Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Gross flow estimation in dual frame surveys

by Yan Lu and Sharon Lohr

SURVEY
METHODOLOGY

[illegible]

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Canada

June 2010

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2010
Vol. 36, No. 1, pp. 13-22
Statistics Canada, Catalogue No. 12-001-X

Gross flow estimation in dual frame surveys

Yan Lu and Sharon Lohr [1]

Abstract

Gross flows are often used to study transitions in employment status or other categorical variables among individuals in a population. Dual frame longitudinal surveys, in which independent samples are selected from two frames to decrease survey costs or improve coverage, can present challenges for efficient and consistent estimation of gross flows because of complex designs and missing data in either or both samples. We propose estimators of gross flows in dual frame surveys and examine their asymptotic properties. We then estimate transitions in employment status using data from the Current Population Survey and the Survey of Income and Program Participation.

Key Words: Complex surveys; Dual frame surveys; Jackknife; Longitudinal estimation; Missing data.

1. Introduction

Many current surveys follow the same individuals at regular time intervals so that longitudinal quantities such as transitions in employment status and poverty status can be studied. The U.S. Current Population Survey (CPS; United States Census Bureau 2006), for example, uses a rotating panel design in which persons in a housing unit selected for the survey are interviewed for four consecutive months, rested for eight months, and then interviewed again for four consecutive months. This design allows estimation of quantities related to individuals' changes over time. Since many survey responses are categorical, gross flows, which are transitions among states of a categorical variable over time, are particularly important.

Table 1 displays the counts of a categorical variable measured at two times in a population of N units. At time 1, the variable can be in one of r states and at time 2, the variable can be in one of c states. To illustrate Table 1, we give the following example. In studying changes in employment status, we might have $r = 2$ and $c = 2$, with state 0 representing unemployment and state 1 representing employment. Then X_{00} gives the count of persons in the population who are unemployed at both times, X_{10} is the number of persons who are employed at time 1 but unemployed at time 2, X_{0+} is the total number of persons who are unemployed at time 1, and so on. It is of interest to obtain estimates and standard errors of the gross flows X_{kl}, $k = 0, ..., r-1$, $l = 0, ..., c-1$, using survey data. This can be complicated in practice because of missing data and other problems.

While successive cross-sectional estimates can assess a change in unemployment rates over time, only a longitudinal survey addresses issues such as persistence of unemployment in individuals. Gross flow estimation using survey data has been studied by many authors, including Chambers, Woyzbun and Pillig (1988), Hocking and Oxspring (1971), Blumenthal (1968), Chen and Fienberg (1974), Stasny (1984, 1987), and Stasny and Fienberg (1986). Most of this work considered methods for obtaining maximum likelihood (ML) estimators for expected cell values in contingency tables with partially cross-classified data. Pfeffermann, Skinner and Humphreys (1998) proposed estimators that account for misclassification in survey data. All of this work has assumed that a probability sample, usually a simple random sample, has been taken from a single sampling frame.

Table 1
Gross flow table for population

		Time 2					
		0	**1**	**2**	$\cdots$	$c-1$	
Time 1	**0**	X_{00}	X_{01}	X_{02}	$\cdots$	$X_{0,c-1}$	X_{0+}
	1	X_{10}	X_{11}	X_{12}	$\cdots$	$X_{1,c-1}$	X_{1+}
	2	X_{20}	X_{21}	X_{22}	$\cdots$	$X_{2,c-1}$	X_{2+}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$r-1$	$X_{r-1,0}$	$X_{r-1,1}$	$X_{r-1,2}$	$\cdots$	$X_{r-1,c-1}$	$X_{r-1,+}$
		X_{+0}	X_{+1}	X_{+2}	$\cdots$	$X_{+,c-1}$	N

A number of longitudinal surveys, such as the Canadian National Longitudinal Survey of Children and Youth and the Canadian Household Panel Survey, have now started or are considering implementation of a dual frame or multiple frame design. In a multiple frame survey, probability samples are selected independently from two or more frames. Using more than one frame often gives better coverage of the population, and can achieve considerable cost savings in some populations. For example, the Assets and Health Dynamics Survey (Heeringa 1995), with the goal of estimating characteristics of the population aged over 65, used a dual frame survey in which frame A was

[1] Yan Lu, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001. E-mail: yljenlu@gmail.com; Sharon Lohr, School of Mathematical and Statistical Sciences, Arizona State University, Tempe AZ 85287-1804. E-mail: sharon.lohr@asu.edu.

the frame for a national general population survey and frame B was a list of Medicare enrollees. The structure of this survey is illustrated in Figure 1. Frame A covered the entire population but required extensive screening to identify individuals in the target population and was thus expensive to sample from; frame B was less expensive to sample, but did not include the entire population. Kalton and Anderson (1986) described uses of dual frame surveys to sample rare populations; Blair and Blair (2006) argued that dual frame surveys can take advantage of less expensive sampling modes such as internet sampling when sampling rare populations.

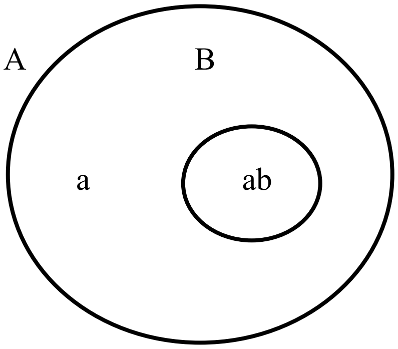

Figure 1 Frame B is a subset of frame A

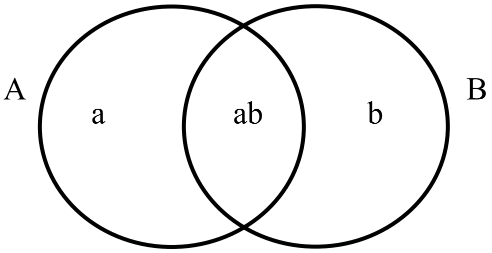

Figure 2 Frames A and B are both incomplete but overlapping

In other situations, both frames may be incomplete, as depicted in Figure 2. Hartley (1962, 1974) first proposed estimators for the dual frame survey design in Figure 2, when independent samples are taken from each frame. Subsequent developments are given in Bankier (1986), Fuller and Burmeister (1972), Skinner and Rao (1996), and Lohr and Rao (2000). Lohr and Rao (2006) summarized methods for estimating population quantities in cross-sectional multiple frame surveys.

