Handling Imperfect Overlap Determination in a Dual-Frame Survey

Jay Clark, Marianne Winglee, Benmei Liu
Jay Clark, Westat, 1650 Research Boulevard, Rockville, MD 20850

Abstract

The analysis of dual frame surveys requires identification of which sampled units are included on both frames. However, this identification is imperfect when only limited and nonidentical matching data are available (except in the case of units sampled from both frames). This situation is encountered in the National Incidence Study of Child Abuse and Neglect. That study employs a dual-frame design that combines a list frame of all maltreated children investigated by Child Protective Services agencies and another sample frame compiled from maltreated children reported by sources such as the police and school staff. This paper compares a logistic regression procedure and a record linkage procedure for determining overlap in a way that minimizes the misclassification rate for matches and nonmatches. This paper also explores the impact of misclassification on five alternative dual-frame estimators in a simulation study.

Keywords: Dual-frame estimation, record linkage, threshold selection, domain misclassification, single-frame estimator, pseudo maximum likelihood estimator

1. Introduction

The National Incidence Study of Child Abuse and Neglect (NIS) is treated as a dual frame survey for estimation purposes. A complexity for estimation is that the classification of observations into estimation domains is not clear cut. This paper describes the methods used for domain classification and examines the effects of domain misclassification in a simulation study. Five estimation methods are compared: 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 SF method, and a modified SF method developed for estimation with the NIS.

Section 2 presents background on the NIS sample design. Section 3 describes the framework of estimation domains for a dual-frame survey. Section 4 describes the NIS approach for determining domain membership for sample observations and section 5 describes the simulation study. The paper concludes with a discussion and assessment of the results.

2. The NIS Sample Design

The NIS is a national survey conducted to estimate the number of maltreated children in the United States. A complex multistage and multiple frame sample design is employed to cover a number of possible reporting sources for maltreated children. The primary sampling units (PSUs) are either individual counties or county clusters. Within sampled PSUs, Child Protective Services (CPS) agencies are the primary source of data for maltreated children. However, the coverage from the CPS agencies is incomplete because these agencies may not investigate all forms of maltreatment. For broader coverage, the NIS constructs list frames in sampled PSUs for 10 different agency categories including police, juvenile probation, hospitals (children and general), public schools, day care centers, shelters, public housing, social service, and mental health agencies. Agencies are sampled from these list frames, staff rosters are constructed within sampled agencies, and staffs are sampled to serve as informants (sentinels) for maltreated children.

For estimation purposes, the NIS is treated as a dual-frame design within each sampled PSU: frame *A* is a list frame of the maltreated children investigated by CPS agencies and frame *B* comprises possibly maltreated children observed by professional staffs in the non-CPS agencies. A self-weighting sample is selected from list frame *A*. However, there is no list of possibly maltreated children for frame *B* and the size of frame *B* is unknown. Children are sampled from frame *B* by a two-stage sampling process within each PSU, first sampling agencies and then sampling sentinels within sampled agencies. The sample from frame *B* is a non-self-weighting sample.

This paper addresses the development of sampling weights within sampled PSUs. It therefore focuses on a single PSU. Overall weights are computed in a straightforward way by multiplying the within-PSU weight by the inverse of the PSU selection probability.

3. Dual Frame Estimation Domains

With the NIS dual frame design, samples S_A and S_B are selected independently from the two frames, *A* and *B*, with sample sizes n_A and n_B. Figure 1 is a pictorial representation of the NIS dual frame design. There are three estimation domains with two frames: U_a for observations only in frame *A*; U_b for observations only in frame *B*; and

U_{ab} for observations common to both frames. It is assumed that the union of the two frames covers the population of interest.

The key to estimation is that domain membership is known for every observation in the sample. Each sample observation can be classified into one of five distinct segments. S_1 are members in domain U_a and are sampled from frame *A*. S_2, S_3, and S_4 are members in domain U_{ab}: S_2 are sampled only from *A*, S_3 are sampled from both *A* and *B*, and S_4 are sampled only from *B*. S_5 are members in domain U_b and are sampled from frame *B*.

