Journal of Official Statistics, Vol. 27, No. 4, 2011, pp. 633–650

Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys

Avinash C. Singh[1] *and Fulvia Mecatti*[2]

The available multiple frame estimation methods do not deal with the case of mixed frame level information where units from the same sample are allowed to have mixed information. That is, some units may have only basic (possibly due to privacy concerns or lack of memory on the part of the respondent) while others may have more than basic information, where basic is defined as having known selection probability for each unit from the sampled frame and the number of frames the unit could have been selected from but not knowing the frame identification except, of course, for the sampled frame. To address this new problem, we first propose a unified approach based on multiplicity-adjusted estimation which encompasses all the proposed estimators (classified in this article as either combined or separate) as well as new estimators obtained by combining simple and complex multiplicity estimators. We also propose hybrid multiplicity estimators to account for mixed information. The methods discussed here are limited to the combined frame approach only because of their ability to deal with the case of mixed information. Simulation results are presented to compare various methods in terms of relative bias and relative root mean squared error of point and variance estimators.

Key words: Basic, partial and full frame level information; multiplicity adjustments; separate and combined frame approaches; variance estimation.

1. Introduction

In a Multiple Frame (MF) survey, a set of at least two frames is used instead of a traditional single frame of units from the target population. Each frame by itself may or may not be complete but the union is assumed to be complete. The cost of sampling from different frames may vary quite a bit and typically the complete frame (if available) is more expensive to sample than incomplete frames. Even if this were not the case, it might not be economical or practical to create a single frame from multiple frames by removing duplicate units from overlapping parts. The main purpose of using MF surveys is to reduce

[1] Center for Excellence in Survey Research, NORC at the University of Chicago, Chicago, IL 60637, U.S.A. Email: singh-avi@norc.org

[2] Department of Statistics, University of Milan-Bicocca, Italy. Email: fulvia.mecatti@unimib.it

Acknowledgments: This work was presented as an invited paper at the First Italian Conference on Survey Methodology, Siena, Italy, June 10–12, 2009. The research was partially supported by an individual operating grant from NSERC of Canada to the first author held at Carleton University, Ottawa, under an Adjunct Research Professorship. The second author's research was also supported by an FAR grant from University of Milan-Bicocca. This research was conducted when the first author was at Statistics Canada, Ottawa, and sponsored the visit of the second author during the summers of 2007 and 2008. The authors would like to thank referees for their valuable comments.

cost while maintaining estimation efficiency almost at par with single frame surveys. They are also useful for *difficult-to-sample* populations (rare, elusive, or hidden), improving the target population coverage and response rates.

For the estimation problem of combining samples from multiple frames, there are several methods available in the literature that adjust the sampling weights for multiplicity so that no selected unit from overlapping frames is counted more than once. It is assumed that the frame level information for any selected unit is of three types: basic (known selection probability from the sampled frame and the number of frames from which the unit could have been selected but without the frame identification), partial (basic plus identification of the frames from which the unit could have been selected), and full (partial plus the selection probabilities from all the relevant frames). Moreover, units from the same sample are allowed to have mixed frame level information in that some may have only basic (possibly due to privacy concerns such as drug-use behavior or elusive-status behavior in the case of homeless, illegal immigration and ex-imprisonment, or memory lapse on the part of the respondent) while others may have more than basic. We also assume no classification error in the frame membership when it is reported.

It is interesting to note that if only basic frame level information is available for all sampled units, all available MF estimators become inapplicable except for the one recently proposed by Mecatti (2007) based on the idea of multiplicity counting rule of Sirken (1972) in network sampling and of Casady and Sirken (1980) in multiple frame sampling; see also Singh and Wu (1996), who used a dual frame simple multiplicity estimator as input for weight calibration. In this article, the Mecatti estimator will be termed as simple multiplicity (SM) because the multiplicity factor for each unit is simply constant for all frames to which the unit may belong. The class of MF estimators can be broadly classified into two types termed as separate frame and combined frame approaches, as explained in the next section. The separate frame approach requires basic or partial frame level information about sampled units, and encompasses methods of Hartley (1962; 1974), Fuller and Burmeister (1972), Skinner (1991), Skinner and Rao (1996), Singh and Wu (1996; 2003), Lohr and Rao (2006), and Mecatti (2007) among others. The combined frame approach, on the other hand, requires full frame level information about sampled units, and consists of methods, among others, proposed by Bankier (1986, to be termed as a nonmultiplicity (NM) estimator because it behaves like a single frame estimator without requiring any multiplicity adjustment); and Kalton and Anderson (1986), to be termed as proportional multiplicity (PM) because the multiplicity adjustment factor for each unit is proportional to the frame-specific sample inclusion probability of the unit. However, none of the above methods (except SM) deal with the case of mixed frame level information for units in the same sample.

In this article, we propose new estimators in the combined frame approach obtained by compositing simple and complex multiplicity estimators. We also propose hybrid multiplicity (HM) estimators under the combined frame approach to handle the case of mixed information. The methods discussed here are, however, limited mainly to the combined frame approach (except for SM) because of their ability to deal with the case of mixed information. An example of the application of HM might arise in a three-frame survey (e.g., three homeless shelter lists in a city), where we know the frame identification of sampled individuals who report the number (and not the identification of frame membership) as one or three, but not so for those reporting only two. For units having only basic information, the HM estimator uses the multiplicity factor as in SM. For units having full information it can use, in particular, the multiplicity factor as in PM.

To motivate and understand the proposed HM estimators in relation to other estimators, we first propose a unified formulation of MF estimators, termed as generalized multiplicity-adjusted Horvitz-Thompson (GMHT) estimators. This class encompasses all MF estimators (based on separate frame and combined frame approaches) listed above and includes, of course, Horvitz-Thompson (HT) estimators like the NM estimator which does not require multiplicity adjustment factors because it does not allow duplicate sampled units in the combined sample. The NM estimator does not preserve the identity of samples from different frames, but in effect combines them as a single frame sample (after deduplication of sampled units, if any) with corresponding inclusion probabilities adjusted for multiple selection from overlapping frames. The GMHT class of estimators is analytically simple and can be readily implemented for any number of frames. In addition, it allows for general HT-type unbiased variance estimation and Sen-Yates-Grundy form, in particular, for nonrandom sample sizes.

The organization of this article is as follows. Section 2 presents the proposed GMHT class of estimators while Section 3 shows how the existing and proposed methods under the combined frame approach and SM under the separate frame approach can be obtained as special cases of GMHT. In Section 4, a limited simulation study is performed to compare various estimators in terms of relative bias and relative root mean squared error of point and variance estimators. A number of different scenarios in a three-frame set-up are considered by varying the frame-overlapping pattern, the data generation model, and the population sizes. Three schemes of over- and under-sampling are considered, including the option of complex designs (e.g., simple random in one frame and probability proportional to size in others). The impact of small inclusion probabilities on precision of point and variance estimators is also explored. Finally, Section 5 contains concluding remarks.

