Submitted to the Annals of Applied Statistics

VARIANCE ESTIMATION FOR NEAREST NEIGHBOR IMPUTATION FOR U.S. CENSUS LONG FORM DATA

By Jae Kwang Kim*, Wayne A. Fuller and William R. Bell

Iowa State University and U.S. Bureau of Census

Variance estimation for estimators of state, county, and school district quantities derived from the Census 2000 long form are discussed. The variance estimator must account for (1) uncertainty due to imputation, and (2) raking to census population controls. An imputation procedure that imputes more than one value for each missing item using donors that are neighbors is described and the procedure using two nearest neighbors is applied to the Census long form. The Kim and Fuller (2004) method for variance estimation under fractional hot deck imputation is adapted for application to the long form data. Numerical results from the 2000 long form data are presented.

1. Introduction . In Census 2000 income data were collected on the long form that was distributed to about one of every 6 households in the United States. These data were used to produce various income and poverty estimates for the U.S., and for states, counties, and other small areas. The state and county income and poverty estimates from the Census 2000 long form sample have been used in various ways by the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program. The poverty estimates produced by SAIPE have been used by the U.S. Department of Education in allocating considerable federal funds each year to states and school districts. In 2008 the Department of Education used SAIPE estimates, directly and indirectly, to allocate approximately $16 billion to school districts.

The Census 2000 long form had questions for eight different types of income for each individual in a household. (For details, see Table 1 in Section 5.) If there was nonresponse for an income item, a version of nearest neighbor imputation (NNI) was used, where the nearest neighbor was determined by several factors such as response pattern, number of household members, and other demographic characteristics. NNI is a type of hot deck imputation that selects the respondent closest, in some metric, to the nonrespondent, and

* The research was supported by a contract with the U.S. Census Bureau. We thank George McLaughlin and George Train for computational support and Yves Thibaudeau for discussion on the imputation methods.

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35

Keywords and phrases: Fractional imputation, Hot deck imputation, Nonresponse, Replication variance estimation

inserts the respondent value for the missing item. Most imputation rates for income items in the Census 2000 long form data were more than double the corresponding imputation rates from the 1990 census (Schneider 2004, pp. 17-18 and Table 1, p. 27). For example, the Census 2000 imputation rate for wage and salary income was 20% while in 1990 it was 10%, and for interest and dividend income the imputation rates were 20.8% in 2000 and 8.1% in 1990. Overall, 29.7% of long form records in 2000 had at least some income imputed, compared to 13.4% in 1990. Given the 2000 imputation rates, it is important that variance estimates for income and poverty statistics reflect the uncertainty associated with the imputation of income items.

The Census Bureau performed nearest neighbor imputation for eight income items in producing the long form estimates. The estimation procedure had been implemented and the estimates were not subject to revision. Our task was to estimate the variances of the existing long form point estimates that are used by the SAIPE program. The problem is challenging because of the complexity of the estimates. While total household income is a simple sum of the income items for persons in a household, and average household income (for states and counties) is a simple linear function of these quantities, our interest centers on (i) median household income, and (ii) numbers of persons in poverty for various age groups. Poverty status is determined by comparing total family income to the appropriate poverty threshold, with the poverty status of each person in a family determined by the poverty status of the family. For such complicated functions of the data, the effects of imputation on variances are difficult to evaluate.

It is well known that treating the imputed values as if they are observed and applying a standard variance formula leads to underestimation of the true variance. Variance estimation methods accounting for the effect of imputation have been studied by Rubin (1987), Rao and Shao (1992), Shao and Steel (1998), and Kim and Fuller (2004), among others. Sande (1983) reviewed the NNI approach, Rancourt, Särndal, and Lee (1994) studied NNI under a linear regression model, and Fay (1999) and Rancourt (1999) considered variance estimation in some simple situations. Chen and Shao (2000) gave conditions under which the bias in NNI is small relative to the standard error and proposed a model-based variance estimator. Chen and Shao (2001) described a jackknife variance estimator. Shao and Wang (2008) discussed interval estimation and Shao (2009) proposed a simple nonparametric variance estimator.

Our approach to estimating variances under NNI is based on the fractional imputation approach suggested by Kalton and Kish (1984) and studied by Kim and Fuller (2004). In fractional imputation, multiple donors, say M, are

chosen for each recipient. We combine fractional imputation with the nearest neighbor criterion of selecting donors, modifying the variance estimation method described in Kim and Fuller (2004) to estimate the variance due to nearest neighbor imputation. Replication permits estimation of variances for parameters such as median household income and the poverty rate. Also, replication is used to incorporate the effect of raking, another feature of the estimation from the Census 2000 long form sample.

It should be noted that the official estimation and imputation procedures for the long form were fixed and production was completed before the research described here was even started. Hence, our objective was to develop variance estimates, accounting for imputation and raking, for the production point estimates, not to explore alternative imputation procedures in an attempt to improve the point estimates. Thus, we used $M = 2$ nearest neighbor imputations in developing variance estimates for the production long form estimates that used $M = 1$ nearest neighbor imputation.

The paper is organized as follows. In Section 2, the model for the NNI method and the properties of the NNI estimator are discussed. In Section 3, a variance estimation method for the NNI estimator is proposed. In Section 4, the proposed method is extended to the stratified cluster sampling. In Section 5, application of the approach to the Census 2000 long form income and poverty estimates is described.

2. Model and estimator properties. Our finite universe U is the census population of the United States. The Census Bureau imputation procedure defines a measure of closeness for individuals. Let a neighborhood of individual g be composed of individuals that are close to individual g, and let B_g be the set of indices for the individuals in the neighborhood of individual g. We assume that it is appropriate to approximate the distribution of elements in the neighborhood by

$$y_j \overset{i.i.d.}{\sim} \left(\mu_g, \sigma_g^2\right) \qquad j \in B_g, \tag{1}$$

where $\overset{i.i.d.}{\sim}$ denotes independently and identically distributed. Chen and Shao (2000) have given conditions such that it is possible to define a sequence of samples, populations and neighborhoods so that the distribution of y_i can be approximated by that of (1). See also Section B of the online supplemental document for an alternative justification of (1). These conditions do not necessarily hold for our population because the neighbors are defined by discrete variables. If response is independent of y and if the value of the discrete variables are the same for all elements in B_g, then (1) holds when the original observations are independent. We feel (1) is reasonable because

the sample is large relative to a neighborhood composed of three sample individuals. We assume that response is independent of the y-values so that the distribution (1) holds for both recipients and donors.

