A Case Study in Dual-Frame Estimation Methods

Marianne Winglee[1], Inho Park[2], Keith Rust[1], Benmei Liu[1], Gary Shapiro[1],
Marianne Winglee, Westat, 1650 Research Boulevard, Rockville, MD 20850[1]
Inho Park, Bank of Korea, Namdaemun-Ro 106, Jung-Gu, Seoul, Korea[2]

Keywords: Dual-frame design, rare population, domain estimation, sampling weights, single frame (SF) estimator, pseudo maximum-likelihood (PML) estimator

1. Introduction

This paper compares five methods to estimate a population total in a dual-frame survey when one of the two samples is not self-weighting. The estimation methods compared are: the pseudo maximum-likelihood (PML) method (Skinner and Rao, 1996), two classic single frame (SF) methods (Kalton and Anderson, 1986; and Bankier, 1986), a pseudo single-method frame method, and a modified single frame method developed for the Third National Incidence Study of Child Abuse and Neglect (NIS-3, 1997).

Section 2 describes the NIS and the reasons for adopting a modified estimation method. Section 3 defines the estimation domains in a dual-frame survey. Section 4 reviews the PML and the classic SF methods. Section 5 discusses the estimation issues in the NIS: the difficulties with the classic methods (5.1), inefficiencies of a pseudo SF method (5.2), and the motivation for a modified SF method. Section 6 describes a simulation study. The results suggest that the modified SF method performed reasonably well for the NIS.

2. The National Incidence Study of Child Abuse and Neglect (NIS)

The NIS is a national survey to estimate the number and the characteristics of maltreated children in the United States. This study uses a multistage and multiple frame design to broaden coverage of possible reporting sources for maltreated children. The primary sampling units (PSUs) are counties and county clusters. Within sampled PSUs, Child Protective Services (CPS) agencies are the primary data source for maltreated children. However, the coverage of CPS agencies is incomplete because some possibly maltreated children may not be investigated by the CPS agencies.

This paper presents the NIS as a simple one stage dual-frame survey inside a single PSU. Frame *A* is simply stated as a list frame of maltreated children investigated by CPS agencies and a self-weighting sample is selected from this frame. Frame *B* is a second frame of maltreated children, those observed by professionals in non-CPS agencies as possibly maltreated children. The NIS constructs list frames of agencies for police, hospitals, schools, shelters, day cares, and other agencies for a total of 10 agency categories. Agencies were sampled, a roster of professional staff constructed, and then staffs were sampled to serve as informants (sentinels) for maltreated children. There is not a complete list frame *B* and the number of maltreated children in frame *B* is unknown.

The estimation issues in the NIS are as follows: (1) the intersection domain is not fully defined, (2) the sample in frame *B* is not self weighting, and (3) the NIS study design requires eliminating overlapping observations such that each maltreated child will be counted once for incidence estimation.

3. Estimation Domains in Dual-frame Surveys

The basic assumption in dual-frame estimation is that the union of the frames covers the population of interest. With two frames, there are three estimation domains: those units common to both frames; those unique to one frame; and those unique to the second frame. The key to unbiased estimation is that one can correctly identify the domain membership for each sample observation and account for the selection probabilities of every member in both frames (not only the frame from which the observation is sampled). Lohr and Rao (2000, 2006) describe methods of inference from dual-frame surveys and estimation methods in multiple-frame surveys.

Consider a dual-frame survey where the population sizes of frames *A* and *B* are known and both frames are incomplete. Let U_A and U_B denote the two frames *A* and *B* with population size N_A and N_B, and U_{ab} denote the frame intersection with size N_{ab}, then the frames can be expressed as the union of two distinct sets $U_A = U_a \cup U_{ab}$ and $U_B = U_b \cup U_{ab}$ where $N_A = N_a + N_{ab}$ and $N_B = N_a + N_{ab}$. (Note that $U_a = U_A \cap U_{ab}^c$, $U_b = = U_B \cap U_{ab}^c$, where U_{ab}^c is the complement of U_{ab}).

The samples S_A and S_B are selected independently from frames *A* and *B* with sample sizes n_A and n_B. Using S_{ab} to denote the sample selected from the frame intersection with size n_{ab}; the samples can again be described as two distinct sets $S_A = S_a \cup S_{ab}$ and $S_B = S_b \cup S_{ab}$ where $S_a = S_A \cap S_{ab}^c$ and $S_b = S_B \cap S_{ab}^c$ with sample sizes $n_A = n_a + n_{ab}$ and $n_B = n_a + n_{ab}$. Following the notation in Skinner and Rao (1996), let S'_{ab} denote the overlap sample from frame *A* and S''_{ab} denote the overlap sample from frame *B*, then S_A and S_B can be expressed as $S_A = S_a \cup S'_{ab}$ and $S_B = S''_{ab} \cup S_b$ with sizes $n_A = n_a + n'_{ab}$ and $n_B = n''_{ab} + n_b$.