2. Generalized Multiplicity-adjusted Horvitz-Thompson (GMHT) Class of Multiple Frame Estimators

Let $U_1 \cdots U_q \cdots U_Q$ denote the collection of frames whose union is assumed to cover the target population $U = \cup_q U_q$. The frames are generally overlapping in practice, and in fact some of them may be complete by themselves too. Independent samples $s_1 \cdots s_q \cdots s_Q$ from the Q frames are selected under possibly different designs which may be simple (simple random sample without replacement) or complex (stratified multistage cluster unequal probability sample, for example). We focus on the estimation of the population total $T_y = \sum_{i \in U} y_i$ of a study variable y. Estimation in an MF survey is essentially a problem of combining data from the Q samples. The target parameter T_y can be alternatively expressed as the sum over (possibly overlapping) frames

$$T_y = \sum_{i \in U} y_i = \sum_{q=1}^{Q} \sum_{i \in U_q} y_i \alpha_{q(i)} \tag{1}$$

where $\alpha_{q(i)} \in [0, 1]$, in general but not necessarily, and $\sum_q \alpha_{q(i)} = 1$, are the multiplicity adjustment factors corresponding to frames q for each unit i and sum to one to ensure that the unit is not counted more than once. Note that the simplest choice of $\alpha_{q(i)}$ is inverse of the unit multiplicity $m_i = \#(U_q \ni i)$ although in general it may depend on first and second order inclusion probabilities under the MF design. Next introduce an observable random variable $\delta_{i(q)}$ for each unit i from frame q under the design-randomization; a common choice of which, analogous to the HT estimation, is $\delta_{i(q)} = \mathbf{1}_{i \in s_q}$ i.e., the sample membership indicator of unit i in the sample s_q from frame U_q. Note that the subscript order in $\alpha_{q(i)}$ helps to interpret it as being defined for each frame q that contains the unit i, while the reversed subscript order in $\delta_{i(q)}$ helps to interpret it as being defined for each unit i that is contained in the frame q.

Now the Generalized Multiplicity-adjusted Horvitz-Thompson (GMHT) class of MF estimators corresponding to various choices of (α, δ) is defined by

$$t_{y(GMHT)} = \sum_{q=1}^{Q} \sum_{i \in U_q} y_i \alpha_{q(i)} \frac{\delta_{i(q)}}{E(\delta_{i(q)})} \tag{2}$$

which is design-unbiased by construction and clearly analytically simple regardless of the number Q of frames. The HT estimator is a special case when $Q = 1$ and $\alpha_{q(i)} = 1$. Observe that the GMHT estimators are linear combinations of independent HT-type estimators, and therefore the exact design-based variance is easily obtained as

$$Var(t_{y(GMHT)}) = \sum_{q=1}^{Q} \left\{ \sum_{i \in U_q} z_{i(q)}^2 Var(\delta_{i(q)}) + \sum_{i \neq j} \sum_{\in U_q} z_{i(q)} z_{j(q)} Cov(\delta_{i(q)}, \delta_{j(q)}) \right\} \tag{3a}$$

which for fixed sample designs reduces to

$$Var(t_{y(GMHT)}) = \sum_{q=1}^{Q} \sum_{i<j} \sum_{\in U_q} - Cov(\delta_{i(q)}, \delta_{j(q)}) (z_{i(q)} - z_{j(q)})^2 \tag{3b}$$

where

$$z_{i(q)} = \frac{y_i \alpha_{q(i)}}{E(\delta_{i(q)})}$$

collects all the nonrandom components of $t_{y(GMHT)}$. The Sen-Yates-Grundy form of unbiased variance estimators for fixed sample size designs in all frames, assuming positive joint design-expectation, $E\left(\delta_{i(q)} \delta_{j(q)}\right) > 0$ for all pair of units $i \neq j \in U_q$, $q = 1 \cdots Q$, is given by

$$v(t_{y(GMHT)}) = \sum_{q=1}^{Q} \sum_{i<j} \sum_{\in U_q} \frac{-Cov(\delta_{i(q)}, \delta_{j(q)})}{E(\delta_{i(q)} \delta_{j(q)})} (z_{i(q)} - z_{j(q)})^2 \delta_{i(q)} \delta_{j(q)} \tag{4}$$

Note that $Cov(\delta_{i(q)}, \delta_{j(q')}) = 0$ for independent frame samples. For $\delta_{i(q)} = \mathbf{1}_{i \in s_q}$ Equations (3b) and (4) reduce to the usual variance formulae for the sum of independent

HT-type estimators for fixed sample sizes and are given by

$$Var\left(t_{y(GMHT)}\right)=\sum_{q=1}^{Q}\sum_{i<j}\sum_{\in U_q}\left(\pi_{i(q)}\pi_{j(q)}-\pi_{ij(q)}\right)\left(\frac{y_i\alpha_{q(i)}}{\pi_{i(q)}}-\frac{y_j\alpha_{q(j)}}{\pi_{j(q)}}\right)^2 \tag{5}$$

and

$$v\left(t_{y(GMHT)}\right)=\sum_{q=1}^{Q}\sum_{i<j}\sum_{\in s_q}\frac{\left(\pi_{i(q)}\pi_{j(q)}-\pi_{ij(q)}\right)}{\pi_{ij(q)}}\left(\frac{y_i\alpha_{q(i)}}{\pi_{i(q)}}-\frac{y_j\alpha_{q(j)}}{\pi_{j(q)}}\right)^2 \tag{6}$$

Finally, for the special case $\delta_{i(q)}=\mathbf{1}_{i\in s_q}$ and simple random sampling in each frame, we have $E(\delta_{i(q)})=n_q/N_q$ ($=f_q$, the sampling fraction from frame q) for all $i\in U_q$, $E(\delta_{i(q)}\cdot\delta_{j(q)})=f_q(n_q-1)/(N_q-1)$, for all $i\neq j\in U_q$, and Equations (3) and (4) simplify to expressions without the double sum:

$$Var\left(t_{y(GMHT)}\right)=\sum_{q=1}^{Q}N_q^2\frac{\left(1-f_q\right)}{n_q}\frac{1}{N_q-1}\left[\sum_{i\in U_q}y_i^2\alpha_{q(i)}^2-N_q\left(\frac{1}{N_q}\sum_{i\in U_q}y_i\alpha_{q(i)}\right)^2\right] \tag{7}$$

and

$$v\left(t_{y(GMHT)}\right)=\sum_{q=1}^{Q}N_q^2\frac{\left(1-f_q\right)}{n_q}\frac{1}{n_q-1}\left[\sum_{i\in s_q}y_i^2\alpha_{q(i)}^2-n_q\left(\frac{1}{n_q}\sum_{i\in s_q}y_i\alpha_{q(i)}\right)^2\right] \tag{8}$$

The GMHT class encompasses all the MF estimators available in the literature by suitably specifying $\alpha_{q(i)}$ and $\delta_{i(q)}$. The specification of $\alpha_{q(i)}$ depends on the available frame level information for each sampled unit. As mentioned in the introduction, this information could be basic, partial or full. In the case of *basic* frame level information for a sampled unit, we have

1. The unit multiplicity m_i for the sampled unit, that is, the number of frames in which the unit appears, where $i\in s_q$ and $q=1\cdots Q$; here the frame identification is not assumed to be available except, of course, for the frame from which it was actually sampled.
2. The inclusion probability $\pi_{i(q)}=P(s_q\ni i)$ for the sampled unit only for the frame(s) from which it was actually sampled, where $i\in s_q$ and $q=1\cdots Q$.

