THEORY AND APPLICATION OF NEAREST NEIGHBOR IMPUTATION IN CENSUS 2000

Robert E. Fay*
U.S. Census Bureau, Washington, DC 20233-9001

Key words: missing data, variance estimation, replication, replicate weights.

1. Introduction

1.1 Changing Census Plans The evolving plans for the next decennial census in the U.S., Census 2000, have been widely reported in the press and remain a focus of political debate. Last year, Thompson and Fay (1998) reviewed the milestones in the development of the decennial program at the Joint Statistical Meetings, particularly with respect to statistical sampling and estimation, which was to have been one of the cornerstones of the plans. While the 1970 census provided a precedent for inclusion of sampling and estimation in producing the population count (Wright 1999), Census 2000 was to have been the first in the U.S. designed with the goal of achieving an optimal combination of counting, assignment, and estimation in order to obtain population totals.

The 1980 and 1990 censuses primarily employed a mail strategy in which households in most parts of the country were mailed or delivered census forms. The majority of households responded by mail, but nonresponding households were followed up by personal visit to complete the enumeration. In some cases, interviewers had to obtain information about nonrespondents from neighbors or other sources, although precise figures on such proxy response are unavailable.

Both demographic analysis (Robinson, Ahmed, Das Gupta, and Woodrow 1993) and coverage studies using sample surveys have documented a persistent undercount of some groups, including Blacks. In both 1980 and 1990, the issue of a potential adjustment to the census counts to compensate for differential undercoverage became a matter of both debate and litigation. Results from coverage evaluations could not be produced until months after the release of the apportionment counts used to allocate the number of representatives among the states. No official numbers were adjusted in 1980. In 1990, Secretary of Commerce Mosbacher decided in July, 1991, against the Census Bureau recommendation for adjustment. The only official figures incorporating an adjustment for undercount in the 1990 census are postcensal population estimates at the state and national level used as controls for some demographic surveys, such as the Current Population Survey (CPS), the monthly labor force survey in the U.S. In particular, the official postcensal population estimates, published and used for other purposes including allocation of funds, do not incorporate an adjustment for census undercount, unlike current Canadian practice (Germain and Julian 1993, Dick 1995).

As of August 1998, the Census Bureau's plan for Census 2000, developed over several years, employed statistical sampling and estimation in two primary ways (Thompson and Fay 1998):

1) A sample of nonresponding households was to be selected for Nonresponse Followup (NRFU), and the results used to form estimates for nonsample nonresponding households. Sampling for NRFU offered savings in both time and money. Nonetheless, this sample was to have been extremely large, on the order of tens of millions of households, enabling conventional survey estimates down to very low levels of geography.
2) A large coverage study, called the Integrated Coverage Measurement (ICM), would be based on approximately 750,000 housing units and viewed as an integral part of the census. Its purpose was to measure differential undercoverage. The results would be incorporated into all official results, including the state population counts delivered to the President on December 31, 2000, to be used for the apportionment of the U.S. House of Representatives among the 50 states.

In January, 1999, however, the U.S. Supreme Court upheld lower court rulings that the current census legislation did not permit the use of either form of sampling for the apportionment. The court's ruling did not resolve the constitutionality of such a census if current law were revised to permit it. Plans for Census 2000, however, were changed to conform with the existing legislation and the court's interpretation of it.

Thus, the effect of the Supreme Court ruling was to eliminate the first type of sampling, sampling for NRFU, from Census 2000. In other words, once NRFU sampling is excluded for purpose of apportionment, it would have no practical use for any other purpose. Revised plans for the census now reflect the increased workload and time requirements to follow up roughly 15,000,000 more nonresponding housing units. State counts without the use of sampling and estimation will be produced by December 31, 2000, for the apportionment.

Although NRFU sampling was entirely eliminated by the court's decision, the ICM has been redesigned as the Accuracy and Coverage Evaluation (A.C.E.). Indeed, the

court's ruling did not exclude the use of sampling for uses other than apportionment, and to some degree suggested that sampling would be appropriate if feasible. The A.C.E. sample size has been reduced from 750,000 to approximately 300,000 housing units, and the time schedule has been adjusted for the more extensive period required for NRFU. Current plans are to obtain A.C.E. estimates by approximately February 15, 2001, relatively earlier than previous studies in 1980 and 1990. The A.C.E. estimates, in the form of estimated percent undercounts for a set of poststrata, are planned to be incorporated into the detailed official counts down to the block-level for release by April 1, 2001. This timing permits their potential use in redistricting. (In the U.S., states allocated more than one representative in the House of Representatives are divided into Congressional Districts on the basis of population. *Redistricting* is the process of defining these boundaries. Although court rulings now tightly limit variation in the population sizes of congressional districts within a state, states are given some latitude in other respects in determining their boundaries.) Under the revised plan, all official data products from Census 2000, except the apportionment counts, will incorporate the results of the A.C.E.