Figure 1 is a pictorial representation of a dual-frame design with two incomplete frames. The domains are: S_1 if an observation $k \in S_a$; S_2 if $k \in S'_{ab} \cap S''^{C}_{ab}$; S_3 if $k \in S'_{ab} \cap S''_{ab}$; S_4 if $k \in S''_{ab} \cap S'^{C}_{ab}$, and S_5 if $k \in S_b$. The intersection domain comprises of $S'_{ab} = S_2 \cup S_3$, $S''_{ab} = S_3 \cup S_4$, and S_3 contains the sample observations selected from both frames *A* and *B*. The domains for estimation are: $S_a = S_1$ for units unique to frame *A*, $S_{ab} = S_2 \cup S_3 \cup S_4$ for those units common to both frames, and $S_b = S_5$ for those units unique to frame *B*. The domain sizes are $n_a = n_1$, $n_{ab} = n_2 + n_3 + n_4$, and $n_b = n_5$.

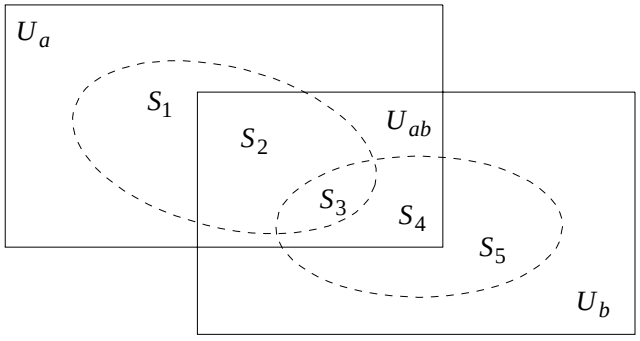

Figure 1. A Dual-Frame Design with Two Incomplete Frames

4. Dual-frame Estimation Methods

For an observation k, let π_{Ak} be the selection probability of the observation in frame *A* according to specified probability sampling design $p'(S_A)$. Likewise, let π_{Bk} be the selection probability in frame *B* with probability sample design $p''(S_B)$. The sample weights for the two samples are $w_{Ak} = \pi_{Ak}^{-1}$ and $w_{Bk} = \pi_{Bk}^{-1}$. An unbiased estimator of the population size for each frame is $\hat{N}_{A,p'} = \sum_{S_a \cup S'_{ab}} w_{Ak}$ and $\hat{N}_{B,p''} = \sum_{S''_{ab} \cup S_b} w_{Bk}$.

The issue in dual-frame estimation is how best to derive the sample weights for observations in the intersection domain such that the sample weights for observations in the three estimation domains can be used to provide unbiased estimates of the union of the two frames:

$$w_k = \begin{cases} w_{Ak} & if\ k \in S_a, \\ w_{ABk} & if\ k \in S'_{ab}, S''_{ab}, \\ w_{Bk} & if\ k \in S_b. \end{cases} \quad (1)$$

4.1 PML Estimation

The pseudo maximum-likelihood (PML) method adopts the maximum likelihood principles for estimation with simple random samples (Skinner, 1991) and applies them for samples with complex survey designs (Skinner and Rao, 1996). The PML weighting scheme for observations in the two samples can be summarized as follows:

$$\hat{w}'_k = \begin{cases} \left(\dfrac{N_A - \tilde{N}_{ab,PML}}{\tilde{N}'_a} \right) w_{Ak} & if\ k \in S_a, \\ \left(\dfrac{\tilde{N}_{ab,PML}}{\tilde{N}'_{ab}} \right) \left(\dfrac{\tilde{n}'_{ab}}{\tilde{n}'_{ab} + \tilde{n}''_{ab}} \right) w_{Ak} & if\ k \in S'_{ab}, \end{cases} \quad (2a)$$

$$\hat{w}''_k = \begin{cases} \left(\dfrac{\tilde{N}_B - \tilde{N}_{ab,PML}}{\tilde{N}''_b} \right) w_{Bk} & if\ k \in S_b, \\ \left(\dfrac{\tilde{N}_{ab,PML}}{\tilde{N}''_{ab}} \right) \left(\dfrac{\tilde{n}''_{ab}}{\tilde{n}'_{ab} + \tilde{n}''_{ab}} \right) w_{Bk} & if\ k \in S''_{ab}, \end{cases} \quad (2b)$$

where $\tilde{N}'_a = \sum_{S_a} w_{Ak}$, $\tilde{N}''_b = \sum_{S_b} w_{Bk}$, $\tilde{N}'_{ab} = \sum_{S'_{ab}} w_{Ak}$, $\tilde{N}''_{ab} = \sum_{S''_{ab}} w_{Bk}$, $\tilde{N}_B = \sum_{S_B} w_{Bk}$, $\tilde{n}'_{ab} = n_A \tilde{N}'_{ab} / N_A$, $\tilde{n}''_{ab} = n_B \tilde{N}''_{ab} / \tilde{N}_B$, $\tilde{N}_{ab,PML}$ is the smallest root of $px^2 - qx + r = 0$, with $p = n_A + n_B$, $q = n_A \tilde{N}_B + n_B N_A + n_A \tilde{N}'_{ab} + n_B \tilde{N}''_{ab}$ and $r = n_A \tilde{N}'_{ab} \tilde{N}_B + n_B \tilde{N}''_{ab} N_A$ and $\tilde{N}_{ab,PML} = (2p)^{-1} \left[q - \left(q^2 - 4pr \right)^{0.5} \right]$. For a complex sample design, the sample sizes n_A and n_B are replaced by their effective sample sizes that account for the design effects of the domain sizes. Lohr and Rao (2006) further extended the method to multi-frame situations, and suggested the use of a compromise value of the design effect that works well for the most important variables.

