A COMPARISON OF IMPUTATION METHODS FOR SAMPLING FOR NONRESPONSE FOLLOW-UP

Jim Farber, Bureau of the Census[1]
Decennial Statistical Studies Division, Bureau of the Census, Washington, DC 20233

Key Words: Bias, Variance, Simulation

Abstract

In previous mail-out Decennial Censuses, enumerators were sent to conduct personal interviews at all households that did not return census questionnaires. This massive undertaking has become prohibitively expensive, however, and has led the Census Bureau to plan to visit only a sample of these households in Census 2000. Though it will save money, this sampling for nonresponse follow-up will also create an unprecedented amount of missing data. In particular, no data will be available for the households that do not mail back their census forms and are not chosen in the follow-up sample.

Traditionally, the Census Bureau has imputed missing data for an entire household using the responses from a nearby household. However, with sampling for nonresponse follow-up, the nearest housing unit may be quite far and thus quite different from the nonrespondent household. A number of methods have been developed to cope with this problem. This paper gives a review of these methods, and an assessment of their performance in a simulation study. The simulations yield estimates of bias and variance, which allow for comparison of the methods. This information will assist in the selection of the imputation method that will best meet the goals of improved accuracy and efficiency in Census 2000.

I. Introduction

Sampling for nonresponse follow-up (NRFU) is an innovation in the decennial census process that will help the Census Bureau achieve its primary goals of a faster, less costly, and more accurate census in 2000. Sampling for NRFU will certainly be faster and less costly than exhaustive NRFU, since a fraction of the housing units that are most difficult to count will be personally enumerated. These households, which can require several enumerator visits, cost approximately 18 times more to enumerate than a mail return housing unit in the 1990 Census (The Plan for Census 2000, 1996). At the same time, sampling may improve the accuracy of the census. In the 1990 Census, population undercounts varied greatly across factors such as race. Sampling will enable the Census Bureau to target certain population groups and geographic areas that traditionally have high nonresponse rates. The additional sample allocated to these groups and areas may yield the information needed to reduce their persistent undercounts.

The extent to which the Census Bureau's goals are met depends on the actual implementation of sampling for NRFU. The details of NRFU sampling have not been finalized, but the general outline is as follows. All addresses from which census forms have not been received by a certain date will comprise the NRFU universe. A sample will be taken from this universe, and census enumerators will personally visit the sampled addresses. The enumerators will determine if a housing unit physically exists at the address, and if so, will collect data about the unit and any residents. The responses from the sampled and mail-return (MR) addresses will then be used in a model or procedure to impute characteristics for the nonrespondent, nonsampled addresses. Finally, the MR, NRFU sampled, and imputed NRFU nonsampled data will be combined to create a traditional Census roster of households and persons in those households.

II. NRFU Sampling Options

There are several alternatives available for certain components of the general NRFU sampling procedure described above. First, the primary sampling unit can consist of either a census block or an individual address. Census blocks are geographic areas defined by visible landmarks. Under block sampling, all nonrespondent addresses in selected blocks would be visited by enumerators for data collection. Blocks could be stratified by some factor, such as race, to ensure that traditionally undercounted groups are represented in the sample. Under unit sampling, individual addresses are selected for personal enumeration. Unit sampling would likely lead to lower variability in the imputed data, since a unit sample would reach nearly all geographic areas. However, block sampling would be easier to implement for the Census

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau. The author wishes to thank William Bell, Joseph Schafer, Elaine Zanutto, and Alan Zaslavsky for their assistance with implementing the imputation methods in simulation.

Bureau. Operations such as Integrated Coverage Measurement will be carried out at the block level, so the logistics and personnel requirements under block NRFU sampling would be familiar to Census officials. The current plan for Census 2000 specifies that unit sampling will be used.