1.2 Nearest Neighbor Imputation in the Dress Rehearsal The paper will focus on the nearest neighbor imputation as an estimation procedure for NRFU in the Census 2000 Dress Rehearsal in Sacramento and on an associated variance estimator. Thus, the paper concerns methodological aspects of an application obviated by the Supreme Court's ruling. Nonetheless, this paper, and one in preparation (Fay and Farber 1999), will focus on methodological findings from the Dress Rehearsal effort.

The 2000 Dress Rehearsal was conducted in three sites. Through agreement with Congress, the Census Bureau implemented its plan for Census 2000, combining sampling for NRFU with Integrated Coverage Measurement, only in Sacramento, California. A site in Columbia, South Carolina, and surrounding counties was enumerated without the statistical methods, although an accompanying Post-Enumeration Survey, similar in design to the ICM, was employed as an evaluation. A smaller site in Wisconsin did not use sampling for NRFU but incorporated the ICM corrections.

The selection of a probability sample for NRFU theoretically permits the use of standard survey weighting procedures. Instead, a nearest neighbor/hot-deck imputation method was implemented in the Sacramento site, in effect treating nonsample nonresponding housing units as a problem in unit nonresponse. The subject of section 2 and a principal focus of this paper is the rationale for this methodological selection. The section also describes the Sacramento application in more detail.

1.3 Variance Estimation A second purpose of the paper is to present a suitable variance estimator for the NRFU imputation. The variance estimator has potential application to other nearest neighbor imputation situations.

As often noted, the term *hot-deck imputation* has been applied to a variety of similar methodologies. For purposes of discussion in this paper, it is useful to group most of these into three broad categories:

1. *The sequential hot deck.* This original form appears to have been substantially shaped by available computer resources and practice at the time of its development. In its simplest form, a characteristic x is available for all units but y is subject to possible nonresponse. Units are classified on the basis of x into prespecified cells. Typically, an array is loaded with a *cold deck* of initial values based on an earlier survey or some other suitable source. The units are processed sequentially, often in a sort reflecting geographic proximity or another measure of similarity. New units with observed y, termed *donors,* are used to replace old values in the hot-deck array; units with missing y are assigned values from the hot deck. For example, the empirical study of Rizvi (1983) considered only this form of hot deck, whereas the overview by Ford (1983) considered this form of hot deck as well as nearest neighbor imputation below.
2. *Statistical matching with fixed cells.* The ability to sort moderate or large files and other forms of data access removed the restriction that the hot deck be tied to the sequential order of the data file. For example, the variance estimator developed by Rao and Shao (1992) was for a hot deck unconstrained by the order of the file. Specifically, they consider units cross-classified into a potentially large number of cells, and each unit requiring imputation can receive a value from any of the donors falling in the corresponding cell. In Rao and Shao (1992), observations requiring an imputation are independently assigned values with probability proportional to each donor's survey weight. The asymptotic argument in Rao and Shao (1992) permitted an increasing number of cells but required that generally each cell have an increasing number of eligible donor cases. Fay (1996) contrasted the Rao and Shao (1992) estimator for this situation with a multiple imputation (Rubin 1987) approach, finding a clearer frequentist interpretation for the Rao and Shao approach than for multiple imputation.
3. *Nearest neighbor imputation.* This form extends the logic of statistical matching further, searching for either a unique best match or small number of

equivalent matches on the basis of x among units with observed y. The cells in this form of imputation are not predetermined. For example, College, Johnson, Paré, and Sande (1978) describe an application to economic data in which a specific set of nearest neighbors were identified for each case requiring imputation. I.G. Sande (1983) reviewed the general features of the nearest neighbor approach, and G.T. Sande (1983) provided additional comments on its features and computational feasibility. Lee, Rancourt, and Särndal (1994) presented and studied a variance estimator for y using nearest neighbor imputation on the basis of a continuous x and an assumed regression model for y on x. The extension of multiple imputation to this form of nearest neighbor imputation is less clear, and the version of multiple imputation investigated by Lee, Rancourt, and Särndal did not perform as satisfactorily in an empirical study as their own estimator.

This grouping is useful for categorizing variance estimators for the hot deck, but the distinctions among the three groups are not always precise, particularly between the second and third form. Bankier, Luc, Nadeau, and Newcombe (1996) characterize the New Imputation Methodology (NIM) developed at Statistics Canada (Bankier, Houle, Luc, and Newcombe 1997) as a minimum change hot deck, and it appears appropriate to categorize the NIM in the third group. The current imputation methodology for work history and income items for the Annual Demographic Supplement ("March Supplement") to the Current Population Survey (CPS) in the U.S. employs a series of sorts to achieve a statistical match between donors and cases requiring imputation. (Coder 1978 and David, Little, Samuhel, and Triest 1986 provide a more detailed summary of the algorithm.) In effect, several different cross-classifications of observed characteristics are considered for each case requiring imputation. In many cases, a large number of donors may be available for particular cases requiring imputation, but in others cases imputations may be made by selection from a relatively small set of donors. Thus, the CPS application is best categorized as belonging to the third group but having ties to the second. Other versions of the hot deck, such as the use of the nearest neighbor approach with distance determined in whole or in part by predicted values from a parametric model, do not fit neatly into any of the three groups.