The PML method provides two weights $\hat{w}'_k$ and $\hat{w}''_k$ for observations in the sample overlap segment S_3 . For estimation of a population total Y , let Y_a, Y_{ab}, Y_b be the population totals in the dual-frame domains a , ab , and b (for sets U_a , U_b and U_{ab}). Let y_k be the value of observation k . Skinner and Rao (1996) showed that the PML estimator of the population total $Y = Y_a + Y_{ab} + Y_b$ can be obtained as follows:

$$\hat{Y}_{PML} = \sum_{S_a} \hat{w}'_k y_k + \sum_{S'_{ab}} \hat{w}'_k y_k + \sum_{S''_{ab}} \hat{w}''_k y_k + \sum_{S_b} \hat{w}'' y_k \quad (3)$$

4.2 Single frame Estimation (SF)

The classic single frame (SF) methods (Kalton and Anderson, 1986; Bankier, 1986) estimate the population total by treating all observations as though they had been sampled from a single frame and the sampling weights of observation in the intersection domain are modified according to their inclusion probability in each sample (see Lohr and Rao, 2000, 2006). Kalton and Anderson (1986) discussed ways to apply this approach. One option (SF_1) uses the following weighting scheme:

$$w_{k,SF_1} = \begin{cases} w_{Ak} = \pi_{Ak}^{-1} & if\ k \in S_a, \\ w_{ABk} = (\pi_{Ak} + \pi_{Bk})^{-1} & if\ k \in S'_{ab}, S''_{ab} \\ w_{Bk} = \pi_{Bk}^{-1} & if\ k \in S_b, \end{cases} \quad (4)$$

An estimate of the population total is obtained again by using all observations in both samples:

$$\hat{Y}_{SF_1} = \sum_{k \in S_a \cup S'_{ab}} w_k y_k + \sum_{k \in S_b \cup S''_{ab}} w_k y_k \quad (5)$$

An alternative application, also proposed by Bankier (1986) (SF_2), is to first eliminate overlapping observations in the sample overlap segment, and the sampling weights for distinct units in the sample are derived as follows:

$$w_{k,SF_2} = \begin{cases} w_{Ak} = \pi_{Ak}^{-1} & if\ k \in S_a, \\ w_{ABk} = \left(1 - (1 - \pi_{Ak})(1 - \pi_{Bk})\right)^{-1} & if\ k \in S_{ab}, \\ w_{Bk} = \pi_{Bk}^{-1} & if\ k \in S_b, \end{cases} \quad (6)$$

An unbiased estimate of the population total under this scheme is:

$$\hat{Y}_{SF_2} = \sum_{k \in S_a} w_k y_k + \sum_{k \in S_{ab}} w_k y_k + \sum_{k \in S_b} w_k y_k \quad (7)$$

Both implementations are comparable when $\pi_{Ak} * \pi_{Bk}$ is small and when there are relatively few overlapping observations in the sample.

5. Estimation in the NIS

The key estimates of interest in the NIS are the total number of maltreated children in the union of the two frames $\hat{N}$, the total number of maltreated children investigated by CPS agencies $\hat{N}_A$, and the number of the maltreated children not investigated by the CPS agencies $\hat{N}_b$. Another subgroup of interest is children measurable by the NIS endangerment standards ($\hat{E}$).

5.1 Difficulties in Applying the Classic Estimation Methods

The SF estimation methods are not easily applicable in the NIS because (1) frame *B* uses a non self-weighting sample, and (2) the assignment of domain membership is problematic. The single frame method SF_2 is the most applicable in the NIS. This approach can be applied in small PSUs where all maltreated children investigated by CPS agencies were sampled with certainty. In this special case, $\pi_{Ak} = 1$ where $k \in S_a \cup S_{ab}$ and the sample weights are $w_k = 1$ for all children investigated by CPS agencies. It is not necessary to know π_{Bk}. However, in all other PSUs where $\pi_{Ak} < 1$, this SF method is not possible because π_{Bk} are unknown.

The PML method can bypass this problem. However, the issues are how best to estimate the overall design effects and how to derive a composite weight for observations after overlapping observations are removed in the NIS.