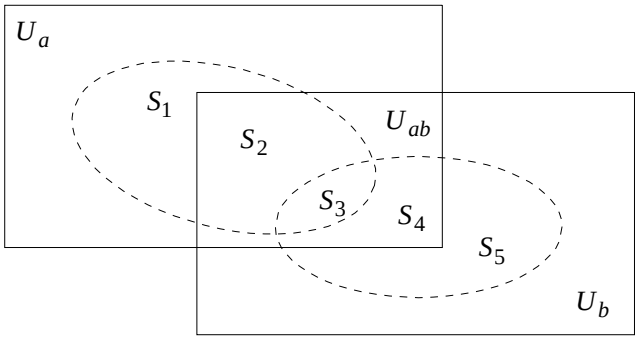

Figure 1. Dual-frame estimation domains and sample segments

4. Domain Determination in the NIS

Determination of domain membership in the NIS is limited by the data available. Section 4.1 describes the manual review process used to identify the observations selected in both samples, i.e., the observations in segment S_3. Section 4.2 discusses the difficulty in separating the remaining observations in S_A between segments S_1 and S_2. Section 4.3 describes two alternative methods—logistic regression and record linkage—that can be used to allocate the remaining observations in S_B between segments S_4 and S_5 and to estimate the extent of classification errors.

4.1 Members in S_3

A manual review process was employed to identify observations selected in both NIS samples. All sampled observations in a PSU were compared with respect to a child's first name, last initial, and birth date. Reviewers examined all observations in a PSU that agreed on any two of these three fields and also all observations with the same age if any missing data occurred in these fields. The decision of whether two observations were related to the same child or not was then based on reviews of all reported data in the survey questionnaires.

The combined sample for the most recent NIS round, the NIS-4 conducted in 2005, comprised 29,565 records. Reviewers examined over 5,000 data forms in the first review cycle and determined that duplicate forms across the samples were received for 378 children.

4.2 Members in S_1 and S_2

The classification of the remaining observations in S_A (after removing observations in S_3) into either S_1 or S_2 is problematic because of the lack of a list frame for frame *B*. Without such a list, reliance must be placed on the information in the CPS agency records about the sources of the maltreatment reports. A possible procedure is to classify the personnel who reported a maltreated child to CPS agencies into two groups: those who are surveyed through frame *B* in the NIS (e.g., personnel in police departments and schools); and others (e.g., neighbors and relatives). The sampled children from CPS agencies can then be classified into segments S_1 and S_2 according to whether or not the personnel who reported them were from agencies surveyed in the NIS. The limitation of this classification is that the CPS information about the reporting personnel is for the most part relatively general (e.g., medical personnel) and not easily mapped into the agency categories that are covered by frame *B*. Also, since sampling in frame *B* is not self-weighting, the selection probabilities of observations classified in S_2 are not known.

4.3 Members in S_4 and S_5

The classification of observations into S_4 or S_5 can be made by reference to the list frame *A*. The limitation is that while frame *A* provides data on personal identifiers—first name, last initial, sex, birth date, city of residence, and number of children in the household—that can be used to separate the observations into S_4 or S_5, no survey data exists to validate the classification. For this classification, the NIS used the personal identifiers in a logistic regression approach in the previous round, the NIS-3 conducted in 1993, and a probability record linkage approach in the NIS-4.

