VARIANCE ESTIMATION FOR THE 1998 CENSUS DRESS REHEARSAL

Robert E. Fay and Machell Kindred Town[1]
Robert Fay, U.S. Bureau of the Census, Washington, DC 20233-9001

Key words: Nearest-neighbor imputation, generalized variances, replication

1. Introduction.

There are two primary stages of sampling and estimation for Census 2000 (U.S. Bureau of the Census 1998, Wright 1998, Farber, Fay, and Schindler 1998). The first stage occurs during the initial phase of the enumeration. As in previous censuses, there will be an attempt to contact all households in the country. In most areas, the contact will be through the mailing or dropping off of a questionnaire to each housing unit, in a manner similar to the 1980 or 1990 censuses. It is expected that the majority of households will again respond by mail, as requested. Instead of following up all nonrespondents, however, the Census Bureau will select a sample for nonresponse followup (NRFU) approximately two weeks after Census Day. Sampling rates will be set to yield a 90% total response to the census in each census tract. For example, if the initial response is 75% in a tract, the sampling rate for nonresponse followup will be 60%. If the initial response is above 85% in a tract, however, a 1-in-3 sample will be selected.

Sampling will also be used in a related manner for a second component of the initial phase. In most of the country, the U.S. Postal Service will deliver census forms, returning undeliverable ones, including those for units believed by the carrier to be vacant. In the 1995 census tests, approximately 28% of these undeliverable - as - addressed (UAA) vacants were occupied. All such units were assigned to followup in 1980 and 1990; in 2000, these units will be sampled at 3-in-10.

Estimates from these samples, combined with mail returns, Be Counted forms, and other concurrent census operations, will constitute the initial phase of the enumeration. The initial phase is expected to cover the population approximately as well as the 1990 census.

The Integrated Coverage Measurement (ICM) is the second major stage of sampling and estimation. The ICM is to be based on a sample of approximately 750,000 housing units and is designed to estimate the population missed by the initial phase. The sample will be composed of about 25,000 block clusters averaging approximately 30 housing units each. The ICM parallels the 1990 Post-Enumeration Survey (PES) in many respects, but it is both larger and designed to permit its results to be incorporated into the census products, including the apportionment counts due on December 31, 2000. The current plans are to produce direct estimates of population at the state level and to distribute the ICM corrections down to the block level.

Thus, both forms of sampling will affect estimates at all levels. This paper describes the variance estimation approach to be implemented in the Dress Rehearsal, as the basis for the methodology in Census 2000. The approach employs replication to reflect the variance from both the NRFU/UAA vacant estimation and the ICM component.

More specifically, the replication will be implemented through the creation of replicate weights. Replicate weights assigned to each observation are analogous to survey weights but specify how the observation is to be weighted to create the respective replicate estimates. Suppose a census estimate of total, Y, is expressed as a weighted sum,

$$Y = \sum_i w_{i0} y_i . \qquad (1)$$

In census production for the short-form data, usually all the weights are 1's. Expressing the census estimate in the form of (1), however, permits some necessary generalizations for the purposes of variance estimation. Suppose a replicate estimate,

$$Y_r = \sum_i w_{ir} y_i , \qquad (2)$$

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. A generalized variance estimate of the variance of Y is given by

$$Var(Y) = \sum_{r=1}^{n} b_r (Y_r - Y)^2 , \qquad (3)$$

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

The expression of a replication method through replicate weights greatly facilitates the calculation of direct variance estimates. The Census Bureau currently uses replicate weights for the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), and the American Community Survey (ACS).

Replicate weights were not used in the 1995 Census Tests because appropriate methodology was unavailable to reflect the combination of the block sample drawn for NRFU and the NRFU estimation methodology below the site level. Site level variance estimates were obtained through replication without replicate weights, through an adaptation of the stratified jackknife (Town and Fay 1995).