It is assumed that the basic information is available for all sampled units. In the case of *partial* frame level information for a sampled unit, we have

1. Identification of all frame membership for the sampled unit, that is, all the frames the sampled unit could have come from, where $i\in s_q$ and $q=1\cdots Q$;
2. The inclusion probability $\pi_{i(q)}=P(s_q\ni i)$ for the sampled unit only for the frame(s) from which it was actually sampled, where $i\in s_q$ and $q=1\cdots Q$.

Clearly, having partial information implies basic plus frame identification. In the case of *full* frame level information for a sampled unit, we have

1. Identification of all frame membership for the sampled unit, that is, all the frames the sampled unit could have come from, where $i \in s_q$ and $q = 1\cdots Q$;
2. The inclusion probabilities $\pi_{i(q)} = P(s_q \ni i)$ for the sampled unit for all the frames from which it could have been sampled (regardless of which frame it was actually sampled), where $i \in s_q$ and $q = 1\cdots Q$.

Clearly having full information implies partial information plus sample inclusion probabilities from all the applicable frames. We also consider the case where all units in the same sample may not have identical frame level information. This leads to the case of mixed frame level information which may arise, as mentioned in the introduction, when dealing with respondents' privacy concerns or sensitivity of information about frame membership.

For the proposed composite multiplicity estimators (next section), it would be useful to define a partition of the target population U into disjoint domains and an alternative expression of the GMHT class as follows. The knowledge of frame membership allows the classification of each sampled unit into $2^Q - 1$ disjoint domains $U_K = \left(\bigcap_{q\in K} U_q\right) \cap \left(\bigcap_{q\notin K} U^c_q\right)$ following the MF notation of Lohr and Rao (2006), where $K \subseteq \{1, \cdots, q, \cdots, Q\}$ is an index set denoting an unordered subset of the collection of frame indices (except $\varnothing$). In Figure 1 the domain classification and the index sets $K \subseteq \{1, \cdots, q, \cdots, Q\}$ are exemplified for the case of three frames.

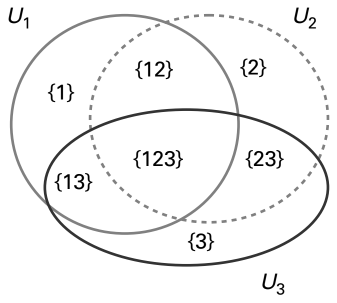

Fig. 1. Classification into seven domains for the three-frame set up

Each population unit is included in a unique single domain, that is, $i \in U_K$ for some K, to be denoted by $U_{K(i)}$ to indicate that it corresponds to the unit i. Now an alternative expression for the GMHT class is given by

$$t_{y(GMHT)} = \sum_{q=1}^{Q} \sum_{i\in U_q} y_i \alpha_{q(i)} \frac{\delta_{i(q)}}{E\left(\delta_{i(q)}\right)} = \sum_{i\in U} y_i \sum_{q\in K(i)} \alpha_{q(i)} \frac{\delta_{i(q)}}{E\left(\delta_{i(q)}\right)} = \sum_{i\in U} y_i \psi_i \qquad (9)$$

where $\psi_i = \sum_{q\in K(i)} \alpha_{q(i)} \frac{\delta_{i(q)}}{E\left(\delta_{i(q)}\right)}$, and $E\left(\psi_i\right) = \sum_{q\in K(i)} \alpha_{q(i)} = 1$ ensures design-unbiasedness

Remarks (*Separate and Combined Frame Approaches*): As mentioned in the introduction, all the available MF estimators can be classified into two approaches depending on what frame level information is needed to compute the multiplicity adjustment factors $\alpha_{q(i)}$. If only basic (i.e., number of frames from which the unit could have been sampled but only identification and inclusion probability for the frame from which the unit was actually sampled) or partial information (i.e., basic information plus identification of all the frames from which the unit could have been sampled) is required, then the estimator is classified under the separate frame approach. On the other hand, if full information (i.e., partial information plus inclusion probabilities for all the frames from which the unit could have been sampled) is required, then the estimator is classified under the combined frame approach. The SM estimator belongs to the separate frame approach because it only requires the basic information. The separate frame approach may be preferable in practice to the combined frame approach because there is more likelihood of having partial than full information for all sampled units.

For the separate frame approach, the multiplicity adjustment factor $\alpha_{q(i)}$ is common to all units belonging to the same domain classification and depends only on the domain $K(i)$ as defined above for each unit i. It could also depend on the study variable and the underlying sampling designs of multiple frames (first and second order inclusion probabilities) through variance minimization of the linear combination of estimators obtained from different samples for the common domain or the overlapping frame. Thus, for each frame q intersecting with the domain $K(i)$, the adjustment factor $\alpha_{q(i)}$ depends only on q and can be obtained using optimal regression (for example) as inversely proportional to the variance of the corresponding estimator under the constraint $\sum_q \alpha_{q(i)} = 1$, which implies that each factor is between 0 and 1 (here the assumption of independent samples from different frames for the common domain is used). The above optimal regression approach in the case of two frames was used in the pioneering papers of Hartley (1962; 1974) and later by Fuller and Burmeister (1972), who also included, for the sake of increased efficiency, an extra predictor defined by the difference of the estimated population counts from the two samples from the common domain. The Fuller-Burmeister estimator can be viewed as a GMHT estimator because it is equivalent to performing regression on the extra predictor after obtaining the multiplicity adjusted estimator; that is calibration following GMHT estimation. Although the above GMHT-type estimators were developed from the initially calibrated (such as raking-ratio adjusted) estimators from each frame, the alternative of first defining a GMHT estimator using the initial uncalibrated weights and then performing calibration on multiplicity-adjusted weights might be more appealing as it generalizes in a natural way the calibration of single frame HT estimators to multiple frame GMHT estimators. This topic will be addressed in a separate paper.