Let $\hat{\theta}_n$ be an estimator based on the full sample. We write an estimator that is linear in y as

$$\hat{\theta}_n = \sum_{i \in A} w_i y_i,$$

where A is the set of indices in the sample and the weight w_i does not depend on y_i. An example is the estimated total $\hat{T}_y = \sum_{i \in A} \pi_i^{-1} y_i$, where π_i is the selection probability. Let $V(\hat{\theta}_n)$ be the variance of the full sample estimator. Under model (1) we can write

$$y_i = \mu_i + e_i,$$

where the e_i are independent $(0, \sigma_i^2)$ random variables and μ_i is the neighborhood mean. Thus, $\mu_i = \mu_g$ and $\sigma_i^2 = \sigma_g^2$ for $i \in B_g$. Then, under model (1) and assuming that the sampling design is ignorable under the model in the sense of Rubin (1976), the variance of a linear estimator of the total $T_y = \sum_{i \in U} y_i$ can be written

$$V\left\{\sum_{i \in A} w_i y_i - T_y\right\} = V\left\{\sum_{i \in A} w_i \mu_i - \sum_{i \in U} \mu_i\right\} + E\left\{\sum_{i \in A} \left(w_i^2 - w_i\right) \sigma_i^2\right\}.$$

Assume that y is missing for some elements and assume there are always at least M observations on y in the neighborhood of each missing value, where in the Census long form application, $M = 2$. Let an imputation procedure be used to assign M donors to each recipient. Let w_{ij}^* be the fraction of the original weight allocated to donor i for recipient j, where $\sum_i w_{ij}^* = 1$. If we define

$$d_{ij} = \begin{cases} 1 & \text{if } y_i \text{ is used as a donor for } y_j \\ 0 & \text{otherwise,} \end{cases}$$

then one common choice for w_{ij}^* is $w_{ij}^* = M^{-1} d_{ij}$ for $i \neq j$. Then

$$\alpha_i = w_i + \sum_{j \neq i} w_j w_{ij}^* = \sum_{j \in A} w_j w_{ij}^*$$

is the total weight for donor i, where it is understood that $w_{ii}^* = 1$ for a donor donating to itself. Thus the imputed linear estimator is

$$\hat{\theta}_I = \sum_{j \in A} w_j y_{Ij} = \sum_{i \in A_R} \alpha_i y_i,$$

where A_R is the set of indices for the n_R respondents and the mean imputed value for recipient j is

$$y_{Ij} = \sum_{i \in A} w_{ij}^* y_i. \tag{2}$$

Note that $y_{Ij} = y_i$ if j is a respondent. Then, under model (1),

$$V\left(\hat{\theta}_I - T_y\right) = V\left\{\sum_{i \in A} w_i \mu_i - \sum_{i \in U} \mu_i\right\} + E\left\{\sum_{i \in A_R} \left(\alpha_i^2 - \alpha_i\right) \sigma_i^2\right\}, \tag{3}$$

where A_R is the set of indices of respondents. The variance expression (3) is smaller for larger M, $1 \leq M \leq n_R$, as long as model (1) holds for the M nearest neighbors. See Kim and Fuller (2004).

3. Variance estimation. Let the replication variance estimator for the complete sample be

$$\hat{V}\left(\hat{\theta}\right) = \sum_{k=1}^{L} c_k \left(\hat{\theta}^{(k)} - \hat{\theta}\right)^2, \tag{4}$$

where $\hat{\theta}$ is the full sample estimator, $\hat{\theta}^{(k)}$ is the k-th estimate of θ_N based on the observations included in the k-th replicate, L is the number of replicates, and c_k is a factor associated with replicate k determined by the replication method. Assume that the variance estimator $\hat{V}(\hat{\theta})$ is design unbiased for the sampling variance of $\hat{\theta}$. If the missing y_i are replaced in (4) with y_{Ij} of (2), the resulting variance estimator $\hat{V}_{naive}(\hat{\theta})$ satisfies

$$E\left\{\hat{V}_{naive}(\hat{\theta})\right\} = V\left\{\sum_{i \in A} w_i \mu_i - \sum_{i \in U} \mu_i\right\} + E\left\{\sum_{k=1}^{L} \sum_{i \in A_R} c_k (\alpha_{i1}^{(k)} - \alpha_i)^2 \sigma_i^2\right\}, \tag{5}$$

where $\alpha_{i1}^{(k)} = \sum_j w_j^{(k)} w_{ij}^*$ and $w_j^{(k)}$ is the weight for element j in replicate k. The weights $\alpha_{i1}^{(k)}$ are called the naive replication weights.

We consider a procedure in which the individual w_{ij}^* are modified for the replicates, with the objective of creating an unbiased variance estimator. Let $w_{ij}^{*(k)}$ be the replicated fractional weights of unit j assigned to donor i at the k-th replication. Letting

$$\hat{\theta}_I^{(k)} = \sum_{i \in A_R} \alpha_i^{(k)} y_i,$$

where $\alpha_i^{(k)} = w_i^{(k)} + \sum_{j \neq i} w_j^{(k)} w_{ij}^{*(k)} = \sum_{j \in A} w_j^{(k)} w_{ij}^{*(k)}$, define a variance estimator by

$$\hat{V}\left(\hat{\theta}_I\right) = \sum_{k=1}^{L} c_k \left(\hat{\theta}_I^{(k)} - \hat{\theta}_I\right)^2.$$