In this paper, we propose estimators for gross flows that can be applied to dual frame surveys in which longitudinal information is collected in one or both samples. Units sampled in one or both surveys are followed over time; in some cases, additional units are sampled at later times to incorporate new population units or compensate for attrition. A longitudinal dual frame survey presents additional challenges to those found in longitudinal single frame surveys or in cross-sectional dual frame surveys. Missing data can occur in the sample from either frame, and units may change frame membership between interviews in the survey. In addition, either sampling design may be complex, with stratification and clustering. In an overlapping dual frame survey such as that depicted in Figure 2, one wishes to use the information in the overlap as efficiently as possible. The problem studied in this article is to use all the information sampled from frame A and frame B to estimate the transition probabilities of the population.

The article is organized as follows. In Section 2, we set up the research problem. In Section 3, we derive gross flow estimators in dual frame surveys for complex samples with possibly missing data. In Section 4, we derive asymptotic properties and discuss variance estimation. An application of our research to the Current Population Survey and Survey of Income and Program Participation is given in Section 5. Finally, we give our conclusions in Section 6.

2. Notation and sample quantities

Suppose there are two sampling frames, frame A and frame B, which together cover the population of interest $A \cup B$ as shown in Figure 2. In Hartley's (1962) notation, there are three nonoverlapping domains: $a = A \cap B^c$, $b = A^c \cap B$, and $ab = A \cap B$, where c denotes complement of a set. The population sizes for frames A and B are N_A and N_B, with domain population sizes N_a, N_b, and N_{ab}. We assume that N_A and N_B are known, but the population size $N = N_A + N_B - N_{ab}$ may be unknown. In this article, we assume that both the population and the frames are fixed over time. These are strong assumptions but in many longitudinal surveys the population of interest and the frames may be defined for time 1.

Assume for this section that domain membership is constant over time. For simplicity of notation in this paper we assume that $r = 2$ and $c = 2$ so that there are two possible categories at each time; the general case is similar. Since the three domains are nonoverlapping, each population count X_{kl}, $k = 0, 1$, $l = 0, 1$, can be written as $X_{kl} = X_{kla} + X_{klab} + X_{klb}$, where X_{kld} is the number of population units in domain d that are in state k at time 1 and state l at time 2. The corresponding population and domain probabilities are $p_{kl} = X_{kl}/N$ and $p_{kld} = X_{kld}/N_d$ for $d \in \{a, ab, b\}$.

Independent probability samples, S_A and S_B, with sample sizes n_A and n_B, are taken from frames A and B. Let w_i^A be the weight of sampled unit i for the sample from frame A and let w_j^B be the weight of sampled unit j for the sample from frame B. We may take w_i^A to be the sampling weight $[P(i \in S_A)]^{-1}$ or a Hájek-type weight $[P(i \in S_A)]^{-1} N_A /$ (sum of sampling weights in S_A). Other weighting schemes for longitudinal data, discussed in Verma, Betti and Ghellini (2007) and Lavallée (2007), might also be used. Let $\mathbf{y}_i = (y_{i1}, y_{i2})$ be the response for unit i in S_A, with $y_{i1}, y_{i2} \in \{0, 1, M\}$ where M denotes that the value is missing. Then $\hat{X}^A_{kla} = \sum_{i \in S_A} w^A_i \, I(y_{i1} = k) \, I(y_{i2} = l) \, I(i \in a)$ and $\hat{X}^A_{klab} = \sum_{i \in S_A} w^A_i \, I(y_{i1} = k) \, I(y_{i2} = l) \, I(i \in ab)$ estimate the population counts for the (k, l) cell in domains a and ab from S_A, for $k, l \in \{0, 1, M\}$. Let $\mathbf{y}_j = (y_{j1}, y_{j2})$ be the response for unit j in S_B, and let $\hat{X}^B_{klb} = \sum_{j \in S_B} w^B_j \, I(y_{j1} = k) \, I(y_{j2} = l) I(j \in b)$ and $\hat{X}^B_{klab} = \sum_{j \in S_B} w^B_j I(y_{j1} = k) I(y_{j2} = l) \, I(j \in ab)$ be the corresponding estimators from S_B.

In this paper, we assume that domain membership can be determined for every sample unit and that the responses $\mathbf{y}_i$ have no classification error. Thus, we assume that we know whether each unit in the frame A or frame B sample belongs to the other frame or not. We also assume that there is no measurement error for $\mathbf{y}_i$ and $\mathbf{y}_j$ – in the employment example, this means that every respondent gives the correct response for his or her employment status. Thus, the methods we proposed in our article are sensitive to misclassification of observations into domains and into cells. If the domain means differ or if observations are classified incorrectly, the estimators of gross flows could be biased; Pfeffermann *et al.* (1998) discussed methods of accounting for misclassification in single frame surveys.

The estimators from S_A are displayed in Table 2. A similar table may be constructed for the estimators from S_B. We assume that each unit is sampled during one or both time periods. If there is no missing data, then all the estimated counts for cells (k, M) and (M, l) are zero. Using the exact or approximate unbiasedness of the estimators, depending on whether the sampling or Hájek weights are used, when there is no missing data, $E[\hat{X}^A_{kla}] \approx X_{kla}$, $E[\hat{X}^A_{klab}] \approx E[\hat{X}^B_{klab}] \approx X_{klab}$ and $E[\hat{X}^B_{klb}] \approx X_{klb}$.

Table 2
Estimators from the frame A sample

			Time 2			
			0	**1**	**Missing**	
		0	$\hat{X}^A_{00a}$	$\hat{X}^A_{01a}$	$\hat{X}^A_{0Ma}$	$\hat{X}^A_{0+a}$
	domain *a*	**1**	$\hat{X}^A_{10a}$	$\hat{X}^A_{11a}$	$\hat{X}^A_{1Ma}$	$\hat{X}^A_{1+a}$
Time 1		**Missing**	$\hat{X}^A_{M0a}$	$\hat{X}^A_{M1a}$		$\hat{X}^A_{M+a}$
		0	$\hat{X}^A_{00ab}$	$\hat{X}^A_{01ab}$	$\hat{X}^A_{0Mab}$	$\hat{X}^A_{0+ab}$
	domain *ab*	**1**	$\hat{X}^A_{10ab}$	$\hat{X}^A_{11ab}$	$\hat{X}^A_{1Mab}$	$\hat{X}^A_{1+ab}$
		Missing	$\hat{X}^A_{M0ab}$	$\hat{X}^A_{M1ab}$		$\hat{X}^A_{M+ab}$
			$\hat{X}^A_{+0}$	$\hat{X}^A_{+1}$	$\hat{X}^A_{+M}$	$\hat{N}_A$

3. Gross flow estimators in dual frame surveys

In this section, we derive gross flow estimators for complex samples in dual frame surveys. A dual frame pseudo-likelihood approach is used to account for the sampling designs and missing data mechanism. A dual frame approach can improve precision of the estimators and provide more flexibility to model the missing data mechanism. Methods in current use for handling missing data are based on standard statistical methods and fall into four general categories (Little and Rubin 2002): complete-case analysis, weighting methods, imputation methods and model-based methods. We adopt a model-based approach for the missing data. In this section, we first consider a simple setup with simple random samples from a population with no missing data. Then we add a model for the missing data mechanism. Finally, we discuss estimators for more complex survey designs.