Fay developed a variance estimator especially for the third group, based on different assumptions than used by Rao and Shao. Unlike the variance estimator proposed by Lee, Rancourt and Särndal, the method does not require a parametric model. A key feature of the method is the use of data from a second nearest neighbor for purposes of variance estimation. Various forms of the variance estimator have appeared earlier (Town and Fay 1995, Steel and Fay 1995, Fay and Town 1996, 1998). Section 3 describes the rationale of the estimator in more detail.

2 Selection of Nearest Neighbor Imputation in the Dress Rehearsal

2.1 Methodological Background

2.1.1 Block vs. Unit Sampling The Census Bureau's planned use of sampling for both NRFU and the correction of differential census undercoverage through ICM led to a number of separate research efforts on how to estimate census results under NRFU sampling. The design for NRFU sampling was constrained, however, to be consistent with the plans for the ICM.

The design strategy for the ICM involved sampling census blocks or block clusters and typically including all of the housing units in the sampled blocks in the ICM sample. (The largest blocks were to be subsampled. The average size of ICM clusters was projected to be about 30 housing units after subsampling.) Because of the required matching of ICM sample cases to their initial census counterparts, the design called for 100% NRFU, rather than NRFU sampling in ICM blocks. NRFU sampling in ICM blocks would have induced complexities arising from attempting to match an ICM household to a nonsample household in NRFU.

Two primary candidate designs were available for NRFU sampling of non-ICM blocks:

- Block sampling, where a sample of blocks with mail nonresponse would be selected. All nonresponding housing units in sampled NRFU blocks would be followed up.
- Unit sampling of nonresponding housing units in an unclustered manner. One form, implemented in one of two panels in Oakland, CA, in the 1995 Census Test, assured selection of sample units in all blocks with nonresponse. (For example, a single nonresponding unit in a block was always included.) In Sacramento, a systematic sample of nonresponding units was selected, resulting in sampled NRFU units in most but not all blocks with nonresponse.

The block-based design logically fit better with the ICM design, since NRFU would then be completed on a block basis in both ICM and non-ICM blocks. In other words, if

$$E(y \mid \text{unit sampling}) \neq E(y \mid \text{block sampling})$$

then statistical adjustment of counts from NRFU using unit sampling on the basis of an ICM using block sampling would be problematic. A distinct potential disadvantage, however, later confirmed by empirical studies, was that sampling blocks for NRFU could yield much higher variances.

Initially, research efforts had concentrated on NRFU estimation under block sampling. Tsay, Isaki, and Fuller (1996) and Zanutto and Zaslavsky (1996) investigated block-level models. Estimates at the block level would then be used as constraints by statistical procedures to estimate the nonsampled households and persons within each block to be added to the results of direct enumeration. Schaefer (1995) investigated a different procedure, which modeled at the household and person level. In Schaefer's approach, estimates for blocks emerged by summation. Schaefer's approach also attempted to integrate estimation for NRFU with imputation for item nonresponse for data on characteristics.

In the 1995 Census Tests, a split panel experiment compared unit and block sampling designs for NRFU sampling in Oakland, CA. The two panels produced post-NRFU estimates within sampling error of each other, but variances were substantially less for unit sampling (Fay and Town 1996). In other words, the evidence indicated

$$E(y \mid \text{unit sampling}) \doteq E(y \mid \text{block sampling})$$

$$Var(y \mid \text{unit sampling}) \ll Var(y \mid \text{block sampling})$$

In general, cost is an additional factor to consider in selecting among survey designs. In this respect, block sampling was thought to have a small advantage in terms of slightly reduced travel, but the differences were considered marginal because both designs, with their high sampling fractions, would be very densely distributed.

The findings became the impetus to move to unit sampling. A related study based on matching ICM blocks to similar non-ICM blocks for the 1998 Dress Rehearsal showed no systematic differences (U.S. Census Bureau 1999a).

The effect of unit sampling is suggested by the following example: for a typical block of 30 housing units, a 70% response rate leaves 9 nonresponding units, of which an expected 6 would be sampled for NRFU, leaving an expected 3 to be estimated. Use of sampling rates defined within each block, or, as in the Dress Rehearsal, systematic sampling, are available to limit random variation in the realized sample size within blocks. Although the previous research had shown that models can make effective use of block-level variables in forming estimates under block sampling, unit sampling reduces these gains by largely eliminating the effect of between-block variability on the estimate.

2.1.2 Weighting vs. Imputation With the selection of unit sampling and the accompanying feasibility of using simpler design-based estimators, a traditional weighting approach appeared to deserve consideration. Although the design-based rationale for weighting is clear, there were a few significant constraints. For the sake of ease of use, weights could only be integers. (Indeed, the Census Bureau has consistently used integer weights even for the census "long-form" sample data. In Census 2000, sample data will be collected from approximately 1 out of 6 households.) For NRFU, most weights would thus be either 1 or 2, except in areas where the response rate exceeded 80%.