The dual frame estimators of Skinner (1991) for simple random samples and of Skinner and Rao (1996) for complex designs are also based on regression estimation, but for combining estimates of the mean for the common domain instead of totals, although the ultimate goal is to estimate totals. In this case, the estimator of the common domain mean can be expressed as a ratio of GMHT estimators of the study variable total in the numerator and the domain subpopulation count in the denominator. It follows that for estimating domain totals, the above estimator can be expressed as a GMHT estimator by noting that the multiplicity adjustment factors involve as a denominator the GMHT estimator of the common domain subpopulation count. To avoid the problem of having to define the multiplicity adjustment factor for each study variable as encountered by the estimators of

Hartley and Fuller-Burmeister, Skinner-Rao suggested a compromise using the value corresponding to the counting variable in estimating the common domain population count – an important extension of which for the case of multiple frames was provided by Lohr and Rao (2006). The estimator proposed by Singh and Wu (1996), on the other hand, provides a generalized regression alternative to the optimal regression approach in estimating totals for multiple frames in order to avoid the well-known problem of instability of the optimal regression estimator in the presence of many auxiliary predictors used for calibration. For this purpose, the simple multiplicity estimators for common domains were used as the initial estimators before calibration. Later, Singh and Wu (2003) proposed an optimal modification (to render it somewhat closer to the optimal regression but without the problem of instability) involving choosing the combining coefficients (or multiplicity adjustment factors) that minimize the generalized variance over a set of key study variables.

Having provided above a brief review of methods under the separate frame approach, it is observed that this approach (except for SM) is generally not applicable to the case of mixed frame level information (i.e., when some units in the sample may provide only basic information while others may have more information) which is one of the focal points of this article. It is for this reason that we limit our study in this article to only the combined frame approach except for SM of the single frame approach, a detailed discussion of which is given in the next section.

3. Existing and Proposed Methods Based on Basic or Full Frame Level Information

In this section, in order to propose new estimators based on mixed frame level information, we review SM requiring basic information under the separate frame approach and other existing methods requiring full information under the combined frame approach. We first show how all of them belong to the GMHT class and then propose new estimators including ones that deal specifically with the case of mixed frame level information.

3.1. Proportional Multiplicity (PM) Adjusted Estimator (Kalton and Anderson 1986)

The Kalton-Anderson estimator requires full information. It belongs to the GMHT class and can be expressed as

$$t_{y(PM)} = \sum_{q=1}^{Q}\sum_{i\in U_q} y_i\alpha_{q(i)}\frac{\mathbf{1}_{i\in s_q}}{\pi_{i(q)}} = \sum_{q=1}^{Q}\sum_{i\in U_q}\frac{y_i\mathbf{1}_{i\in s_q}}{m_i\bar{\pi}_i},\ \bar{\pi}_i = m_i^{-1}\sum_{q'\in K(i)}\pi_{i(q')} \tag{10a}$$

where $\alpha_{q(i)} = \pi_{i(q)}/\sum_{q'\in K(i)}\pi_{i(q')}$ (hence the name PM; i.e., $\alpha_{q(i)}$ is proportional to $\pi_{i(q)}$) and $\delta_{i(q)} = \mathbf{1}_{i\in s_q}$. It can also be equivalently expressed as

$$t_{y(PM)} = \sum_{i\in U} y_i\psi_{i,PM},\ \text{where } \psi_{i,PM} = \sum_{q\in K(i)}\alpha_{q(i)}\frac{\mathbf{1}_{i\in s_q}}{\pi_{i(q)}} \tag{10b}$$

Formulae (3) and (4) for the variance and the variance estimator are directly applicable by setting $z_{i(q)} = y_i/\sum_{q'\in K(i)}\pi_{i(q')}$. Under simple random sampling with proportional allocation to all frames, we have $\pi_{i(q)} = f$ for all q in $K(i)$, and hence $\sum_{q \in K(i)} \pi_{i(q)} = fm_i$ where m_i denotes the cardinality of the set $K(i)$ and f is the common sampling fraction over all frames. Hence for this sampling scheme, the PM estimator coincides with the SM estimator of Subsection 3.2 defined below. In fact, this is true more generally whenever $\pi_{i(q)} = \pi_i$.

3.2. *Simple Multiplicity (SM) Adjusted Estimator (Mecatti 2007)*

The Mecatti estimator requires only basic information and leads to the following estimator

$$t_{y(SM)} = \sum_{q=1}^{Q} \sum_{i \in U_q} \frac{y_i \mathbf{1}_{i \in s_q}}{m_i \pi_{i(q)}} \tag{11a}$$

which can also be equivalently expressed as

$$t_{y(SM)} = \sum_{i \in U} y_i \psi_{i,SM}, \text{ where } \psi_{i,SM} = \sum_{q \in K(i)} \frac{1}{m_i} \frac{\mathbf{1}_{i \in s_q}}{\pi_{i(q)}} \tag{11b}$$

The SM estimator is clearly a member of the GMHT class with $\alpha_{q(i)} = m_i^{-1}$ and $\delta_{i(q)} = \mathbf{1}_{i \in s_q}$ so that $E\left(\delta_{i(q)}\right) = \pi_{i(q)}$. Formulae (3) and (4) for variance and variance estimation are easily applicable with $z_{i(q)} = y_i / \left(m_i \pi_{i(q)}\right)$.

3.3. *Composite Multiplicity (CM) Adjusted Estimator (Proposed)*

We observe from (11a) that although the SM estimator requires less information, it may be unstable (in the sense of high coefficient of variation of point and variance estimators) when some units may have very small inclusion probabilities–this is likely when dealing with large populations. On the other hand, the PM estimator requires more information, but is expected to be more stable because the average of inclusion probabilities for each unit over the frames is in the denominator of (10a). However, since the choice of $\alpha_{q(i)}$ is not based on optimality considerations, it would be of interest to consider alternatives in the GMHT class. In particular, assuming the availability of full information, we consider a composite of the two such as $\lambda_i \psi_{i,SM} + (1 - \lambda_i)\psi_{i,PM}$ minimizing the variance for a suitable choice of $\lambda_i \in \left(0, 1\right)$ as shown in Appendix A. The resulting estimator, termed composite multiplicity (CM), is obtained as

$$t_{y(CM)} = \sum_{q=1}^{Q} \sum_{i \in U_q} y_i \alpha_{q(i)}^{CM} \frac{\mathbf{1}_{i \in s_q}}{\pi_{i(q)}} \tag{12}$$

where

$$\alpha_{q(i)}^{CM} = \lambda_i^{CM} \frac{1}{m_i} + \left(1 - \lambda_i^{CM}\right) \frac{\pi_{i(q)}}{m_i \bar{\pi}_i} \tag{13a}$$

and

$$\lambda_i^{CM} = \frac{\sum_{q \in K(i)} \left(1 - \pi_{i(q)}^{-1} \bar{\pi}_i\right) \pi_{i(q)} \left(1 - \pi_{i(q)}\right)}{\sum_{q \in K(i)} \left(1 + \pi_{i(q)}^{-2} \bar{\pi}_i^2 - 2\pi_{i(q)}^{-1} \bar{\pi}_i\right) \pi_{i(q)} \left(1 - \pi_{i(q)}\right)} \tag{13b}$$

