Survey Methodology, June 1998
Vol. 24, No. 1, pp. 79-88
Statistics Canada

Combining Multiple Frames to Estimate Population Size and Totals

DAWN E. HAINES and KENNETH H. POLLOCK[1]

ABSTRACT

Efficient estimates of population size and totals based on information from multiple list frames and an independent area frame are considered. This work is an extension of the methodology proposed by Hartley (1962) which considers two general frames. A main disadvantage of list frames is that they are typically incomplete. In this paper, we propose several methods to address frame deficiencies. A joint list-area sampling design incorporates multiple frames and achieves full coverage of the target population. For each combination of frames, we present the appropriate notation, likelihood function, and parameter estimators. Results from a simulation study that compares the various properties of the proposed estimators are also presented.

KEY WORDS: Incomplete frame; Capture-recapture sampling; Screening estimator; Dual frame methodology; Multiple frame estimation.

1. INTRODUCTION

In classical sampling theory, it is assumed that a complete frame exists. In practice, however, this assumption is often violated. Frame imperfections such as omissions, duplications, and inaccurate recordings are almost inevitable in any large data collection operation (Hansen, Hurwitz and Madow 1953). Information collected from list and area frames is used to obtain estimates of the unknown population size and totals. For example, an ecologist or wildlife biologist may use one list and one area frame sample to estimate the number of bald eagle nests in a given region. The U.S. Bureau of the Census uses dual system estimation to measure decennial census undercounts. Darroch, Fienberg, Glonek and Junker (1993) describe a three-sample multiple-capture approach to estimating population size when inclusion probabilities are heterogeneous. In addition, state agriculture officials may be interested in estimating the number of hog farms and the total number of hogs in North Carolina. Typically, information from multiple information sources is combined to estimate population sizes and totals.

List frames are physical listings of sampling units in the target population. These are constructed over the years using information from scientists as well as city, county, state, and federal agencies. Items found on a list frame can include, but are not limited to, names, addresses, telephone numbers, social security numbers, or physical descriptions of location. These and other miscellaneous stratification variables are used to identify persons, animals, businesses, or other establishments. When estimating the number of bald eagle nests in a region, we construct this year's list frame using information from last year's list frame. With the addition of new eagle nests, last year's list frame becomes quickly outdated and incomplete. Because of this incompleteness, estimates based solely on list frames typically underestimate the true population size. Supplementing available information with an area frame sample may provide an efficient estimation of the population size and totals.

An area frame is a collection of geographical areas defined by identifiable boundaries. The entire area in which data are collected is divided into mutually exclusive and exhaustive sampling units called segments. The segments are usually stratified according to a characteristic of interest. Once a stratified random sample of segments is drawn, enumerators visit the sampled segments and record measurements on all reporting units contained therein.

The National Agricultural Statistics Service (NASS) currently employs a multi-frame approach for its sampling and estimation of numerous agricultural commodities. Fecso, Tortora and Vogel (1986) provide a review of sampling frames for the agricultural sector of the United States while Nealon (1984) details the multiple and area frame estimators used by the U.S. Department of Agriculture. Kott and Vogel (1995) provide a general overview of multiple frame surveys.

In Section 2, we consider estimation based on information from two or more independent list frames. We show how these methods are related to capture-recapture methods. In Section 3, we consider more efficient estimators of population size and totals when information from an independent area frame sample is available. We extend these methods to the case of dependent list frames in Section 4. Results from a simulation study that compare different estimators are summarized in Section 5. Finally,

[1] Dawn E. Haines, U.S. Bureau of the Census, Washington, DC 20233; Kenneth H. Pollock, North Carolina State University, Department of Statistics, Box 8203, Raleigh, NC 27695-8203, U.S.A.

Section 6 summarizes our results and discusses future directions for research.

2. MULTIPLE LIST FRAMES

2.1 Population Size Estimation

List frames used to estimate population size are usually incomplete and do not cover the entire population. One solution to the incomplete list frame problem is to merge two or more incomplete list frames. Combining multiple list frames may result in improved coverage of the target population, and thus, may provide better estimators. In the case of multiple list frames, it is commonly assumed that each element in the population has the same probability of being included on a given list frame. Hence, the list frame elements themselves constitute our "samples." For example, individuals may decide independently whether or not to list their telephone numbers in the telephone directory with equal probability. In the case of bald eagle nests, this year's list frame is constructed based on last year's nest sightings. If we assume that the probability of a nest being sighted is the same for all nests, then the above assumption is valid. Finally, the assumption is also valid in capture-recapture experiments where the first list frame consists of all animals captured on the first sampling occasion and the second list frame consists of all animals captured on the second sampling occasion. This scenario corresponds to Model M_t in the capture-recapture literature. See Otis, Burnham, White and Anderson (1978) for details. Model M_t assumes all animals in the population are equally at risk to capture on each sampling occasion, but this probability can vary over different sampling occasions.

To begin, we consider the case of two independent list frames, B_1 and B_2. Suppose B_1 has size N_{B_1} and B_2 has size N_{B_2}. Let domain $b_1(b_2)$ consist of those $N_{b_1}(N_{b_2})$ elements that belong only to frame $B_1(B_2)$ and domain $b_1 b_2$ contain $N_{b_1 b_2}$ units that belong to both frames. The final domain includes existing target population elements that are not included on either list frame. Its size is $N - N_{b_1} - N_{b_2} - N_{b_1 b_2}$. Domain notation for list frames B_1 and B_2 is presented in Table 1. Note that every element in every frame must be categorized into a domain without error. Errors in domain determination are serious and cannot be corrected at a later time. These errors are not considered in the estimation phase and thus are regarded as nonsampling errors. Nealon (1984) claims that domain determination is the single largest source of nonsampling error in multiple frame designs (Kott and Vogel 1995).