For higher levels of geography such as census tracts, application of integer weights would have provided estimates of population plausibly consistent with housing unit totals. Use of weights at the block level, however, could have produced marked inconsistencies between population counts directly affected by NRFU sampling and counts of housing units, which were almost free from random variation.

Use of nearest neighbor imputation, with nonsampled nonresponse housing units imputed from nearby sampled NRFU units, maintained the integer nature of the census data and linkage between housing units and population. Farber and Griffin (1998) compared weighting vs. nearest neighbor imputation, and found almost equal performance at most geographic levels, but awarded the advantage to nearest neighbor imputation for its ready handling of blocks lacking sampled NRFU units.

2.2 Basic Implementation Strategy in the Sacramento Dress Rehearsal

As in previous U.S. censuses, part of the overall task of taking a census is to obtain an inventory of housing units. Sampling was not to be employed at this stage; rather, the Master Address File (MAF) of all housing units was the frame for sampling. (Sampling enters slightly, in that the result of NRFU is occasionally to determine that a unit in the MAF must be deleted. For example, the unit might be a commercial address or demolished. Consequently, the imputation also classified some nonsample units to delete status.)

Sampling for NRFU was to be used for two types of incomplete data. The first of these was housing units not responding by mail: nonresponding housing units represent a mixture of households not returning their forms, vacant units, and units that should have been deleted from the MAF. Consequently, NRFU was to determine the occupancy status of sampled nonresponding units as well as other characteristics. The sampling rates were set so that statistical estimation would be used to represent only 10% (or less) of the

housing units in each census tract (a publication area generally representing about 3000 housing units). With mail nonresponse in Sacramento around 50%, overall about 4-in-5 nonresponding units were sampled for NRFU -- a very large sample compared to usual statistical practice.

The second type of incomplete data was entirely distinct from the first. During the mail delivery of the census questionnaires, carriers were allowed to return forms to the Census Bureau with the designation "Undeliverable as Addressed-Vacant" (UAA-vacant). From past experience, the majority of such units are vacant, but not all of them. (This again proved to be the case in Sacramento.) A sample of 3-in-10 of the UAA-vacant units was selected for followup visit by an enumerator, and units found to be occupied were enumerated. This sampling rate required that most UAA sample cases be used in imputation 2 or 3 times.

Each of the two groups was treated as an independent estimation problem, since the respective universes did not overlap. The following discussion will be primarily in terms of NRFU, although parallel operations were carried out for UAA. Farber, Fay, and Schindler (1998) describe the sampling and estimation in more detail, but a summary will be included here.

Based on the overall approach of section 2.1, the basic strategy was to identify a neighboring sample NRFU case as the basis for imputation for each nonsample nonresponse housing unit. In terms of the previous notation, characteristics x, including sort order in the MAF (which is related to location), census block, and basic address (generally enabling identification of units in the same building), are available for both sample and nonsample nonresponse units. Wherever possible, the matching algorithm selected donors from the same address for nonsample units at multi-unit addresses. Otherwise, the nearest neighbor of a nonsample case was defined on the basis of the distance when the file was sorted by block and MAF order within block. Thus, the algorithm favored using units in the same block, taking into consideration the closeness indicated by the MAF (U.S. Census Bureau 1999b). The overall design was that donor units would provide characteristics, y, including whether the unit should be deleted, occupancy status, and for occupied units, number and demographic characteristics of the residents, owner/renter status, *etc.*

In order not to depart from a design-based rationale too substantially, the number of times each eligible sample case was used was constrained to be consistent with the weight it would have had under a weighting approach. At the tract level, this weight would have been

$$w_{NRFU} = \frac{NRFU\ sample\ units + NRFU\ nonsample\ units}{NRFU\ sample\ units}$$

omitting ICM blocks. A UAA weight, w_{UAA}, was defined similarly. For example, for a UAA sample case in a tract with $w_{UAA} = 3.33$, then the sample case was constrained to be used as a donor either 2 or 3 times (in addition to representing itself). Consequently, when the weights in a tract were an integer, then the nearest neighbor imputation was constrained to use each donor the same number of times and thus to reproduce at the census tract level the results of a weighting approach.

As a remark, had the Supreme Court ruled differently, and had sampling for NRFU remained part of Census 2000 plans, further empirical investigation of this approach would have been warranted. The procedure implemented for Dress Rehearsal insured a fairly high level of agreement between the imputation and weighting at aggregate levels, while effectively leaving the nearest neighbor approach some latitude in allocating whatever fractional weight was available. Because of high sampling rates for NRFU, in most tracts the weight did not reach 2, so the effect of the constraint was to allow the measure of distance to select donors but to constrain each donor to a single use at maximum. Alternative approaches, such as randomly determining which donors to use to distribute any fractional weight, would have provided even further protection against bias, possibly at the expense, however, of somewhat increased variance. Further empirical research on such points would have been warranted if NRFU sampling had remained in Census 2000 plans.

