

The motivation behind this method is to provide more realistic small area estimates. This is accomplished by using replication, which spreads population estimates out over geography, as opposed to weighting, which places population estimates only in census blocks that contain sampled units. The hot deck is not affected by lack of sample in a block: nonsampled units in blocks with no sample will still receive imputed household sizes that will be realistic since they are substituted directly from nearby sampled units. It is this serial correlation, the tendency of nearby housing units to be similar on census items like household size and race, that gives support to the hot deck as an initial phase estimation method. However, the hot deck does suffer from the potential problem of influential large households.

C. Schafer Method

The Schafer method (Schafer, 1995) uses statistical modeling to produce population estimates. Like hot deck imputation, Schafer will produce more realistic small area estimates than simple weighting. In addition, Schafer uses more information than the previous two methods in estimation: data from respondents in addition to sampled nonrespondents are included in model fitting. However, Schafer is significantly more complicated than the other methods, and has no history of use at the Census Bureau.

Individual models are built for each housing unit and person item on the census form, including household size, and rely on the monotone[1] missingness pattern in the data for nonsampled housing units. The basic model consists of a sequence of logistic regressions, which yield an estimated probability of each outcome for a missing item. For example, householder race/origin is modeled first as Non-Hispanic Black or not, then if the first result is not Black, the second model is Hispanic or Non-Hispanic Other. The Non-Hispanic Other category for race/origin could be further divided into White, Asian and Pacific Islander, and American Indian if these groups have enough data to support modeling. The current models, their outcomes, and their order are:

- Status: Occupied, Vacant, or Nonexistent
- Tenure: Own or Rent
- Race/Origin: Black, Hispanic, or Other
- Household Size: 1, 2, 3, 4, or 5 or more

For the Household Size model, a Poisson regression is used to estimate the size for the 5 or more category.

The data for model fitting is provided by all of the respondent and sampled nonrespondent housing units in an entire local census office (LCO), which vary in size from 100,000 to more than 300,000 housing units. The models include parameters for trends at the levels of the LCO, census tract, and census block, so that geographic heterogeneity, the tendency for housing units to vary increasingly over large geographic areas, and serial correlation are considered in the estimation. Thus a large amount of information is incorporated into producing an estimated household size for each nonsampled address. The modeling also follows probabilistic trends in the observed data, meaning that values are imputed only as frequently as they occur in the observed data.

[1] A monotone pattern means that a data item missing a value for one item will have missing values for all items that follow it in the model-fitting sequence.

Like the hot deck, Schafer's method produces an estimated household size for each nonsampled unit, but uses significantly more information for estimation than the previous methods. Thus population estimates from Schafer will likely be more realistic and accurate than those provided by the hot deck method. Census block estimates are not at high risk to the problems of influential large households or lack of sample in small blocks. Large household sizes will be imputed only as frequently as they occur in the observed data. And since imputation is based on data from an entire LCO, small blocks will have enough data for estimation.

IV. Empirical Comparison of the Methods

To identify the method that will provide optimal initial phase population estimates in Census 2000, we conducted a simulation study that compared the census block, census tract, and LCO estimates produced by two of the three methods: we did not have time to implement Schafer. We selected 1990 Census data from one LCO from Sacramento, CA, one of the sites in the 1998 Census 2000 Dress Rehearsal. Some summary statistics about this LCO from the 1990 Census are presented in Table 1.

Table 1. Sacramento LCO 1990 Census Results

Housing Units	227,606
Respondent Housing Units	149,496
Nonrespondent Housing Units	73,312
UAA Housing Units	4,798
Population	562,029
Black Population	64,934
Hispanic Population	92,829
Renter Population	261,673
Number of Census Tracts	124
Number of Census Blocks	6,111

Using the data from this LCO, we used the following simulation methodology to compare the estimation methods. We first identified the universe of UAA vacants. One limitation from using 1990 Census data is that the reason for a UAA designation was not recorded; that is, we cannot identify UAA vacants from other types of UAA units. Based on an evaluation of the 1990 Census, we randomly designated half of the 1990 UAA units as UAA vacant and the remainder as UAA other (Treat 1992). The second step was to compute the initial completion rates for each census tract and select the NRFU and UAA vacant samples. Finally, we computed total population and three demographic estimates, Black, Hispanics, and Renters, for census blocks, census tracts, and the LCO. A total of 1,000 simulations were conducted. The results of each simulation were compared to the 1990 Census counts to obtain measures of mean squared error in the population estimates produced by each method.