Formulae (3) and (4) for variance and variance estimation remain easily applicable. In the case of simple random sampling with proportional allocation to all frames, that is, when $\pi_{i(q)} = f$ for all $i \in U_q$, $q = 1 \cdots Q$, we have from (13b) $\lambda_i^{CM} = 0$; which implies that all the three estimators, PM, SM, and CM become identical. As an alternative to the above CM, we also considered the combination $\sum_{q \in K(i)} \alpha^{*}_{q(i)} \mathbf{1}_{i \in s_q} / \pi_{i(q)}$ such that the variance is minimized. This gives rise to $\alpha^{*}_{q(i)}$ being proportional to the inverse of the variance of $\mathbf{1}_{i \in s_q} / \pi_{i(q)}$, which is $\left(\pi_{i(q)}^{-1} - 1\right)$. The resulting estimator, however, tends to be more unstable than PM (mainly because the coefficient of variation of $\alpha^{*}_{q(i)}$ turns out to be higher than that of $\alpha^{PM}_{q(i)}$) and was not considered further. This is possible since in the composite multiplicity estimation, instead of minimizing the variance of an overall estimator for the study variable, we are simply optimizing for the multiplicity factor from different frames for each unit *i*.

3.4. Hybrid Multiplicity (HM) Adjusted Estimators (Proposed)

Going back to the case of mixed frame level information not covered by estimators considered so far except for SM, it would be of interest to explore whether SM could be improved upon. Let U_{full} denote a domain for which full information would be available for any unit if it were selected, and U_{basic} denote a domain for which we would have only basic information; typically for units in U_{basic}, $2 \le |K(i)| \le Q - 1$ and otherwise for units in U_{full}. A Hybrid Multiplicity (HM1) estimator based on SM and PM, for example, can be defined as

$$t_{y(HM1)} = \sum_{q=1}^{Q} \sum_{i \in U_q} y_i \alpha_{q(i)}^{HM1} \frac{\mathbf{1}_{i \in s_q}}{\pi_{i(q)}} \tag{14}$$

where $\alpha_{q(i)}^{HM1} = \left(\alpha_{q(i)}^{PM} \cdot \mathbf{1}_{i \in U_{full}} + \alpha_{q(i)}^{SM} \cdot \mathbf{1}_{i \in U_{basic}}\right)$. Clearly HM1 belongs to the GMHT class and SM is a special case of it. Similarly, HM2 can be defined as a hybrid of SM and CM. The HM estimators are expected to improve upon the possible instability of SM when the inclusion probabilities of some of the units in U_{full} could be very small.

3.5. Non-Multiplicity (NM) Adjusted Estimator (Bankier 1986)

As mentioned in the introduction, the Bankier estimator of the combined frame approach belongs trivially to the GMHT class as it behaves like an HT estimator. The reason is that this estimator does not require multiplicity adjustment factors because it does not allow duplicate sampled units in the combined sample. It follows that besides requiring full information, it also requires extra information at the estimation stage in order to be able to identify duplicate sampled units. Using the notation of the GMHT class, it is possible to express the NM estimator in a simple form for any number of frames. First, different frame samples $s_1, \cdots, s_q, \cdots, s_Q$ are combined into a single sample s* (with random size n^*) of *distinct* units by first classifying sampled units into disjoint domains U_K, and then discarding duplicate units if any. Thus, for any population unit i, let the indicator variable that the sampled unit is in at least one of the frames $q \in K(i)$ be $\delta_i = \left[1 - \coprod_{q' \in K(i)} \left(1 - 1_{i \in s_{q'}}\right)\right]$. Therefore, the corresponding expectation or the probability of being selected in at least one of the frames is given by

$$E(\delta_i) = \left[1 - \prod_{q \in K(i)} \left(1 - \pi_{i(q)}\right)\right] \tag{15}$$

The NM estimator can be defined as

$$t_{y(NM)} = \sum_{i \in U} \frac{y_i \delta_i}{E(\delta_i)} = \sum_{i \in s^*} \frac{y_i}{1 - \prod\limits_{q \in K(i)} \left(1 - \pi_{i(q)}\right)} \tag{16}$$

For the above estimator, being HT (a rather special case of GMHT), the variance and its estimate can be derived as shown in Appendix B. Note that the Sen-Yates-Grundy form of the variance estimator is not applicable because of the sample size n^* being random. It may be of interest to note that if the product $\pi_{i(q)}\pi_{i(q')}$ of inclusion probabilities is very small (this could happen if at least one of the two is small enough), the sample is not likely to have any duplicate units, in which case the NM estimator is approximately equivalent to the PM estimator.

4. Simulation Study

To compare the six estimators (PM, SM, CM, HM1, HM2, and NM) described in the previous section, a simulation study was designed involving three frames with simple random sampling in one frame and unequal probability sampling in the remaining two; in particular, Rao-Sampford (Rao 1965; Sampford 1967) probability proportional to size (πps) was used. The πps sampling was introduced by using an auxiliary variable x designed to be approximately proportional to the study variable y and was generated by inverting the model $y = 5x + N\left(0, 1\right)$. The population y-values were generated under six different models as shown in Figure 2. With three frames, there are a total of seven disjoint domains (not all of them necessarily nonempty) partitioning the target frame made up of possibly different patterns of coverage (N_q/N) and overlap (N_K/N) among the three frames. Figure 3 shows twelve patterns of coverage and overlap that were considered such that

$$\text{Coverage} : 1 \leq \sum\nolimits_{q=1}^{Q} N_q/N \leq 3 \,\text{and Overlap} : \; 0 \leq \sum_{|K| \geq 2} N_K/N \leq 1$$

Figure 4 presents a visual display of examples of different patterns of coverage and overlap. Note that the measures of coverage and overlap are positively correlated, and therefore the selected patterns of interest are concentrated around a line sloping upward.