Let the probability that a population element is included on frame $B_1(B_2)$ be $p_{B_1}(p_{B_2})$. Since list frames B_1 and B_2 are assumed to be independent, the probability of an element belonging to domain b_1 is $p_{b_1} = p_{B_1}(1 - p_{B_2})$. The remaining domain probabilities are defined similarly. The population size N and the inclusion probabilities p_{B_1} and p_{B_2} are unknown parameters. The likelihood function is given by

$$\mathcal{L}(p_{B_1}, p_{B_2}, N \mid N_{b_1}, N_{b_2}, N_{b_1 b_2}) = \binom{N}{N_{b_1}, N_{b_2}, N_{b_1 b_2}} * p_{B_1}^{N_{B_1}} p_{B_2}^{N_{B_2}} (1 - p_{B_1})^{N - N_{B_1}} (1 - p_{B_2})^{N - N_{B_2}}. \quad (1)$$

Table 1
Domain Notation for List Frames B_1 and B_2

Domain Size	Domain Probability
N_{b_1}	$p_{b_1} = p_{B_1}(1 - p_{B_2})$
N_{b_2}	$p_{b_2} = (1 - p_{B_1})p_{B_2}$
$N_{b_1 b_2}$	$p_{b_1 b_2} = p_{B_1} p_{B_2}$
$N - N_{b_1} - N_{b_2} - N_{b_1 b_2}$	$1 - p_{b_1} - p_{b_2} - p_{b_1 b_2} = (1 - p_{B_1})(1 - p_{B_2})$

Maximum likelihood estimators (MLEs) of the frame inclusion probabilities are obtained by maximizing the logarithm of the likelihood (1). This procedure yields

$$\hat{p}_{B_1} = \frac{N_{B_1}}{\hat{N}} \quad \text{and} \quad \hat{p}_{B_2} = \frac{N_{B_2}}{\hat{N}}, \quad (2)$$

where the MLE $\hat{N}$ is substituted for N. Rather than differentiating the log-likelihood function to approximate the value of N, we employ the "ratio method" of maximizing the likelihood which equates $\mathcal{L}(N)$ to $\mathcal{L}(N-1)$ (Darroch 1958). This process accounts for the discrete parameter N and yields the equation

$$\frac{\mathcal{L}(N)}{\mathcal{L}(N-1)} = \frac{\hat{N}}{(\hat{N} - N_{b_1} - N_{b_2} - N_{b_1 b_2})} * (1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) = 1. \quad (3)$$

Here we assume that N is large so that

$$\frac{N_{B_1}}{N-1} \approx \frac{N_{B_1}}{N} \quad \text{and} \quad \frac{N_{B_2}}{N-1} \approx \frac{N_{B_2}}{N}.$$

Substituting the estimators in (2) into (3) yields

$$\hat{N}_1 = \hat{N} = \frac{N_{B_1} N_{B_2}}{N_{b_1 b_2}}. \quad (4)$$

Sekar and Deming (1949) derive an estimate of the variance of (4), given by

$$\hat{V}(\hat{N}_1) = \frac{N_{B_1} N_{B_2} N_{b_1} N_{b_2}}{(N_{b_1 b_2})^3}.$$

Substituting (4) into (2) yields the MLEs of p_{B_1} and p_{B_2},

$$\hat{p}_{B_1} = \frac{N_{b_1 b_2}}{N_{B_2}} \quad \text{and} \quad \hat{p}_{B_2} = \frac{N_{b_1 b_2}}{N_{B_1}}.$$

The estimator $\hat{N}_1$ of N in (4) is called the Lincoln-Petersen estimator in closed population capture-recapture models. The elements on list frame B_1 may be considered as the units captured in the first sampling occasion and the elements on list frame B_2 may be viewed as the units captured in the second sampling occasion. The elements in domain $b_1 b_2$ correspond to recaptured elements. With this correspondence, it is easy to see that the likelihood for the population size and capture probabilities for two occasions will be the same as that given in (1). Hence, the MLEs derived for two independent list frames will be the same as the corresponding MLEs for the capture-recapture model with two sampling occasions.

Extending these ideas, we contend that combining k independent list frames is directly related to having k sampling occasions under Model M_t in closed population capture-recapture models, where $t = k$ (Otis *et al.* 1978). The general likelihood function for k independent list frames, $B_1, B_2, ..., B_k$, has the form

$$\mathcal{L}(p_{B_1}, ..., p_{B_k}, N \mid N_{b_1}, ..., N_{b_1 \ldots b_k}) =$$

$$\binom{N}{N_{b_1}, ..., N_{b_1 \ldots b_k}} \prod_{l=1}^{k} p_{B_l}^{N_{B_l}} (1 - p_{B_l})^{N - N_{B_l}}, \qquad (5)$$

which has exactly the same structure as the likelihood introduced by Darroch (1958) and is discussed in great detail by Otis *et al.* (1978) and Seber (1982). The form of the estimated frame inclusion probabilities is

$$\hat{p}_{B_l} = \frac{N_{B_l}}{\hat{N}}, \quad l = 1, ..., k. \qquad (6)$$

Values of $\hat{N}$ are obtained by numerically solving the $(k - 1)$ degree polynomial in $\hat{N}$ resulting from the equality

$$\frac{\mathcal{L}(N)}{\mathcal{L}(N-1)} = \frac{\hat{N}}{(\hat{N} - N_{b_1} - \cdots - N_{b_1 \ldots b_k})} *$$

$$(1 - \hat{p}_{B_1}) \cdots (1 - \hat{p}_{B_k}) = 1. \qquad (7)$$

We then select as $\hat{N}$ as the root that maximizes the value of the likelihood function (5). Substituting this root into (6) yields MLEs of the k frame inclusion probabilities.