In addition to the different primary sampling units available for NRFU, there are several alternative sampling plans under consideration. The first is called the 90% Truncation Plan. When the cutoff date for acceptance of MR census forms is reached, enumerators would visit as many households as required to achieve at least a 90% response rate in each census tract in the country (a tract is an area larger than a census block but no larger than a county). A 1-in-10 sample of the remaining nonrespondent addresses in each tract would then be selected for field enumeration, with data imputed for the addresses that are not mail-returns, not in the initial follow-up, and not in the NRFU sample. The second sampling plan available for use in Census 2000 is called the 70% Truncation Plan. This plan is identical to 90% Truncation, except the initial follow-up will attain a 70% response rate, and the sampling rate for remaining addresses is 1-in-6. The final sampling plan, Direct Sampling, includes no initial follow-up. Instead, the sampling frame consists of all addresses not responding by the cutoff date. Samples are selected from each tract, with the sampling rate determined by the response rate of the tract, as follows:

Response Rate	Sampling Rate
less than 60%	1 in 2 addresses
60% to 70%	1 in 3
70% to 80%	1 in 6
80% to 90%	1 in 9
more than 90%	1 in 12

A tract that lies on the border of two response rate classes will be given the higher sampling rate. Currently, the 90% Truncation Plan is the Census Bureau's choice for Census 2000.

III. Imputation Methods

Regardless of which primary sampling unit and sampling plan the Census Bureau uses in Census 2000, the amount of missing data will be greater than that of any modern decennial census. A number of new imputation methods have been developed to address this situation. There are currently five imputation methods available for implementation in Census 2000. In this paper, we refer to these methods as Isaki (Isaki, Tsay, and Fuller, 1994), ZZ (Zanutto and Zaslavsky, 1995), Bell (Bell and Otto, 1994), Schafer (Schafer, 1995), and the 1990 hot deck (Treat, 1993). The first three methods listed follow a "top-down" approach, in that some aggregate characteristics of each household are modeled first, followed by imputation of individual census responses using some additional procedure. The latter two methods use a "bottom-up" approach: the missing items for each household and person are imputed individually using sequential models or procedures. Brief descriptions of each of the five methods are given below.

A. The Isaki Method

This method involves modeling and estimation of twenty different household types, formed by the cross-classification of householder race, household number of persons, and tenure. The householder races are Non-Black Hispanic, Black, and Other. The number of persons is collapsed into three groups: 1 or 2, 3 or 4, and 5 or more persons. Tenure is either Own or Rent. These 18 types are combined with types for vacant households and delete/kills (addresses where a physical housing unit does not exist) to yield the total of twenty household types.

Currently, the Isaki method is applicable only to a block sample. A ratio model is used to estimate the number of each household type (except delete/kill and vacant, as discussed below) in each block: the block-level estimate of each type is the number of nonrespondent occupied housing units in that block (assumed to be known) multiplied by the District Office (DO)-level proportion of that household type. The DO proportions are calculated using data from the mail-return and NRFU sampled housing units. After the block-level counts of each household type have been determined, an imputation procedure is used to fill in responses for each census question. For vacants and delete/kills, estimates are derived from a logistic regression model fit to the MR and NRFU sampled addresses in the DO. The probabilities of the two types obtained from the model are multiplied by the block-level counts of nonsampled, nonrespondent addresses to yield estimates for these household types.

In addition to a block NRFU sample, this method also requires the assumption that each household type is approximately equally prevalent in each block of a DO. The validity of this assumption is suspect, since block characteristics can vary greatly across a DO (Schafer, 1995). Violation of this assumption would introduce bias into estimates for small areas such as blocks and perhaps even tracts. This bias would be most severe for blocks and tracts where the household type distribution differs greatly from that of the entire DO.

B. The ZZ Method

The ZZ method uses the same twenty household types as Isaki, but in a more complex model. ZZ treats vacants and delete/kills identically, using logistic regression to estimate the block counts of these types. For the remaining 18 types, ZZ models a large three-way contingency table of block number by response status by

household type for each DO. Response status is either Mail Return or Not. Like Isaki, the current version of ZZ requires a block NRFU sample.

