A BLOCK BASED NONRESPONSE FOLLOWUP SURVEY DESIGN

Julie H. Tsay, Cary T. Isaki, Bureau of the Census
Wayne A. Fuller, Iowa State University
Julie H. Tsay, Room 3130-4, Washington, D.C. 20233

KEY WORDS: Small area estimation, Transparent Census file, Raking, Logistic regression, Controlled rounding

I. Introduction

Sampling for nonresponse followup (NRFU) as a potential procedure for use in the 2000 Census of Population and Housing was conducted in the 1995 Integrated Coverage Measurement (ICM) Test. In the following, we provide a detailed description of the block based sample design and the housing unit estimation method used to provide a transparent Census data file (transparent to the application of sampling and estimation) of nonrespondents. We also discuss an extension of the procedure to provide a final Census file that utilizes the coverage measurement survey data.

An important motivation for using a NRFU sample is the potential reduction in costs over previous Census procedures. NRFU activities in the 1990 Census were estimated to have cost $560 million. In previous Censuses, NRFU was conducted in every block containing at least one nonrespondent. As part of that procedure, enumerators were able to identify housing units not previously known to the Census Bureau. Such housing units are termed NRFU adds. In the block based sample design NRFU enumeration would be done only in a sample of selected blocks. Estimation of such NRFU adds under a block design is straightforward. An alternative NRFU sample design is to sample from NRFU addresses in each and every block. From a large area point of view, such a design is preferred over a block design given the inherent properties of a unit versus cluster of units sample design.

However, from the point of view of NRFU adds and the associated effects on coverage and estimation, a sample design that selects a sample of nonresponding addresses in every block may be less effective than a block design. Specifically, it is not clear as to how one estimates NRFU adds under an address design. Arguments for lessening the importance of NRFU adds coverage stress the use of the coverage measurement vehicle and/or the sampling from a NRFU universe consisting of only ten percent of the addresses. At the same time, arguments are also made for needed flexibility in the 2000 Census. It is possible, due to budget constraints, that the ten percent figure will increase to 35 percent and/or due to Congressional mandate the coverage measurement vehicle may be eliminated.

Descriptions of a block based NRFU sample design with housing unit estimation and some preliminary results can be found in Fuller, Isaki and Tsay [1994] and Zanutto and Zaslavsky [1995]. In simulations, (based on 1990 Census data,) Fuller, et. al. found that the mean square errors of the block based NRFU design are comparable to those of the unit based NRFU design.

If a coverage measurement survey is to be used to modify the results of NRFU sampling, then a transparent Census data file may be created after implementation of the coverage measurement survey. In the final section, we discuss parallels between NRFU estimation and coverage measurement estimation with regard to transparent file production. We first describe the housing unit estimation procedure and the block sample design used to produce the transparent Census data file based on the NRFU methodology of the 1995 ICM Test.

II. Methodology - Sample Design

Without constraints, the block sample design for NRFU in the 1995 ICM would have consisted of a stratified random sample of blocks. Strata would have consisted of blocks with similar percent of minority households defined using 1990 Census data, and formed with regard to closeness of geographic vicinity as identified through address register areas (similar to Census tracts). The original design called for stratum sample sizes of about 30 blocks.

For the 1995 ICM Test, an external constraint was placed on the NRFU sample design. It was required that blocks that were selected for the ICM block sample be in the NRFU sample and that NRFU block sampling strata be constructed to follow ICM stratification as much as possible. Chronologically, the ICM sample of blocks was selected in the fall of 1994 and the NRFU sample was selected in the Spring of 1995. The ICM sample design was a stratified sample of block clusters with strata formed on the basis of size of block clusters (number of housing units in the 1990 Census) and concentration of race (again based on 1990 Census data).

Consequently, the strata in the NRFU block design were subdivisions of ICM strata with many block clusters and, conversely, NRFU strata were composed of several ICM strata when the ICM strata contained few blocks.