2.2 Population Total Estimation

Suppose the measured y_i values are available for all units on the k independent list frames. The estimated probability that the first element is included on at least one of the k list frames is

$$\hat{\pi}_1 = \hat{P}\left[\cup_{l=1}^{k} B_l\right] = 1 - (1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) \cdots (1 - \hat{p}_{B_k}),$$

where $\hat{p}_{B_l} = N_{B_l}/\hat{N}$ and $\hat{N}$ is the MLE of N obtained from (7). From equation (7),

$$\frac{\hat{N}}{(\hat{N} - N_{b_1} - \cdots - N_{b_1 \ldots b_k})} (1 - \hat{\pi}_1) = 1$$

which simplifies to

$$\hat{\pi}_1 = \frac{N_{b_1} + \cdots + N_{b_1 \ldots b_k}}{\hat{N}}.$$

An estimated Horvitz and Thompson (1952) estimator of the population total is

$$\hat{\hat{Y}}_{\text{H-T}} = \frac{1}{\hat{\pi}_1} \sum_{i \in B_1 \cup \ldots \cup B_k} y_i$$

$$= \frac{\hat{N}}{N_{b_1} + \cdots + N_{b_1 \ldots b_k}} \sum_{i \in B_1 \cup \ldots \cup B_k} y_i = \hat{N} \bar{Y}_L,$$

where $\bar{Y}_L$ is the mean of distinct elements on the list frames. Thus, for k independent list frames, the estimated Horvitz-Thompson estimator coincides with the population total estimator proposed by Pollock, Turner and Brown (1994).

In some situations, values of the variable of interest, y_i, are not available for all units on the list frames. If the list frames are large in size, random samples are selected from each list frame and data are collected on those subsampled elements. If there are k list frames, it is possible to define 2^k domains. We consider an extension of Lund's (1968) estimator for the total of all units on the list frames,

$$\hat{Y}_{L,L} = \sum_{l=1}^{2^k - 1} N_l \bar{y}_l,$$

which is a weighted sum of $2^k - 1$ domain means, $\bar{y}_l$. The weights are given by the domain sizes. Further, the population total estimator is

$$\hat{Y} = \hat{N} \frac{\hat{Y}_{L,L}}{\sum_{l=1}^{2^k - 1} N_l}.$$

3. MULTIPLE LISTS PLUS AN AREA FRAME

3.1 Population Size Estimation

Joining multiple, individual list frames with an area frame sample is a solution to overcoming list frame deficiencies. Assume that the geographical area of interest is

subdivided into U_A segments. Also, assume that a simple random sample of u_A segments is selected from U_A segments that cover the entire population. Therefore, the probability of a segment being selected is $p_A = u_A/U_A$. In some surveys, it is possible to subdivide the region into approximately equally-sized segments. In such cases the segment selection probability corresponds approximately to the proportion of area sampled. The inclusion of an area frame provides completeness of the target population (Hartley 1962). We assume that each reporting unit belongs to exactly one segment. Once a segment is selected, all reporting units within the segment are observed. For example, when estimating the number of bald eagle nests, each nest belongs to one and only one segment. However, this assumption is not always valid. Consider the case where a hog farm crosses segment boundaries. In this case, population elements may be associated with more than one segment. To address this problem, association rules linking population elements to segments are established at the estimation stage. See Faulkenberry and Garoui (1991) for more detail. The National Agricultural Statistics Service implements three correspondence rules that map elements in the population to sampled segments. The open, closed, and weighted segment estimators are described in Nealon (1984). Another related reference is Sirken (1970).

Consider the case of k independent list frames plus an area frame. The population size, N, and the list frame inclusion probabilities, p_{B_i}, $i = 1, ..., k$, are unknown parameters. The area frame inclusion probability $p_A = u_A/U_A$ is known. The likelihood function has the form

$$\mathcal{L}(p_{B_1}, ..., p_{B_k}, N | p_A, n_a, n_{ab_1}, ..., n_{ab_1...bk}, N_{b_1}, ..., N_{b_1...b_k})$$

$$= \binom{N}{n_a, n_{ab_1}, ..., n_{ab_1...b_k}, N_{b_1}, ..., N_{b_1...b_k}} p_A^{n_A}(1 - p_A)^{N - n_A}$$

$$\prod_{l=1}^{k} p_{B_l}^{N_{B_l}}(1 - p_{B_l})^{N - N_{B_l}},$$

where n_A is the total number of elements in the u_A sampled area segments and n_a is the number of elements in the u_A sampled area segments which do not belong to any list frames. Similarly, $n_{ab_1}, ..., n_{ab_1...b_k}, N_{b_1}, ..., N_{b_1...b_k}$ are defined as the sizes of different domains. It is important to emphasize that the inclusion of an area frame may cause the value of N_{b_1} to change. N_{b_1} now corresponds to the number of elements on list frame B_1 which are not in the u_A selected area segments and not on any other list frame.

The MLEs of the parameters are given by $\hat{p}_{B_l} = N_{B_l}/\hat{N}$, where $\hat{N}$ is a solution to the k-th degree polynomial

$$\hat{N}(1 - p_A)(1 - \hat{p}_{B_1}) ... (1 - \hat{p}_{B_k}) =$$

$$(\hat{N} - n_a - n_{ab_1} - \cdots - n_{ab_1...b_k} - N_{b_1} - \cdots - N_{b_1...b_k}). \quad (8)$$

Numerical methods are essential for solving (8) for the MLE $\hat{N}$ of N. Among the k roots of (8), we select $\hat{N}$ that maximizes the likelihood.