3.1 Simple random samples with complete data

To motivate the estimator in the general case, we first study estimation of gross flows when there is no missing data and when the sample from each frame is a simple random sample. Then $x^A_{kld} = n_A \hat{X}^A_{kld} / N_A$, for $d = a, ab$, is the observed sample count in cell kl and domain d from S_A; $x^B_{kld} = n_B \hat{X}^B_{kld} / N_B$ for $d = b, ab$ is the corresponding observed sample count from S_B.

If the sampling fractions are small, a multinomial approximation may be used for the likelihood. For the sample from frame A, there are eight cells with associated probabilities $P^A_{kld} = p_{kld} N_d / N_A$, for $k, l \in \{0, 1\}$ and $d \in \{a, ab\}$. The related probabilities for the sample from frame B are $P^B_{kld} = p_{kld} N_d / N_B$ for $k, l \in \{0, 1\}$ and $d \in \{b, ab\}$. Using the multinomial distribution and the assumption that the samples from the two frames are selected independently, the likelihood function is

$$L(\mathbf{p}, N_{ab}) \propto \prod_{k,l,d} (P^A_{kld})^{x^A_{kld}} \times \prod_{k,l,d} (P^B_{kld})^{x^B_{kld}}.$$

Although the likelihood is written for simplicity in terms of P^A_{kld} and P^B_{kld}, the underlying parameters of interest are $\mathbf{p} = (p_{00a}, p_{01a}, \ldots, p_{11b})$ and N_{ab}.

Setting the partial derivatives of the loglikelihood with respect to the parameters equal to zero, the maximum likelihood estimators are $\hat{p}_{kla} = x_{kla} / n^A_a$, $\hat{p}_{klb} = x_{klb} / n^B_b$ and $\hat{p}_{klab} = (x^A_{klab} + x^B_{klab}) / (n^A_{ab} + n^B_{ab})$, where $n^A_{ab} = \sum_{i \in S_A} I(i \in ab)$, $n^B_{ab} = \sum_{j \in S_B} I(j \in ab)$, $n^A_a = n_A - n^A_{ab}$ and $n^B_b = n_B - n^B_{ab}$. The MLE for N_{ab}, $\hat{N}_{ab}$, is the smaller root of the quadratic equation

$$[n_A + n_B] \, \hat{N}^2_{ab} - [n_A N_B + n_B N_A + n^A_{ab} N_A + n^B_{ab} N_B] \, \hat{N}_{ab} + [n^A_{ab} + n^B_{ab}] \, N_A N_B = 0. \quad (1)$$

Finally, using the above results, we construct the MLEs for X_{kl} and p_{kl}:

$$\hat{X}_{kl} = (N_A - \hat{N}_{ab})\hat{p}_{kla} + \hat{N}_{ab}\hat{p}_{klab} + (N_B - \hat{N}_{ab})\hat{p}_{klb}.$$

$$\hat{p}_{kl} = \frac{(N_A - \hat{N}_{ab})\hat{p}_{kla} + \hat{N}_{ab}\hat{p}_{klab} + (N_B - \hat{N}_{ab})\hat{p}_{klb}}{N_A + N_B - \hat{N}_{ab}}.$$

These estimators are the same as those obtained by Skinner (1991). However, Skinner used the approximate normal distribution of the response mean $\bar{y}$ in each domain to obtain the MLEs, while our estimators come from a multinomial model. The multinomial model allows us to include partially classified information from units observed at only one time period, as shown in the next section.

3.2 Simple random samples with missing data

In practice, individuals may appear in the sample at only one of the times. This can occur due to sample attrition (when members of the sample drop out during the course of a study) or other causes. In a rotating panel survey such as the CPS, persons rotating out of the survey at time 1 will not be contacted for time 2 and thus their time-2 employment status will be unknown. In other situations, one of the samples may be cross sectional, in which case all observations are measured at exactly one time.

3.2.1 Model for missing data

Blumenthal (1968), Chen and Fienberg (1974), Stasny (1984, 1987) and Stasny and Fienberg (1986) used a two-phase procedure to model the missing data in a single sample. A model is proposed for the complete data, and then the missing data mechanism is modeled. We extend this procedure to our dual frame structures. One advantage of a dual frame survey is that it provides more flexibility for the missing data models.

First, we assume that if all units were measured at both times, the model in Section 3.1 could be used. For the nonresponse mechanism, assume that each observation in cell (k, l) and domain d from S_A has probability ϕ^A_{kld} of being missing at time 1 and probability ψ^A_{kld} of being missing at time 2. We assume the unit cannot be missing at both times.

This formulation assumes a constant probability that an observation will be missing within a given cell, domain, and frame. If data could be missing for different reasons, additional parameters could be used to distinguish observations that have partial classification because of, say, the rotating panel design, and observations that have partial classification because of nonresponse. In Section 5, we discuss an alternative approach that might be used with multiple mechanisms for missing data.

For $k, l \in \{0, 1\}$, the probability that a unit from S_A is observed in cell (k, l) and domain d is

$$Q^A_{kld} = P^A_{kld}(1 - \phi^A_{kld} - \psi^A_{kld}).$$

The probability that a unit from S_A is observed in cell (k, M) and domain d is

$$Q^A_{kMd} = \sum_{l=0}^{1} P^A_{kld}\psi^A_{kld}.$$

Similarly, the probability that a unit from S_A is observed in cell (M, l) and domain d is

$$Q^A_{Mld} = \sum_{k=0}^{1} P^A_{kld}\phi^A_{kld}.$$

The probabilities for frame B are defined similarly with $Q^B_{kld} = P^B_{kld}(1 - \phi^B_{kld} - \psi^B_{kld})$, $Q^B_{kMd} = \sum_{l=0}^{1} P^B_{kld}\psi^B_{kld}$ and $Q^B_{Mld} = \sum_{k=0}^{1} P^B_{kld}\,\phi^B_{kld}$.