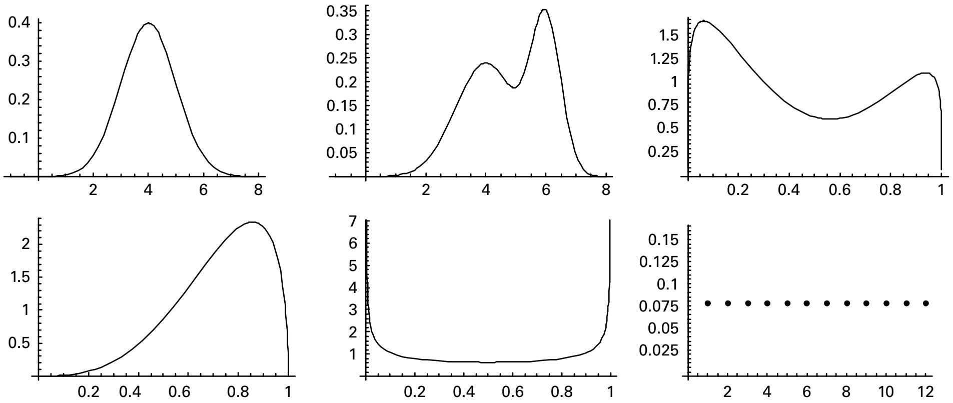

Fig. 2. Six Different models used to simulate y-data

The choice of the population size N was limited to 50, 100, and 120 to reduce the computational burden. For a fixed overall sample size, $n = \sum_q n_q$ we considered three sample allocation schemes for the three frames:

i) Proportional allocation by selecting 10% in each frame, that is, $f_q = n_q/N_q = 0.10$, $q = 1, 2, 3$

ii) A disproportional allocation by selecting a sample of constant size $n_q = n/3$ in all the three frames;

iii) Another type of disproportional allocation by over-sampling in one frame (f_q around 20%) and under-sampling in the remaining two (f_q around 5%).

In all, we considered a total of 32 scenarios for each of the above three sample allocation schemes, out of which twelve scenarios correspond to two choices of coverage/overlap

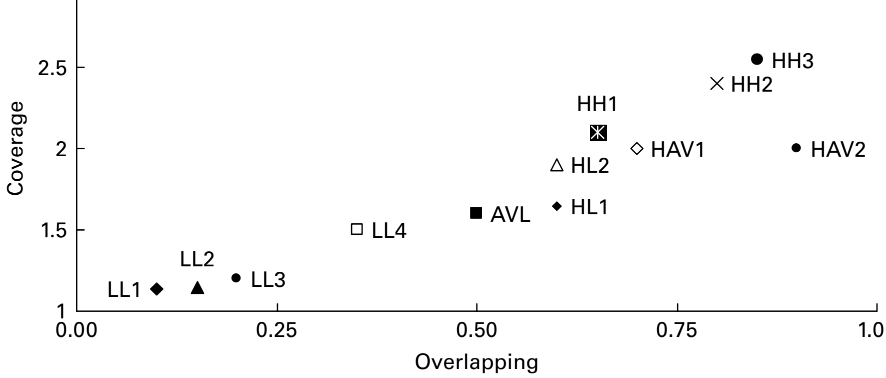

Fig. 3. Twelve Different patterns in coverage and overlap for Multiple Frames. LL: Low overlap (under 0.5) and Low coverage (under 2); AVL: Average overlap (0.5) and Low coverage ($<$ 2); HL: High overlap ($>$ 0.5) and Low coverage ($<$ 2); HAV: High overlap ($>$ =0.5) and Average coverage ($\sim$ 2); HH: High overlap and High coverage.

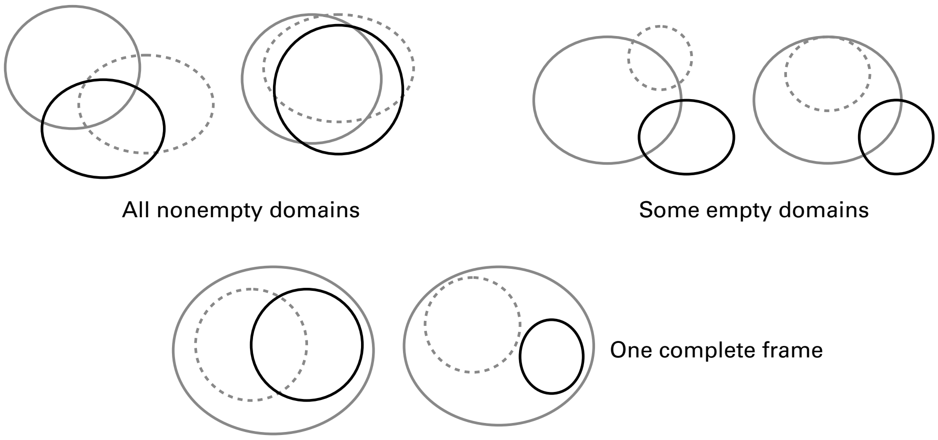

Fig. 4. Cases of some empty domains and presence of one complete frame in a three-frame set-up

pattern (LL1, HL1), six different y-models, and $N = 100$; eight scenarios (actually one scenario out of nine was dropped due to its demanding computational time) correspond to three patterns (LL1, LL3, LL4), one y-model (asymmetric unimodal Beta), and $N = 50$, 100, and 120; and the remaining twelve correspond to each of the twelve patterns of coverage and overlap with the y-model chosen as the asymmetric unimodal Beta, and N at 100. The above choice of different sample allocations was somewhat ad hoc but was motivated from realistic scenarios which depend on field implementation and cost considerations. In practice, however, a cost-variance optimization is usually performed to obtain suitable sample allocations to different frames.

For the empirical study considered here, it is assumed that we do not have frame identification whenever the frame multiplicity as reported by the respondent is two out of a total of three. However, if it is three, then we do have the frame identification because the total number of possible frames considered here is three. If it is one, then by default, the frame identification is known. As mentioned earlier, with mixed frame level information, only three methods, SM, HM1, and HM2, out of six are applicable. However, other methods (PM, CM, and NM) belonging to the combined frame approach are also included for the sake of interest although they require full information. For evaluating the performance of various point estimators, we consider two measures RB and RRMSE as defined below.

1. (Monte Carlo) Relative Bias: $RB\big(E_{mc}(t_y) - T_y\big)/T_y$ as a measure of simulation accuracy since all the estimators are unbiased, and
2. Monte Carlo Relative Root Mean Squared Error: $RRMSE = \sqrt{E_{mc}(t_y - T_y^2}/T_y$ as a measure of stability like the coefficient of variation.

The total number of simulations was chosen such that $|RB| \leq 0.01$ for all the point estimators. This resulted in the number of simulation runs being between 10,000 and 15,000. We also compared the RRMSE for the variance estimator for each point estimator to check the stability of variance estimators. In this case the number of simulation runs was chosen such that the absolute relative bias was less than 0.05 for all the variance estimators.