Empirical evidence from 1995 Census Tests supported the combined use of unit sampling in NRFU and block sampling for the ICM. In other words, the NRFU sample can be selected on the basis of housing units rather than whole blocks. The alternative design, using block sampling for both NRFU and ICM, was substantially less effective for NRFU estimation (Fay and Town 1996).

In addition to the variance gains for NRFU for the Dress Rehearsal and for Census 2000, the choice of unit sampling also makes nearest-neighbor imputation an effective methodology for NRFU estimation (Farber and Griffin 1998). In turn, this estimation strategy permits application of the variance estimator for nearest neighbor imputation discussed in Fay and Town (1996), which is based on replicate weights. The ICM variance contribution may also be estimated through a stratified jackknife implemented through replicate weighting, thus giving a unified computational approach to the variance estimation.

2. Defining Replicate Weights for NRFU/UAA Estimation

2.1 General Approach The NRFU universe will be established approximately two weeks after Census Day, on the basis of response up to that date. A systematic sample of units will be selected at the rate determined by the initial response rate at that point. After data are collected for the NRFU sample units, each non-sample NRFU housing unit will be imputed from some sample NRFU unit ("donor") nearby in a systematic sort of the file (Farber and Griffin 1998). This is consequently a form of "nearest neighbor" imputation, where the standard for nearness in this case is the sorted order of the census file, which is primarily geographic. In other words, for each nonsample case, *j*, an imputation will be made from its nearest neighbor, $nn1(j)$. For multi-unit addresses, the algorithm will give preference to matches between units at the same street address.

The proposed variance estimator requires that a "second nearest neighbor," $nn2(j)$, always different from the nearest neighbor (donor), be identified for each imputation. Basically, the second nearest neighbor is the imputation that one would have made had the actual donor been excluded from the universe.

The variance estimator for nearest neighbor imputation is appropriate for a number of sampling conditions and for multiple use of the same donor. In the simplest case, however, in which a donor is used at most once (likely for sampling rates above 1-in-2 for NRFU, which are planned wherever the initial response is at or below 80%), the variance estimator is simply a replication approximation of

$$Var(Y) = \sum_{j \in A_{nr}} (y_{nn1(j)} - y_{nn2(j)})^2 , \qquad (4)$$

where A_{nr} represents nonsample units. In other words, the estimator in effect is the sum of squared differences over nonsample units between the imputation actually made and an equivalent imputation that might have been made.

As an aside, (4) provides us a preview of the variance effect for NRFU sampling. If the number of persons in a followup unit has approximately a unit standard deviation equal to its expected value, i.e., a c.v. of 1.0 on a unit level, then we would expect a block of 30 units with 3 nonsample units to have a c.v. of about 8% and for a block of 90 units, about 5%.

In the more general case, the variance estimator implemented through replication includes three types of replicates, grouped into the ranges 1-100, 101-200, and 201-300. By some form of systematic assignment, each actual donor used will be assigned to some replicate *r*, between 1 and 100. For each nonsample case using this donor, the data from the second nearest neighbor will be substituted in replicate *r*. In other words w_{ir}= 1 for the imputations from the second nearest neighbors and w_{ir}= 0 for the actual imputation. For all the remaining replicates between 1 and 100, the full sample weights are used. If the donor is used more than once, the second nearest neighbors may be the same or different. The replicate weights for replicates 1 through 100 create the term from Fay and Town (1996)

$$\begin{aligned} Y^c(-k) &= Y \\ &+ \sum_{j \in nn1^{-1}(k)} (y_{nn2(j)} - y_k) \quad \text{if } k \in A_r \qquad (5) \\ &= Y \qquad\qquad \text{if } k \in A_{nr} \end{aligned}$$

where A_r is the set of NRFU sample cases and A_{nr} represents the nonsample NRFU units. In this formula, $nn1^{-1}(k)$ denotes the set of imputed cases with donor *k*.