It should be noted that we actually implemented two varieties of the hot deck. The first is the method described above, in which nonsampled addresses receive donor data from the nearest sampled address. We also computed estimates for a large block hot deck, which is identical to the nearest-neighbor hot deck except for large census blocks, those that contain at least 100 housing units. Under the large block hot deck, nonsampled units in these census blocks could receive donor data only from a sampled address in the same census block. Anecdotal evidence suggests that housing units in large blocks tend to be more alike than those in other census blocks, and hence estimation may be improved by the constraint added under the large block hot deck method.

V. Results

First, there were no differences between the nearest-neighbor hot deck and the large block hot deck for census blocks. This occurs because in large blocks the nearest-neighbor hot deck nearly always satisfies the large block hot deck constraint of within-block donation, and therefore formalizing this constraint has little effect on the estimates. We will therefore limit further discussion to comparing simple weighting and the hot deck.

The mean relative root mean squared errors (RRMSE) for simple weighting and the nearest-neighbor hot deck by geographic area for the total population are presented in Table 2. We omit the results of estimation for the demographic groups since they parallel the total population estimation results.

Table 2. Mean Total Population RRMSE's

Area	Weighted Mean RRMSE	
	Simple Weighting	Hot Deck
Block	3.99%	3.91%
Tract	0.59%	0.67%
LCO	<0.01%	<0.01%

For very large geographic levels, such as the LCO, there is no difference between the two methods.

For census tract population estimates, simple weighting does slightly better, which is expected. The hot deck integerizes the weights, since hot deck replication is an all-or-nothing procedure. That is, either the entire sampled unit is replicated some number of times or it is not. This differs from simple weighting, which allows sampled units to count in the final estimates a fractional number of times. This situation increases the variance of the hot deck estimates, leading to the slightly higher RRMSE for hot deck census tract estimates.

One might expect a similar occurrence for census block estimates, but in fact the hot deck does slightly better than simple weighting at this level. This is due to the problems with simple weighting and census block estimation discussed above, particularly the problem of lack of sample in some census blocks. The simple weighting RRMSE is inflated from an increase in the variance because of this problem, while the hot deck RRMSE is not affected. Thus the hot deck RRMSE is slightly lower than that of simple weighting. Figure 1 contains a mean difference plot that better illustrates how the hot deck does better than simple weighting for census blocks. For "good" census blocks, those with a relatively large population, the methods produce equitable results. But for "problem" blocks, those with small populations or with a wide distribution of household size, the hot deck generally does better. A greater number of problem blocks have positive RRMSE differences, indicating that the hot deck has a lower RRMSE for these blocks.

Since the nearest-neighbor hot deck is the planned estimation method for Census 2000, a more detailed examination of its results is of interest. Figure 2 gives a breakdown of census tract hot deck RRMSE by response rate. Note that the sample design and estimation method generally satisfy the primary goal of treating census tracts equitably. Most of the RRMSE's lie in a band between 0.5 percent and 1.5 percent, showing that the accuracy of the estimates is nearly equal whether the census tract has a 30 percent initial completion rate or an 85 percent initial completion rate. However, the three outliers indicate weaknesses in the sample design. The maximum RRMSE occurs in a census tract in which 25 percent of the addresses are UAA vacant; in most census tracts, no more than 10 percent are UAA vacant. In such a census tract, the 3-in-10 UAA vacant sampling rate is simply too low, leading to an increased RRMSE. The other two outliers are very small census tracts that have average initial completion rates. In these census tracts, fewer mail return and sample units are available to offset the variability in the estimation for nonsampled units, which results in higher RRMSE. In all three cases, higher sampling rates would bring the RRMSE into line with those of the other census tracts.

VI. Conclusions and Limitations

For all three geographic levels, the census block, census tract, and LCO, both the nearest-neighbor hot deck and simple weighting produce estimates of nearly equal accuracy. However, it is clear that simple weighting does suffer from the problem of lack of sample in some census blocks, leading to slightly greater uncertainty in the simple weighting estimates overall for census blocks. Since the two methods are nearly equal for all other areas and since the hot deck does not suffer from this problem, the nearest-neighbor hot deck is the optimal estimation method for Census 2000. Under the current sample designs, the hot deck produces equitable estimates for most census tracts, although greater uncertainty can result for outlier tracts due to weaknesses in the sample design. But changing the sample design would not be efficient at this point since the gains in accuracy would likely be small: the uncertainty of the outlier census tracts is only slightly greater than that of other census tracts.