Table 1. % RRMSE of t_y (Summary Measures over 96 Simulation Scenarios)

Estimator	Average	Min	25th Quantile	Median	75th Quantile	Max	Standard deviation
SM	24.57	9.63	16.63	22.86	31.31	50.34	9.83
PM	22.90	9.62	16.27	20.26	29.21	50.09	9.26
CM	23.02	9.63	16.42	20.50	29.24	50.09	9.23
HM1	23.45	9.61	16.51	20.49	29.96	50.34	9.19
HM2	23.48	9.62	16.46	20.57	29.96	50.34	9.19
NM	23.39	10.03	16.93	20.66	29.81	50.19	9.14

Tables 1 and 2 provide, respectively, summary measures of RRMSE of t_y and $v(t_y)$ for the six estimators. It is observed that the SM estimator shows most instability, especially in the case of variance estimation. This is as expected, because it requires the least amount of frame level information. In terms of point estimators, however, SM performs reasonably well in comparison to others. All other estimators perform almost at par. With respect to the median and the 75th quantile, it is seen that PM and CM perform most favorably (in terms of small RRMSE) compared to others both with respect to point and variance estimators. On the other hand, the NM estimator, although requiring the most information, including deduplication of sampled units, performs slightly worse for point estimation but better for variance estimation. It also appears from the behavior of HM estimators that the hybridization approach succeeds in improving the stability of the SM estimator.

We investigated further how SM and others would behave as the percentage of small $\pi_{i(q)}$ increases over the 96 simulation scenarios considered. It is known that the presence of small inclusion probabilities would tend to make any estimator unstable. Using categories with increasing percentage of small $\pi_{i(q)}$ (defined as being less than 1%), out of 96 scenarios we have 27 with no small $\pi_{i(q)}$, 39 with less than 10% small $\pi_{i(q)}$, 26 cases between 10% and 20% and only four cases with more than 20% of small $\pi_{i(q)}$ in the three frames. Table 3 presents average RRMSE of point and variance estimators as the percentage of small $\pi_{i(q)}$ increases. It is seen that all the estimators get worse as the percentage of small inclusion probabilities increases. However, SM seems to be affected most.

Table 2. % RRMSE of $v(t_y)$ (Summary Measures over 96 Scenarios)

Estimator	Average	Min	25th Quantile	Median	75th Quantile	Max	Standard deviation
SM	371.04	40.89	116.22	224.63	546.10	1,272.62	907.91
PM	275.83	31.23	75.66	170.40	349.71	1,286.24	687.67
CM	274.64	31.16	74.57	170.33	344.42	1,284.87	685.33
HM1	309.65	40.70	111.57	192.27	408.52	1,287.08	702.33
HM2	309.21	40.77	111.58	191.76	408.17	1,286.99	697.65
NM	272.93	37.99	85.19	167.48	335.65	1,292.89	654.11

Table 3. Average % RRMSE of t_y and $v(t_y)$ as % Small $\pi_{i(q)}$ over Simulation Scenarios Increases

	Average %RRMSE of t_y				Average %RRMSE of $v(t_y)$			
	% of small $\pi_{i(q)}$				% of small $\pi_{i(q)}$			
Estimator	0	0–10%	10–20%	>20%	0	0–10%	10–20%	>20%
SM	18.53	23.97	29.08	41.81	144.32	362.83	567.27	706.09
PM	18.78	22.17	25.88	38.43	141.50	273.91	356.21	678.89
CM	18.90	22.30	25.99	38.50	140.87	272.70	354.68	676.16
HM1	18.53	23.08	26.77	38.75	145.31	326.43	400.83	662.61
HM2	18.55	23.11	26.82	38.83	145.27	326.11	400.13	660.07
NM	19.20	22.73	26.76	38.66	147.16	267.78	349.48	674.40

5. Concluding Remarks

We proposed a unified approach to multiple frame estimation using a generalized multiplicity-adjusted Horvitz-Thompson (GMHT) class of estimators. In this article we restricted our attention to mainly the combined frame approach (except for one from the separate frame approach) so that the case of mixed frame level information available for sampled units could be dealt with. Besides the known estimators PM (Kalton and Anderson 1986) and SM (Mecatti 2007) that belong to the GMHT class, and the estimator NM (Bankier 1986) being essentially Horvitz-Thompson (and hence trivially belonging to the GMHT class), three new estimators, composite multiplicity (CM) and hybrid multiplicity (HM1 and HM2) in the GMHT class, were considered. Based on a limited simulation study, it was observed that in terms of relative bias (RB) and relative root mean squared error (RRMSE) of point estimation, all six estimators are generally at par. However, in terms of RB and RRMSE of variance estimation, SM could suffer most if some of the units have very small sample inclusion probabilities. The new estimator CM, as expected, behaves almost at par with PM. In the case of mixed frame level information, only SM and the proposed HM estimators are applicable. In practice, however, the HM estimators may be preferable as they are less subject to the instability problem of SM. In the near future it is planned to extend our study to include other methods based on the separate frame approach when at least partial frame level information is assumed to be available for all sampled units.

Appendix A

We consider the optimal combination of the two unbiased estimators $\psi_{i,SM}$ and $\psi_{i,PM}$ of 1 by minimizing $Var\left[\lambda_i\psi_{i,SM} + (1-\lambda_i)\psi_{i,PM}\right]$ for a suitable $\lambda_i \in \left(0,1\right)$, i.e., find λ_i that

$$\min_{\lambda_i \in \left(0,1\right)} \left\{\lambda_i^2 Var\left(\psi_{i,SM}\right) + (1-\lambda_i)^2 Var\left(\psi_{i,PM}\right) + 2\lambda_i(1-\lambda_i)Cov\left(\psi_{i,SM},\psi_{iPM}\right)\right\} \tag{A1}$$

Notice that $\lambda_i\psi_{i,SM} + (1-\lambda_i)\psi_{i,PM} = \psi_{i,PM} + \lambda_i\left(\psi_{i,SM} - \psi_{i,PM}\right)$. Hence the solution of the problem (A1) is obtained as (minus) the regression coefficient of $\psi_{i,PM}$ on

$\left(\psi_{i,SM} - \psi_{i,PM}\right)$; that is, with $\bar{\pi}_i = m_i^{-1} \sum_{q' \in K(i)} \pi_{i(q')}$ we have

$$\begin{aligned}
\lambda_i^{CM} &= -\frac{Cov\left(\psi_{i,PM}, \psi_{i,SM} - \psi_{i,PM}\right)}{Var\left(\psi_{i,SM} - \psi_{i,PM}\right)} \\
&= \frac{\sum\limits_{q \in K(i)} \frac{\pi_{i(q)}\left(1 - \pi_{i(q)}\right)}{(m_i \bar{\pi}_i)^2} - \sum\limits_{q \in K(i)} \frac{\pi_{i(q)}\left(1 - \pi_{i(q)}\right)}{m_i \pi_{i(q)} m_i \bar{\pi}_i}}{\sum\limits_{q \in K(i)} \frac{\pi_{i(q)}\left(1 - \pi_{i(q)}\right)}{(m_i \bar{\pi}_i)^2} + \sum\limits_{q \in K(i)} \frac{\pi_{i(q)}\left(1 - \pi_{i(q)}\right)}{m_i^2 \pi_{i(q)}^2} - 2\sum\limits_{q \in K(i)} \frac{\pi_{i(q)}\left(1 - \pi_{i(q)}\right)}{m_i^2 \pi_{i(q)} \bar{\pi}_i}} \\
&= \frac{\sum\limits_{q \in K(i)} \left(1 - \pi_{i(q)}^{-1} \bar{\pi}_i\right) \pi_{i(q)} \left(1 - \pi_{i(q)}\right)}{\sum\limits_{q \in K(i)} \left(1 + \pi_{i(q)}^{-2} \bar{\pi}_i^2 - 2\pi_{i(q)}^{-1} \bar{\pi}_i\right) \pi_{i(q)} \left(1 - \pi_{i(q)}\right)}
\end{aligned} \tag{A2}$$