Both methods employ a "truth set" containing the pairs examined previously by manual reviewers to determine S_3 membership, including the decision of matched or nonmatched status. For the NIS-4, the first review cycle identified 2,939 candidate pairs of children in the truth set, of which 1,140 pairs were determined to be true matched pairs and the remaining 1,799 were determined to be true nonmatched pairs. The 1,140 matched pairs included the 378 matched pairs identified in S_3 as selected from both frames

A and *B*, as well as 762 matched pairs of children selected more than once within one of the frames, a result of multiple incidents in the time period and multiple reports of the same child. The 1,799 true nonmatched pairs included 554 pairs containing one record from frame *A* and one record from frame *B*, and 1,245 pairs of children within one of the frames.

For the logistic regression method, the personal identifiers available on frame *A* were compared between the two members of each pair. Each comparison was then coded as 1 if there was full agreement and 0 if not. A logistic regression model was then run with true matched status as the dependent variable and the 0-1 agreement indicators as the independent variables.

Figure 2 plots the cumulative distribution for record pairs in the true matched set and 1 minus the cumulative distribution for record pairs in the true nonmatched set by the predicted probability of a match. A probability close to 1 indicates a high level of agreement in all match fields and a probability close to zero indicates disagreement in the match fields. Since most of the pairs either fully agree or fully disagree on their match fields, logistic regression is able to clearly separate most of the pairs. For the few hundred pairs in the middle of the graph that have some agreement and some disagreement, however, the separation is not as clear.

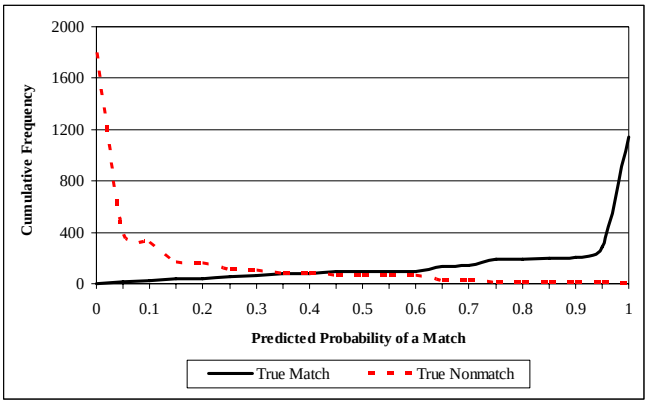

Figure 2. Cumulative distributions by predicted probability of a match

Probability-based record linkage provides an alternative approach for classifying pairs as matched or nonmatched (Fellegi and Sunter, 1969; Winkler, 1995). This approach uses a match weight to measure the likelihood of a correct match and a decision rule to classify record pairs. Methods to determine the selection threshold and estimate the linkage error rates are discussed in Belin and Rubin (1995), Lawson (2006), and Winglee, Valliant, and Scheuren (2005).

The match weight assigned to a pair of records is derived from a likelihood ratio that accounts for the closeness of the fields being compared for each pair. Assuming that the fields are independent, the probability of agreement/disagreement on a field is modeled as a Bernoulli random variable. Using r for a record pair, v for a match field (or variable) where there are $v = 1,\ldots,V$ fields, the match weight w_r is:

$$w_r = \log_2\left[\frac{\prod_{v=1}^{V} m_v^{y_{rv}}\left(1-m_v\right)^{1-y_{rv}}}{\prod_{v=1}^{V} u_v^{y_{rv}}\left(1-u_v\right)^{1-y_{rv}}}\right]$$

where $m_v = P\left(\text{field } v \text{ agrees in pair } r \,\middle|\, r \in M\right)$, M is the true set of matched pairs, $u_v = P\left(\text{field } v \text{ agrees in pair } r \,\middle|\, r \in U\right)$, U is the true set of nonmatched pairs, and $y_{rv} = 1$ if field v agrees and 0 otherwise. The weight w_r is a type of log-odds or log-likelihood ratio. The conditional agreement probabilities m_v and u_v are the match parameters for each field.

Estimates of the probabilities of agreement m_v for record pairs in the true matched set and for record pairs in the true nonmatched set are shown in Table 1 for each of the personal identifiers used for matching. For example, the probability of agreement on date of birth was 0.86 in the true matched set and only 0.06 in the true nonmatched set.