Applying this methodology to one list frame and one area frame, we obtain

$$\hat{N} = N_{B_1} + \frac{n_a}{p_A}. \quad (9)$$

This estimator is also known as the screening estimator (Kott and Vogel 1995). The screening estimator categorizes elements into two distinct groups. The first group contains elements which belong to both the list and area frames and is called the overlap domain. Since it is assumed that all elements on a list frame belong to the area frame, the size of the overlap domain coincides with the number of elements on frame B_1 and has the value N_{B_1}. The second group contains elements in the area frame not included on the list frame(s) and is referred to as the nonoverlap domain. The size of the nonoverlap domain is an unobserved random quantity, N_a. The term n_a is the number of elements found in the u_A area segments which are not included on the list frame(s) following a specific association rule. An estimated value of N_a is n_a/p_A. Hence, an estimate of the population size is given by $\hat{N}$ in (9). The resulting MLE of p_{B_1} is

$$\hat{p}_{B_1} = \frac{N_{B_1}}{N_{B_1} + \frac{n_a}{p_A}}.$$

When multiple list frames are available, it is possible to combine them into a single list frame and use the above estimator to obtain an estimate of N. That is, consider the screening estimator

$$\hat{N}_2 = \hat{N} = N_{B_1 \cup ... \cup B_k} + \frac{n_a}{p_A} = N_{b_1} + \cdots + N_{b_k} +$$

$$N_{b_1 b_2} + \cdots + N_{b_1...b_k} + \frac{n_a}{p_A}. \quad (10)$$

Note that the screening estimator $\hat{N}_2$ is appropriate even when the list frames are *not* independent of each other. We discuss this further in Section 4.

Using this methodology for one area and two independent list frames yields the likelihood

$$\mathcal{L}(p_{B_1}, p_{B_2}, N | p_A, n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1 b_2}, n_{ab_1 b_2}) =$$

$$\binom{N}{n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1 b_2}, n_{ab_1 b_2}} p_A^{n_A} p_{B_1}^{N_{B_1}} P_{B_2}^{N_{B_2}}$$

$$(1 - p_A)^{N - n_A}(1 - p_{B_1})^{N - N_{B_1}}(1 - p_{B_2})^{N - N_{B_2}}.$$

The MLE of N is

$$\hat{N}_3 = \hat{N} = (2p_A)^{-1} * \left[(N_{B_1} + N_{B_2})p_A + (n_a - N_{b_1b_2} - n_{ab_1b_2})\right] + (2p_A)^{-1} \sqrt{\left[(N_{B_1}+N_{B_2})p_A+(n_a-N_{b_1b_2}-n_{ab_1b_2})\right]^2 + 4p_A(1-p_A)N_{B_1}N_{B_2}}, \quad (11)$$

where $n_{ab_1b_2}$ denotes the number of elements included in the u_A sampled area segments that belong to both list frames. An estimate of the variance of $\hat{N}_3$ may be obtained using the Taylor series approximation of (11) and the asymptotic distribution of $(N_{B_1}, N_{B_2}, n_a, N_{b_1b_2}, n_{ab_1b_2})$.

3.2 Population Total Estimation

When y_i's are available for all elements on k independent list frames and for a sample of segments from an area frame, we consider an estimated Horvitz-Thompson estimator to estimate the population total. Recall that we assume the following:

1. The probability that a unit is included on the i-th list frame, p_{B_i}, is the same for all units.
2. The event that a unit is included on one frame is independent of its inclusion on another frame.
3. The probability that a unit is included in the area frame sample of u_A segments is $p_A = u_A/U_A$.

Since we consider the case where population units belong to exactly one area segment and all units within a sampled segment are observed, the third assumption is valid. Hence, the probability the i-th element is on at least one of the k list frames and/or the area frame sample is

$$\tilde{\pi}_1 = 1 - (1 - p_A)(1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) \cdots (1 - \hat{p}_{B_k}) = \frac{n_a + n_{ab_1} + \cdots + N_{b_1 \ldots b_k}}{\hat{N}}.$$

The estimated Horvitz-Thompson population total estimator is

$$\hat{\bar{Y}}_{\text{H-T}} = \frac{\hat{N}}{n_a + n_{ab_1} + \cdots + N_{b_1 \ldots b_k}} \sum_{i \in \text{ sample}} y_i = \hat{N}\,\bar{y}_L,$$

where $\bar{y}_L$ is the mean of the distinct elements on list frames $B_1, \ldots, B_k$ and the elements in the area frame sample.

We can also use the screening estimator to estimate the population total. The known overlap domain total is combined with an estimator of the nonoverlap domain (NOL) total to yield $\hat{Y}_S = Y_L + \sum_{i \in \text{NOL}} y_i/p_A$. The NOL domain consists of elements on the area frame that are not on any of the list frames and $Y_L = Y_{B_1 \cup \ldots \cup B_K}$ is the total of the distinct units on the k list frames. In the subsampling case, we may replace Y_L in $\hat{Y}_S$ by Lund's estimator, given by

$$\hat{Y}_{L,L} = N_{b_1}\bar{y}_{b_1} + \cdots + N_{b_k}\bar{y}_{b_k} + N_{b_1b_2}\bar{y}_{b_1b_2} + \cdots + N_{b_1 \ldots b_k}\bar{y}_{b_1 \ldots b_k}.$$

4. DEPENDENT LIST FRAMES

We now consider the case where dependencies exist among list frames but where area and list frames remain independent. In capture-recapture experiments, for example, the probability an animal is captured on the second sampling occasion may depend on whether it was captured on the first sampling occasion. See Fienberg (1972), Cormack (1989), Wolter (1990), Pollock, Hines, and Nichols (1984), Huggins (1989), and Alho (1990) for specific examples.