The full form of the variance estimator, including the remaining terms, is:

$$
\begin{aligned}
V(Y) \;=\; & \sum_{k \in A_r} \left[Y^c(-k) - Y \right]^2 \\
& - \frac{1}{2} \sum_{k, k' \in A_r} \left[\sum_{j \in nnp^{-1}(k,k')} \left(y_{k'} - y_k \right) \right]^2 \\
& + \frac{1}{2} \sum_{j \in A_{nr}} \left[y_{nn2(j)} - y_{nn1(j)} \right]^2
\end{aligned}
\qquad (6)
$$

In (6), $nnp^{-1}(k, k')$ represents the set of nonsample cases with nearest and second-nearest neighbors k and k', respectively. The second term on the right-hand side of (6) is realized with replicate weights 101-200, and the third term with replicate weights 201-300. Replicates 101-200 are organized by pairings of first and second nearest neighbors, and employs $b_r = -1/2$. This negative term compensates for some potential overestimation of the variance from the same pairing of first and second nearest neighbors. The third term concerns errors in prediction, with $b_r = 1/2$. In the simple case of only one use of a donor, the second and third terms cancel and reduce the entire expression, in effect, to (4).

Estimator (6) relies on model ξ assumptions:

$$
\begin{aligned}
E_\xi(y_k) &= E_\xi(y_{nn1(k)}) \\
V_\xi(y_k \mid x_k) &= 1/2 \; E_\xi(y_{nn1(k)} - y_{nn2(k)})^2 \qquad (7) \\
Cov_\xi(y_k, y_{k'} \mid x_k, x_{k'}) &= 0 \,, \; k \neq k'
\end{aligned}
$$

Note that (6) does not assert a specific functional relationship between the x's and the y's or their variance. This assumption might be called local exchangeability, since the missing nonsample case and first and second nearest neighbors are assumed to come from the same distribution.

In the decennial application, the x's represent the geographic information reflected in the sort of the census frame, including the information on whether the units are multi-units at the same address.

2.2 Critical Issues The preceding account divides NRFU estimation neatly into sampled and nonsampled units. In fact, some households return their forms after the initial date. As a matter of policy, the Census Bureau is committed to accepting this information for several additional weeks. To compensate for the somewhat nonrandom nature of these responses, a modification to the hot deck imputation is planned to compensate for the fact that most late respondents will be from occupied units (Farber, Fay, and Schindler 1998). We hope to conduct empirical investigations to investigate the effect of these modifications on the underlying assumptions and their consequence for the performance of the estimator.

Dress Rehearsal NRFU is now complete, and we will soon receive the data to begin calculations. Some mail returns and followup outcomes were lost during data capture, and cannot be recovered. Lost mail returns have been imputed from other mail returns through nearest-neighbor imputation; lost NRFU forms will be treated as nonsample units. Fortunately, we are able simply to expand the nearest-neighbor treatment to include these cases without any significant modification to the planned approach.

3. Defining Replicate Weights for ICM Estimation

3.1 General The ICM component is based on the 1990 PES (Hogan 1992, 1993). Incompleteness of the initial phase of the census is estimated through dual-system estimation (DSE). The ICM employs an overlapping sample of blocks for the two basic survey components:

- A *population* or *P sample,* an independent listing of housing units and interviewing of residents, which provides information on census omissions, and
- An *enumeration* or *E sample* selected from the initial phase in order to estimate erroneous enumerations and enumerations with insufficient information to match to the P sample.