Table 1. Agreement probability in the true matched and true nonmatched sets by match field

	Probability of agreement	
Match field	Match set	Nonmatch set
Sex	0.97	0.96
First name	0.94	0.76
Last initial	0.94	0.82
Age	0.93	0.29
Date of birth	0.86	0.06
City of residence	0.83	0.39
Children in household	0.68	0.29

Figure 3 plots the cumulative distribution of record pairs in the true matched set and 1 minus the cumulative distribution of record pairs in the true nonmatched set by match weight. A positive and high match weight indicates agreement in the match fields and a negative or smaller weight means more disagreement in the match fields. As with logistic regression, this graph shows that record linkage is able to separate the true matched pairs from the true nonmatched pairs well, with the added feature of a clearly defined point where the two lines cross.

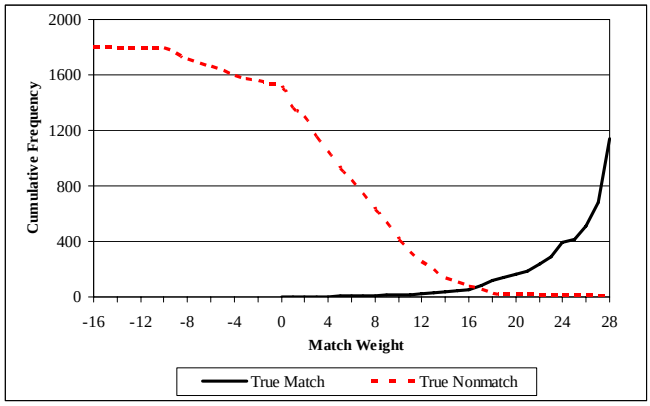

Figure 3. Cumulative distributions by match weight

For classification, the regression and the record linkage parameters developed from the "truth set" can be applied to candidate pairs of observations in S_4 and in S_5. Selecting a threshold to separate the matched and nonmatched pairs is necessary: a high threshold will result in omitting some true matched pairs, and a low threshold will result in accepting nonmatched pairs as matched pairs. This study set the threshold at the intersection of the cumulative distributions for matched and nonmatched pairs as displayed in figures 2 and 3. This threshold minimizes the total number of misclassified pairs.

Table 2 shows the number of record pairs in the truth set that would be misclassified using this threshold decision, for both the logistic regression and record linkage approaches. The omission error for the regression method is 79 pairs that were true matches, and for record linkage 55 pairs that were true matches, with only 36 pairs in common in both methods. The false positive error for the regression method is 77 nonmatched pairs, and for record linkage 74 nonmatched pairs, with 54 of the pairs in common.

Table 2. Number of record pairs in the "truth sets" by classification method and status

Classification Method	Status	Truth sets	
		Matched	Nonmatched
Logistic regression	Match	1,061	77
	Nonmatch	79	1,722
Record linkage	Match	1,085	74
	Nonmatch	55	1,725

5. A Simulation Study

A simulation study was conducted to evaluate the performance of the NIS estimation method when close to 6 percent of the observations in sample segments S_4 and S_5 were misclassified. Section 5.1 describes the simulation data sets, section 5.2 outlines the estimation methods, and section 5.3 presents the simulation results.

5.1 Simulation Data with Domain Misclassification

The simulation study involved constructing two sampling frames and repeating the processes of sampling, weighting, and estimation over 10,000 iterations. Simple random samples of n_A units were selected from frame *A* using a sampling fraction $f_A = n_A / N_A$. The samples of n_B units from frame *B* were 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 from the N_j units in each sampled cluster *j*. The design sampling weights were $w_{Ak} = N_A / n_A$ and $w_{Bjk} = (M / m)(N_j / n_O)$ for unit *k* in cluster *j*.