2.3 Modifications for Late Mail Returns

The conceptual model just described is based on a simple dichotomy between sample and nonsample nonresponse units. In practice, a cutoff date approximately 2 weeks after the due date was set during Dress Rehearsal, and the sample of nonresponding units was identified at that time. Some returns continued to be received thereafter. When a late mail return was received from a unit sampled for NRFU, a computer algorithm selected between the late return and the followup form that may have been obtained by personal visit on the basis of completeness, with a preference for the mail return. Late returns were incorporated into the census for nonsampled units, eliminating the need for any imputation for them. The decision to accept late returns was made on the basis of policy, to avoid rejection of data

from the public within the time span that the information could actually be processed and tabulated.

Superficially, late mail returns appear simply to reduce the size of the remaining followup somewhat. In fact, however, late mail returns cannot be regarded simply as a random sample from the outstanding units. Late mail returns, like mail returns in general, were almost exclusively from occupied housing units. Ignoring late returns, the nonsample nonrespondents are a probability sample of all nonrespondents as of the cutoff date, but when those returning late returns are removed, the remaining units should be somewhat skewed towards a population of vacants and deletes. Since the occupancy status belonged among the sample characteristics, *y*, rather than among the characteristics, *x*, available for matching, the use of nearest neighbor imputation would not correctly represent this aspect of response. In other words, although selecting a probability sample for NRFU would have led to nonresponse (where nonresponse in the general theory is equated to not a member of the NRFU sample in this instance) *unconfounded* with *y*, the subsequent exclusion of late mail returns from the nonsample cases leaves a set of nonsample cases with nonresponse *confounded* with *y*, particularly through occupancy status.

Lee, Rancourt, and Särndal (1994) investigated the performance of their proposed form of nearest neighbor imputation under both unconfounded and confounded response situations, but proposed no modification to the procedure in instances in which response is confounded.

A modification to the imputation to compensate for confounded response was implemented for the Dress Rehearsal. The approach was to reduce proportionately the number of available donors in accord with the number of late mail returns received for nonsample units. For each tract, a ratio of sampled to nonsampled units was computed; specifically (Farber, Fay, and Schindler 1998),

$$r_{NRFU} = \frac{\textit{NRFU sampled addresses in non-ICM blocks}}{\textit{NRFU nonsampled addresses in non-ICM blocks}} = \frac{1}{w_{NRFU} - 1}$$

Separately for the three categories, *c*, of sampled 1) occupied, 2) vacant, and 3) deleted units, a number of units,

$$Rm_c = r_{NRFU} \times LR_c = \frac{LR_c}{w_{NRFU} - 1}$$

were removed from the hot deck from *c*, where LR_c is the number of nonsampled late returns in *c*. Within *c*, sampled cases with late mail returns were targeted first for removal, followed by the remaining sample cases.

For $r_{NRFU} < 1$, that is, $w_{NRFU} > 2$, some sample units would be used more than once for imputation, and further research could have investigated an alternative that reduced the allowed imputations by 1 over a larger group of donors instead of completely eliminating some from any use. Again, circumstances no longer justify further pursuit of this option at this point.

2.4 NRFU/UAA Variances for Sacramento Although the official Dress Rehearsal result for Sacramento incorporated the ICM component, the Census Bureau has also released a population total without the ICM component of 377,741 for Sacramento. Using the methodology reported in the next section, the estimated standard error for this estimate was 321 people, or a c.v. of less than .09%. This standard error pertains to the effect of NRFU and UAA estimation. By contrast, when the ICM correction is also included, the official population total is 403,312 with estimated standard error 4,810 or a c.v. of 1.2%.

For a block of approximately 30 housing units and 75 people, a simple scaling of the site-level results suggests a standard error of approximately 4.5 people, or a c.v. of roughly 6%, from the NRFU/UAA component of estimation.

According to the ICM, the census count, 377,741, without the ICM for Sacramento is subject to an estimated 6.3% undercount (s.e. 1.1%).

2.5 Concluding Remarks on the Application Available evidence suggests that the implementation of nearest neighbor imputation essentially achieved its primary objective, namely, to produce statistical estimates comparable to conventional enumeration while agreeing closely in aggregate with traditional survey estimation results. Further empirical details will be provided by Fay and Farber (1999).

Changes in legislation could open the possibility of sampling for NRFU in 2010; if so, the 2000 Dress Rehearsal experience will represent a potential starting point for research efforts.

The application also illustrates consideration of the effect of confounded response in applying nearest

neighbor imputation. Although the procedure presented here could be further refined, the modification of the hot deck to compensate for the effects of confounded response may be a potential approach useful in other applications.

3 A Variance Estimator for Nearest Neighbor Imputation

3.1 Basic Rationale As noted in the first section, forms of the variance estimator had been reported in several previous collaborations. Although basic assumptions were stated, the previous accounts were considerably truncated. The discussion here is intended to provide an expanded account of the estimator. Section 3.1 will deal with a special case, which leads to a simplified, almost obvious, form. Section 3.2 provides the full estimator used in the Dress Rehearsal.

As noted in Section 1.3, the Rao-Shao variance estimator was developed relying on a random selection from a set of possible donors in the imputation cell. Forms of nearest neighbor imputation emphasizing use of an optimal nearest neighbor depart from conditions required for the Rao-Shao variance estimator.