The expectation of the variance estimator $\hat{V}(\hat{\theta}_I)$ is

(6)
$$E\left\{\hat{V}(\hat{\theta}_I)\right\} = E\left[\sum_{k=1}^{L}\left\{\sum_{i \in A_R}(\alpha_i^{(k)} - \alpha_i)\mu_i\right\}^2\right] + E\left[\sum_{i \in A_R}\left\{\sum_{k=1}^{L} c_k(\alpha_i^{(k)} - \alpha_i)^2\right\}\sigma_i^2\right].$$

Because the $w_{ij}^{*(k)}$ satisfy

$$\sum_{i \in A_R} w_{ij}^{*(k)} = 1 \tag{7}$$

for all j, then, under the model (1), ignoring the smaller order terms,

$$\begin{aligned} E\left\{\sum_{k=1}^{L}\left[\sum_{i \in A_R}\left(\alpha_i^{(k)} - \alpha_i\right)\mu_i\right]^2\right\} &= E\left\{\sum_{k=1}^{L}\left[\sum_{i \in A}\left(w_i^{(k)} - w_i\right)\mu_i\right]^2\right\} \\ &= V\left(\sum_{i \in A} w_i \mu - \sum_{i \in U}\mu_i\right). \end{aligned}$$

Thus, the bias of the variance estimator $\hat{V}(\hat{\theta}_I)$ is

$$Bias\left\{\hat{V}\left(\hat{\theta}_I\right)\right\} = E\left\{\sum_{i \in A_R}\left[\sum_{k=1}^{L} c_k\left(\alpha_i^{(k)} - \alpha_i\right)^2 - \left(\alpha_i^2 - \alpha_i\right)\right]\sigma_i^2\right\}.$$

If the replicated fractional weights were to satisfy

$$\sum_{k=1}^{L} c_k\left(\alpha_i^{(k)} - \alpha_i\right)^2 = \alpha_i^2 - \alpha_i, \tag{8}$$

for all $i \in A_R$, then the bias would be zero. However, it is difficult to define replicate weights that satisfy (8). Therefore we consider the requirement

$$\sum_{k=1}^{L} c_k\{(\alpha_i^{(k)} - \alpha_i)^2 + \sum_{t \in D_{Ri}}(\alpha_t^{(k)} - \alpha_t)^2\} = \alpha_i^2 - \alpha_i + \sum_{t \in D_{Ri}}(\alpha_t^2 - \alpha_t), \tag{9}$$

where $D_{Ri} = \left\{ t \, ; \sum_{j \in A_M} d_{ij} d_{tj} = 1, \ t \neq i \right\}$ is the set of donors, other than i, to recipients from donor i. Under assumption (1), the recipients in the neighborhood of donor i have common variance and (9) is a sufficient condition for unbiasedness.

We outline a replication variance estimator that assigns fractional replicate weights such that (7) and (9) are satisfied. There are three types of observations in the data set: (1) respondents that act as donors for at least one recipient, (2) respondents that are never used as donors and (3) recipients. The naive replicate weights defined in (5) will be used for the last two types. For donors, the fractional weights w_{ij}^* in replicate k will be modified to satisfy (7) and (9).

We first consider jackknife replicates formed by deleting a single element. The next section considers an extension to a grouped jackknife procedure. Let the superscript k denote the replicate where element k is deleted. First the replicates for the naive variance estimator (5) are computed, and the sum of squares for element i is computed as

$$\sum_{k=1}^{L} c_k \left(\alpha_{i1}^{(k)} - \alpha_i \right)^2 = \phi_i, \qquad i \in A_R,$$

where $\alpha_{i1}^{(k)}$ is defined following (5).

In the second step the fractions for replicates for donors are modified. Let the new fractional weight in replicate k for the value donated by k to j be

$$w_{kj}^{*(k)} = w_{kj}^* \left(1 - b_k\right), \tag{10}$$

where b_k is to be determined. Let t be one of the other $M-1$ donors, other than k, that donate to j. Then, the new fractional weight for donor t is

$$w_{tj}^{*(k)} = w_{tj}^* + (M-1)^{-1} b_k w_{kj}^*. \tag{11}$$

For $M = 2$ with $w_{kj}^* = w_{tj}^* = 0.5$, $w_{kj}^{*(k)} = 0.5\,(1 - b_k)$ and $w_{tj}^{*(k)} = 0.5\,(1 + b_k)$.

For any choice of b_k, condition (7) is satisfied. The variance estimator will

be unbiased if b_k satisfies

$$
\begin{aligned}
(12) \qquad & c_k \left(\alpha_{k1}^{(k)} - \alpha_k - b_k \sum_{j \in A_M} w_j^{(k)} w_{kj}^* \right)^2 - c_k \left(\alpha_{k1}^{(k)} - \alpha_k \right)^2 \\
& + \sum_{t \in D_{Rk}} c_k \left[\alpha_{t1}^{(k)} - \alpha_t + b_k \left(M - 1 \right)^{-1} \sum_{j \in A_M} w_j^{(k)} w_{kj}^* d_{tj} \right]^2 \\
& - \sum_{t \in D_{Rk}} c_k \left(\alpha_{t1}^{(k)} - \alpha_t \right)^2 = \alpha_k^2 - \alpha_k - \phi_k,
\end{aligned}
$$

where D_{Rk} is defined following (9). The difference $\alpha_k^2 - \alpha_k - \phi_k$ is the difference between the desired sum of squares for observation k and the sum of squares for the naive estimator. Under the assumption of a common variance in a neighborhood and the assumption that the variance estimator $\hat{V}(\hat{\theta})$ of (4) is unbiased for the full sample, the resulting variance estimator with $w_{ij}^{*(k)}$ defined by (10)-(12) is unbiased for the imputed sample.