Under this two phase model, and using the assumption of independence of the samples, the likelihood function for the two samples is:

$$\begin{aligned} L(\mathbf{p}, \boldsymbol{\psi}, \boldsymbol{\phi}, N_{ab}) \propto & \prod_{k\in\{0,1\}} \prod_{l\in\{0,1\}} \prod_{d\in\{a,ab\}} (Q^A_{kld})^{x^A_{kld}} \\ & \times \prod_{k\in\{0,1\}} \prod_{l\in\{0,1\}} \prod_{d\in\{b,ab\}} (Q^B_{kld})^{x^B_{kld}} \\ & \times \prod_{k\in\{0,1\}} \prod_{d\in\{a,ab\}} (Q^A_{kMd})^{x^A_{kMd}} \\ & \times \prod_{l\in\{0,1\}} \prod_{d\in\{a,ab\}} (Q^A_{Mld})^{x^A_{Mld}} \\ & \times \prod_{k\in\{0,1\}} \prod_{d\in\{b,ab\}} (Q^B_{kMd})^{x^B_{kMd}} \\ & \times \prod_{l\in\{0,1\}} \prod_{d\in\{b,ab\}} (Q^B_{Mld})^{x^B_{Mld}}, \qquad (2) \end{aligned}$$

where $\boldsymbol{\psi}$ is the vector of ψ^A_{kld}'s and ψ^B_{kld}'s and $\boldsymbol{\phi}$ is the vector of ϕ^A_{kld}'s and ϕ^B_{kld}'s.

The expression in (2) is for the most general model, in which both surveys are longitudinal and both have missing data at each time period. If frame A uses a rotating panel survey, for example, then all of the probabilities Q^A_{kld} are nonzero: the units in the panels measured at both time periods will be included in the estimators x^A_{kld} for $k, l \in \{0, 1\}$, the units in the panels leaving the survey after time 1 will be included in the estimators x^A_{kMd}, and the units in the incoming panels will be included in the estimators x^A_{Mld}. Depending on the structure of the surveys, some of the factors in (2) may be omitted. For example, if the survey from frame B is a repeated cross-sectional survey with small sampling fraction, the probabilities Q^B_{kld} for $k, l \in \{0, 1\}$ will be close to zero, and we would omit those factors from the likelihood.

The likelihood in (2) can be written as a product of a factor with N_{ab} and a factor containing the remaining parameters. As a consequence, the MLE for N_{ab} is again the smaller root of the equation in (1). We discuss the estimators of the remaining parameters in the next section.

3.2.2 Model identifiability and reduced models

A problem with maximizing the likelihood in (2) is that under the general model there are a total of 42 parameters while the two samples have only 32 observed cell counts. Thus we cannot estimate all the parameters under the most general model. But we can consider models with reduced parameterizations, as done in Chen and Fienberg (1974) for single frame surveys. The dual frame situation, in fact, gives much more flexibility for modeling the missing data because of the independent information from the two samples about domain ab.

We first state conditions for a reduced model to be locally identifiable. Let $\boldsymbol{\theta}$ denote the s-vector of parameters of interest; in our case, $\boldsymbol{\theta}$ would include linearly independent components of $\mathbf{p}$, N_{ab}/N, and parameters for the missing data mechanism. In the likelihood in (2), the probabilities from the independent multinomial samples are Q^A_{kld} and Q^B_{kld}. These probabilities may be written as functions of $\boldsymbol{\theta}$, with $\mathbf{Q}^A(\boldsymbol{\theta}) = (Q^A_{00a}, \ldots, Q^A_{lMab})$ a g-vector of the nonzero Q^A_{kld} 's and $\mathbf{Q}^B(\boldsymbol{\theta}) = (Q^B_{00b}, \ldots, Q^B_{lMab})$ a q-vector of the nonzero Q^B_{kld} 's. When all cells in Table 2 and the analogous table for frame B have nonzero probabilities, $g = q = 16$. Let $\mathbf{D} = (\mathbf{D}'_A, \mathbf{D}'_B)'$ be the derivative matrix of the transformation, with $\mathbf{D}_{A(\alpha\beta)} = \partial \mathbf{Q}^A_{\alpha} / \partial \theta_{\beta}$ and $\mathbf{D}_{B(\delta\beta)} = \partial \mathbf{Q}^B_{\delta} / \partial \theta_{\beta}$ for $\alpha = 1, \ldots, g-1$, $\delta = 1, \ldots, q-1$, and $\beta = 1, \ldots, s$. Then, using Theorems 3, 4 and 5 in Catchpole and Morgan (1997), the model is locally identifiable if the matrix $\mathbf{D}$ is of full rank. The proof for the dual frame situation is given in Lu (2007).

In a dual frame survey, we consider two types of models for the missing data. In a Type (1) model, the probabilities of missing time-1 or time-2 information for cell (k, l) is the same for each domain within a frame, *i.e.*, $\phi^A_{kla} = \phi^A_{klab} = \phi_{klA}$, $\psi^A_{kla} = \psi^A_{klab} = \psi_{klA}$, $\phi^B_{klb} = \phi^B_{klab} = \phi_{klB}$ and $\psi^B_{klb} = \psi^B_{klab} = \psi_{klB}$. In this type of model, we estimate the ϕ 's and ψ 's separately from each sample. It might be considered when the samples from the two frames are collected using different modes. For example, if the frame A sample is a mail survey and the frame B sample is a cell phone survey, one might expect different probabilities of dropout from the two samples.

In a Type (2) model, the probabilities of having missing data are the same in each domain, *i.e.*, $\phi^A_{klab} = \phi^B_{klab} = \phi_{klab}$. This type of model might be considered when nonresponse is expected to be related to the cell membership, and frame membership is thought to have little effect on nonresponse. For example, if the two surveys have similar types of designs and administrative procedures, a Type (2) model might be appropriate.