For each case i requiring imputation, let y_i be the true but unobserved value and $y_i^* = y_{nn1(i)}$ the imputed value from the nearest neighbor, *nn1(i),* of *i.* The scope of the variance estimator includes both applications in which the nearest neighbor, *nn1(i),* is defined on the basis of distance only or is subject to constraints on the number of times each donor may be used, as in the Dress Rehearsal. The variance estimator employs the value from a second nearest (responding) neighbor, *nn2(i).* In general, a second nearest neighbor, *nn2(i),* for *i* may be defined by omitting the first nearest neighbor, *nn1(i),* and applying the nearest neighbor definition/algorithm. (In the Dress Rehearsal application, the constraint on the number of uses of the donor was not employed in defining the second nearest neighbor, nor is it theoretically necessary to do so.)

The variance estimator is developed under a population model, ξ, for a population of y conditional on x (Fay and Town 1998). The variance estimator relies on model ξ assumptions:

$$E_\xi(y_i) = E_\xi(y_{nn1(i)}) = E_\xi(y_{nn2(i)}) \qquad (1)$$

$$Var_\xi(y_i|x_i) = 1/2\ E_\xi(y_{nn1(i)} - y_{nn2(i)})^2 \qquad (2)$$

$$Cov_\xi(y_i, y_{i'}|x_i, x_{i'}) = 0,\ i \neq i' \qquad (3)$$

To the extent that $x_{nn1(i)}$ and $x_{nn2(i)}$ are close to but not necessarily equal to x_i, assumption (1) represents an ideal that practical applications will generally only approximate. Assumption (3) follows from an assumption of independence of the *y's* given the *x.* Assumption (2) follows from (1) and (3) if nonresponse is unconfounded with *y* given *x.*

Let Y denote the sum of the *y's* in the given finite population, which is a realization under ξ. Suppose Y^* denotes the sum with the imputed values for *y.* If each nearest neighbor is used in imputation at most once, then,

$$\begin{aligned} E_\xi(Y - Y^*)^2 &= \sum_{i \in A_{nr}} E_\xi(y_i - y_{nn1(i)})^2 \\ &= \sum_{i \in A_{nr}} E_\xi(y_{nn2(i)} - y_{nn1(i)})^2 \end{aligned} \qquad (4)$$

where A_{nr} denotes the set of cases requiring imputation. The variance estimator is derived by replacing the expectation over the model in (4) with the observed values,

$$Var_\xi^*(Y^*) = \sum_{i \in A_{nr}} (y_{nn2(i)} - y_{nn1(i)})^2 \qquad (5)$$

Under these simple conditions, the interpretation of (5) is straightforward. The variance under the model in estimating the unobserved y_i by $y_i^* = y_{nn1(i)}$ is approximated by the squared differences of first and second nearest neighbors.

Expression of a replication method through replicate weights facilitates subsequent analysis. For the sake of generality let, Y^*, be expressed as a weighted sum,

$$Y^* = \sum_i w_{i0} y_i^*. \qquad (6)$$

Suppose a replication-based estimator of variance can be written in the following form,

$$Var^*(Y^*) = \sum_{r=1}^{n} b_r (Y_r^* - Y^*)^2, \qquad (7)$$

where the b_r, $r=1,\ldots, n$, are an appropriate set of coefficients independent of the choice of characteristic Y, and for replicate estimates,

$$Y_r^* = \sum_i w_{ir} y_i^*, \quad (8)$$

is defined for replicate r, on the basis of a set of replicate weights, w_{ir}, where n replicate weights are assigned to each i.

Estimator (5) can be approximated through a replicate weighting of the form (6) - (8). The Dress Rehearsal implementation used $n = 100$ and $b_r = 1$, $r=1,\ldots,100$. For each $i \in A_{nr}$, a second record based on the second nearest neighbor, with $x = x_i$, and $y^* = y_{nn2(i)}$, can be incorporated into the data set with $w_{i0} = 0$. When the number of imputed cases is less than n, each imputed case, i, may be assigned to a unique r. Setting

$$\begin{aligned} w_{ir'} &= 1 \quad \text{for } y_i^* \text{ record, } r' \neq r,\ 0 \le r' \le n \\ &= 0 \quad \text{for } nn2 \text{ record, } r' \neq r,\ 0 \le r' \le n \end{aligned} \quad (9)$$

$$\begin{aligned} w_{ir} &= 0 \quad \text{for } y_i^* \text{ record,} \\ &= 1 \quad \text{for } nn2 \text{ record,} \end{aligned}$$

exactly implements (5).

For more than 100 imputations, the cases can be serially assigned to $r=1,\ldots,100$, losing some available precision but nonetheless producing a reasonable variance estimator with effective degrees of freedom approaching 100.