The cells of the table contain the observed number of housing units in each three-way category. The table will contain numerous zero cells since not all blocks are in the NRFU sample. A loglinear model with iterative proportional fitting is used to estimate the number of housing units in these zero cells. The response variable is the probability that an occupied, nonrespondent housing unit falls into a particular cell on the table. Estimation of the block-level household type counts then requires simple multiplication of the probability of each type in each block by the number of nonsampled, nonrespondent units in each block. Some procedure, such as the hot deck, would be required for imputation of detailed census items.

The model is flexible with respect to the choice of explanatory variables, which correspond to main effects and interactions for various housing unit characteristics and various levels of geography. The covariates are necessarily coarse at low geographic levels, such as blocks, but can be very fine at higher levels where the data can support greater detail. The loglinear model also has the advantage of providing nearly unbiased estimates, since the predicted values do not change the covariate patterns observed in the MR and sampled blocks.

However, this method is computationally intensive, and requires acceptance of complex statistical methods in the decennial census on the part of governmental and public data users.

C. The Bell Method

Instead of estimating household types, the Bell method focuses on population estimation. The block-level mean number of persons per nonrespondent household is modeled as a function of selected block-level characteristics of the respondents. The estimate of the nonrespondent population in nonsampled blocks is then simply the number of nonrespondent housing units in the block multiplied by the predicted mean number of nonrespondents per nonrespondent housing unit for the block. The population estimate for any aggregate of blocks is a sum of the block-level estimates. Since the Bell method produces only population estimates, it requires some procedure, such as the hot deck, to impute individual responses.

Bell assumes that the distribution of the household number of persons is Poisson, and therefore can be estimated using Generalized Linear Models (GLM's). This assumption, however, is neither unrealistic nor overly strong, since GLM's are robust to mis-specification of the distribution. Previous research (Bell and Otto, 1994) has suggested that the only significant covariate is the block mean number of persons per respondent household, although any other aggregated characteristics, such as the mean number of respondents in owned housing units, can be included as well. The population is stratified by race, with separate models for Blacks, non-Black Hispanics, and Others.

Problems with this method arise in situations where a block contains no nonrespondent housing units, no respondent housing units, or no respondent persons of a given race. In the first case of no nonrespondents, the model will always perfectly predict no nonrespondents, since none can be in the sample for that block, and hence modeling is unnecessary. Attempts to modify this method to allow nearby blocks to share information have not been successful, which means that the only solution currently is to eliminate respondent data from the model. This is not the ideal use for data that we believe to be "true." A similar situation arises when a block contains no respondent housing units. Without information from nearby blocks, parameter estimation and prediction cannot occur if there is no respondent data for the covariates. As before, the only solution is to remove these blocks from model fitting and prediction. The situation of no respondents of a given race in a block is less serious than the previous two problems. Predictions can still be made for the nonrespondents of the given race using the models from the other two races in the block, although some bias may be introduced.

D. The Schafer Method

Unlike the previous methods, the Schafer method and the hot deck use the "bottom-up" approach, in which imputation is carried out at the level of individual households and persons. But where the hot deck uses heuristic procedures, Schafer builds sequential logistic regression models for each item. Since all census items except household number of persons and respondent age are discrete, each item can be collapsed into a series of binary choices and therefore modeled by logistic regressions. For example, the first model for marital status could be Currently Married against Not Currently Married. The Not Currently Married branch could then be divided into Divorced against Not Divorced, and so on until all five possible outcomes for marital status are defined. The response variables of the logistic regressions are the probabilities of the outcomes in these binary choices. Thus, imputation of each census item is simply binomial simulation using the probabilities yielded by the logistic regressions for that item.