5.2 Pseudo Single-Frame Estimation

A pseudo single-frame weighting scheme that can circumvent the estimation difficulties with the classic methods is the following:

$$w_{k,SF_3} = \begin{cases} w_{Ak} & if\ k \in S_1 \cup S_2 \cup S_3, \\ 0 & if\ k \in S_4, \\ w_{Bk} & if\ k \in S_5. \end{cases} \quad (8)$$

With this scheme, an estimate of the population total and the frame *A* total are:

$$\hat{Y}_{SF_3} = \sum_{k \in S_A} w_k y_k + \sum_{k \in S_b} w_k y_k \quad (9)$$
$$\hat{Y}_{A,SF_3} = \sum_{k \in S_A} w_k y_k$$

This scheme is analogous to the situation where frame overlaps are removed through a prescreening process. For example, in the National Survey of America's Families (Waksberg et al, 1997), the random digit dialing (RDD) survey covering the households that have a telephone can be viewed as frame *A*, and the area probability sample of households as frame *B*. In the area sample, households with telephones were screened out. Prescreening, however, is not possible for the NIS. Its effect, however, is approximately the same as if one ignores observations in S_4 and assigns zero weight to sampled observations in this segment. An obvious disadvantage of this approach is the loss of data.

5.3 Modified Single Frame Estimation Method

A practical weighting scheme used in the Third NIS (NIS-3, 1997) is the following:

$$w_{k,SF_4} = \begin{cases} w_{Ak} & if\ k \in S_1 \cup S_2, \\ 1 & if\ k \in S_3 \cup S_4, \\ w_{Bk} & if\ k \in S_5. \end{cases} \quad (10)$$

With this scheme, an estimate of the population total is computed in the same way as the SF_2 method. The frame *A* total is estimated somewhat differently, using all observations found in frame *A* as follows:

$$\hat{Y}_{SF_4} = \sum_{k \in S_a} w_k y_k + \sum_{k \in S_{ab}} w_k y_k + \sum_{k \in S_b} w_k y_k \text{ , and}$$
$$\hat{Y}_{A,SF_4} = \sum_{k \in S_a \cup S_{ab}} w_k y_k \ . \quad (11)$$

The rationale for this scheme is as follows. By definition, any domain estimator of the form $\hat{Y}_{d,S} = \sum_S w_k y_k d_k$ is unbiased for the domain total $Y_d = \sum_U y_k d_k$, where $U = U_A \cup U_B$

and d_k is a domain indicator, that is, $d_k = 1$ if k is in the domain and =0 otherwise.

Consider the expectation of $\hat{Y}$ taken over all possible observations realized in the survey is $E(\hat{Y}) = E\left(\sum_{k\in U} w_k y_k d_k\right)$. The estimator $\hat{Y}$ is unbiased for Y when $E\left(\sum_{k\in U} w_k y_k d_k\right) = Y$ and this condition is satisfied when $E(w_k) = 1$ for all k.

By domain, the estimator $\hat{Y}$ is unbiased for Y when

$$\begin{cases} E\left(\sum_{k\in U_a} w_k y_k\right) = \sum_{k\in U_a} y_k & , \\ E\left(\sum_{k\in U_{ab}} w_k y_k\right) = \sum_{k\in U_{ab}} y_k & , \\ E\left(\sum_{k\in U_b} w_k y_k\right) = \sum_{k\in U_b} y_k & . \end{cases} \quad (12)$$

In domain U_a, w_k is the inverse of the probability of selection of case k where

$$w_k = \begin{cases} \pi_{Ak}^{-1} & if\ k \in S_a, \\ 0 & if\ k \notin S_a. \end{cases} \quad (13)$$

When $E(w_k \mid k \in U_a) = \pi_{Ak} * \frac{1}{\pi_{Ak}} + (1-\pi_{Ak}) * 0 = 1$, the unbiased condition is satisfied. Likewise, for domain U_b, one can define $w_k = \pi_{Bk}^{-1}$ for $k \in S_b$ and 0 otherwise. Again when $E(w_k \mid k \in U_b) = \pi_{Bk} * \frac{1}{\pi_{Bk}} + (1-\pi_{Bk}) * 0 = 1$, the unbiased condition is satisfied.

For the intersection domain U_{ab}, the same is true that expectation $E\left(\sum_{k\in U_{ab}} w_k y_k\right) = \sum_{k\in U_{ab}} y_k$ is satisfied when $E(w_k \mid k \in U_{ab}) = 1$. For the sample segments within this domain, the weights are:

$$w_{2k} = \begin{cases} \pi_{Ak}^{-1} * (1-\pi_{Bk})^{-1} & if\ k \in S_2 \\ 0 & if\ k \notin S_2 \end{cases}$$

$$w_{3k} = \begin{cases} \pi_{Ak}^{-1} * \pi_{Bk}^{-1} & if\ k \in S_3 \\ 0 & if\ k \notin S_3 \end{cases} \quad (14)$$

$$w_{4k} = \begin{cases} (1-\pi_{Ak})^{-1} * \pi_{Bk}^{-1} & \text{if } k \in S_4 \\ 0 & \text{if } k \notin S_4 \end{cases}$$

Then as before,

$$E(w_k \mid k \in U_{ab}) = \pi_{Ak}(1-\pi_{Bk})w_{2k} + \pi_{Ak}\pi_{Bk}w_{3k} + (1-\pi_{Ak})\pi_{Bk}w_{4k} \quad (15)$$

In the NIS there is no reliable way to classify observations in segments S_1 and S_2. For children reported by CPS agencies, data are available on the informant source that reported the maltreated children. One can distinguish those informant sources that are surveyed independently in the NIS (e.g., police, school, hospitals, etc.) and those informant sources that are not surveyed in the NIS (e.g., neighbors, victims, etc.) and use this distinction to assign children into segments. This option, however, is imperfect because the coverage in the NIS is not always captured by the CPS data (e.g., among school personnel, the NIS coverage includes only public school personnel and not all school personnel).