4. Extension. The proposed method in Section 3 was described under the situation where the jackknife replicates are formed by deleting a single element. In practice, grouped jackknife is commonly used where the jackknife replicates are often created by deleting a group of elements. The group can be the primary sampling units (PSU) or, as in the Census long form case, groups are formed to reduce the number of replicates. In the discussion we use the term PSU to denote the group. To extend the proposed method, assume that we have a sample composed of PSUs and let PSU k be deleted to form a replicate. Let $\mathcal{P}_k$ be the indices of the set of donors in PSU k that donate to a recipient in a different PSU. For fractional imputation of size M, let the fractional replication weight in replicate k for the value donated by element i in PSU k to j be

$$
(13) \qquad w_{ij}^{*(k)} = w_{ij}^* \left(1 - b_k \right) \qquad \text{if } i \in \mathcal{P}_k \text{ and } M \neq M_{jk},
$$

where b_k is to be determined and $M_{jk} = \sum_{i \in \mathcal{P}_k} d_{ij}$ is the number of donors to recipient j that are in PSU k. Note that (13) is a generalization of (10). The corresponding replication fraction for a donor to a recipient j, where the donor is not in PSU k, is

$$
w_{tj}^{*(k)} = w_{tj}^* \left(1 + \Delta_{jk} b_k d_{ij} \right) \qquad \text{for } t \in \mathcal{P}_k^c \text{ and } i \in \mathcal{P}_k,
$$

where

$$
\Delta_{jk} = \frac{\sum_{i \in \mathcal{P}_k} w_{ij}^*}{\sum_{i \in \mathcal{P}_k^c} w_{ij}^*}.
$$

The determining equation for b_k is

$$\sum_{i\in\mathcal{P}_k} c_k \left\{ \left(\alpha_{i1}^{(k)} - \alpha_i - b_k \sum_{j\in A_M} w_j^{(k)} w_{ij}^* \right)^2 - \left(\alpha_{i1}^{(k)} - \alpha_i \right)^2 \right\}$$
$$+ \sum_{i\in\mathcal{P}_k} \sum_{t\in\mathcal{P}_k^c} c_k \left[\left\{ \alpha_{t1}^{(k)} - \alpha_t + b_k \sum_{j\in A_M} w_j^{(k)} d_{ij} \Delta_{jk} w_{tj}^* \right\}^2 - \left(\alpha_{t1}^{(k)} - \alpha_t \right)^2 \right]$$
$$= \sum_{i\in\mathcal{P}_k} \left\{ \alpha_i^2 - \alpha_i - \phi_i \right\}$$

which generalizes (12). Here, we assume common variances for the units in the same PSU.

We extend the fractional nearest neighbor imputation to the case of M_1 fractions for point estimation and $M_2 (> M_1)$ fractions for variance estimation. The motivation for this extension is the application to the Census long form where the official estimates are based on a single imputed value. A second imputed value was generated to be used only in variance estimation. Let d_{1ij} and d_{2ij} be the donor-recipient relationship indicator function used for point estimation and for variance estimation, respectively. Also, let w_{1ij}^* and w_{2ij}^* be the fractional weights of recipient j from donor i that are computed from d_{1ij} and d_{2ij}, respectively. For missing unit j, one common choice is $w_{1ij}^* = d_{1ij} M_1^{-1}$ and $w_{2ij}^* = d_{2ij} M_2^{-1}$. Of particular interest is the case where $M_1 = 1$ and $M_2 = 2$.

If $M_1 \neq M_2$, the variance estimator is defined by

$$\hat{V}\left(\hat{\theta}_I\right) = \sum_{k=1}^{L} c_k \left(\hat{\theta}_I^{(k)} - \hat{\theta}_I \right)^2, \tag{14}$$

where

$$\left(\hat{\theta}_I^{(k)}, \hat{\theta}_I \right) = \left(\sum_{i\in A_R} \alpha_{i2}^{(k)} y_i, \sum_{i\in A_R} \alpha_{i1} y_i \right)$$

with $\alpha_{i2}^{(k)} = \sum_j w_j^{(k)} w_{2ij}^{*(k)}$ and $\alpha_{i1} = \sum_j w_j w_{1ij}^*$. Here, $w_{2ij}^{*(k)}$ is the replicated fractional weight of unit j assigned to donor i in the k-th replication. Note that $\hat{\theta}_I$ is based on the point estimation weights and $\alpha_{i2}^{(k)}$ is based on the variance estimation weights. If $w_{2ij}^{*(k)}$ satisfy (7), the bias of the variance estimator (14) is

$$Bias\left\{\hat{V}\right\} = E\left\{ \sum_{i\in A_R} \left[\sum_{k=1}^{L} c_k \left(\alpha_{i2}^{(k)} - \alpha_{i1} \right)^2 - \left(\alpha_{i1}^2 - \alpha_{i1} \right) \right] \sigma_i^2 \right\}.$$

Thus, condition (9) for the unbiasedness of the variance estimator is changed to

(15)
$$\sum_{k=1}^{L} c_k \left\{ \left(\alpha_{i2}^{(k)} - \alpha_{i1}\right)^2 + \sum_{t \in D_{Ri}} \left(\alpha_{t2}^{(k)} - \alpha_{t1}\right)^2 \right\} = \alpha_{i1}^2 - \alpha_{i1} + \sum_{t \in D_{Ri}} \left(\alpha_{t1}^2 - \alpha_{t1}\right).$$

To create the replicated fractional weights satisfying (7) and (15), the sum of squares of the naive replication weights is first computed,

$$\sum_{k=1}^{L} c_k \left(\alpha_{i1}^{(k)} - \alpha_{i1}\right)^2 = \phi_{i1}, \qquad i \in A_R,$$

where $\alpha_{i1}^{(k)} = \sum_{j \in A} w_j^{(k)} w_{1ij}^*$. In the second step the fractions for replicates for donors in the point estimation are modified. Let the new fractional weight in replicate k for the value donated by $i \in \mathcal{P}_k$ to j be

$$w_{2ij}^{*(k)} = w_{1ij}^* \left(1 - b_k\right), \quad \text{if } i \in \mathcal{P}_k \text{ and } M_2 \neq M_{2jk},$$

where b_k is to be determined and $M_{2jk} = \sum_{i \in \mathcal{P}_k} d_{2ij}$. Now, $M_2 (> M_1)$ donors are identified for variance estimation. The new fractional weight for the other $M_2 - 1$ donors to recipient j, denoted by t, is