We summarize some features of the block sample design in Table 1.

Table 1. Block Sample NRFU Design Features in the 1995 ICM Test

Item	Paterson, N.J.	Oakland, CA	NW LA
1. Number of Blocks in NRFU Universe	990	1578	4066
2. Number of Sample ICM Blocks	150	122	318
3. Current Survey Blocks	6	6	0
4. Total Number of Sample NRFU Blocks	293	542	941
5. Overall Sampling Rate	.296	.343	.231
6. Number of Analysis Strata	11	19	33

The sample number of blocks per analysis stratum ranged between 25 and 35 except for a few strata where the sample size dipped as low as ten for a stratum of blocks with an anticipated small number of HUs. The number of analysis strata was dictated by the desire to achieve overall stratum block sample sizes of about 30. In all cases, initial ICM selected block samples were supplemented by further selection of blocks. Analysis stratum construction followed ICM block cluster stratification and produced finer strata by i) assigning blocks within sample block clusters to analysis strata and ii) using 1990 Census race proportion of blocks. When ICM strata were too small, the strata were combined within race. For example, one ICM stratum consisted of six large blocks of which one was selected. In this instance we combined the stratum with the medium block stratum of the same race before selecting the NRFU sample.

The details of NRFU stratification and sample selection are given in Isaki [1995]. The external constraints led to the formation of NRFU block strata (which we termed analysis strata) by assigning all blocks with at least one nonresponding address to an analysis stratum. In the simplest case, the sample in an analysis stratum, was composed of the blocks in the ICM sample, plus a simple 1 in K systematic sample of blocks selected from the remaining blocks. If Π denotes the ICM inclusion probability, we used an overall probability of $P_i = \Pi + (1-\Pi)\,K^{-1}$ for every block in the analysis stratum. Complications arose if a block j was not eligible for the ICM sample but was eligible for the NRFU sample. For example, the ICM sample excluded certain blocks that were already selected for the Bureau's ongoing current demographic surveys such as Current Population Survey (CPS). If a CPS block contained nonrespondents, it was eligible for the NRFU sample. Hence, such blocks were systematically sampled with inclusion probability P_i. We treated the sample as a stratified block design for variance computation. In summary, the goal of a stratified sampling design with equal weighting within stratum was achieved. As will be illustrated later, equal weighting was a desirable feature for small area estimation.

III. Methodology - Estimation

We used a housing unit estimation strategy to produce a transparent NRFU Census data file. The strategy consisted of estimating the total number of housing units by categories in each analysis stratum. Then a ratio was applied to the estimated number of occupied NRFU addresses in each nonsampled block to estimate the NRFU housing units in blocks not in the NRFU sample.

Nonresponding addresses were placed in categories. We first determined if the address was a delete (fictitious) or a housing unit. Next, if it was a housing unit, it was determined whether it was vacant or occupied. Finally, if the address is occupied, the address is placed in a category on the basis of the race of the householder, the number of persons in the unit and the tenure (rented/owned).

For estimation of housing units at the block level we used the ratio estimator $\hat{Y}_{hij}$ where

$$\begin{aligned} \hat{Y}_{hij} &= \hat{R}_{hj}\, M^{\circ}_{hi} \quad \text{if block } i \notin \text{ sample} \qquad (1) \\ &= Y_{hij} \qquad \text{otherwise} \end{aligned}$$

where Y_{hij} is the number of nonresponding addresses in category j in block i in analysis stratum h; M°_{hi} is an estimated number of nonresponding occupied addresses in block i in analysis stratum h,

$$\hat{R}_{hj} = \sum_{i}^{n_h} Y_{hij} / \sum_{i}^{n_h} M^{\circ}_{hi},$$

and n_h = sample size of blocks in analysis stratum h. It is easy to show that the sum of $\hat{Y}_{hij}$ over all blocks i in the analysis stratum is the usual ratio estimator of total, $\hat{R}_{hj} \sum_{i} M^{\circ}_{hi}$.