When observations in segments S_1 and S_2 are inseparable, the default is that members in both segments are assigned a sample weight: $w_{2k} = \pi_{Ak}^{-1}$. Furthermore, even if it is possible to distinguish observations in S_2 and S_1, π_{Bk} is unknown for units in S_2. In this case also, it is natural to use $w_{2k} = \pi_{Ak}^{-1}$. By applying this constraint, equation (15) becomes:

$$E(w_k \mid k \in U_{ab}) = \pi_{Ak}(1-\pi_{Bk})\pi_{Ak}^{-1} + \pi_{Ak}\pi_{Bk}w_{3k} + (1-\pi_{Ak})\pi_{Bk}w_{4k}$$

and the unbiased condition $E(w_k \mid k \in U_{ab}) = 1$ means:

$$(1-\pi_{Bk}) + \pi_{Ak}\pi_{Bk}w_{3k} + (1-\pi_{Ak})\pi_{Bk}w_{4k} = 1 \quad (16)$$

There is no unique solution to this equation. To avoid loss of data in S_4, it is reasonable to impose the constraint that both w_{3k}, w_{4k} should not be smaller than 1, and this leads to the solution $w_{3k} = w_{4k} = 1$. Note that if one accepts the loss of data in S_4 and set $w_{4k} = 0$, this leads to the solution $w_{3k} = w_{Ak}$, that is the pseudo single-frame method (SF_3).

6. Simulation Study

A simulation study was conducted to compare the five estimation methods: the PML method, the two classic SF methods, a pseudo single frame method, and a modified single frame method for the NIS. The basic methodology followed Skinner and Rao (1996). Instead of a superpopulation, this study constructed two finite population frames, and then drew independent samples from each frame, repeating the sample selection 10,000 times. This section summarizes the process of frame construction (6.1), sample selection, weighting, and estimation (6.2), and the results (6.3).

6.1 Dual-frame Construction

Frame A was $\{(y_k, N_a), k = 1,..., N_A\}$ where y_k was the value associated with the kth element, N_a was the number of elements belonging to domain a, N_A was the frame A size. Frame B was $\{(y_{jk}, N_{bj}), j = 1,..., M, k = 1,..., N_j\}$ with M clusters, N_{bj} elements in domain b in the jth cluster, $N_{abj} = N_j - N_{bj}$ elements in domain ab, y_{jk} was the value associated with the kth element. The sizes were $N_B = \sum_{j=1}^{M} N_j$, $N_b = \sum_{j=1}^{M} N_{bj} = N_B - N_{ab}$ and $N_{ab} = \sum_{j=1}^{M} N_{abj}$. The clusters in frame B resembled the clusters of informant agencies and informants in the NIS.

To generate the frames, the first step was to specify γ_a and γ_b as targeted relative sizes of U_a and U_b where $\gamma_a = N_a / N$ and $\gamma_b = N_b / N$. The parameters used to generate frame B were $\{(N_j, N_{bj}, N_{abj}), j = 1,..., M\}$ where N_j was the cluster size, and within-cluster domain sizes N_{bj} and N_{abj} of domains b and ab. Frame B was created in five steps:

1. Generate a cluster size $N_j \sim Gamma(\tau_1, \tau_2)$ for specified τ_1 and τ_2. The expected size for frame B is $E(N_B) = M\tau_1\tau_2$.
2. Generate a probability $\gamma_{bj} \sim Beta(\alpha_1, \alpha_2)$ for a specified $\sigma_\gamma^2 = V(\gamma_{bj})$ where $E(\gamma_{bj}) = \gamma_b / (1 - \gamma_a)$. It follows that $\alpha_1 = \mu_\gamma \left[\sigma_\gamma^{-2} \mu_\gamma^2 (\mu_\gamma^{-1} - 1) - 1\right]$ and $\alpha_2 = \alpha_1 (\mu_\gamma^{-1} - 1)$, where $\mu_\gamma = E(\gamma_{bj})$.
3. Generate $N_{bj} \sim binomial(N_j, \gamma_{bj})$ within-cluster domain size, for given values of N_j and γ_{bj}.
4. Compute $N_{abj} = N_j - N_{bj}$, for given values of N_j and N_{bj}.
5. Repeat Steps 1-4 independently to get $\{(N_j, N_{bj}, N_{abj}), j = 1,..., M\}$.

$U_{abj} = \{j : j = 1,...N_{abj}\}$, $U_{bj} = \{j : j = N_{abj} + 1,..., N_j\}$ denote the index sets for the two domains ab and b in cluster j. Then $U_b = \bigcup_{j=1}^{M} U_{bj}$ and $U_{ab} = \bigcup_{j=1}^{M} U_{abj}$ were the index sets of the two domains for the entire frame B.