(16)
$$w_{2tj}^{*(k)} = w_{1tj}^* + \Delta_{jk} b_k d_{1ij} w_{2tj}^* \quad \text{for } t \in \mathcal{P}_k^c \text{ and } i \in \mathcal{P}_k,$$

where

$$\Delta_{jk} = \frac{\sum_{i \in \mathcal{P}_k} w_{1ij}^*}{\sum_{i \in \mathcal{P}_k^c} w_{2ij}^*}.$$

Then the b_k that gives the correct sum of squares is the solution to the quadratic equation

$$\begin{aligned}
&\sum_{i \in \mathcal{P}_k} c_k \{(\alpha_{i1}^{(k)} - \alpha_{i1} - b_k \sum_{j \in A_M} w_j^{(k)} w_{1ij}^*)^2 - (\alpha_{i1}^{(k)} - \alpha_{i1})^2\} \\
&+ \sum_{i \in \mathcal{P}_k} \sum_{t \in \mathcal{P}_k^c} c_k [\{\alpha_{t1}^{(k)} - \alpha_{t1} + b_k \sum_{j \in A_M} w_j^{(k)} \Delta_{jk} d_{1ij} w_{2tj}^*\}^2 - (\alpha_{t1}^{(k)} - \alpha_{t1})^2] \\
&= \sum_{i \in \mathcal{P}_k} (\alpha_{1i}^2 - \alpha_{1i} - \phi_{1i}).
\end{aligned}$$

If $M_1 = 1$, the adjustment in the replication fractional weights can be made at the individual level. Let the new fractional weight in replicate k for the value donated by $i \in \mathcal{P}_k$ to j, $j \in \mathcal{P}_k^c$, be

$$w_{2ij}^{*(k)} = w_{1ij}^* \left(1 - b_i\right), \quad \text{if } i \in \mathcal{P}_k \text{ and } M_2 \neq M_{2jk},$$

where b_i is to be determined. The new fractional weight for each of the other $M_2 - 1$ donors to recipient j, denoted by t, is

$$w_{2tj}^{*(k)} = w_{1tj}^{*} + \Delta_{jk} b_i d_{1ij} w_{2tj}^{*} \quad \text{for } t \in \mathcal{P}_k^c \text{ and } i \in \mathcal{P}_k,$$

where Δ_{jk} is defined following (16). Then the b_i that gives the correct sum of squares is the solution to the quadratic equation

$$\begin{aligned} & c_k \left\{ \left(\alpha_{i1}^{(k)} - \alpha_{i1} - b_i \sum_{j \in A_M} w_j^{(k)} w_{1ij}^{*} \right)^2 - \left(\alpha_{i1}^{(k)} - \alpha_{i1} \right)^2 \right\} \\ & + \sum_{t \in \mathcal{P}_k^c} c_k \left[\left\{ \alpha_{t1}^{(k)} - \alpha_{t1} + b_i \sum_{j \in A_M} w_j^{(k)} \Delta_{jk} d_{1ij} w_{2tj}^{*} \right\}^2 - \left(\alpha_{t1}^{(k)} - \alpha_{t1} \right)^2 \right] \\ & = \alpha_{1i}^2 - \alpha_{1i} - \phi_{1i}. \end{aligned}$$

5. Application to U.S. Census long form data.

5.1. *Introduction.* We use long form data from the states of Delaware and Michigan to provide examples of the variance estimation methods. Table 1 shows the individual income items and their state level imputation rates for Delaware and Michigan.

TABLE 1

Imputation rate and the person-level average income for each income item (age $\geq$ 15) for two states, Delaware ($n = 87,280$) and Michigan ($n = 1,412,339$).

Income Item	Delaware		Michigan	
	Imputation Rate (%)	Average Income	Imputation Rate (%)	Average Income
Wage	20	21,892	21	20,438
Self Employment	10	1,286	10	1,234
Interest	22	1,989	22	1,569
Social Security	20	1,768	20	1,672
Supplemental Security	20	125	20	148
Public Assistance	19	38	19	47
Retirement	20	2,018	20	1,664
Other	19	543	19	529
Total	31	29,659	31	27,301

The sampling design for the Census 2000 long form used stratified systematic sampling of households, with four strata in each state. Sampling rates varied from 1 in 2 for very small counties and small places to 1 in 8 for

very populous areas. Details of the long form sampling design can be found in Hefter (1999).

The weighting procedure for the Census 2000 long form was performed separately for person estimates and for housing unit estimates. For the income and poverty estimates considered here, the person weights are needed.

The census long form person weights are created in two steps. In the first step, the initial weights are computed as the ratio of the population size (obtained from the 100% population counts) to the sample size in each cell of a cross-classification of final weighting areas (FWAs) by person types (Housing unit person, Service Based Enumeration (SBE) person, other Group Quarters (GQ) person). Thus, the initial weights take the form of post-stratification weights. The second step in the weighting is raking where, for person weights, there are four dimensions in the raking. The dimensions are household type and size (21 categories), sampling type (3 categories), householder classification (2 categories), and Hispanic origin/race/sex/age (312 categories). Therefore, the total number of possible cells is 39,312, although many cells in a FWA will be empty. The raking procedure is performed within each FWA. There are about 60,000 FWAs in the whole country and the FWAs are nested within counties. Details of the long form weighting procedure can be found in Hefter (2002a).

5.2. *Computational Details.* The variance estimation methodology is based on the grouped jackknife, where the method described in Section 3 is used to estimate the variance due to imputation. We summarize the main steps of variance estimation and then discuss the steps in more detail:

[Step 1] Create groups and then define initial replication weights for the grouped jackknife method. The elements within a stratum are systematically divided into groups. A replicate is created by deleting a group.

[Step 2] Using the initial replication weights, repeat the weighting procedure to compute the final weights for each replicate.

[Step 3] Using fractional weighting, modify the replicate weights to account for the imputation effect on the variance. In the process, a replicate imputed total income variable is created for each person with missing data.

[Step 4] Using the replicate total income variables, compute the jackknife variance estimates for parameters such as the number of poor people by age group and the median household income.