The development of M°_{hi} utilized a logistic regression which varied by test site. The separate regression models were used to estimate the proportions of NRFU addresses that were vacant and delete for each block. The model used in estimating proportion vacant in Paterson was

$$P_{hi} = [1 + \exp(\mathbf{W}'_{hi}\,\boldsymbol{\beta})]^{-1} \exp(\mathbf{W}'_{hi}\,\boldsymbol{\beta})$$

where the $\mathbf{W}_{hi}$ is a vector of explanatory variables, P_{hi} is the proportion of nonrespondents in block i in analysis stratum h that are vacant and $\boldsymbol{\beta}$ is to be estimated. The explanatory variables W_{hi} in the model were

i) $W_{0hi} = 1$
ii) $W_{1hi} = [C_{hi} + M_{hi} + 10]^{-1}\,[10\,R_{1h} + M_{hi}]$

where
C_{hi} = Census respondent HUs in block hi
M_{hi} = nonrespondent addresses in block hi

$$R_{1h} = \left[\sum_{i}^{N_h} (C_{hi} + M_{hi})\right]^{-1} \sum_{i}^{N_h} M_{hi}$$

iii) $W_{2hi} = W^{2}_{1hi}$
iv) W_{3hi} = fraction of minority renters among Census respondents in block hi
v) W_{4hi} = fraction of single HUs in block hi
vi) W_{5hi} = stratum fraction of nonrespondents that are vacant

$$= \left(\sum_{i}^{n_h} M_{hi}\right)^{-1} \sum_{i}^{n_h} V_{hi}$$

vii) W_{6hi} = fraction of precanvass addresses that were deleted after field canvass in block hi
viii) W_{7hi} = fraction of nonresponding addresses in block hi that are deleted by the Post Office prior to nonresponse follow-up or Post Office deletes of nonresponse questionnaires

The estimated $\boldsymbol{\beta}$ in the order given above with their estimated standard errors below were

β = (-4.138, 3.888, -2.779, -.0743, -1.381, 9.155, 1.098, .0482)
.0607 1.870 1.482 0.197 .404 1.805 0.279 .265

Using the logistic regression models, we were able to obtain separate estimates of proportion vacant and delete for each block hi. Denote the proportions $\hat{P}_{vhi}$ and $\hat{P}_{Dhi}$. Then we obtained $M^{\circ}_{hi} = M_{hi}\,(1 - \hat{P}_{vhi} - \hat{P}_{Dhi})$. For a few cases, $\hat{P}_{vhi} + \hat{P}_{Dhi}$ exceeded one. For these cases only we utilized a logistic regression model on the proportion vacant and delete.

The categories are delete, vacant, and occupied, with occupied broken down by two to four race of householder types by renter/owner by person size groups (1 to 2, 3 to 4, 5 or more persons). The race of householder categories varied among the three sites in the 1995 Test. For example, in NW Louisiana, the races were Black and nonBlack while in Oakland, the races were Black, Hispanic, Asian, and Other.

In previous work on the use of auxiliary variables in ratio estimation, Fuller, Isaki and Tsay [1994], suggested that the mail respondents to the Census as the auxiliary variable. Subsequent research using the 1995 Test nonrespondent distribution together with 1990 Census simulated nonrespondent characteristics indicated that M°_{hi} was a better choice. The simulation is described in Isaki and Tsay [1996]. That report also describes alternative auxiliary variables and several alternative estimators such as regression, Horvitz Thompson and hybrid estimators, alternating auxiliary variables. For the 95 Test, however, there did not appear to be any advantage to using M°_{hi} versus M_{hi}. Consequently, the test results suggest using M_{hi} as the auxiliary variable.