The P-sample results are summarized as an estimate $\hat{M}$, of cases matched to the initial enumeration out of the weighted P-sample total of $\hat{N}_p$. The E sample provides an estimate of erroneous enumerations and enumerations with insufficient information to match (incomplete or missing names, *etc.*), $\hat{EE}$, out of its weighted total $\hat{N}_e$. The initial phase estimate IP includes II imputations for persons without the minimum number of characteristics to be considered data defined in the initial enumeration. Since the II imputations are not allowed to match to the P sample, they are excluded from the E sample. The DSE is

$$
D\hat{S}E \;=\; (IP - II) \times \frac{(\hat{N}_e - \hat{EE})}{\hat{N}_e} \times \frac{\hat{N}_p}{\hat{M}} \,. \qquad (8)
$$

The DSE is based on the assumption of statistical independence of the initial enumeration and P-sample coverage. As in previous studies, the estimator will be computed separately within poststrata defined by age, sex, race/ethnicity, tenure, and geography. In the Dress Rehearsal and potentially in Census 2000, the direct DSE's will be adjusted by a raking procedure. An array of the initial phase estimates will be adjusted through iterative proportional fitting to the marginal totals of an array of DSE estimates. In the Dress Rehearsal, for example, the marginal totals will be for age by sex by race/ethnicity as one dimension, and for tenure as the second.

In the 1990 PES, the stratified jackknife (Krewski and Rao 1981) was applied to estimate the variance of the DSE, both for the direct DSE's based on 1392 poststrata in 1991, which were input to a "smoothing" procedure (Hogan 1993) and for 357 poststrata in 1992. Both applications employed the same replicate design, which resulted in one replicate for each sample block cluster, or over 5000 replicates. The purpose of this large number of replicates was to obtain as precise an estimates of variance as possible for the PES, particularly since variance estimates were a component of the 1991 smoothing.

The stratified jackknife is appropriate for stratified sampling with equal or unequal probabilities of selection, with replacement. Differences between sampling with and without replacement will be trivial in 2000, since the ICM sampling fraction will be negligible. Finite population corrections will be ignored for the Dress Rehearsal. The stratified jackknife does not reflect any internal stratification resulting from systematic sampling of the ICM sample.

In the Dress Rehearsal, the stratified jackknife will be independently implemented at each site, creating several hundred replicates. Specifically, suppose the ICM strata are ordered and sample block clusters, the unit of sampling, ordered within the strata. In the stratified jackknife, a cluster from stratum k with n_k sample block clusters is deleted from one replicate, while the remaining block clusters in the stratum are multiplied by $n_k/(n_k - 1)$. The coefficient b_r is $(n_k - 1)/n_k$.

If this procedure is extended in 2000, it may be implemented separately in each state. There may be some drawbacks to this approach, however. Variance for estimates involving more than one state may be computed by summing both state-level contributions to the estimate and to the total variance, but this approach does not permit the convenience of working with replication. The current allocation provides over 2000 block clusters to California, which would require an equal number of replicates.

A possible alternative for 2000 would be based on a *modified stratified jackknife*. The purpose of the modification is to create replicates similar to the stratified jackknife but where the appropriate b_r are all 1. In place of the deletion of one cluster from stratum k with n_k sample block clusters, this cluster should be multiplied by $(1 - ((n_k - 1)/n_k)^{1/2})$ instead. In place of multiplying the remaining block clusters in the stratum by $n_k/(n_k - 1)$, they should be multiplied by $(1 + (n_k(n_k - 1))^{-1/2})$. (This device may be a reinvention on our part of an approach already taken by others.)

With $b_r = 1$, it is possible to confound replicates from different strata, and consequently impose a maximum number of replicates per state. For example, suppose a limit of 400 replicates is set on the number of replicates per state, but that a given state has 500 block clusters. Considering the block clusters to be sorted by sampling stratum, one may create modified stratified replicates for the first 400 replicates in the usual manner. For the 401*st*, however, instead of creating a replicate 401, one may instead modify the first replicate. In this way, what would have been replicates 401-500 may be wrapped onto replicates 1-100. Although this confounding reduces the precision of the variance estimate compared to the results from the full use of 500 replicates, it is approximately unbiased unless a stratum wraps over itself (for example, if one stratum had more than 400 sample clusters).