3.2 Donor Reuse The approach of 3.1 must be modified to account for use of donors more than once, which was to be required for NRFU sampling in tracts with mail response rates over 80% and for UAA sampling in general. If imputed cases i and i' share the same nearest neighbor, that is, $nn1(i) = nn1(i')$, then the expected value of the cross product is

$$\begin{aligned} &E_\xi((y_{nn1(i)} - y_i)(y_{nn1(i')} - y_{i'})) \\ &= Var_\xi(y_{nn1(i)} \mid x_{nn1(i)}) \end{aligned} \quad (10)$$

using (3). In other words, reuse of a donor contributes additional covariance affecting the variance of the estimated sum, Y^*. Estimator (5) does not incorporate the effect of this covariance.

In addition to the first 100 replicates as described, the full variance estimator incorporated two more sets of 100 replicates each that, when used jointly in (8) and (7), represented the effect of (10). To understand the general case, it is helpful to consider first a special case. For each donor, k, let $nn1^{-1}(k)$ denote the set of imputed cases with donor k. Donor k is associated with a replicate r in 1-100, and (9) is implemented for this assignment of r for each imputed case in $nn1^{-1}(k)$.

Special Case: Suppose that, for each donor, k, whenever $nn1^{-1}(k)$ comprises $c_k > 1$ elements, each imputed case in $nn1^{-1}(k)$ has a different assigned second nearest neighbor. For this special case,

$$\begin{aligned} &E_\xi((\sum_{i \in nn1^{(-1)}(k)} (y_i - y_i^*))^2) \\ &= \sum_{i \in nn1^{(-1)}(k)} E_\xi(y_i - y_{nn1(i)})^2 \\ &\quad + 2\, c_k (c_k - 1)\, V_{ix}(y_k \mid x_k) \\ &= E_\xi((\sum_{i \in nn1^{(-1)}(k)} (y_{nn2(i)} - y_{nn1(i)}))^2) \end{aligned} \quad (11)$$

since for

$$\begin{aligned} &E_\xi((y_{nn1(i)} - y_i)(y_{nn1(i')} - y_{i'})) \\ &= Var_\xi(y_k \mid x_k) \end{aligned} \quad (12)$$

Thus, replicates 1-100 provide a consistent variance estimate.

When two imputed cases share both the first and second nearest neighbors, (12) no longer holds; in fact,

$$\begin{aligned} &E_\xi((y_{nn1(i)} - y_{nn2(i)})(y_{nn1(i)} - y_{nn2(i)})) \\ &= 2\, Var_\xi(y_k \mid x_k) \end{aligned} \quad (13)$$

To address the resulting variance estimate in the general case, replicates 101-200 and 201-300 are included in the calculation. For any order pair of donors, k, k', let $nnp^{-1}(k, k')$ be the set of imputations, i, with $nn1(i) = k$, and $nn2(i) = k'$, and let $c_{k,k'}$ be the number of imputations associated with the pair. For $c_{k,k'} > 1$, assign the pair to a replicate $101 \le r \le 200$. As in (9), for $101 \le r' \le 200$

$$\begin{aligned} w_{ir'} &= 1 \quad \text{for } y_k^* \text{ record, } r' \neq r \\ &= 0 \quad \text{for } nn2 \text{ record, } r' \neq r \\ w_{ir} &= 0 \quad \text{for } y_k^* \text{ record,} \\ &= 1 \quad \text{for } nn2 \text{ record,} \end{aligned} \quad (14)$$

These replicates are used with $b_r = -1/2$ in (7). These replicates correct effect of overestimation of the variance from the cross-product terms in (13). Unfortunately, they also subtract too much from the diagonal. To compensate, a third series of replicates $201 \le r \le 300$ with $b_r = 1/2$ is constructed similarly for imputed cases with $c_{k,k'} > 1$. Each imputed observation is assigned to a distinct replicate, r, for which (14) is again applied.

The presence of negative terms in the variance estimate risks negative variance estimates to a generally small degree, and it appears to increase the variance of the variance estimate compared to some alternatives. When the conditions of the Rao and Shao variance approach are met, earlier simulation of the variance estimator showed higher variance than the Rao-Shao variance formula.

The less obvious advantage to this approach is that it is directed to estimating the variance of subdomains as well as the domain total. The issue of variance estimation for subdomains was previously raised (Fay 1996) with respect to multiple imputation.

Previously cited work on this approach covers additional types of applications, including to sample surveys with negligible sampling fractions and those where the finite population correction was important. Extensions to other replication methods besides the jackknife remains an open question.

---

* 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. Research results and conclusions expressed are those of the author and do not necessarily indicate concurrence by the Census Bureau. It is released to inform interested parties of current research and to encourage discussion.

The author would like to thank Aref Dajani and Cary Isaki for helpful comments and Mary Ann Cochran for editorial assistance.

References

Bankier, M., Houle, A.-M., Luc, M., and Newcombe, P. (1997), "1996 Canadian Census Demographic Variables Imputation," *1997 Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 389-394.

Bankier, M., Luc, M., Nadeau, C., and Newcombe, P. (1996), "Imputing Numeric and Qualitative Variables Simultaneously," *1996 Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 90-99.

Coder, J. (1978), "Income Data Collection and Processing for the March Income Supplement to the Current Population Survey," *Proceedings of the Data Processing Workshop: Survey of Income and Program Participation,* U.S. Department of Health, Education, and Welfare, Washington, DC.