In Step 1, the sample households in a final weighting area are sorted by their identification numbers, called MAFIDs. Let n be the sample number

of households in a final weighting area. The first $n/50$ sample households are assigned to variance stratum 1, the next $n/50$ sample households are assigned to variance stratum 2, and so on, to create 50 variance strata. Within each variance stratum, the sample households are further grouped into two groups by a systematic sample of households arranged in a half-ascending-half-descending order based on the MAFID. Using the two groups in each of the 50 strata, $L = 100$ replication factors are assigned to each unit in the sample. For unit i in variance stratum h, $(h = 1, 2, \cdots, 50)$, the replication factor for the replicate formed by deleting group k in variance stratum h is

$$F_i^{(hk)} = \begin{cases} 1 & \text{if unit } i \text{ does not belong to variance stratum } h \\ 2 - \delta_i & \text{if unit } i \text{ belongs to variance stratum } h \text{ and } i \notin \mathcal{P}_{hk} \\ \delta_i & \text{if unit } i \in \mathcal{P}_{hk}, \end{cases}$$

where $\delta_i = 1 - \{(1 - 1/IW_i)\, 0.5\}^{1/2}$, IW_i is the initial weight of unit i, and $\mathcal{P}_{hk}$ is the set of sample indices in group k in variance stratum h. With this replication factor, c_k of (4) is one.

In Step 2, the Step 1 replication weights are modified using the production raking operation. The weighting procedure consists of two parts. The first part is a poststratification in each final weighting area and the second part is raking ratio estimation using the short form population totals as controls. If the raking was carried to convergence the estimated variance for controls would be zero. In the actual operation, the replicated final weights produce very small variance estimates for the estimates of the population controls.

In Step 3, a second nearest neighbor is identified for each nonrespondent for each income item. There are eight income items – see Table 1 given earlier. A fractional weight of one is assigned to the imputed value from the first donor and a fractional weight of zero is assigned to the imputed value from the second donor for production estimation. The fractional weights are changed for the replicate, when the jackknife group containing the first donor is deleted. The amount of change is determined so that conditions (7) and (9) are satisfied. Replicate fractional weights are constructed separately for each income item.

Once the replicated fractional weights are computed, replicates of the person-level total income are constructed. Let Y_{tis} be the s-th income item for person i in family t and let R_{tis} be the response indicator function for Y_{tis}. For the k-th replicate, the replicated total income for person i in family t is

$$TINC_{ti}^{(k)} = \sum_{s=1}^{8} \left\{ R_{tis} Y_{tis} + (1 - R_{tis})\, Y_{tis}^{*(k)} \right\} \tag{17}$$

where $Y_{tis}^{*(k)}$ is the k-th replicate of the imputed value for Y_{tis}, defined by

$$Y_{tis}^{*(k)} = w_{tisa}^{*(k)} Y_{tisa}^{*} + w_{tisb}^{*(k)} Y_{tisb}^{*},$$

$\left(w_{tisa}^{*(k)}, w_{tisb}^{*(k)}\right)$ is the vector of the two k-th replicate fractional weights, one for the first donor and one for the second donor, for the s-th income item, and $(Y_{tisa}^{*}, Y_{tisb}^{*})$ is the vector of the imputed values of Y_{tis} from the first and second donor, respectively. The k-th replicate of total family income for family t is

$$TINC_t^{(k)} = \sum_{i=1}^{m_t} TINC_{ti}^{(k)}, \tag{18}$$

where m_t is the number of people in family t and $TINC_{ti}^{(k)}$ is defined in (17).

For the age group poverty estimates, a poverty status indicator function is defined for the family, and applies to all family members. That is, all family members are either in poverty or all are not in poverty. The poverty status indicator for family t is defined as

$$\zeta_t = \begin{cases} 1 & \text{if } TINC_t < c_t \\ 0 & \text{if } TINC_t \geq c_t, \end{cases}$$

where, as with the replicates in (17),

$$TINC_t = \sum_{i=1}^{m_t} \sum_{s=1}^{8} \left\{ R_{tis} Y_{tis} + (1 - R_{tis}) Y_{tisa}^{*} \right\}$$

is the total income of family t, where Y_{tisa}^{*} is the imputed value for Y_{tis} using the first nearest donor, and c_t is the poverty threshold value for family t. The threshold is a function of the number of related children under 18 years of age, the size of the family unit, and the age of the householder. (Poverty thresholds for all recent years are available on the Census Bureau web site at http://www.census.gov/hhes/www/poverty/threshld.html.)

To compute the replicate of ζ_t, we use the following procedure.

1. For person i in family t, compute two total incomes, $TINC_{tia}$ and $TINC_{tib}$, by

$$\begin{aligned} TINC_{tia} &= \sum_{s=1}^{8} \left\{ R_{tis} Y_{tis} + (1 - R_{tis}) Y_{tisa}^{*} \right\} \\ TINC_{tib} &= \sum_{s=1}^{8} \left\{ R_{tis} Y_{tis} + (1 - R_{tis}) Y_{tisb}^{*} \right\}. \end{aligned}$$

Also, compute the two total family incomes

$$(TINC_{ta}, TINC_{tb}) = \sum_{i=1}^{m_t} (TINC_{tia}, TINC_{tib}) .$$

Using the replicated total family income $TINC_t^{(k)}$ defined in (18), define

$$\alpha_t^{(k)} = \begin{cases} \frac{TINC_t^{(k)} - TINC_{tb}}{TINC_{ta} - TINC_{tb}} & \text{if } TINC_{ta} \neq TINC_{tb} \\ 1 & \text{otherwise.} \end{cases} \tag{19}$$

The $\alpha_t^{(k)}$ is the weight satisfying

$$TINC_t^{(k)} = \alpha_t^{(k)} TINC_{ta} + \left(1 - \alpha_t^{(k)}\right) TINC_{tb}.$$

2. The replicated poverty status variable is now computed by

$$\zeta_t^{(k)} = \alpha_t^{(k)} POV_{ta} + \left(1 - \alpha_t^{(k)}\right) POV_{tb} \tag{20}$$

where POV_{ta} is computed by

$$POV_{ta} = \begin{cases} 1 & \text{if } TINC_{ta} < c_t \\ 0 & \text{if } TINC_{ta} \geq c_t \end{cases}$$

and POV_{tb} is computed similarly using $TINC_{tib}$.