For each type of model, we may need to place additional restrictions on the parameters in order to solve the likelihood equations. Following Stasny and Fienberg (1986) the following are possible restrictions:

$$\text{Model 1: } \phi_{kl} = \lambda_{t-1(l)}, \ \psi_{kl} = \lambda_{t(k)} \tag{3}$$

$$\text{Model 2: } \phi_{kl} = \lambda_{t-1}, \ \psi_{kl} = \lambda_{t}$$

$$\text{Model 3: } \phi_{kl} = \lambda_{l}, \ \psi_{kl} = \lambda_{k}$$

$$\text{Model 4: } \phi_{kl} = \lambda_{t-1(l)}, \ \psi_{kl} = \lambda_{t}$$

$$\text{Model 5: } \phi_{kl} = \lambda_{t-1}, \ \psi_{kl} = \lambda_{t(k)}.$$

Under model 1, the probability that an individual is a nonrespondent in a given time period depends on the given time period and the individual's classification in the observed time period. Under model 2, the probability that an individual is a nonrespondent in a given time period depends only on the given time period. Under model 3, the probability that an individual is a nonrespondent in a given time period depends only on the individual's classification in the observed time period. Under model 4, the probability that an individual is a nonrespondent at time 1 depends on that time period and the individual's classification in the observed month, and the probability that an individual is a nonrespondent at time 2 depends only on the time period 2. Under model 5, the probability that an individual is a nonrespondent at time 1 depends only on the time period, and the probability that an individual is a nonrespondent at time 2 depends on the time period and the individual's classification in the observed month. Many other models are possible in addition to these five models for each type. Using the derivative matrices, it is easily shown that Models 1-5 are all identifiable.

In general, we will not have closed form solutions for the parameter estimates and the parameters must be estimated using an iterative method. We use the function 'nlm' in R (www.r-project.org) to calculate parameter estimates; the code is available from the authors.

3.3 Estimators from complex samples

When either or both samples are collected with a complex design, using the cell counts directly in the likelihood in (2) will give estimators that are not design-consistent. Skinner and Rao (1996) used a pseudo-maximum likelihood (PML) method to obtain design-consistent estimators in cross-sectional dual frame surveys. They showed that, unlike the estimators of Hartley (1962) and Fuller and

Burmeister (1972), the PML estimators for different response variables used the same set of modified weights and thus were internally consistent.

We propose to study estimators inspired by the PML method for gross flows in dual frame longitudinal complex surveys that allow for missing data at either time period in either sample. The basic idea is to use a working assumption of a multinomial distribution from a finite population to give the form of the estimators and use a design effect to adjust the cell counts to reflect the complex survey design.

In the simple random sampling case, x^A_{kld}/n_A is a design-consistent estimator of Q^A_{kld}. To obtain a pseudo-likelihood for general sampling designs, we replace x^A_{kld}/n_A by $\hat{X}^A_{kld}/N_A$, a design-consistent estimator of Q^A_{kld} under the complex sampling design, in the likelihood (2). Define $\overline{x}^A_{kld} = \overline{n}_A \hat{X}^A_{kld}/N_A$ and $\overline{x}^B_{kld} = \overline{n}_A \hat{X}^B_{kld}/N_B$, where, following Skinner and Rao (1996), we allow $\overline{n}_A$ and $\overline{n}_B$ to be arbitrary constants. Note that if N_A or N_B is unknown, it may be estimated by $\hat{N}_A$ or $\hat{N}_B$ instead.

The pseudo-likelihood has the same form as (2), with x^A_{kld}, x^B_{kld}, n_A and n_B replaced by $\overline{x}^A_{kld}$, $\overline{x}^B_{kld}$, $\overline{n}_A$ and $\overline{n}_B$, respectively. Iterative procedures are then used to find the pseudo-MLEs of the quantities of interest p_{kld}, $\boldsymbol{\phi}$, $\boldsymbol{\psi}$ and N_{ab}. By the fact that the pseudo-likelihood factors, $\hat{N}_{ab}$ is found to be the smaller of the roots of

$$\begin{aligned}[\overline{n}_A + \overline{n}_B]\,\hat{N}^2_{ab,\,PML} & \\ -[\overline{n}_A N_B + \overline{n}_B N_A + \overline{n}_A \hat{N}^A_{ab} + \overline{n}_B \hat{N}^B_{ab}]\,\hat{N}_{ab,\,PML} & \\ +[\overline{n}_A \hat{N}^A_{ab} N_B + \overline{n}_B \hat{N}^B_{ab} N_A] = 0. & \qquad (4)\end{aligned}$$

In a complex survey, particularly when clustering is involved, the actual sample sizes n_A and n_B do not necessarily reflect the relative amounts of information from the samples. We thus suggest taking $\overline{n}_A$ and $\overline{n}_B$ to be the effective sample size for each sample, with $\overline{n}_A = n_A/$ (design effect of S_A) and $\overline{n}_B = n_B/$ (design effect of S_B). The design effect of an estimator $\hat{\mu}$ is the ratio

$$\frac{[V(\hat{\mu}) \text{ from complex survey design}]}{[V(\hat{\mu}) \text{ from SRS of same size}]}.$$

The design effect is usually different for different variables. For estimating gross flows, however, the only estimators used from the component surveys are estimated cell counts, and we might expect that in many surveys the design effects for the estimators $\hat{X}^A_{kld}$ would all be similar, and would also be similar to the design effect of the estimator $\hat{N}^A_{ab}$. We thus, as in Skinner and Rao (1996), suggest using the design effect for the estimator $\hat{N}^A_{ab}$ in determining $\overline{n}_A$, and the design effect for the estimator $\hat{N}^B_{ab}$ in determining $\overline{n}_B$. If the design effects of the other variables are indeed identical, then the resulting PMLEs will minimize the variances of the estimated quantities; if they differ, the PMLEs will not be optimal but they will be consistent and in most situations will be close to the optimal values (Lohr and Rao 2006). If the design effect for $\hat{N}^A_{ab}$ is unavailable, as would occur, for example, if the survey were poststratified to N^A_{ab}, then we suggest using a generalized design effect, computed by taking an average or weighted average of design effects from other variables in the survey.

4. Properties of the estimators

In this section, we will investigate properties of the estimators. We derive asymptotic variances, discuss jackknife variance estimators, and perform a small simulation study to explore the properties.

4.1 Properties

We consider the general case in which stratified multistage samples are taken from each frame. The estimators of population totals are the standard Horvitz-Thompson or Hájek estimators from complex surveys. From frame A, the parameter vector $\boldsymbol{\eta}_A = [(\mathbf{Q}^A)', N_{ab}/N_A]'$ is estimated by $\hat{\boldsymbol{\eta}}_A = [(\hat{\mathbf{Q}}^A)', \hat{N}^A_{ab}/N_A]'$, where $\hat{Q}^A_{kld} = \hat{X}^A_{kld}/N_A$; similarly, $\boldsymbol{\eta}_B = [(\mathbf{Q}^B)', N_{ab}/N_B]'$ is estimated by $\hat{\boldsymbol{\eta}}_B = [(\hat{\mathbf{Q}}^B)', \hat{N}^B_{ab}/N_B]'$ with $\hat{Q}^B_{kld} = \hat{X}^B_{kld}/N_B$.