We consider the case where we have two list frames, B_1 and B_2, that are dependent. Let p_{11} denote the probability of being included on both list frames. If B_1 and B_2 are independent, then $p_{11} = p_{B_1}p_{B_2}$ where p_{B_1} and p_{B_2} are inclusion probabilities for B_1 and B_2, respectively. Define $p_{10}(p_{01})$ as the probability of being included on frame $B_1(B_2)$ but not on frame $B_2(B_1)$. The probability of exclusion from both list frames is denoted by $p_{00} = 1 - p_{B_1} - p_{B_2} + p_{11}$.

The likelihood function is given by

$$\begin{aligned}\mathcal{L}(p_{B_1}, p_{B_2}, p_{11}, N \mid p_A, n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1b_2}, n_{ab_1b_2}) &= \binom{N}{n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1b_2}, n_{ab_1b_2}} p_A^{n_A}(1-p_A)^{N-n_A} \\ &\quad (p_{B_1} - p_{11})^{N_{b_1}+n_{ab_1}} (p_{B_2} - p_{11})^{N_{b_2}+n_{ab_2}} p_{11}^{N_{b_1b_2}+n_{ab_1b_2}} \\ &\quad (1 - p_{B_1} - p_{B_2} + p_{11})^{N-N_{b_1}-N_{b_2}-n_{ab_1}-n_{ab_2}-N_{b_1b_2}-n_{ab_1b_2}}. \end{aligned} \quad (12)$$

Maximizing (12) with respect to p_{B_1}, p_{B_2}, p_{11} and N leads to the approximate solution

$$\hat{N} = N_{b_1} + N_{b_2} + n_{ab_1} + n_{ab_2} + N_{b_1b_2} + n_{ab_1b_2} + \frac{n_a}{p_A},$$

which coincides with the screening estimator $\hat{N}_2$. That is, $\hat{N}$ is also the estimator that is obtained by pooling the two list frames into a single list frame where the duplications are eliminated and the nonoverlap domain size is estimated using the area frame sample. Also, it can be shown that the two-stage maximum likelihood procedure of Sanathanan (1972) leads to:

$$\hat{N} = \frac{n_a + N_{B_1 \cup B_2}}{p_A + (1 - p_A)\dfrac{N_{B_1 \cup B_2}}{\hat{N}_2}} = \hat{N}_2.$$

Thus, the maximum likelihood estimator and Sanathanan's estimator both coincide with the screening estimator. If information from two dependent list frames is available and the nature of the dependency is unknown, then we cannot estimate the individual parameters. When information from an independent area frame is available, all parameters are estimable. However, for estimating N, $N_{B_1 \cup B_2}$ is sufficient and no additional information is gained from N_{B_1}, N_{B_2}, and $N_{b_1 b_2}$.

Methods are available for modeling the dependence among k list frames when estimating population size and totals. Additional population information or information from an independent area frame is needed to accurately model the dependence. Fienberg (1972) and Cormack (1989) consider constrained log-linear models to model the dependence. On the other hand, Wolter (1990) uses external constraints such as a known sex ratio to estimate the population size in the dependence case. Another technique used is to model the inclusion probabilities as a function of the covariates. Alho, Mulry, Wurdeman and Kim (1993) use a conditional logistic regression model to estimate the probability of being enumerated in a census and apply the model to the 1990 Post-Enumeration Survey. The role of auxiliary variables in capture-recapture experiments with unequal capture probabilities is addressed in Pollock *et al.* (1984), Huggins (1989), and Alho (1990).

5. SIMULATION STUDY

We conduct a simulation study to assess the overall efficiency of different population size estimators for the special case of two list frames plus an area frame. This is the most feasible combination of sampling frames for real survey problems.

5.1 Design of the Study

In order to study both dependent and independent cases, we define the parameter θ that reflects the dependence structure between list frames B_1 and B_2. It has the same form as the odds ratio and is written formally as

$$\theta = \frac{p_{00} p_{11}}{p_{01} p_{10}}.$$

In the case of two list frames, the value of θ determines a unique solution for p_{11}. Our study varies the following factors:

Factor	Levels	Definition
N	500, 5000	Population size
p_A	0.05, 0.10, 0.20	Inclusion probability for area frame A
$p_{B_1}(=p_{B_2})$	0.7, 0.9	Inclusion probability for list frame $B_1(B_2)$
θ	0.5, 1.0, 1.5, 2.0	Odds ratio

For each parametric combination, we generate data (n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, $N_{b_1 b_2}$, $n_{ab_1 b_2}$). One thousand Monte Carlo replications are generated for each parametric combination.

5.2 Estimators

We compare four population size estimators, $\hat{N}_1$, $\hat{N}_2$, $\hat{N}_3$, and $\hat{N}_4$. $\hat{N}_1$ is the Lincoln-Petersen estimator which does not incorporate area frame information. The estimator $\hat{N}_1$ is suitable when the list frames are independent. Since the estimator ignores information from the area frame sample, it is expected to be inefficient when information from an area frame is available. The screening estimator, $\hat{N}_2$, sums the overlap and nonoverlap domain estimates and is particularly suitable for the dependent list frame case. The third estimator, $\hat{N}_3$, is derived from the full, independent sampling frame likelihood function. This estimator exploits the information contained in the area and list frames and the fact that the list frames are independent ($\theta = 1$).