The explanatory variables in these regression models include known or previously imputed household or person characteristics, plus terms for geographic heterogeneity and serial dependence. The geography covariate is included since household and person characteristics may vary widely across blocks, tracts, and DO's. The serial dependence term is included since empirical research (Schafer, 1995) has demonstrated that

strong trends exist across the blocks and tracts in a DO for many census items, such as race and tenure. The estimation of the parameters in these logistic regressions is more complicated than the imputation itself. Schafer uses a combination of Gibbs Sampling and the Metropolis-Hastings algorithm to perform random draws of the parameter estimates until they have reached a stationary distribution. Gibbs and Metropolis are necessary since the joint distribution of the regression parameters is intractable.

Like the ZZ method, this method requires intensive computation, as well as acceptance of advanced statistical theory by government officials and public data users.

E. The 1990 Hot Deck

Unlike the previous four methods, which use statistical models to improve imputation, the hot deck is a collection of heuristic procedures that have been developed over many years by Census Bureau officials. In the context of sampling for NRFU, the 1990 Census procedure filled in data missing for an entire household with data from a previous household that had the same number of persons. In 1990, a number of persons was known for every household, since such information could be obtained from neighbors or other sources. In Census 2000, however, this will likely not be the case. Thus, the 1990 procedure is unrealistic under sampling for NRFU. Furthermore, even if the matching number of persons criterion is eliminated, the definition of a donor unit is still problematic. If we require that the donor unit be a sampled non-MR unit, then that unit might be quite far and thus quite different from the household that has missing data. If we allow any previous unit to be a donor regardless of MR status, then we would potentially be ignoring the inherent differences between MR and non-MR housing units. Furthermore, the hot deck consists of non-statistical procedures. Data users will need to know the levels and sources of sampling error in Census 2000; the hot deck does not allow reliable estimation of imputation error.

IV. Research Methodology and Data

Our research includes testing of the Bell, ZZ, and Schafer methods. We did not have access to software for the Isaki method, and did not include the hot deck since our data, which are from the 1990 Census, had undergone a hot deck procedure during census production. We used simulation to produce fitted values for each method, which were then compared to the corresponding true values to obtain bias and variance estimates. These bias and variance results allow us to compare the methods empirically. Previous comparisons were based solely on theoretical considerations.

However, we cannot directly compare all three methods since their estimates are not equivalent. The Bell method produces population estimates, while the ZZ method yields estimated counts of household types. These different estimates mean that the bias and variance results of the two methods are incompatible. But we can compare all three methods indirectly, using results from the Schafer method. The sequential regression models used in Schafer allow estimation of the numbers of each household type in each block, making Schafer comparable to ZZ. The addition of a Poisson regression model in Schafer's method provides us with the means to estimate the actual number of persons in each household, making Schafer comparable to Bell. And through Schafer, we may be able to compare ZZ and Bell, despite their different estimates. The only situation that would not allow us to draw a conclusion would be if Schafer were "worse" than both of the other methods.

We used 1990 Census data from two DO's: 2309 in Paterson, NJ, and 3305 in Oakland, CA. These were selected from a limited number of DO's available because they contain large proportions of minorities, they have relatively high nonresponse rates, and they were used in the 1995 Census Test and are therefore widely familiar. Some characteristics of these DO's are listed below:

	DO 2309	DO 3305
Number of Blocks	4768	3205
Number of Tracts	69	80
Number of Housing Units	170759	128334
Occupied Housing Units	155269	122159
Occupied Mail Return HU's	108365	87389
Occupied Non-Mail Returns HU's	46904	34770
Owned Housing Units	86638	59202
Rented Housing Units	68631	62957
Total Population	442797	280294
Pop. in Mail Returns	303224	199133
Pop. in Non-Mail Returns	139573	81161
Pop. in Other Race HU's	290294	179370
Pop. in Black Race HU's	63596	57113
Pop. in Hispanic HU's	88907	5403

At each iteration of the simulation, a random sample of blocks was drawn, as required by the current ZZ method. All data for non-MR households in nonsampled blocks were deleted. In the sampled blocks, data collection was assumed to be perfect, so that there was no nonresponse in the NRFU sample. (In Census 2000, however, not all sampled non-MR addresses will be resolved. The optimal way to handle these cases is a topic for further research.) The sampled and MR data were used with each method to create the appropriate models, which we then applied to the nonsampled blocks to produce fitted values. There were one thousand iterations for each method.