For each cluster j, the y_{jk} values were generated independently following the same nested error model used in Rao and Skinner (1996):

$$y_{jk} = \begin{cases} \mu_b + \alpha_{bj} + \varepsilon_{jk} & \text{for } k \in U_{bj}, \\ \mu_{ab} + \alpha_{abj} + \varepsilon_{jk} & \text{for } k \in U_{abj}, \end{cases}$$

where the domain means μ_b and μ_{ab} were specified, ε_{jk}'s were independent of $(\alpha_{bj}, \alpha_{abj,} N_{bj})$ with $\varepsilon_{jk} \sim iid\ N[0, \sigma^2(1-\rho)]$ for specified σ^2 and ρ, and $(\alpha_{bj}, \alpha_{abj})$ were drawn from a bivariate normal distribution with mean vector 0, common variance $\rho\sigma^2$ and covariance $\rho\delta\sigma^2$ for a specified value δ. This model allows one correlation, ρ, within domains ab and b and a different correlation, δ, across domains.

Given that $E(y_{jk} | U_{abj}) = \mu_{ab}$ and $V(y_{jk} | U_{abj}) = \sigma^2(1-\rho) + \sigma^2\rho = \sigma^2$, one can view a data value y_k in U_{ab} as being generated from the model $y_k = \mu_{ab} + \varepsilon_k$. Hence, one only needs to create the dataset for domain U_a to complete frame A. The steps involved first determining N_a from γ_a, γ_b, N_{ab} and N_B using $N_A = N_a + N_{ab}$ and $N_A / N_B \cong (1-\gamma_b)/(1-\gamma_a)$. Then, the dataset $\{y_k, k = 1,..., N_a\}$ was generated for U_a from the model $y_k = \mu_a + \varepsilon'_k$, where μ_a was specified and $\varepsilon'_k \sim iid\ N(0, \sigma^2)$.

In addition, a dataset $\{E_k : k \in U\}$ of a 0-1 variable determined from y_k as $E_k = I[y_k > E_0]$ was specified for a constant E_0. This was created as a variable to measure incidence by the NIS endangerment standard.

6.2 Sampling, Weighting, and Summary Statistics

Sampling involved drawing simple random samples of n_A units from frame A by specifying a constant sampling fraction f_A. The samples of n_B units from frame B was selected in two-stages using simple random sampling at both stages. In the first stage, m_B clusters were selected out of a total of M clusters and then n_O sample units were selected within each sampled cluster where $n_O = N_B f_B / m_B$ (i.e., the clusters have equal sample sizes but unequal selection probabilities per cluster). The design sampling weights were: $w_{Ak} = N_A / n_A$ and $w_{Bk} = (M / m) \times (N_j / n_O)$ for $k \in j$ cluster.

Sampling weights were constructed for five estimation methods. The sampling weights for the PML method were computed using equations (2a) and (2b) with the following adaptation to the NIS situation. The NIS design is one where N_A is known, N_B and N_{AB} are unknown. The sample S_A is a self-weighting sample and the sample S_B is a complex non-self-weighting sample. To simulate this design, equation 2(a) used $\tilde{w}_{Ak} = (N_A / \hat{N}_{A,p'}) * w_{Ak}$, the ratio-adjusted weight post-stratified to the known population total.

The sample size n_B in $\tilde{n}''_{ab}$, p, q, r and $\tilde{N}_{ab,PML}$ was replace by n_B^* the effective sample size where $n_B^* = n_B / deff_B$. Following Lohr and Rao (2006), the design effect $deff_B$ was set to the average design effects of the two domain sizes $\tilde{N}''_{ab}$ and $\tilde{N}_b = \tilde{N}_B - \tilde{N}''_{ab}$ coming from frame *B*. This study assumed a design effect $deff_B$ = 2.0.

The sampling weights for the SF_1 method used equation (4) and those for SF_2 method used equation (6). Note that these classic SF methods are not applicable in the NIS because members in frame *B* are selected with unequal probabilities. They are included in the simulation for comparative purposes. The sampling weights for SF_3 the pseudo single-frame method used equation (8), and the weights for SF_4 the modified SF method for the NIS used equation (10). For all these four SF related methods, the sampling weights were again ratio adjusted to the known population total for frame *A*.