We expect $\hat{N}_3$ to be the best estimator when list frames B_1 and B_2 are independent whereas we expect $\hat{N}_2$ to be the best estimator in the dependent case. As a result, we also consider a pre-test estimator that tests for independence of the list frames. We define $\hat{N}_4$ to be $\hat{N}_2$ if there is strong evidence to believe that frames B_1 and B_2 are not independent. Otherwise, we take $\hat{N}_4 = \hat{N}_3$. Formally,

$$\hat{N}_4 = \begin{cases} \hat{N}_2 & \text{if GOF} > \chi^2_{1,0.05} = 3.84 \\ \hat{N}_3 & \text{otherwise,} \end{cases}$$

where GOF is the chi-square goodness-of-fit test statistic for testing H_0: $\theta = 1$ and is derived from the following two-way table.

	In B_1	Not In B_1	
In B_2	$n_{ab_1 b_2}$	n_{ab_2}	$n_{A \cap B_2}$
Not In B_2	n_{ab_1}	n_a	$n_{A \cap B_2'}$
	$n_{A \cap B_1}$	$n_{A \cap B_1'}$	n_A

Figure 1. Classification of Sampled Area Frame Elements

Figure 1 categorizes the n_A elements according to their presence on or absence from list frames B_1 and B_2.

5.3 Comparing the Estimators

Tables 2 and 3 display the percent relative bias and the percent relative root mean square error of the estimates $\hat{N}_1, \hat{N}_2, \hat{N}_3$, and $\hat{N}_4$ for population sizes of 500 and 5000, respectively. We scale the bias and the root mean square error by N in order to directly compare estimators based on different population sizes. A comparison of $\hat{N}_1$ with $\hat{N}_3$ shows the benefit of drawing an area frame sample. In practice, these benefits depend on the relative cost of the area frame sample. In this study, we do not take sampling costs into account. The probability of being included on both list frames, p_{11}, is given in parentheses in the θ column. When $p_B = p_C = .9, p_{11}$ must lie between .8 and .9. However, for θ ranging from .5 to 2, p_{11} varied only from .806 to .817.

The estimator $\hat{N}_2$ is unbiased for N and has the smallest percent relative bias. The estimators $\hat{N}_1$ and $\hat{N}_3$ are asymptotically consistent for N and yield biases close to 0 when $\theta = 1$. On the other hand, $\hat{N}_1$ and $\hat{N}_3$ have large biases when $\theta \neq 1$. The percent relative bias of $\hat{N}_4$ is smaller than that of $\hat{N}_3$ but it is not close to zero. The bias does not change significantly as p_A increases from .05 to .10 to .20.

When $N = 500$ and $p_B = p_C = .9$, $\hat{N}_3$ has the smallest percent relative root mean square error (% RRMSE). This is partly due to the fact that the limited range of p_{11} values is similar to the p_{11} value for the independence case (.810). The % RRMSE for $\hat{N}_3$ is 40 - 50 % smaller than that of $\hat{N}_2$. On the other hand, the % RRMSE of $\hat{N}_3$ is only 15 - 30 % smaller than that of $\hat{N}_1$. Therefore, when the list frames have very high inclusion probabilities, both $\hat{N}_1$ and $\hat{N}_3$ are much better than $\hat{N}_2$. Additionally, if area frame sampling costs are high, $\hat{N}_1$ may be a reasonable alternative estimator to $\hat{N}_3$. When $N = 500$ and $p_B = p_C = .7$, $\hat{N}_3$ has the smallest % RRMSE for the independence case. When $\theta = 2$, $\hat{N}_2$ has the smallest % RRMSE. If $N = 5000$ and $p_B = .7$, $\hat{N}_3$ has the smallest % RRMSE for only $\theta = 1$. For all other θ values, $\hat{N}_2$ yields the smallest % RRMSE. In all cases, $\hat{N}_3$ has very small variance and most of the % RRMSE is due to the bias in $\hat{N}_3$. For $\theta < 1$, $\hat{N}_3$ tends to have positive bias while for $\theta > 1$, $\hat{N}_3$ has negative bias. For the case of $N = 5000$ and $p_B = .9$, $\hat{N}_3$ has the smallest % RRMSE for $\theta = 1$. $\hat{N}_2$ has the smallest % RRMSE for $\theta = .5$ and 2. For $\theta = 1.5$, there is no best estimator with respect to % RRMSE.

As expected, the percent relative root mean square errors of $\hat{N}_2$, $\hat{N}_3$, and $\hat{N}_4$ decrease as the value of p_A increases. Thus, as the area frame information increases, the % RRMSE decreases. Also, as the population size increases from 500 to 5000, the % RRMSE decreases. Since the values of p_A in our simulation are small, $\hat{N}_2$ has a large variance. On the other hand, even though $\hat{N}_3$ is biased, it has a very small standard error and results in a smaller % RRMSE. The estimator $\hat{N}_4$ reduces the bias of $\hat{N}_3$ but has a large standard error. Hence, $\hat{N}_4$ is not a particularly beneficial estimator. For larger values of θ and p_A, we expect $\hat{N}_2$ to perform better than $\hat{N}_3$. For the values of θ and p_A we considered, we recommend $\hat{N}_3$ over other estimators.

The value of % RRMSE for $\hat{N}_4$ is between that of $\hat{N}_2$ and $\hat{N}_3$ in most cases. We write the estimator $\hat{N}_4$ as $\hat{N}_4 = \delta\hat{N}_2 + (1 - \delta)\hat{N}_3$, where $\delta = 0$ or 1 based on the results of the goodness-of-fit test. The % RRMSE and % RBias of $\hat{N}_4$ need not be between those of $\hat{N}_2$ and $\hat{N}_3$ because δ is not independent of $\hat{N}_2$ and $\hat{N}_3$.