Winglee et al. (2007) describe the details of frame construction, sampling, weighting, and estimation for the simulation study. The simulated frames comprised 42,828 units in frame *A*, 52,005 units in frame *B*, and 77,451 units in the union of the two frames. A 0-1 variable *E* was generated as the maltreatment measure according to the NIS endangerment standard, assuming a proportion of 0.473 endangerment countable children in the survey universe. Another 0-1 variable was generated to index domain classification. The true domain membership in S_4 and S_5 was randomly misclassified for approximately 6 percent of the units in these segments.

The sample fractions used were $f_A = 0.022$, $f_B = n_B / N_B = 0.011$, and $m_B = 10$. The sample sizes realized over the 10,000 draws ranged between 1,498 and 1,512 observations.

5.2 Estimation Methods

Sampling weights were constructed for five estimation methods. The pseudo-maximum likelihood (PML) sampling weights followed the approach proposed by Skinner and Rao (1996) with the following adaptation to the NIS situation. The NIS design has N_A known, N_B and N_{AB} unknown, the sample S_A is a self-weighting sample, and the sample S_B is a complex non-self-weighting sample. A ratio-adjusted weight was used to post-stratify the weights for the sample from frame *A* to the known population total for that frame. The sample size n_B was replaced by n_B^*, the effective sample size, defined as $n_B^* = n_B / d_B$, where d_B is the average design effect of the two domain sizes coming from frame *B*. A design effect of $d_B = 2.0$ was used in this study.

Table 3. Example of sample sizes, estimates of totals, and Kish's design effect factor for sample weights for five estimation methods by five segments (6 percent of samples S_4 and S_5 misclassified)

	Domains						Kish's factor
	S_1	S_2	S_3	S_4 †	S_5 †	Total	$(1+cv_w^2)$
Sample size	*602*	*337*	*3*	*213 (197)*	*354 (370)*	*1,509*	
*PML**	25,976	10,245	176	6,431	36,600	79,428	1.31
SF_1	26,140	9,766	171	6,751	33,355	76,183	1.26
SF_2	26,120	9,830	86	6,792	33,355	76,183	1.26
SF_3 (Pseudo SF)	27,370	15,322	136	0	33,355	76,183	1.33
SF_4 (Modified SF)	27,319	15,293	3	213	33,355	76,183	1.34

* Effective sample size in frame *B* was computed assuming a design effect of 2.0
† True sample size is shown in parentheses

The classic single frame (SF) methods, SF_1 and SF_2, followed Kalton and Anderson (1986) and Bankier (1986), respectively. These methods treat all observations as though they had been sampled from a single frame and adjust the weights in the intersection domains according to their inclusion probability in each sample. The sampling weights for the SF_1 method were computed for all observations in segment S_3. The SF_2 method retained only the distinct units in S_3, reducing the weight for this segment by half. Although these SF methods are not applicable in the NIS because the probability of selection in frame *B* is unknown for observations selected from frame *A*, they are included in the simulation study for comparative purposes.

The pseudo single frame method SF_3 was computed by ignoring observations in S_4 (i.e., setting sampling weights of observations in S_4 to zero). This method is analogous to the situation where frame overlaps are removed through a prescreening process. In the NIS, prescreening was not possible. The obvious disadvantage of ignoring the S_4 segment is the loss of data.

A modified SF method SF_4 was developed for the NIS-3. This method essentially follows the SF_2 approach, where duplicate observations in S_3 are eliminated. When observations in S_1 and S_2 are inseparable, one weighting scheme is to assign unit weight for observations in S_3 and S_4. Winglee et al. (2007) describe the rationale of this approach. They found that this method compared favorably against the alternatives under the simulation conditions used in this study but without domain misclassification.