Sampling and weighting were repeated for $R = 10{,}000$ times. For each dataset $S(r)$ say, at the *r*th selection (or iteration), an estimate of the total $Y = \sum_U y_k$ was

$$\hat{Y}(r) = \sum_{S(r)} \omega_k(r) y_k ,$$

where $\omega_k(r)$ denotes the sampling weight for the unit *k* in the sample. When estimating the domain total for $D \subset U$ say, $S(r)$ can be replaced by $S(r) \cap D$ in the summation above. The percent relative bias (RelBias) of the estimator $\hat{Y}$ is computed as follows:

$$\mathrm{Re}\,lBias(\%) = \frac{\bar{\hat{Y}} - Y}{Y} \text{ *100, where } \bar{\hat{Y}} = R^{-1} \sum_{r=1}^{R} \hat{Y}(r).$$

The empirical mean squared error (EMSE) and its Monte Carlo standard error were computed as

$$EMSE = \frac{1}{R} \sum_{r=1}^{R} \left[\hat{Y}(r) - Y\right]^2 \text{ and}$$

$$s(EMSE) = \sqrt{\frac{1}{R(R-1)} \sum_{r=1}^{R} \left(\hat{Z}(r) - \bar{\hat{Z}}\right)}, \text{ where}$$

$$\hat{Z}(r) = \left[\hat{Y}(r) - Y\right]^2 \text{ and } \bar{\hat{Z}} = \sum_r \hat{Z}(r) / R .$$

6.3 Simulation Parameters and Results

The simulation parameters were selected based on the experience in one large PSU in the NIS. The relative sizes for U_a and U_b were set at $\gamma_a = N_a / N = 0.32$ and $\gamma_b = N_b / N = 0.44$. The parameters for frame *B* were $\tau_1 = 50,\ \tau_2 = 50,\ \alpha_1 = 14.13,\ \alpha_2 = 7.707$, $V(r_{bj}) = 0.1^2$, $E(r_{bj}) = 0.647$, and $M = 20$ clusters. For each cluster j, y_{jk} was generated using $\mu_a = 11.3,\ \mu_{ab} = 11.2,\ \mu_b = 10.7,$ $\sigma = 1,\ \rho = 0.1$ and $\delta = 0.5$. The parameter for endangerment standard was E_0 =-0.3, an average proportion of 0.473 endangerment children in the survey universe.

Table 1 shows the finite population size N, population total Y, and total endangerment children E by estimation domains. Samples were selected from the finite populations using the sampling fractions $f_A = n_A / N_A = 0.022,$ $f_B = n_B / N_B = 0.011$ and $m_B = 10$. The sample size realized over the 10,000 iterations ranged between 1,498 and 1,512 observations. Table 2 shows the minimum, maximum, and median sample sizes for each sample segment over the 10,000 iterations.

Using the samples from one simulation cycle, table 3 shows the sample size and estimates of the population size for each of the five estimation methods by domain. The Kish design effect factor was computed as $1 + cv_w^2$. The NIS-3 method and the pseudo SF method are similar in that a weight of 1 or 0 for observations in domain S_3 and S_4 makes no real difference as compared with large weights otherwise. However, the cases preserved by the NIS method would have an impact when their conditional weights within PSUs are multiplied by the PSU selection probability.

For estimates on population size ($\hat{N}$), table 4 shows the percent relative bias (RelBias), empirical mean square error (EMSE) and their standard errors S(EMSE) for each of the five estimation methods. Tables 5 and 6 show the same statistics for estimates of a population total ($\hat{Y}$) and for a subpopulation total on the number of maltreated children by endangerment standard ($\hat{E}$). While the PML method is best for $\hat{N}$ and $\hat{Y}$, the SF methods performed equally well for $\hat{E}$, the maltreatment standard measurement.

Table 1. True population size (*N*), population total (*Y*) and subpopulation total (*E*-endangerment)

Estimates	Domains			Frames		
	a	*ab*	*b*	*A*	*B*	*U*
Population Size N (percent)	25,446 (32.9)	17,382 (22.4)	34,623 (44.7)	42,828 (55.3)	52,005 (67.1)	77,451 (100.0)
Population Total Y (mean)	287,539 (11.3)	192,284 (11.1)	365,920 (10.6)	479,823 (11.2)	558,204 (10.7)	845,743 (10.9)
Subpopulation Size E (mean)	15,838 (0.622)	9,144 (0.526)	11,662 (0.337)	24,982 (0.583)	20,806 (0.400)	36,644 (0.473)

Table 2. Sample size over 10,000 iterations (minimum, maximum and median)

Distribution	Estimation domains				
	S_1	S_2	S_3	S_4	S_5
Minimum	505	321	0	122	311
Maximum	616	435	14	253	446
Median	559	378	4	187	379
(%)	(37.1)	(25.1)	(0.3)	(12.4)	(25.2)

Table 3. Population size (sum of weights) and Kish's design effect factor for samples in one simulation cycle

Estimation method	Domains						Kish's Factor $\left(1+cv_w^2\right)$
	S_1 (n_1 =567)	S_2 (n_2 =372)	S_3 (n_3 =3)	S_4 (n_4 =212)	S_5 (n_5 =355)	Total	
PML *	25,457	13,385	166	3,820	32,101	74,929	1.25
SF_1	25,348	11,101	180	6,199	30,761	73,589	1.22
SF_2	25,325	11,172	91	6,240	30,761	73,589	1.22
SF_3 --Pseudo SF	25,779	16,913	136	0	30,761	73,589	1.29
SF_4 --Modified SF	25,731	16,882	3	212	30,761	73,589	1.30

* Effective sample size in frame *B* was computed assuming a design effect=2.0.