The ratio estimator in (1) yields estimates of occupied housing unit counts by categories,estimates of vacants, and estimates of deletes. A requirement placed on the block estimates was that for each block, the sum of the estimates in the three groups of addresses (in NRFU we sample from a list of addresses that were originally sent a census form but from which we have not received a form) must equal the number of NRFU addresses, denoted M_{hi}, for the block. To meet this requirement we used a two way raking procedure. The columns of a matrix corresponded to the various occupied housing unit/vacant/delete categories while the rows corresponded to the blocks. For example, in Paterson we used a matrix with 696 rows and 20 columns. The entries of the matrix are the block 'housing unit' estimates. The raking procedure was composed of three full cycles. After raking the row totals differed from M_{hi} by .1 to .3 addresses.

Recall that the NRFU procedure has the potential to obtain housing units not on the NRFU list used for sampling (NRFU adds). In the test, only a few such cases were identified. In the Oakland test site 39 NRFU adds were obtained from the block sample design in a study area consisting of half of the test site while 18 NRFU adds were obtained in Paterson. To put these figures in perspective, we note that 1990 Census NRFU adds for 12 district offices (DOs) in Massachusetts (data used for some earlier work on NRFU design) ranged from .41 to 2.34 with a state percent of 1.27. Two of the district offices for Boston, a large urban city, experienced 1.96 and .96 percent NRFU adds, respectively. A one percent NRFU add figure for the Oakland and Paterson sites would be approximately 400 and 70 sample NRFU adds, respectively.

The housing unit estimation procedure treated NRFU adds separately from identified NRFU addresses (those that are part of the M_{hi} count). The block estimation of NRFU adds is similar to that in (1) except that the Y_{hij} value, the number of nonresponding addresses in category j in block i in analysis stratum h is replaced by Y_{Ahij} which is the number of NRFU adds in category j in block i in analysis stratum h. In this process, the category 'delete' is not used. The NRFU adds block estimator, denoted by $\hat{Y}_{Ahij}$, is added to the raked $\hat{Y}_{hij}$ values.

The matrix of estimates is then control rounded to give integer estimates. The output is a block data file with integer counts of housing units by category. The next estimation step is the selection of donor households for the estimated NRFU housing units in each nonsample block. We selected donors by analysis stratum using as donors the NRFU sample respondents in the NRFU sample blocks. Recall that the NRFU sample design was constructed to yield common inclusion probabilities of blocks within analysis strata. This enabled development of a simple donor selection procedure.

The donor selection procedure can be described by representing each NRFU sample respondent household by a coin and stacking together all coins in a particular housing unit category, e.g., all renter households with a Black householder containing one to two persons. If the first nonsample block requires two renter households with a Black householder with one to two persons, the donor procedure is to remove two coins from the top of the stack, assign the households to the block, and place the two coins at the bottom of the stack. The procedure is repeated for all other categories for the block and is then repeated it for the remaining blocks.

This donor allocation procedure produces block totals of persons that are close to the direct ratio estimator of total persons at the site level. This is because the approximate equal use of each NRFU sample respondent as a donor reflects the equal sampling weight.

IV. Some Results of NRFU Block Sampling

The estimated total NRFU HUs and their estimated coefficient of variation (C.V.) for each site by various HU categories are provided in Table 2 (see at end). Estimated C.V.'s for estimates of total persons by categories were similar. The variances of HU and person total estimates were estimated assuming a stratified block design.

One modification in the definition of housing unit categories that would be fruitful for future application is to dichotomize renter type housing units by single versus multi-units at the same street address. Doing this, will provide a control within renter categories where the proportions of single and multi unit are not negligible and when it may be important to discriminate between donors.

V. Extension to Quality Check Sample (QCS) Estimation

In the 1990 Census a coverage survey (CS) was conducted and a quality check survey (QCS) is planned for the 2000 Census. The CS consisted of samples of blocks within which extensive field work is conducted to measure coverage of the Census. In the CS a dual system estimation (DSE) procedure for estimating persons (and their characteristics such as race by sex by age) was used. No estimates of housing units were made in the CS. Hence, a transparent Census data file could not be constructed.