Using the samples from one of the 10,000 simulation cycles, table 3 presents the sample sizes and the estimates of totals by domain for each estimation method. The numbers in parentheses report the sample sizes in S_4 and S_5 before the domain membership was changed for approximately 6 percent of the observations in S_4 and S_5. Kish's design effect factor is also included to show that the estimation methods have similar variances. The weights in the first 4 segments were post-stratified to the known frame *A* total, so the sum of those weights is the same for each estimation method.

The pseudo and modified SF methods are similar in that a weight of 0 or 1 for observations in domain S_4 makes no real difference compared to large weights otherwise. However, for national estimates, the cases preserved by the modified SF method would have an impact when their conditional weights within PSUs are multiplied by the PSU selection probability.

5.3 Simulation Results

To compare the PML estimation method to the SF methods, estimates of the percent relative bias (Relbias) and the empirical mean square error (EMSE) were computed for estimates of the total population sizes $\hat{N}$ for three domains: *U* (union of the two frames or total maltreated children), *A* (the number of maltreated children investigated by CPS agencies), and *b* (the number of children not investigated by CPS agencies). The PML method had a Relbias of -0.7% for *U* and -1.5% for *b*, while the SF methods had a Relbias of -2.3% for *U* and -5.1% for *b*. The PML also had a smaller EMSE (2.68) than the SF methods (6.16) for *U* and *b*. The Relbias and EMSE for domain *A* were zero because the sampling weights for all five estimation methods were ratio adjusted to the known frame *A* total.

Table 4 shows the same calculations for estimates of the subpopulation total $\hat{E}$ (the number of countable maltreated children under the endangerment standard). The PML method performs relatively well, possibly because the frame *B* sample makes a smaller contribution to the estimates with the PML method than is the case with the other methods. In the NIS, the frame *B* sample has larger variance due to the

complex sampling design and sampling fraction. The SF_3 and SF_4 methods have the lowest errors in table 4, due to their lower weights for observations in segment S_4, which lessens the effect of misclassification.

Table 4. RelBias and EMSE by Domain (*U*, *A*, *b*) for estimates of endangered children ($\hat{E}$)

Estimation method	RelBias (%)			EMSE (10^6)		
	U	*A*	*b*	*U*	*A*	*b*
*PML**	-0.4	-1.0	1.0	2.40	0.47	1.75
SF_1	-1.9	-1.5	-2.7	2.69	0.55	1.68
SF_2	-1.9	-1.5	-2.7	2.69	0.55	1.68
SF_3	-0.9	0.0	-2.7	2.16	0.45	1.68
SF_4	-0.9	0.0	-2.7	2.16	0.45	1.68

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

6. Discussion

Overlap sample observations (i.e., observations in S_3) are based on a "truth set." Two methods were compared to classify observations sampled from frame *B* into those that overlap with frame *A* and those that do not. Both the logistic regression and probability-based record linkage methods performed well; record linkage was preferred for use in NIS-4.

Lohr and Rao (2006) examined the effects of domain misclassification in multiple-frame surveys and concluded that estimation methods are sensitive to misclassification of observations into domains. The simulation study described in this paper shows that the NIS method performed well relative to the other single frame methods. While the PML method performed best in terms of relative bias and EMSE for estimates of the total population, the SF_3 and SF_4 had lower EMSEs—but larger biases—for estimates of subpopulation totals by the endangerment standard.

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.

Belin, T.R., and Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

Fellegi, I.P., and Sunter, A.B. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

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

Lawson, J.S. (2006). Record linkage techniques for improving online genealogical research using census index records, *Proceeding of the Section on Survey Research Methods*. American Statistical Association.

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

Winglee, M., Rust, K., Liu, B., Shapiro, G., and Park, I. (2007). A case study in dual frame estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Winglee, M., Valliant, R., and Scheuren, F. (2005). A case study in record linkage. *Survey Methodology*, 31(1), 3-11.

Winkler, W. E. (1995). Matching and record linkage. In Cox, Binder, Chinnappa, Christianson. Colledge, and Kott (Eds.), *Business Survey Methods*. John Wiley and Sons.