Theorem 1: Let $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\eta}}'_A, \hat{\boldsymbol{\eta}}'_B)'$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}'_A, \boldsymbol{\eta}'_B)'$. Assume that the regularity conditions on the inclusion probabilities in Isaki and Fuller (1982) hold for each sample. Let $\tilde{n}_A$ and $\tilde{n}_B$ be the number of primary sampling units in frames A and B, respectively, and let $\tilde{n} = \tilde{n}_A + \tilde{n}_B$. Assume that $\tilde{n}_A$ and $\tilde{n}_B$ both increase such that $\tilde{n}_A/\tilde{n}_B \to \gamma$ for some $0 < \gamma < 1$. Then $\hat{\boldsymbol{\eta}}$ is consistent for $\boldsymbol{\eta}$, and

$$\tilde{n}^{1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}), \qquad (5)$$

where $\boldsymbol{\Sigma}$ is a block-diagonal matrix with blocks $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$, $\boldsymbol{\Sigma}_A$ is the asymptotic covariance matrix of $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_A$ and $\boldsymbol{\Sigma}_B$ is the asymptotic covariance matrix of $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_B$. If, in addition, it is assumed that $N_{ab}/N \to \kappa$ for some $0 < \kappa < 1$ and that the model is identifiable, then $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$, the parameter of interest, consists of components of $\mathbf{p}$, N_{ab}/N, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, and $\hat{\boldsymbol{\theta}}$ is the pseudo-maximum likelihood estimator of $\boldsymbol{\theta}$. Furthermore, $\tilde{n}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal with mean 0 and asymptotic variance $\mathbf{H}_A \boldsymbol{\Sigma}_A \mathbf{H}'_A + \mathbf{H}_B \boldsymbol{\Sigma}_B \mathbf{H}'_B$, where $\mathbf{H}_F$ is the derivative matrix of the function $\boldsymbol{\theta}$ with respect to the parameters $\boldsymbol{\eta}_F$ for frames $F \in \{A, B\}$.

Proof. With gross flows, observed values of all variables are 0 or 1. Thus the boundedness conditions in Lemmas 1 and 2 of Isaki and Fuller (1982) are met, and the estimators of frame A are consistent and asymptotically normal with

$$\tilde{n}_A^{1/2}(\hat{\boldsymbol{\eta}}_A - \boldsymbol{\eta}_A) \xrightarrow{d} N[0, (\gamma/(1+\gamma))\boldsymbol{\Sigma}_A].$$

The same argument applies to give consistency and asymptotic normality for the vector of estimators from frame B, with

$$\tilde{n}_B^{1/2}(\hat{\boldsymbol{\eta}}_B - \boldsymbol{\eta}_B) \xrightarrow{d} N[0, (1 - (\gamma/(1 + \gamma)))\, \boldsymbol{\Sigma}_B].$$

Combining these two asymptotic results, and using the independence of the sampling designs along with Slutsky's theorem, gives (5). The limiting distribution of $\tilde{n}^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ follows by the delta method, since the parameters in $\boldsymbol{\theta}$ are all twice continuously differentiable functions of those in $\boldsymbol{\eta}$. Since the parameter estimators cannot always be defined explicitly as a function of other statistics from the sample, we may derive the matrices $\mathbf{H}_A$ and $\mathbf{H}_B$ by linearizing the score equations (Binder 1983). The assumption that $N_{ab}/N \to \kappa \in (0, 1)$ guarantees that the linearization is well-defined.

Theorem 1 shows that linearization can be used to estimate the variances of parameters of interest. In many situations, however, the matrices $\mathbf{H}_A$ and $\mathbf{H}_B$ are high-dimensional and the linearized variance estimators have complex form. A practical way to estimate the variances of the estimators is to use the jackknife estimator proposed by Lohr and Rao (2000). Under the regularity conditions in their Theorem 4, the jackknife and linearization variance estimators are asymptotically equivalent. The form of the jackknife variance estimator is $v_{JK}(\hat{\boldsymbol{\theta}}) = v_A(\hat{\boldsymbol{\theta}}) + v_B(\hat{\boldsymbol{\theta}})$, where v_A is a jackknife estimator obtained by deleting one primary sampling unit at a time from frame A while using the full data set for frame B, and v_B is a jackknife estimator obtained by deleting one primary sampling unit at a time from frame B while using the full data set for frame A.

4.2 Simulation study

Theorem 1 shows that the dual frame estimators are consistent for the corresponding population quantities under the modeled missing data mechanism. We performed a small simulation study to investigate properties for moderate sample sizes with overlapping frames. We generated the data following the simulation study in Skinner and Rao (1996), with $\gamma_a = N_a/N$ and $\gamma_b = N_b/N$. A cluster sample from frame A was generated with $\tilde{n}_A$ psus and m observations in each psu, and a simple random sample of n_B observations was generated for frame B. We generated the clustered binary responses for the sample from frame A by generating correlated multivariate normal random vectors and then using the probit function to convert the continuous responses to binary responses.

After generating the sample, we calculated the estimators of the probabilities of the union of frame A and frame B, average of the absolute value of the bias and empirical mean squared error (EMSE) under different settings. The EMSE of a given estimator, $\hat{Y}$ is calculated as:

$$\text{EMSE} = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_r - Y)^2, \tag{6}$$

where $\hat{Y}_r$ is the value of $\hat{Y}$ for the r^{th} simulation run. In our simulation study, we used $R = 100$.

The simulation study was performed with factors: (1) γ_a: 0.2 or 0.4, (2) γ_b: 0.2 or 0.4, (3) clustering parameter ρ: 0.3, (4) missing data mechanism: the probability that an individual is a nonrespondent in a given month depends on the time period and the individual's classification in the observed period; or missing completely at random, (5) amount of missing data: close to 10% or close to 20%, (6) sample sizes: $\tilde{n}_A$: 10, 100 or 500; m: 5, n_B: 100, 1,000 or 5,000. All runs used probability parameters $\mathbf{p}_a$: (0.3, 0.1, 0.2, 0.4), $\mathbf{p}_{ab}$: (0.3, 0.1, 0.1, 0.5), and $\mathbf{p}_b$: (0.4, 0.1, 0.1, 0.4). Table 3 shows the results of the simulation study with missing data generated under Model 1 and fitted with both Model 1 and the model using complete records only.