Using an approach similar to the NRFU housing unit estimation, two types of estimates of HUs based on the ICM are being investigated. One is based on DSE, while the other is called Census Plus (essentially a CS in which a Census is conducted in a sample of blocks and an additional enumeration, the 'Plus,' is applied in the same blocks and used in estimation). As in the NRFU application, the idea is to first estimate the number of housing units by categories.

In addition to race of householder, tenure and number of persons in the household, we will likely need to cross the original categories with sex and age for the householder and also by certain undercounted categories such as young adult Black males for Black householders and young adult Hispanic males for Hispanic householders, etc. (or use modelling techniques). The reason for the added detail is that unlike the NRFU procedure that selected donors from the sample of NRFU cases, the CS procedure may use the Census respondents as the donor pool. Hence, more detailed categories of HUs are required to yield the type of HUs required. The exact nature of the categories are subject to further research.

Clearly, HU estimation in the CS will require much development. The process, however, is quite similar to the NRFU HU approach. Housing unit adjustment factors by category (ratio of CS HU estimate to Census HU estimate) will be constructed at some higher level, say the site, and applied to Census HU estimates at the block level. The adjusted Census HU block estimates would then be control rounded. The resulting integer counts would be compared to the Census HU counts at the block level. If the Census count exceeds the rounded count, we have an overcount and a HU (typically an imputed HU) would be removed from the file. If the Census count is less than the rounded count, we have an undercount and a random Census respondent in the category in the block could be added to the file. In this way, a transparent Census data file is produced.

VI. References

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

2. Hogan, H. (1992), "The 1990 Post-Enumeration Survey: An Overview," The American Statistician, 46, 261-269.

3. Isaki, C.T. and Tsay, J.H. (1996), "Selection of Auxiliary Variables for Ratio Estimation for Nonresponse Followup Estimation and Application of a Housing Unit Estimation Method in the 95 Test HERF", draft 24 pages with attachments.

4. Isaki, C.T. (1995), "Sampling Nonresponse Followup Blocks in the 1995 Test", internal memorandum, 7 pages.

5. Schaefer, J.L. (1995), "Model-Based Imputation of Census Short-Form Items", Proceedings of the 1995 Annual Research Conference - Bureau of the Census, Arlington, VA, pp. 267-299.

6. U.S. General Accounting Office (1992), "Decennial Census - 1990 Results Show Need for Fundamental Reform", Report to Congressional Requesters, 71 pages.

7. Zanutto, E. and Zaslavsky, A.M. (1995), "Models for Imputing Nonsample Households with Sampled Nonresponse Followup", Proceedings of the 1995 Annual Research Conference - Bureau of the Census, Arlington, VA, pg. 673-686.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Table 2. Estimation of the Total and Coefficient of Variation of NRFU Housing Estimates from the 1995 ICM Test

	Paterson, NJ		Oakland, CA		Northwest Louisiana	
	Total		Total		Total	
Categories	HUs	C.V.	HUs	C.V.	HUs	C.V.
Vacant	2,409	.026	4,410	.017	6,811	.013
Occupied	22,558	.005	26,335	.005	14,333	.007
Black	9,309	.014	13,421	.010	6,797	.012
Owners	1,734	.052	3,506	.024	3,493	.025
Renters	7,575	.024	9,914	.014	3,304	.026
Hispanic	9,122	.012	3,431	.022		
Owners	1,690	.030	712	.045		
Renters	7,432	.013	2,718	.026		
Other	4,125	.018	6,354	.018	7,535	.013
Owners	1,427	.035	2,336	.038	5,226	.014
Renters	2,698	.032	4,018	.032	2,309	.030
API			3,128	.028		
Owners			738	.038		
Renters			2,389	.035		