Appendix B

We have

$$Var\left(t_{y(NM)}\right) = \sum_{i \in U} z_i^2 Var(\delta_i) + \sum_{i \neq j} \sum_{\in U} z_i z_j Cov\left(\delta_i, \delta_j\right) \tag{B1}$$

where $z_i = y_i \left[1 - \prod\limits_{q' \in K(i)} \left(1 - \pi_{i(q')}\right)\right]^{-1}$, and

$$\begin{aligned}
Var(\delta_i) &= E\left[1 - \prod_{q' \in K(i)} \left(1 - \mathbf{1}_{i \in s_q}\right)\right]^2 - \left[1 - \prod_{q' \in K(i)} \left(1 - \pi_{i(q')}\right)\right]^2 \\
&= \prod_{q' \in K(i)} \left(1 - \pi_{i(q')}\right) \left[1 - \prod_{q' \in K(i)} \left(1 - \pi_{i(q')}\right)\right] = P_i(1 - P_i)
\end{aligned} \tag{B2}$$

$$\begin{aligned}
E\left(\delta_i \delta_j\right) &= 1 - \prod_{q' \in K(i)} \left(1 - \pi_{i(q')}\right) - \prod_{q' \in K(j)} \left(1 - \pi_{j(q')}\right) \\
&\quad + \prod_{q' \in K(i) \cap K(j)} \left(1 - \pi_{i(q')} - \pi_{j(q')} + \pi_{ij(q')}\right) \cdot \prod_{q' \in K(i) \cap K^c(j)} \left(1 - \pi_{i(q')}\right) \cdot \prod_{q' \in K(j) \cap K^c(i)} \left(1 - \pi_{j(q')}\right) \\
&= 1 - \prod_{q' \in K(i)} \left(1 - \pi_{i(q')}\right) - \prod_{q' \in K(j)} \left(1 - \pi_{j(q')}\right) \\
&\quad + \prod_{q' \in K(i)} \left(1 - \pi_{i(q')}\right) \cdot \prod_{q' \in K(j)} \left(1 - \pi_{j(q')}\right) \cdot \prod_{q' \in K(i) \cap K(j)} \frac{\left(1 - \pi_{i(q')} - \pi_{j(q')} + \pi_{ij(q')}\right)}{\left(1 - \pi_{i(q')}\right)\left(1 - \pi_{j(q')}\right)} \\
&\equiv 1 - P_i - P_j + P_i P_j \left(P_{ij}\right)
\end{aligned} \tag{B3}$$

for $i \neq j \in U$, where $K(i)$ and $K\left(j\right)$ denote two domains which need not be distinct, and P_i, P_j, and P_{ij} denote the corresponding terms in the previous equation. Note that $K(i) \cap K\left(j\right), K(i) \cap K^c\left(j\right)$ and $K\left(j\right) \cap K^c(i)$ are disjoint domains, some of them possibly empty, and therefore using the independence of frame samples, we have, as expected, $P_{ij} = 1$ if $K(i) \cap K\left(j\right) = \phi$. We thus obtain

$$Cov\left(\delta_i, \delta_j\right) = P_i P_j \left(P_{ij} - 1\right) \tag{B4}$$

Finally, since the sample of distinct units s^* has random size n^*, the Sen-Yates-Grundy variance estimator is not applicable, but an unbiased Horvitz-Thompson variance estimator can be obtained from

$$v\left(t_{y(NM)}\right) = \sum_{i \in s^*} z_i^2 \frac{Var(\delta_i)}{E(\delta_i)} + \sum_{i \neq j} \sum_{\in s^*} z_i z_j \frac{Cov\left(\delta_i, \delta_j\right)}{E\left(\delta_i \delta_j\right)} \tag{B5}$$

6. References

Bankier, M.D. (1986). Estimators Based in Several Stratified Samples with Applications to Multiple Frame Surveys. Journal of the American Statistical Association, 81, 1074–1079.

Casady, R.J. and Sirken, M.G. (1980). A Multiplicity Estimator for Multiple Frame Sampling. Proceedings of the American Statistical Association, Section on Survey Research Methods, 601–605.

Fuller, W.A. and Burmeister, L.F. (1972). Estimators of Samples Selected from Two Overlapping Frames. Proceedings of the American Statistical Association, Social Statistics Sections, 245–249.

Hartley, H.O. (1962). Multiple Frame Surveys. Proceedings of the American Statistical Association, Social Statistics Sections, 203–206.

Hartley, H.O. (1974). Multiple Frame Methodology and Selected Applications. Sankhyā, Series C, 36, 99–118.

Kalton, G. and Anderson, D.W. (1986). Sampling Rare Populations. Journal of the Royal Statistical Society, Series A, 149, 65–82.

Lohr, S. and Rao, J.N.K. (2006). Multiple Frame Surveys: Point Estimation and Inference. Journal of the American Statistical Association, 101, 1019–1030.

Mecatti, F. (2007). A Single Frame Multiplicity Estimator for Multiple Frame Surveys. Survey Methodology, 33, 151–158.

Rao, J.N.K. (1965). On Two Simple Schemes of Unequal Probability Sampling without Replacement. Journal of the Indian Statistical Association, 3, 173–180.

Sampford, M.R. (1967). On Sampling without Replacement with Unequal Probabilities of Selection. Biometrika, 54, 499–513.

Singh, A.C. and Wu, S. (1996). Estimation for Multiframe Complex Surveys by Modified Regression. Proceedings of the Survey Methods Section, Statistical Society of Canada, 69–77.

Singh, A.C. and Wu, S. (2003). An Extension of Generalized Regression Estimator to Dual Frame Surveys. Proceedings of the American Statistical Association, Survey Research Methods Section, 3911–3918.

Sirken, M.G. (1972). Stratified Sample Surveys with Multiplicity. Journal of the American Statistical Association, 224–227.

Skinner, C.J. (1991). On the Efficiency of Raking Ratio Estimator 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, 349–435.

Received November 2009
Revised May 2011