Table 3
Results from the simulation study for missing data generated under Model 1. Case (1) fits the correct model: Model 1; Case (2) uses complete records only. Bias is the average absolute bias for the population gross flow proportions p_{kl}; EMSE is the average empirical mean squared error for the p_{kl}; the proportions used to generate the missing data are $\lambda_{(t-1)0} = 0.141$, $\lambda_{(t-1)1} = 0.070$, $\lambda_{(t)0} = 0.137$ and $\lambda_{(t)1} = 0.068$. Here, $\tilde{n}_A$ is the number of psus in sample A with psu size 5 and n_B is the number of elements in sample B

$\tilde{n}_A$	n_B		p_{00}	p_{01}	p_{10}	p_{11}
10	100	Estimator	0.311	0.120	0.149	0.420
Case 1		Bias	0.040	0.029	0.029	0.040
		EMSE	0.002	0.001	0.001	0.002
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{(t)}$
		Estimator	0.159	0.095	0.146	0.094
		EMSE	0.001	0.001	0.002	0.001
10	100	Estimator	0.286	0.120	0.146	0.448
Case 2		Bias	0.048	0.029	0.029	0.041
		EMSE	0.004	0.001	0.001	0.002
100	1,000	Estimator	0.321	0.092	0.138	0.449
Case 1		Bias	0.015	0.011	0.009	0.015
		EMSE	3.337e-04	1.798e-04	1.418e-04	3.256e-04
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{(t)}$
		Estimator	0.145	0.074	0.123	0.068
		EMSE	2.642e-04	9.389e-05	3.917e-04	8.206e-05
100	1,000	Estimator	0.293	0.092	0.135	0.480
Case 2		Bias	0.0280	0.011	0.010	0.040
		EMSE	0.001	1.839e-04	1.711e-04	0.002
500	5,000	Estimator	0.321	0.093	0.135	0.452
Case 1		Bias	0.006	0.008	0.007	0.012
		EMSE	4.960e-05	7.162e-05	6.381e-05	1.857e-04
			$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{(t)}$
		Estimator	0.140	0.071	0.123	0.064
		EMSE	4.466e-05	1.818e-05	2.288e-04	3.545e-05
500	5,000	Estimator	0.292	0.092	0.132	0.483
Case 2		Bias	0.028	0.008	0.008	0.043
		EMSE	8.265e-04	7.642e-05	9.571e-05	1.906e-03

When data are missing at random, all models give estimators of the gross flow proportions p_{kl} that are approximately unbiased so we do not report the results here. From Table 3, both the correct model and the analysis of complete records only produce biased estimators of the p_{kl}'s. With larger sample sizes, however, the bias persists in the analysis that uses complete records only, while it diminishes when Model 1 is fit. This example has relatively small probabilities of missing data. With larger amounts of missing data, the contrast between the estimators is more pronounced.

5. Application

In this section, we apply our results to data from the Survey of Income and Program Participation (SIPP) and the Current Population Survey (CPS) within Arizona. Both CPS and SIPP are longitudinal stratified multistage panel surveys. We treat SIPP and CPS as a dual frame survey with the same target population: the Arizona population 18 years old to 64 years old. Using information from both surveys, we want to model the transition probabilities of employment status changes from January 2001 to January 2002 of people between 18 years old and 64 years old. Note that, strictly speaking, these two surveys are not designed as a dual frame survey. They use different questions for the labor force variables. Although we recoded the variables according to the labor force definitions in CPS, it is possible that these different question wordings and orderings produce bias when combining the information. We use this as an example because a real longitudinal dual frame data is not available. Nevertheless, the example shows the potential gains in efficiency by combining the information from two surveys in estimating gross flows.

Both surveys have target population the noninstitutionalized civilian population of the United States. We consider a subset of the population: the population in the labor force from 18 years old to 64 years old. So $N_A = N_B = N_{ab}$ and the estimation problem is a special case of the theory given in Section 3. The longitudinal file for the 2001 and 2002 SIPP (Westat 2001) uses one panel. We merged Wave 1 (where January 2001 records are stored), Wave 4 (where January 2002 records are stored) and the longitudinal weight file, in which the weights are adjusted to sum to the population count. Since the longitudinal panel weights have been adjusted for the nonresponse, we consider this as a no missing data case. The resulted weighted gross flow table from SIPP is given in Table 4.

For the CPS, the rotation group design introduces partially classified data. January 2001 and January 2002 have 50 percent of the sample in common. We use these 50% of the data together with the partially classified data to perform the analysis. The weight variable we use is a cross-sectional weight with cross-sectional nonresponse and calibration adjustments (United States Census Bureau 2006). For individuals present in the survey for only one of the years, we use the weight from that year. For persons present in both Jan 2001 and Jan 2002, we use the average of the two weights. The rule that we chose the average of the two weights is to minimize the variance of the composite estimator. The population group we used is the 18-64 age group, and we excluded persons who were not in that category during both years. The weighted gross flow table from CPS is in Table 5.

Table 4
Gross flow table for SIPP, in Arizona

		Jan 2002		
		Employed	Unemployed	
January 2001	Employed	2,491,029	73,204	
	Unemployed	30,698	30,160	
				2,625,091

Table 5
Gross flow table for CPS, in Arizona

		January 2002			
		Employed	Unemployed	Missing	
January 2001	Employed	1,129,656	38,848	689,497	
	Unemployed	41,586	8,211	36,041	
	Missing	606,549	57,549		
					2,607,937

Since SIPP is considered as a no missing data case, we assumed $\phi_{kl} = \psi_{kl} = 0$ and use a Type 1 model in the data analysis. We adjusted each weight in the CPS data by the factor 2,625,091/2,607,937 to reach a single population total between the two time periods and a single population total between the two surveys. The number of observations in SIPP (frame A) after combining January 2001 and January 2002 are 551 and the design effect for unemployment is about 1.76, so $\bar{n}_A = 551/1.76 = 313$. The design effect for unemployment in CPS (frame B) is about 1.229, so $\bar{n}_B =$ 1,020/1.229 = 830. Because the likelihood factors, the estimated parameters of probabilities from the five models (3) are all the same. We list the estimated probabilities and the standard errors from SIPP, CPS and data combining these two surveys in Table 6.