The resulting fitted values allow us to estimate the bias and variance of each method. We selected loss functions that yield bias and variance estimates at any level of geography: block, tract, or DO. Bias was measured by

the Root Mean Weighted Squared Bias (Zanutto and Zaslavsky, 1995):

$$\hat{Bias}_j = \sqrt{\frac{\Sigma_i Y_{i+}[Ave_s(d_{ijs}))^2 - \frac{1}{S}Var_s(d_{ijs})]}{\Sigma_i Y_{i+}}}$$

where j=1...18 is the household type (note that j is not applicable to Bell's method); i=1...I, where I is the total number of geographic areas in the DO; s=1...S, where S is the total number of samples; Y_{i+} is the total number of households in area i; Ave_s is the mean over the S samples; Var_s is the variance over the S samples; and

$$d_{ijs} = \frac{\hat{Y}_{ijs} - Y_{ij}}{Y_{i+}}$$

is the relative error for household type j in geographical area i in sample s, where Y_{ij} is the true number of households of type i in area j, and $\hat{Y}_{ijs}$ is the fitted number of households in sample s.

The Root Mean Weighted Mean Squared Error is given by

$$\hat{RMSE}_j = \sqrt{\frac{\Sigma_i Y_{i+}(Ave_s(d_{ijs}^2))}{\Sigma_i Y_{i+}}}$$

using the same definitions as above. The variance of the estimates is then the difference between the MSE squared and the bias squared.

IV. Results

Currently, results have been obtained for the Bell and ZZ methods, which, as stated above, are not directly comparable because of their different levels of estimation. To work around this situation, we extended the ZZ method to produce block-level population estimates. For each DO, we computed the block-level mean non-MR household size for each of nine groups, formed by the cross-classification of household race and coarsened household size as defined in the Isaki and ZZ methods. After summing the ZZ predicted values over tenure, we multiplied the predicted household counts by the mean household sizes and collapsed over coarsened household size to obtain block-level population estimates by race, which are also produced by the Bell method. These population estimates for ZZ are rough and naive, though, since we used information from all non-MR units, not all of which would be in a NRFU sample. Therefore our population estimates most likely underestimate the bias, perhaps giving ZZ an unfair advantage in this comparison. When Schafer's results are available, we will discard these population estimates and use Schafer to compare ZZ and Bell indirectly. It is important to note that this extension of ZZ is not a part of the method developed by Zanutto and Zaslavsky; it is used here only to obtain a comparison between Bell and ZZ.

Tables 1 and 2 below give bias and variance results for the Bell and extended ZZ (denoted ZZ*) methods for the total and Hispanic populations of Oakland. The tables indicate clearly that ZZ* performs better at smaller geographic areas, particularly for blocks. For increasingly larger areas, though, Bell improves at a faster rate than ZZ*, and in fact does better than ZZ* at the DO level. For the total population, Bell has a block bias nearly three times that of ZZ*. This difference is only 50% at the tract level, while at the DO, Bell does more than 50% better than ZZ*. This trend also occurs in the Hispanic population bias estimates.

In addition, Bell is more robust than ZZ* to reduced amounts of data for modeling. The block bias for the ZZ* Hispanic population estimates is more than four times the corresponding total population bias. Hispanic block bias for Bell, however, is just more than twice the total population block bias. The two methods have nearly equal bias increases between the total and Hispanic populations at the tract level. For the DO, however, Bell performs much better than ZZ*. The Hispanic DO bias for Bell is nearly equal to the total population DO bias; for ZZ*, the Hispanic DO bias is more than twice as large as the total DO bias.