It should be noted that these conclusions are based on simulations conducting using only one 1990 LCO. Further simulations should be conducted to make the results more robust. Two other limitations in this research should also be noted, although neither of them affect the results of the comparison of methods since all of the methods were affected equally. First, as described above, the reason for an address being UAA in 1990 was not recorded, so we had to simulate which UAA's were vacant for each iteration in the simulation. Also, we did not simulate ICM sampling, so our sample sizes were lower than they what they would have been with ICM.

VI. References

Cochran, W. G. (1977), *Sampling Techniques*, Third Edition, New York: John Wiley and Sons.

Green, S. L., and Vazquez, V. (1996), "1995 Census Test - Final Report of the Postal Identification of Vacant and Nonexistent Units Evaluation," DMD 1995 Census Test Results Memorandum No. 23, Washington: Bureau of the Census.

Schafer, J. L. (1995), "Model-Based Imputation of Census Short-Form Items," in *Bureau of the Census Proceedings of the 1995 Annual Research Conference*, pp. 267-299.

Treat, J. B. (1992), "Results from the Postmaster Return Questionnaire Delivery Project: Stage II Evaluation," DSSD 1990 REX Memorandum Series #K-16, Washington: Bureau of the Census.

Figure 1. RRMSE Difference by Mean RRMSE

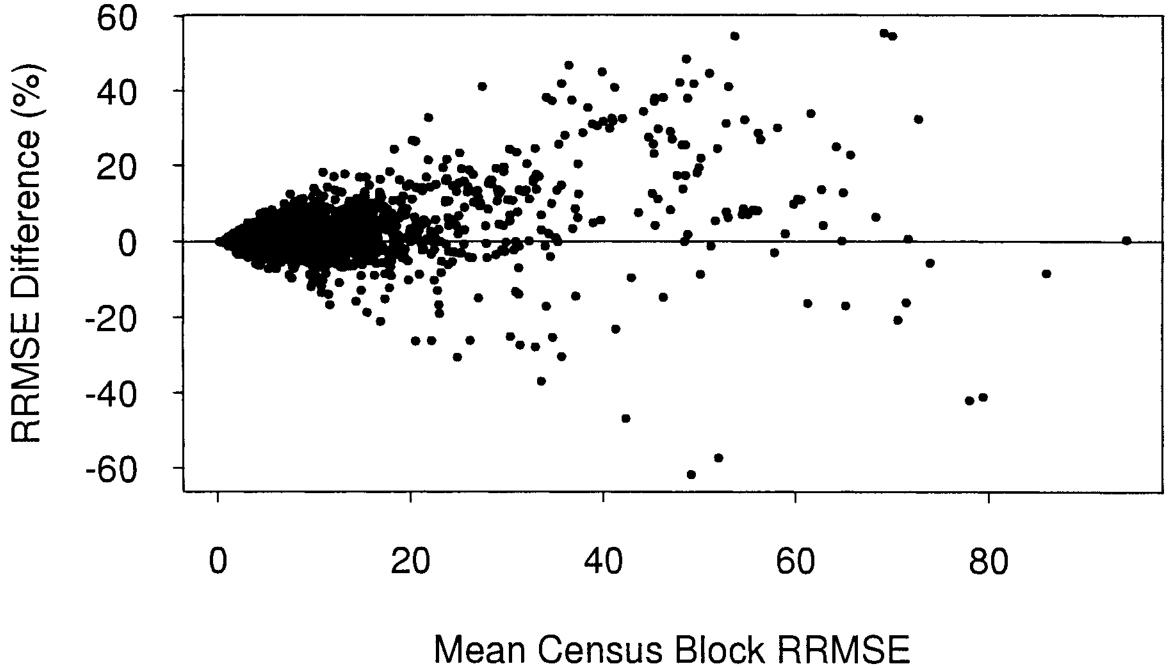

Figure 2. Hot Deck RRMSE by Response Rate