After the replicates samples are generated, the following steps of estimation will be performed and implemented for each replicate:

1. Recalculation of the missing data adjustments for missing P-sample match status and E-sample correctness of enumeration.
2. Calculation of the dual-system estimates.
3. Performing the raking adjustments.
4. Compute adjustment factors as the ratio of the raked adjustment to the original initial phase value.

The results of these calculations will be a set of replicate adjustment factors based on the stratified jackknife. Within each tabulation block, identify each of the ICM poststrata for which post-NRFU persons are found. In small blocks, far more of these poststrata will be present than result in actual adjustment persons. For each poststratum within a block, a record will be created. These records will have noninteger weights. Let the full sample weight for a poststratum be the product of the number of persons and estimated persons in the post-NRFU estimate times the adjustment factor for the poststratum minus 1. For replicates 1-300, the replicate weight for the poststratum will be the full-sample adjustment factor minus 1 times the replicate total of estimated post-NRFU persons in the poststratum. For replicates 301 to 300+n, let the replicate weight for 300+k be the full-sample estimated post-NRFU persons times the replicate adjustment factor for replicate k of the stratified jackknife.

The plan is to construct two sets of replicate weights and associated coefficients b_r,

- 300 replicate weights, defined at the housing unit level, to represent NRFU and UAA imputation. Only imputed cases and associated data will receive replicate weights different from 1. This number of replicates was previously employed in the empirical work reported by Fay and Town (1996).

- A set of replicate weights to represent ICM adjustment, where n' is the number of selected block clusters. Because the replicate weights will be based on a stratified jackknife, the number of replicates proposed is the number, n', of sampled block clusters in the site. Only records representing ICM adds (before controlled rounding) will receive replicate weights different from w_{i0} for replicates above 300. For replicates 1-300, the replicate weights for ICM adds will capture the variance in the estimated ICM correction due to variance in the post-NRFU estimates.

The following components will be included in the variance estimates based on the replicate weights:

1. Variance from using imputation to estimate nonsample NRFU units from NRFU units in sample, and the similar variance for UAA estimation.
2. Variance from estimating the missing data in ICM. A recent decision was to simplify missing data estimation to impute cell proportions for missing P-sample match probabilities and missing E-sample probabilities of correct enumeration, instead of the more complex hierarchical logistic regression procedure in 1990. This decision coincidentally makes estimation of this component of variance through replication straightforward under the assumptions of the missing data model.
3. Variance from applying the estimated adjustment factors from the dual system estimates to the initial enumeration. We will be able to estimate the variance before controlled rounding.

What will be excluded from the replicate weights planned for the Dress Rehearsal

4. Missing data variance from imputing missing item data in the initial enumeration.
5. The effect of local heterogeneity in census undercount. (It may be possible to do a separate estimation of this component averaged over a large number of blocks, but there is no simple way to build this component into the block level estimate. It is even less clear whether estimates can be made for higher geographic levels.)
6. The effect of controlled rounding of ICM estimates.

4. Computing Generalized Variance Estimates

4.1 General The publication of the estimates and their associated variances is an essential part of the 1998 Dress Rehearsal. The decision to construct generalized variances for the small area estimates will allow the variances to be published at the lower levels of geography. The current plan is to report direct variances for State, County, and Congressional Districts and parameters for the generalized variance functions for the tract and block level estimates. For the Dress Rehearsal, we will follow closely the previous work of Krenzke and Navarro (1996) for the 1995 Census Test using the weighted least squares GATT Curve Model. The parameters will be constructed for the tract and block levels by Public Law data items. The data items are categorized by total population, race, age, and Hispanic Origin.