TABLE 1. Results for Total Population in Oakland

	Bias		Variance	
	Bell	ZZ*	Bell	ZZ*
Block	.166	.056	.0083	.0015
Tract	.049	.031	.0005	<.0001
DO	.010	.022	<.0001	<.0001

TABLE 2. Results for Hispanic Population in Oakland

	Bias		Variance	
	Bell	ZZ*	Bell	ZZ*
Block	.342	.228	.0426	.0365
Tract	.115	.068	.0033	.0132
DO	.012	.054	.0004	.0002

The variance results are similar for the two methods. ZZ* provides tighter estimates at the block, but Bell becomes less different from ZZ* at the levels of the tract and DO. Oddly, for the Hispanic population, the ZZ* variance is lower than Bell at the block and DO levels, but much higher at the tract. This is due to a few tracts in Oakland that have large total populations but very few Hispanics. ZZ* produces a wide range of estimates for Hispanics in these tracts, leading to an inflated tract-level variance. Bell is more robust in these tracts; its estimates and its variance are not inflated by the lack of data for Hispanics.

V. Conclusions and Limitations

The simulation results demonstrate that ZZ* produces estimates with lower bias and variance than Bell at many geographic levels. The differences between the two methods are greatest at lower levels, such as blocks. For DO estimation, however, Bell gives less biased estimates with approximately equal variances. In addition, Bell is more robust to less data; the ratios of Hispanic bias and variance to total population bias and variance are greater for ZZ* at all levels of geography.

This research is limited by the assumptions made for the study, by the number of NRFU sampling situations implemented, and by the number of methods tested. For example, we assumed that vacants and delete/kills in our data were perfectly estimated so that each method was applied only to occupied housing units. In Census 2000, this will not be the case. The method chosen for the census will need to determine the numbers of vacant and delete/kill addresses before imputation of occupied housing units can begin. The most accurate model for estimating these types of addresses will be tested in an extension of this research. We also accepted the 1990 Census data as the "truth," meaning we assumed that the hot deck perfectly imputed missing data. This assumption is not critical, though, since the hot deck likely did perform well in the 1990 Census, which had low missingness rates.

Another limitation to this research is the use of only a block NRFU sample and the Direct Sampling plan. It is possible that the methods could provide significantly different results under a different NRFU sampling environment.

Finally, the ideal research would have tested all five of the imputation methods described in section III. The exclusion of the Isaki method and the 1990 hot deck, and the lack of results from the Schafer method prevent us from identifying the method that truly outperforms all of the others currently available.

However, this research does provide a stronger comparison of imputation methods than was previously possible. The consistent data, assumptions, and loss functions used in this research allow us to go beyond theoretical comparison of the methods. In particular, while the bias and variance estimates do not necessarily reflect how well each method will perform under realistic Census 2000 conditions, they do give insight into which method will most likely provide maximum accuracy out of the methods currently available. Continuation of this research to obtain estimates from Schafer will shed even more light on the problematic issue of choosing an imputation method for use in Census 2000.

VI. References

Bell, W.R., and Otto, M.C. (1994), "Investigation of a Model-Based Approach to Estimation under Sampling for Nonresponse in the Decennial Census," paper presented at the 1994 Joint Statistical Meetings, Toronto, Canada.

Bureau of the Census (1996), "The Plan for Census 2000," internal memorandum, Washington, DC.

Isaki, C.T., Tsay, J.H., and Fuller, W.A. (1994), "Design and Estimation for Samples of Census Nonresponse," in *Bureau of the Census Proceedings of the 1994 Annual Research Conference*, pp. 289-305.

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. (1993), "Summary of the 1990 Census Imputation Procedures for the 100% Population and Housing Items," internal memorandum, Decennial Statistical Studies Division, Bureau of the Census, Washington, D.C.

Zanutto, E., and Zaslavsky, A.M. (1994), "Models for Imputing Nonsample Households with Sample Nonresponse Followup," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 236-241.