The replication adjustment $\alpha_t^{(k)}$ is computed from family-level total income and is applied in (20) to get a replicated poverty estimate.

The estimated variance for the estimated total number of people in poverty is

$$\hat{V}_p = \sum_{k=1}^{L} \left(\hat{\theta}_p^{(k)} - \hat{\theta}_p^{(\cdot)}\right)^2, \tag{21}$$

where L is the number of replications (here $L = 100$),

$$\hat{\theta}_p^{(k)} = \sum_{t=1}^{n} \sum_{i=1}^{m_t} w_{tj}^{(k)} \zeta_t^{(k)},$$

$$\hat{\theta}_p^{(\cdot)} = \frac{1}{L} \sum_{k=1}^{L} \hat{\theta}_p^{(k)},$$

$\zeta_t^{(k)}$ is defined in (20), and $w_{ti}^{(k)}$ is the person level replication weight after the raking operation.

The number of people in poverty in a given age group can be estimated by

$$\hat{\theta}_{pz} = \sum_{t=1}^{n} \sum_{i=1}^{m_t} w_{ti} z_{ti}\, \zeta_t,$$

where $z_{ti} = 1$ if the person i in family t belongs to the age group and $z_{ti} = 0$ otherwise. The k-th replicate of the estimate is

$$\hat{\theta}_{pz}^{(k)} = \sum_{t=1}^{n} \sum_{i=1}^{m_t} w_{ti}^{(k)} z_{ti} \zeta_t^{(k)}$$

and the variance is estimated by (21) using $\hat{\theta}_{pz}^{(k)}$ defined above.

The variance estimation for median household income estimates is based on the test-inversion methodology described in Francisco and Fuller (1991). Also, see Woodruff (1952). Let MED be the estimated median household income defined by $MED = \hat{F}^{-1}(0.5)$ where $\hat{F}(\cdot)$ is the estimated cumulative distribution function of total income of the household,

$$\hat{F}(u) = \left(\sum_{t=1}^{n} w_{tt} \right)^{-1} \sum_{t=1}^{n} w_{tt} I\left(TINC_t \leq u\right),$$

w_{tt} is the householder's person weight in household t, and $TINC_t$ is the total income of household t. (Note that households differ from families. The former includes all persons living in a given housing unit; the latter includes only related persons living in a housing unit.)

To apply the test-inversion method, first create the replicated indicator variable

$$INV_t^{(k)} = \alpha_t^{(k)} INV_{ta} + \left(1 - \alpha_t^{(k)}\right) INV_{tb},$$

where $\alpha_t^{(k)}$ is defined in (19) and

$$INV_{ta} = \begin{cases} 1 & \text{if } \sum_{i=1}^{m_t} TINC_{tia} < MED \\ 0 & \text{if } \sum_{i=1}^{m_t} TINC_{tia} \geq MED \end{cases}$$

and INV_{tb} is computed similarly, using $TINC_{tib}$ instead of $TINC_{tia}$ in the above expressions.

The estimated variance of the estimated proportion $\hat{F}(MED) = 0.5$ is computed by applying the variance formula (21) using $INV_t^{(k)}$ instead of $\zeta_t^{(k)}$ to get $\hat{V}_{inv}$. Define

$$(\hat{p}_1, \hat{p}_2) = \left(0.5 - 2\sqrt{\hat{V}_{inv}},\ 0.5 + 2\sqrt{\hat{V}_{inv}}\right)$$

to be an approximate 95% confidence interval for the estimated proportion $\hat{F}(MED) = 0.5$. The estimated variance of the estimated median is

$$\hat{V}_{med} = \left\{ \hat{F}^{-1}(\hat{p}_2) - \hat{F}^{-1}(\hat{p}_1) \right\}^2 / 16.$$

5.3. *Numerical results.* Variance estimates for the long form income and poverty estimates that have been used by SAIPE were computed for all 50 states of the U.S. (plus DC) and their counties. The estimates considered here are the total number of people in poverty, the number of children under age 5 in poverty (state level only), the number of related children age 5 to 17 in families in poverty, the number of children under age 18 in poverty, and the median household income.

TABLE 2
Variance estimation results for Delaware and Michigan

Parameter	Method	Delaware		Michigan	
		Est. SE	Std. SE	Est. SE	Std. SE
θ_1	Naive	870	100	3,217	100
(Total in poverty)	Imputation	1,161	133	4,096	127
θ_2	Naive	221	100	776	100
(0-4 in poverty)	Imputation	260	118	897	116
θ_3	Naive	366	100	1,314	100
(5-17 related in poverty)	Imputation	467	128	1,640	125
θ_4	Naive	458	100	1,608	100
(0-17 in poverty)	Imputation	592	129	2,062	128
Median	Naive	177	100	70	100
HH income	Imputation	207	117	85	121

Table 2 contains variance estimation results (the estimated standard deviations) for the income and poverty statistics for the states of Delaware and Michigan. The variance estimator labeled "naive" treats the imputed values as observed values. The "imputation" variance estimator is that of Section 3 and reflects the imputation effects. Both variance estimators account for the raking in the estimator. Because Michigan is much larger than Delaware its estimated numbers of persons in poverty (not shown) are much larger, and thus, due to the scale effects, so are the corresponding standard errors. The standardized standard errors in the table are computed by dividing the estimated standard error computed by the "imputation" procedure by the estimated standard error computed by the "naive" procedure.

Generally speaking, imputation increases the variance so the naive variance estimator underestimates the true variance. The relative increase is similar for Michigan and Delaware. A result worth noting is that the increase in variance due to imputation is higher for the poverty parameters

TABLE 3
Imputation rates by income level (age ≥ 15)

Total Income	Imputation Rate (%)	
	Delaware	Michigan
0 - 9,999	34	34
10,000 - 19,999	36	35
20,000 - 49,999	28	29
50,000 - 69,999	25	25
70,000 and over	25	25

than for the income parameters. This is because in both states the imputation rate is higher for persons with low imputed income. (See Table 3.)