4.2 Methodology The generalized variance function uses regression models to estimate the relationship between the estimated relative variance and the estimated total. The estimated relative variance is the variance of the estimate divided by the estimate squared. The estimate of interest can be substituted into the generalized variance function equation using the computed parameters to calculate the standard error or the relative variance. The decision to use the weighted least squares model was decided because of timing and the fact that the model was successfully tested against other models (Krenzke and Navarro, 1996). There will be testing of the generalized variance function using the data from the 1998 Dress Rehearsal to determine what models should be used for the 2000 Census. The weighted least squares model:

$$V_x^2 = V_y^2 + b(\frac{1}{x} - \frac{1}{y}) \qquad (9)$$

where

x = the estimated public law item total,
y = the estimated site total,
V_x^2 = the relative variance of x,
V_y^2 = the relative variance of y, and
b = the estimated regression parameter for the model.

Krenzke and Navarro note that (9) is equivalent to

$$V_x = a + \frac{b}{x} \qquad (10)$$

with $a = V_y - \frac{b}{y}$.

The generalized variance function will be computed in SAS using the Regression procedure. The relative variance of y will be forced as the intercept so that the **a** parameter can be computed for publication. There will be 9 iterations run on the weighted least squares model. For each iteration, the weights (the inverse of the squared relative variance of x) will be adjusted and the outliers will be removed. The outlier detection methodology from the Regression procedure in SAS will be used to determine the outliers. In the initial iteration, all of the observations are used. Once the parameters stabilize, the absolute studentized residuals are compared with the given maximum value. The absolute relative deviations (ARD) are calculated for each observation. The ARD is the absolute difference between the predicted relative variance of x and the observed relative variance of x, divided by the observed relative variance of x. The observations are removed as outliers when the absolute residual is greater than the given maximum value or the predicted relative variance is 50 times bigger than the observed relative variance. The absolute standard residuals indicate outliers at the higher end of the curve of residual by predicted values and the ARD indicate outliers at the lower end of the curve of residual by predicted values. The second iteration is run using the observations that were not removed from the model. The process continues to identify and remove outliers. This process continues for 9 iterations by which all the absolute residuals should be less than the given maximum value. After the 9th iteration, the **b** parameters are produced. The **a** parameter will then be computed from (10) using the **b** parameter and the relative variance of y and the estimate of y from the 9th iteration.

The standard error of the estimate of x is computed using:

$$se_x = \sqrt{a\,x^2 + b\,x} \qquad (10)$$

where x is the estimated number of persons, **a** is the estimated regression parameter calculated using the **b** parameter, and **b** is the estimated regression parameter calculated in the 9th iteration. Documentation will accompany the parameters for the data user to be able to compute the needed variances for the tract and block level estimates.

Research will be performed on the Dress Rehearsal data to determine what generalized variance function model should be used to estimate the **a** and **b** parameters for the 2000 Census block and tract level estimates.

[1] 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 authors 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 authors would like to thank Scot Dahl and William Winkler for helpful comments.

REFERENCES

Farber, J. E., Fay, R.E., and Schindler, E.L. (1998), "The Statistical Methodology of Census 2000," unpublished manuscript submitted to the *American Statistician.*

Farber, J. E., and Griffin, R. A. (1998), "A Comparison of Alternative Estimation Methodologies for Census 2000," paper presented at the 1998 Joint Statistical Meetings, Dallas, TX, Aug. 9-13, 1998.

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.

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

______ (1993), "The 1990 Post-Enumeration Survey: Operations and Results," *Journal of the American Statistical Association,* **88**, 1047-1060.

Krenzke, T. R. and Navarro, A., (1996), "Sampling Error Estimation in the 1995 Census Test for Small Areas," *1996 Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Krewski, D. and Rao, J. N. K. (1981), "Inference from Stratified Samples: Properties of the Linearization, Jackknife, and Balanced Repeated Replication Methods," *The Annals of Statistics,* **9**, 1010-1019.

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

U.S. Bureau of the Census (1998), *Census 2000 Operational Plan*, U.S. Department of Commerce, Washington, DC.

Wright, T. (1998), "Sampling and Census 2000: The Concepts, " *American Scientist*, **86**, 245-253.