Table 6
Estimated transition probabilities using SIPP, CPS, and the dual frame method with SIPP and CPS. Standard errors are given in parentheses

	p_{00}	p_{01}	p_{10}	p_{11}
SIPP	0.9489	0.0279	0.0117	0.0115
	(0.0124)	(0.0093)	(0.0061)	(0.0060)
CPS	0.9088	0.0454	0.0353	0.0106
	(0.0100)	(0.0072)	(0.0064)	(0.0035)
SIPP and CPS	0.9230	0.0381	0.0262	0.0127
	(0.0080)	(0.0058)	(0.0050)	(0.0030)

Due to confidentiality issues, no clustering information is available in the CPS public-use data sets. We used a product of the published design effect and the variance from multinomial sampling to estimate the variances from both SIPP and CPS data. The result from Theorem 1 was applied to estimate the variances of $\hat{p}_{kl}$ for $k, l = 0, 1$. In this special situation, the variance estimate from the combination of the two data sets is reduced to $(\overline{n}_A/(\overline{n}_A + \overline{n}_B))^2 V_A + (\overline{n}_B/(\overline{n}_A + \overline{n}_B))^2 V_B$, where V_A denotes the variance estimate from SIPP data and V_B denotes the variance estimate from CPS data. Table 6 shows that the standard errors are reduced by using the dual frame method.

We also performed goodness-of-fit tests, developed in Lu (2007), for the five models in (3). The parameter estimates from the five models and results from the goodness-of-fit tests, are listed in Table 7. All five models fit the data well, so we recommend adopting the simplest model, Model 3, for the data.

Table 7
Estimated parameters and results of goodness of fit tests

	Estimated Parameters				df	Corrected G^2	p-value
Model 1	$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	$\lambda_{t(0)}$	$\lambda_{t(1)}$	3	3.03	0.39
	0.246	0.395	0.277	0.302			
Model 2	λ_{t-1}	λ_t			5	8.58	0.12
	0.255	0.278					
Model 3	λ_0	λ_1			5	6.61	0.25
	0.262	0.353					
Model 4	$\lambda_{t-1(0)}$	$\lambda_{t-1(1)}$	λ_t		4	4.10	0.39
	0.246	0.397	0.278				
Model 5	λ_{t-1}	$\lambda_{t(0)}$	$\lambda_{t(1)}$		4	6.74	0.15
	0.255	0.277	0.313				

With the limited information available on the public-use data sets, we used simple weight adjustments to make the estimated population counts consistent with known totals. The SIPP and CPS weights in the data sets have already been calibrated and adjusted for nonresponse, so that the models for missing data mostly reflect the rotating panel design rather than attrition due to moving and other activities that might be related to employment status.

Future research on these models might include using different weighting adjustments for the longitudinal surveys. In addition, different parameters could be used to distinguish observations that have partial classification because of the rotating panel design, and observations that have partial classification because of nonresponse. To do so, we could introduce a Markov Chain model similar to the one proposed by Stasny (1987). In the complete data model, individuals are allocated to the table according to a single multinomial distribution. At the second step of the process, which is also unobserved, each individual may be chosen to either rotate out of the sample after the interview for month $t-1$ or rotate into the sample before the month t interview according to the sampling plan. Finally, in the third step of the process, each remaining individual may either lose its row classification or lose its column classification by other reasons. Using this model, we can model the nonresponse at both times (*i.e.*, lose both the row and the column classifications).

6. Conclusions

In this article, we developed statistical methods for estimating gross flows from dual frame surveys. These methods are necessary to estimate changes in poverty status or employment status over time. We developed pseudo-maximum likelihood estimators that use the dual frame structure and the properties of the two survey designs. Our models also account for effects of missing data when an individual drops out of the survey or when a rotation panel design is used, so they allow full use of partial information that may be provided by some households. We use a jackknife method to estimate the variance of estimators and examine the properties of the estimators. The results have been applied to real datasets.

In this paper, the categories of the gross flow tables are defined independently from the sample outcomes. It is also possible to define the categories based on values that depend on the sample. For example, in social surveys, the poverty line might be defined using a percentile from the sample and the categories defined as “Below the poverty line” and “Above the poverty line.” Methods from this paper can be used to estimate gross flows if the category definitions depend on the sample, but the variance estimators need to account for the effect of estimating the category boundaries.

Although the results in this paper are for dual frame surveys, the methods are general and could be extended to more than two surveys using PML estimators developed in Lohr and Rao (2006). As the number of frames increases, however, so does the complexity of possible missing data mechanisms. Misclassification error may also be more prevalent with a larger number of frames.

Our research is done in the context of survey sampling, but it also applies to other settings in which data could be combined from two independent sources. As it becomes increasingly difficult for a single survey to cover the entire population of interest, we believe these methods for estimating gross flows can provide better coverage of the population with less expense. They also allow for supplementing a general population survey with surveys of specific subpopulations of interest.

Acknowledgements

This research was partially supported by the National Science Foundation under grants SES-0604373 and

DLS-0909630. The authors thank the associate editor and referees for their insightful and helpful comments.

References

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.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Blair, E., and Blair, J. (2006). Dual frame web-telephone sampling for rare groups. *Journal of Official Statistics*, 22, 211-220.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association*, 63, 542-551.

Catchpole, E.A., and Morgan, B.J.T. (1997). Detecting parameter redundancy. *Biometrika*, 84, 187-196.

Chambers, R.L., Woyzbun, L. and Pillig, R. (1988). Maximum likelihood estimation of gross flows. *Australian Journal of Statistics*, 30, 149-162.

Chen, T., and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

Fuller, W.A., and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. In *Proceedings of the Social Statistics Section*, American Statistical Association, 245-249.

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

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, Series C, 36, 99-118.

Heeringa, S.G. (1995). Technical description of the assets and health dynamics (ahead) survey sample design. Technical Paper, Institute for Social Research, University of Michigan, hrsonline.isr.umich.edu/docs/userg/AHDSAMP.pdf.

Hocking, R.R., and Oxspring, H.H. (1971). Maximum likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, 66, 65-70.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression super-population model. *Journal of the American Statistical Association*, 77, 89-96.

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

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer-Verlag.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

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.

Lu, Y. (2007). Longitudinal estimation in dual frame surveys. *Ph.D Dissertation, Arizona State University*.

Pfeffermann, D., Skinner, C. and Humphreys, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society*, Series A, 161, 13-32.

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.

Stasny, E.A. (1984). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, 25-40.

Stasny, E.A. (1987). Some Markov-Chain models for nonresponse in estimating gross labor force flows. *Journal of Official Statistics*, 4, 359-73.

Stasny, E.A., and Fienberg, S.E. (1986). Estimating gross flows using panel data with nonresponse: An example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.

United States Census Bureau (2006). Current Population Survey: Design and Methodology. Technical Paper 66, U.S. Census Bureau, Washington, DC.

Verma, V., Betti, G. and Ghellini, G. (2007). Cross-sectional and longitudinal weighting in a rotational household panel: application to EU-SILC. *Statistics in Transition*, 8, 5-50.

Westat (2001). Survey of Income and Program Participation Users' Guide (Supplement to the Technical Documentation). Technical report, Washington, DC.