

## ESTIMATION FOR MULTIFRAME COMPLEX SURVEYS BY MODIFIED REGRESSION

A.C. Singh and S. Wu<sup>1</sup>

### ABSTRACT

Estimation from multiframe survey data is essentially a problem of combining domain estimates such that the new auxiliary information obtained from several estimates for overlapping domains is used in addition to the usual auxiliary information. The available approaches to this problem can be classified as either regression-based or Horvitz-Thompson-based. Alternatively, they can be classified as either separate or combined frame approaches, in the sense that a set of final sampling weights is produced either for each sample separately or for the combined sample. The separate frame approach is more flexible and adapts easily to multiframe surveys as well as to multivariate auxiliary data. A new estimator within the separate frame approach, based on the modified regression methodology of Singh (1994,1996), is proposed for combining information from multiframe complex surveys, in a manner analogous to the usual generalized regression estimator for single frame complex surveys. Monte Carlo simulation results on relative performance of various estimators are also presented.

**KEY WORDS:** Combining estimates; Calibration weights; Predictor zero functions; Working covariance.

### RÉSUMÉ

Le problème que pose le calcul d'estimations à partir de données d'enquête à plusieurs bases de sondage consiste essentiellement à combiner des estimations par domaine de façon à utiliser les nouvelles données auxiliaires provenant de plusieurs estimations calculées pour des domaines chevauchants en plus des données auxiliaires habituelles. Les méthodes dont on dispose pour résoudre ce problème se classent en deux catégories, soit les méthodes fondées sur la régression et celles fondées sur l'estimateur d'Horvitz-Thompson. On peut aussi les classer comme des méthodes axées sur les bases de sondage distinctes d'une part, et sur la base de sondage agrégée d'autre part, selon qu'on produit un ensemble de poids d'échantillonnage finals pour les échantillons élémentaires ou pour l'échantillon agrégé. La méthode des bases de sondage distinctes est plus souple et s'adapte plus facilement aux enquêtes à plusieurs bases de sondage ainsi qu'aux données auxiliaires à plusieurs dimensions. Dans le contexte de la méthode des bases de sondage distinctes, on propose d'utiliser un nouvel estimateur, inspiré de la méthode de régression modifiée de Singh (1994, 1996), pour combiner les données tirées d'enquêtes complexes à plusieurs bases de sondage, de façon analogue à l'estimateur de régression généralisé qu'on applique aux enquêtes complexes à base de sondage unique. On présente aussi les résultats d'une simulation de Monte Carlo exécutée pour comparer la performance de divers estimateurs.

**MOTS CLÉS:** Combinaison d'estimations; poids d'étalonnage; fonctions de la variable prédictive nulle; covariance provisoire.

### 1. INTRODUCTION

We consider the problem of estimation by combining information in samples from overlapping frames which together cover the target population. Typically, in practice, a dual frame problem arises when one frame is complete but expensive to sample, while the other frame is incomplete but cheaper to sample. In this paper, a suitable methodology for

estimating parameters and variances of their estimates is developed using the available auxiliary information. This will also be helpful at the design stage when dealing with the cost-variance issues in sample allocation. In general, one may have several frames and different sampling designs (simple or complex) for different frames; see Skinner and Rao (1996) for a recent good review.

The pioneering work in the area of multiple frames

---

1

A.C. Singh and S. Wu, Methodology Research Advisory Group, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

is due to Hartley (1962, 1974). Later, important contributions were made by Lund (1968), Fuller and Burmeister (1972), Bankier (1986), Kalton and Anderson (1986), Skinner (1991) and Skinner and Rao (1996). In this article, we will consider for convenience the case of only two frames A and B, although the proposed method can be easily extended to multiple frames. Let  $s_A, s_B$  be two independent samples drawn respectively from A and B under designs  $p_{sA}$  and  $p_{sB}$ . It is assumed that the impact of possible duplicate units in the two samples is negligible. Let  $N_A, N_B$  denote the population sizes and  $n_A, n_B$  denote the corresponding sample sizes. We will make the usual assumption that the population counts  $N_A, N_B$  are known. In addition, population totals (frame-specific or not) may be available in practice for several auxiliary variables such as demographic and geographic counting variables; the term frame-specific signifies that the auxiliary information is separate for frames A and B. Let domains  $A \cap B^c, A^c \cap B$ , and  $A \cap B$  be denoted respectively by  $a, b$ , and  $c$ . Then, for the study variable  $y$ , parameters of interest are the population total  $\theta_{yd}$  and mean  $\mu_{yd}$  for each domain  $d$ . Denoting by  $N_c$  the unknown population count of domain  $c$ , the component parameters of the population total of  $y$  for the combined frame are given by