Table 4. RelBias, EMSE, and S(EMSE) for population size estimates $\left(\hat{N}\right)$

Estimation method	RelBias (%)			EMSE $\left(10^6\right)$			S(EMSE) $\left(10^6\right)$		
	U	*A*	*b*	*U*	*A*	*b*	*U*	*A*	*b*
PML *	-0.02	0.00	-0.05	2.52	0.00	2.52	0.03	0.00	0.03
SF_1	-0.04	0.00	-0.09	3.43	0.00	3.43	0.05	0.00	0.05
SF_2	-0.04	0.00	-0.09	3.43	0.00	3.43	0.05	0.00	0.05
SF_3 --Pseudo SF	-0.04	0.00	-0.09	3.43	0.00	3.43	0.05	0.00	0.05
SF_4 --Modified SF	-0.04	0.00	-0.09	3.43	0.00	3.43	0.05	0.00	0.05

* Effective sample size in frame *B* was computed assuming a design effect=2.0.

Domains: U = union of the two frames, A = frame A, and b = only frame B.

Table 5. RelBias, EMSE, and S(EMSE) for population total estimates $\left(\hat{Y}\right)$

	RelBias (%)			EMSE $\left(10^6\right)$			S(EMSE) $\left(10^6\right)$		
Estimation Method	*U*	*A*	*b*	*U*	*A*	*b*	*U*	*A*	*b*
PML *	-0.02	0.01	-0.04	267.30	1.70	266.33	3.66	0.02	3.65
SF_1	-0.04	0.01	-0.09	352.37	1.74	352.10	4.95	0.02	4.94
SF_2	-0.04	0.01	-0.09	352.36	1.73	352.10	4.95	0.02	4.94
SF_3 --Pseudo SF	-0.04	0.01	-0.09	353.31	1.85	352.10	4.97	0.03	4.94
SF_4 --Modified SF	-0.04	0.01	-0.09	353.27	1.85	352.10	4.97	0.03	4.94

* Effective sample size in frame *B* was computed assuming a design effect=2.0.
Domains: U = union of the two frames, A = frame *A*, and b = only frame B

Table 6. RelBias, EMSE, and S(EMSE) for subpopulation total estimates on endangerment standards $\left(\hat{E}\right)$

	RelBias (%)			EMSE $\left(10^6\right)$			S(EMSE) $\left(10^6\right)$		
Estimation Method	*U*	*A*	*b*	*U*	*A*	*b*	*U*	*A*	*b*
PML *	0.04	0.03	0.05	2.33	0.41	1.82	0.03	0.01	0.03
SF_1	0.00	0.03	-0.06	2.29	0.41	1.72	0.03	0.01	0.02
SF_2	0.00	0.03	-0.06	2.29	0.41	1.72	0.03	0.01	0.02
SF_3 --Pseudo SF	0.00	0.03	-0.06	2.20	0.45	1.72	0.03	0.01	0.02
SF_4 --Modified SF	0.00	0.03	-0.06	2.20	0.45	1.72	0.03	0.01	0.02

* Effective sample size in frame *B* was computed assuming a design effect=2.0.
Domains: U = union of the two frames, A = frame *A*, and b = only frame B

7. Discussion

This paper compared five estimation options in a dual-frame survey where frame *A* used a simple random sampling design and frame *B* used a complex sample design. The classic single frame (SF) estimation methods cannot apply to the NIS because for members in frame *A* their selection probability in frame *B* is unknown. The NIS-3 has developed a modified single frame method to accommodate this situation and this method compared favorably against the PML method and the classic SF method in initial simulation evaluations.

The modified SF method for NIS has the advantage that it is unbiased, practical, and relatively easy to implement. Further simulation evaluations are needed to test the outcomes when there are (1) misclassification of domain membership in segments S_4 and S_5 (Clark *et al*., 2007), (2) larger sample overlaps in segment S_3, and (3) different design effects for variables in frame *B*.

References

Bankier, M.D. (1986). Estimation based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Clark, J, Winglee, M., and Liu, B. (2007). Handling imperfect overlap determination in a dual-frame survey. *Proceedings of the Survey Research Method Section of the American Statistical Association*.

Kalton, G. and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical* Society, Ser. A, 149, 65-82.

Lohr, S.L. and Rao, J.N.K. (2000). Inference from dual-frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L. and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Skinner, C.J. (1991) On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C.J. and Rao, J.N.K. (1996). Estimation in dual-frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

National Center on Child Abuse and Neglect, Administration on Children, Youth and Families, US Department of Health and Human Services. *Third National Incidence Study of Child Abuse and Neglect* (NIS-3, 1997). Analysis report. Washington, DC 20201: Sedlak, A.J., Broadburst, D., Shapiro, G., Kalton, G, Goksel, H., Burke, J., Brown, J. (Unpublished report).

Waksberg, J., Brick, J. M., Shapiro, G., Flores-Cervantes, I., and Bell, B. (1997). Dual-frame RDD and area sample for household survey with particular focus on low-income population. *Proceedings of the Survey Research Method Section of the American Statistical Association*, pp. 713-718.