5.4 Limitations of the Study

The goal of our study is to compare the bias, standard error, and mean square error of four population size estimators. We assume that inclusion probabilities for both list frames are identical. Future studies may include unequal inclusion probabilities as well as larger values of θ. Clearly the benefit of $\hat{N}_3$ over $\hat{N}_1$ depends on the cost of sampling from an area frame. Our paper considers only small values of p_A. Small p_A values are associated with a high area frame sampling cost. Even in this case, we observe a significant reduction in % RRMSE and % RBias, thereby justifying the use of $\hat{N}_3$ over $\hat{N}_1$. We do not consider an objective function which incorporates sampling costs, % RRMSE, and % RBias.

Throughout this paper, we assume that all units have the same probability of being included on a given list frame. Haines (1997) considers the case where the inclusion probabilities are modeled as a function of a covariate. When inclusion probabilities are heterogeneous, larger units may have a higher list frame inclusion probability than smaller units. Heterogeneous inclusion probabilities play an important role in estimating population totals when the response variable has a highly skewed distribution or has rare values. Haines (1997) also presents two stratification procedures that are useful when area and list frames are stratified on the same variable. These results will be presented in future publications.

6. DISCUSSION

The primary focus of this paper is population size estimation based on several sampling frames. Information from area and/or list frame(s) is collected and combined to obtain various estimators. We derive population size estimators when information is available only on k independent list frames and also when information is available on an area frame sample in addition to the list frames. We conduct a simulation study to compare the performance of the estimators in the special case of two list frames plus an area frame. Based on our simulation study, we recommend the estimator derived from the full, independent likelihood, $\hat{N}_3$, for the case where the list

Table 2
Simulation Results for $N = 500$

			p_A					
			.05		.10		.20	
p_B	θ		% RBias	% RRMSE	% RBias	% RRMSE	% RBias	% RRMSE
.7	.5 (.462)	$\hat{N}_1$	62.30	66.01	60.64	64.04	63.26	66.81
		$\hat{N}_2$	0.30	49.07	-0.75	32.37	0.85	22.58
		$\hat{N}_3$	55.52	58.95	48.15	51.15	40.53	43.32
		$\hat{N}_4$	48.15	58.88	37.88	49.25	24.95	38.80
	1 (.490)	$\hat{N}_1$	0.47	19.26	1.01	19.08	-0.11	19.45
		$\hat{N}_2$	0.45	57.34	0.34	39.61	0.88	27.25
		$\hat{N}_3$	0.43	18.21	0.83	16.93	0.14	15.75
		$\hat{N}_4$	2.40	27.57	1.39	22.94	0.29	17.96
	1.5 (.508)	$\hat{N}_1$	-35.60	40.06	-36.48	40.58	-35.69	40.26
		$\hat{N}_2$	3.11	66.43	-5.08	41.96	0.30	28.79
		$\hat{N}_3$	-32.07	36.79	-31.01	35.28	-24.04	28.88
		$\hat{N}_4$	-22.74	47.62	-26.21	37.57	-17.06	30.38
	2 (.522)	$\hat{N}_1$	-60.07	62.91	-61.31	64.06	-60.41	63.28
		$\hat{N}_2$	-6.12	66.59	-1.15	46.68	1.67	30.99
		$\hat{N}_3$	-55.36	58.35	-51.21	54.19	-40.89	43.99
		$\hat{N}_4$	-41.39	63.79	-34.79	55.45	-18.60	41.35
.9	.5 (.806)	$\hat{N}_1$	5.37	6.79	5.27	6.63	5.59	6.97
		$\hat{N}_2$	0.08	14.78	-0.06	10.17	-0.06	6.55
		$\hat{N}_3$	5.04	6.44	4.62	5.93	4.24	5.53
		$\hat{N}_4$	5.94	9.48	5.03	7.05	4.34	5.72
	1 (.810)	$\hat{N}_1$	0.30	5.01	0.17	5.01	0.25	4.94
		$\hat{N}_2$	0.78	20.72	0.41	14.06	-0.06	9.03
		$\hat{N}_3$	0.33	4.83	0.20	4.68	0.17	4.24
		$\hat{N}_4$	3.23	13.79	1.88	9.35	1.00	5.98
	1.5 (.814)	$\hat{N}_1$	-4.29	7.07	-4.39	7.32	-4.55	7.37
		$\hat{N}_2$	-0.65	21.52	0.35	15.88	0.002	10.27
		$\hat{N}_3$	-4.07	6.78	-3.83	6.73	-3.49	6.15
		$\hat{N}_4$	-0.43	13.77	-1.18	10.92	-1.43	8.20
	2 (.817)	$\hat{N}_1$	-8.28	10.27	-8.40	10.36	-8.33	10.32
		$\hat{N}_2$	-0.29	25.59	0.39	17.66	0.35	11.41
		$\hat{N}_3$	-7.80	9.82	-7.35	9.38	-6.30	8.20
		$\hat{N}_4$	-2.52	17.96	-3.10	14.02	-2.73	10.33