College, M.J., Johnson, J.H., Paré, R., and Sande, I.G. (1978), "Large Scale Imputation of Survey Data," *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Washington, DC, pp. 431-435.

David M., Little, R.J.A., Samuhel, M.E., and Triest, R.K. (1986), "Alternative Methods for CPS Income Imputation," *Journal of the American Statistical Association,* **81**, 29-41.

Dick, P. (1995), "Modelling Net Undercoverage in the 1991 Canadian Census," *Survey Methodology,* 21, 45-54.

Farber, J. E., Fay, R.E., and Schindler, E.L. (1998), "The Statistical Methodology of Census 2000," unpublished manuscript.

Farber, J. E. and Griffin, R. (1998), "A Comparison of Alternative Estimation Methodologies for Census 2000," *1998 Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 629-634.

Fay R.E. (1996), "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association,* **91**, 490-498.

Fay, R. E. and Farber, J. (1999), "The Census 2000 Dress Rehearsal: Methodological Basis for Nonresponse Followup Estimation," to be presented at the Federal Committee on Statistical Methodology Research Conference, Arlington, VA, Nov. 15-17, 1999.

Fay, R. E. and Town, M. K. (1996), "Variance Estimation for the 1995 Census Test: Methodology and Findings," in *Proceedings of the 1996 Annual Research Conference,* U.S. Bureau of the Census, pp. 761-781.

_____ (1998), "Variance Estimation for the 1998 Census Dress Rehearsal," *1998 Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 605-610.

Ford, Barry L. (1983), "An Overview of Hot-Deck Procedures," Madow, W.G., Olkin, I., and Rubin, D.B. (eds.), *Incomplete Data in Sample Surveys, Vol. 2,* National Research Council, Panel on Incomplete Data, Academic Press, New York, pp. 185-207.

Germain, M.-F. and Julien, C. (1993), "Results of the 1991 Census Coverage Error Measurement Program,

Proceedings of the Seventh Annual Research Conference, U.S. Bureau of the Census, pp. 55-70.

Lee, H., Rancourt, E., and Särndal, C.E. (1994), "Experiments with Variance Estimation from Survey Data with Imputed Values," *Journal of Official Statistics,* **10**, 231-243.

Navarro, A., Treat, J., and Mulry, M. (1996), "Nonresponse Followup: Unit vs. Block Sampling," *1996 Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 551-556.

Rao, J.N.K. and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation," *Biometrika*, **79**, 811-822.

Rizvi, M.H. (1983), "An Empirical Investigation of Some Item Nonresponse Adjustment Procedures," in Madow, W.G., Nisselson, H. and Olkin, I. (eds.), *Incomplete Data in Sample Surveys, Vol. 1,* National Research Council, Panel on Incomplete Data, Academic Press, New York, pp. 299-366.

Robinson, J.G., Ahmed, B., Das Gupta, P., and Woodrow, K.A., "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis," *Journal of the American Statistical Association,* **88**, 1061-1071.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys,* New York: John Wiley.

Sande, G.T. (1983), "Replacement for a Ten-Minute Gap," in Madow, W.G. and Olkin, I., *Incomplete Data in Sample Surveys, Vol. 2,* National Research Council, Panel on Incomplete Data, Academic Press, New York, pp. 337-338.

Sande, I.G. (1983), "Hot-Deck Imputation Procedures," in Madow, W.G. and Olkin, I., *Incomplete Data in Sample Surveys, Vol. 2,* National Research Council, Panel on Incomplete Data, Academic Press, New York, pp. 339-349.

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

Steel, P. and Fay, R.E. (1995), "Variance Estimation for Finite Populations with Imputed Data," *1995 Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 374-379.

Thompson, J. H. and Fay, R. E. (1998), "Census 2000: The Statistical Issues," *1998 Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 101-110.

Town, M.K., and Fay, R.E. (1995), "Properties of Variance Estimators for the 1995 Census Test," *1995Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 724-729.

Tsay, J.H., Isaki, C.T., and Fuller, W.A. (1996), "A Block Based Nonresponse Followup Survey Design," *1996 Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 557-562.

U.S. Census Bureau (1999a), "Contamination of Initial Phase Data Collected in ICM Block Clusters," by Sam Hawala, Census 2000 Dress Rehearsal Evaluation Memorandum C2, Executive Summary available from http://www.census.gov/census2000/evaluations/pdf/sumc2.pdf.

_____(1999b), "Specifications for Nonresponse Followup and Undeliverable-as-Addressed Vacant Estimation in the Census 2000 Dress Rehearsal," DSSD Census 2000 Dress Rehearsal Memorandum Series # A-40, from Donna L. Kostanich to Dennis W. Stoudt.

Wright, T. (1999), "A one-number census: some related history," *Science,* **283,** pp. 491-492.

Zanutto, E. and Zaslavsky, A.M. (1996), "Estimating a Population Roster from an Incomplete Census, Using Mailback Questionnaires, Adminstrative Records, and Sampled Nonresponse Followup," *1996 Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, VA, pp. 538-543.