Table 4 contains some numerical results for the estimated standard errors for the county estimates in Delaware. The age groups in the table are those used by SAIPE at the county level, which are fewer than the age groups used by SAIPE at the state level. As with state estimates, imputation increases the variance. However, the effect of imputation is much smaller for county estimates than for state estimates. County level estimation is an example of domain estimation, where the values used for imputation can come from donors outside the domain. Donors from outside the domain contribute less to the imputation variance of the domain total than donors in the domain because the imputed value from outside the domain is uncorrelated with the values observed in the domain. In effect, imputations from outside the domain increase the sample size on which the estimates are based, whereas imputations from inside the domain change the weights given to the observations in the estimates. Because the proportions of outside donors differ across counties, the effect of imputation on county variances is not uniform across counties. In Delaware, the overall imputation rates for total income (the percent of records with at least one income item imputed) are 30.7 %, 29.5%, and 34.5 % for county 1, county 3, and county 5, respectively. Table 5 presents the distribution of donors for wage income in Delaware. In county 1, about 59% of the donors are from outside the county, whereas in county 3, only about 25% of the donors are from outside the county. Thus, the variance inflation due to imputation, as reflected by the standardized standard error, is greater for county 3 than for county 1.

Acknowledgements. The research was supported by a contract with the U.S. Census Bureau. We thank George McLaughlin and George Train for computational support and Yves Thibaudeau for discussion on the long form imputation methods. The online supplemental document and some program codes are available at the website via “http://jkim.public.iastate.edu/nni.html”.

TABLE 4
County variance estimates for Delaware

County	Parameter	Method	Est. SE	Std. SE
001	θ_1 (Total poor)	Naive	409	100
		Imputation	444	109
	θ_3 (5-17 related poor)	Naive	183	100
		Imputation	203	111
	θ_4 (0-17 poor)	Naive	219	100
		Imputation	241	110
	Median HH income	Naive	323	100
		Imputation	336	104
003	θ_1 (Total poor)	Naive	687	100
		Imputation	838	122
	θ_3 (5-17 related poor)	Naive	317	100
		Imputation	351	111
	θ_4 (0-17 poor)	Naive	365	100
		Imputation	417	114
	Median HH income	Naive	200	100
		Imputation	226	113
005	θ_1 (Total poor)	Naive	518	100
		Imputation	608	117
	θ_3 (5-17 related poor)	Naive	197	100
		Imputation	217	110
	θ_4 (0-17 poor)	Naive	270	100
		Imputation	300	111
	Median HH income	Naive	361	100
		Imputation	389	108

References.

[1] CHEN, J. and SHAO, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*. **16**, 113–132.

[2] CHEN, J. and SHAO, J.(2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of American Statistical Association*. **96**, 260–269.

[3] FAY, R.E. (1999). Theory and application of nearest neighbor imputation in Census 2000. *Proceedings of the Section on Survey Research Methods* (pp. 112-121). Alexandria, VA: American Statistical Association.

[4] FRANCISCO, C.A. and FULLER, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.

[5] HEFTER, S. (1999). Long form sampling specifications for Census 2000. *DSSD Census 2000 Procedures and Operations Memorandum Series LL-5*.

[6] HEFTER, S. (2002a). Long form weighting specifications for Census 2000. *DSSD Census 2000 Procedures and Operations Memorandum Series LL-10*.

[7] HEFTER, S. (2002b). Requested output for long form weighting review: Census 2000. *DSSD Census 2000 Procedures and Operations Memorandum Series LL-12*.

[8] KALTON, G. and KISH, L. (1984). Some efficient random imputation methods. *Communications in Statistics A*, **13**, 1919-1939.

[8] Kim, J. and Fuller, W.A. (2004). Inference procedures for hot deck imputation.

TABLE 5
Donor distribution for wage income in Delaware (age ≥ 15)

County	Number of donors from county 1	Number of donors from county 3	Number of donors from county 5
1	1,271	1,512	325
($n = 15,735$)	(41%)	(49%)	(10%)
3	1,142	7,374	1,343
($n = 51,869$)	(11%)	(75%)	(14%)
5	847	1,137	2,045
($n = 19,661$)	(21%)	(28%)	(51%)

Biometrika. 91, 559-578.

[9] RANCOURT, E., SÄRNDAL, C.E., and LEE, H. (1994). Estimation of the variance in the presence of nearest neighbor imputation. *Proceedings of the Section on Survey Research Methods* (pp. 888-893). Alexandria, VA: American Statistical Association.

[10] RANCOURT, E. (1999). Estimation with nearest neighbor imputation at Statistics Canada. *Proceedings of the Section on Survey Research Methods* (pp. 131-138). Alexandria, VA: American Statistical Association.

[11] RAO, J. N. K. and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811-822.

[12] RUBIN, D. B. (1976). Inference and missing data. *Biometrika,* **63**, 581-590.

[13] RUBIN, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

[14] SANDE, I. (1983). Hot-deck imputation procedures, in *Incomplete Data in Sample Surveys, Vol 2.* (pp. 339-349). Academic Press, New York.

[15] SÄRNDAL, C.E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18**, 241-252.

[16] SCHNEIDER, P. J. (2004). Content and Data Quality in Census 2000, *Census 2000 Testing, Experimentation, and Evaluation Program Topic Report* No. 12, TR-12, U.S. Census Bureau, available at http://www.census.gov/pred/www/rpts/TR12.pdf.

[17] SHAO, J. (2009). Nonparametric variance estimation for nearest neighbor imputation. *Journal of Official Statistics* **25**, 55-62.

[18] SHAO, J. and STEEL, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association* **94**, 254-265.

[19] SHAO, J. and WANG, H. (2008). Confidence intervals based on survey data with nearest neighbor imputation. *Statistica Sinica* **18**, 281-297.

[20] WOODRUFF, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, **47**, 635-646.

IOWA STATE UNIVERSITY, AMES, IA 50011, U.S.A.
E-MAIL: jkim@iastate.edu; waf@iastate.edu

U.S. BUREAU OF CENSUS, WASHINGTON DC, 20233, USA,
E-MAIL: william.r.bell@census.gov