Table 3
Simulation Results for $N = 5000$

			p_A					
			.05		.10		.20	
p_B	θ		% RBias	% RRMSE	% RBias	% RRMSE	% RBias	% RRMSE
.7	.5 (.462)	$\hat{N}_1$	61.47	61.82	61.39	61.76	61.69	62.04
		$\hat{N}_2$	-0.18	15.78	0.26	10.65	-0.15	6.72
		$\hat{N}_3$	54.84	55.17	49.06	49.38	39.38	39.65
		$\hat{N}_4$	19.73	38.12	4.77	19.52	-0.01	7.21
	1 (.490)	$\hat{N}_1$	-0.28	6.14	-0.13	5.99	0.35	6.15
		$\hat{N}_2$	0.43	18.14	0.47	12.85	-0.20	8.34
		$\hat{N}_3$	-0.22	5.82	-0.03	5.35	0.16	4.88
		$\hat{N}_4$	0.26	9.82	-0.04	7.44	0.11	5.95
	1.5 (.508)	$\hat{N}_1$	-36.21	36.68	-36.29	36.78	-35.90	36.38
		$\hat{N}_2$	0.41	20.39	-0.16	14.21	0.39	9.55
		$\hat{N}_3$	-32.87	33.37	-29.97	30.49	-24.13	24.66
		$\hat{N}_4$	-19.11	31.15	-11.51	23.92	-3.12	14.03
	2 (.522)	$\hat{N}_1$	-61.04	61.3	-60.53	60.81	-60.64	60.92
		$\hat{N}_2$	0.40	20.09	0.60	15.43	0.31	9.67
		$\hat{N}_3$	-55.69	55.96	-50.24	50.55	-41.46	41.76
		$\hat{N}_4$	-14.10	36.31	-2.34	20.96	0.26	9.84
.9	0.5 (.806)	$\hat{N}_1$	5.56	5.70	5.52	5.67	5.54	5.68
		$\hat{N}_2$	-0.12	4.55	0.11	3.19	-0.03	2.08
		$\hat{N}_3$	5.21	5.35	4.86	5.01	4.22	4.35
		$\hat{N}_4$	4.97	5.41	3.64	4.88	2.26	3.79
	1 (.810)	$\hat{N}_1$	-0.02	1.58	0.08	1.55	0.01	1.57
		$\hat{N}_2$	-0.09	6.16	-0.17	4.08	-0.14	2.79
		$\hat{N}_3$	-0.03	1.53	0.05	1.48	-0.02	1.35
		$\hat{N}_4$	0.37	3.19	0.11	2.18	0.09	1.89
	1.5 (.814)	$\hat{N}_1$	-4.66	5.00	-4.52	4.85	-4.61	4.90
		$\hat{N}_2$	-0.25	7.54	0.11	4.95	-0.09	3.14
		$\hat{N}_3$	-4.39	4.73	-3.96	4.32	-3.55	3.85
		$\hat{N}_4$	-2.50	6.31	-2.26	5.02	-1.84	3.82
	2 (.817)	$\hat{N}_1$	-8.45	8.68	-8.38	8.60	-8.46	8.69
		$\hat{N}_2$	-0.21	7.86	-0.06	5.29	0.01	3.73
		$\hat{N}_3$	-7.95	8.18	-7.39	7.61	-6.49	6.73
		$\hat{N}_4$	-3.76	8.80	-2.77	6.99	-1.25	4.97

frames are independent or nearly independent. For the moderate to strong dependence cases, we recommend the screening estimator, $\hat{N}_2$.

We also study population total estimation. We consider two scenarios for estimating population totals. In the first case, we assume that observations are available on all units that comprise the list frames. In contrast, the second case assumes that information is available only on subsamples from each of the list frames. We consider an estimated Horvitz-Thompson estimator if list frames are independent and a screening estimator to estimate the population total if the list frames are dependent.

In this paper, our focus is on population size estimation. In practice, one may be interested in estimating population totals for several characteristics based on multi-stage samples involving unequal inclusion probabilities. Relevant papers on this topic include Bankier (1986), Skinner (1991), and Skinner, Holmes, and Holt (1994).

7. ACKNOWLEDGEMENTS

The authors thank the editor and two referees for useful comments on an earlier version of the article. This research was partially funded by the U.S. Geological Survey, Biological Resources Division. Christine Bunck is the BEST Program Manager. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

REFERENCES

ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.

ALHO, J.M., MULRY, M.H., WURDEMAN, K., and KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.

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

CORMACK, R.M. (1989). Log-linear models for capture-recapture. *Biometrics*, 45, 395-413.

DARROCH, J.N. (1958). The multiple-recapture census I: estimation of a closed population. *Biometrika*, 45, 343-359.

DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., and JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88, 1137-1148.

FAULKENBERRY, G.D., and GAROUI, A. (1991). Estimating a population total using an area frame. *Journal of the American Statistical Association*, 86, 445-449.

FECSO, R., TORTORA, R.D., and VOGEL, F.A. (1986). Sampling frames for agriculture in the United States. *Journal of Official Statistics*, 2, 279-292.

FIENBERG, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59, 591-603.

HAINES, D.E. (1997). Estimating Population Parameters Using Multiple Frame and Capture-Recapture Methodology. Ph.D. thesis, North Carolina State University.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons.

HARTLEY, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.

KOTT, P.S., and VOGEL, F.A. (1995). Multiple-frame business surveys. *Business Survey Methods* (Ed., B.G. Cox). New York: John Wiley & Sons, 185-203.

LUND, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.

NEALON, J.P. (1984). Review of the Multiple and Area Frame Estimators. Staff Report 80, U.S. Department of Agriculture, Statistical Reporting Service, Washington, DC.

OTIS, D.L., BURNHAM, K.P., WHITE, G.C., and ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-135.

POLLOCK, K.H., HINES, J.E., and NICHOLS, J.D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40, 329-340.

POLLOCK, K.H., TURNER, S.C., and BROWN, C.A. (1994). Use of capture-recapture techniques to estimate population size and population totals when a complete frame is unavailable. *Survey Methodology*, 20, 117-124.

SANATHANAN, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, 43, 1, 142-152.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters*, (2nd Edition). New York: Macmillan.

SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

SIRKEN, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

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., HOLMES, D.J., and HOLT, D. (1994). Multiple frame sampling for multivariate stratification. *International Statistical Review*, 62, 333-347.

WOLTER, K.M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.