$$\theta_y = \theta_{ya} + \theta_{yb} + \theta_{yc} \quad (1.1a)$$

$$= (N_A - N_c) \mu_{ya} + (N_B - N_c) \mu_{yb} + N_c \mu_{yc} \quad (1.1b)$$

where

$$\mu_{ya} = \theta_{ya}/N_a, \quad \mu_{yb} = \theta_{yb}/N_b, \quad \mu_{yc} = \theta_{yc}/N_c,$$

$$N_a = N_A - N_c, \quad N_b = N_B - N_c.$$

Let  $\hat{\theta}_{ya}^{GR}, \hat{\theta}_{yb}^{GR}$  denote the usual generalized regression (GR) estimates of  $\theta_{ya}$  and  $\theta_{yb}$  respectively from samples  $s_A$  and  $s_B$  as defined by Särndal (1980). In particular, they will be simply ratio-adjusted Horvitz-Thompson estimates when  $N_A, N_B$  are the only available auxiliary information. For the common (or overlapping) domain  $c$ , let  $\hat{\theta}_{ycA}^{GR}$  and  $\hat{\theta}_{ycB}^{GR}$  denote the two estimates from  $s_A$  and  $s_B$  respectively.

Now, a naive GR-estimator of  $\theta_y$  can be defined as

$$\hat{\theta}_y^{GR} = \hat{\theta}_{ya}^{GR} + \hat{\theta}_{yb}^{GR} + \left( \hat{\theta}_{ycA}^{GR} + \hat{\theta}_{ycB}^{GR} \right) / 2, \quad (1.2)$$

where the component estimates are not combined optimally. The methods proposed in the literature use, however, some form of optimality considerations.

They can be classified as either regression-based or Horvitz-Thompson-based methods. The regression-based methods approach the  $\theta_y$ -estimation problem either through the linear parametrization (1.1a) in terms of domain totals or through the nonlinear parametrization (1.1b) in terms of domain means. In the nonlinear case, the estimates are nonlinear functions (*e.g.*, ratio-cum-product) of usual regression (*i.e.*, GR) estimates. The methods of Hartley (H), and Fuller-Burmeister (FB) use regression-based approach via domain totals, and are applicable to complex designs. For simple random samples, the methods of Lund (L), and another one due to Fuller-Burmeister (FB\*) use the regression-based approach via domain means. The recently developed method of Skinner-Rao (SR) also uses this approach via domain means but is applicable to complex designs. The Horvitz-Thompson-based methods, on the other hand, approach the  $\theta_y$ -estimation problem by visualizing the combined sample  $s = s_A \cup s_B$  as coming from a single frame, and assign inclusion probabilities to units in the three domains  $a, b$ , and  $c$ . Once this is done, usual expansion methods can be applied for estimating  $\theta_y$ . The methods of Bankier (B) and Kalton-Anderson (KA) fall into this category. In the presence of auxiliary information, Bankier also suggests an important modification of the above expansion estimator, which takes the form of a GR-estimator (such as raking) and is computed in a second stage.

In practice, it is desirable to have an expansion form for any given estimator in terms of final sampling weights. This is useful for producing estimates for all study variables. From this perspective, it may be better to use an alternative classification of existing methods into combined or separate frame approaches. In the combined frame approach, sampling weights for the combined sample,  $s$ , are adjusted so that it can be visualized as a single sample. These weights could be further calibrated in light of the auxiliary information. Finally, the expansion estimates are computed from the calibrated weights in the usual way. The B-, KA-, and SR-methods correspond to the combined frame approach. The L- and FB\*-methods can also be classified in this category when the design is restricted to simple random samples. On the other hand, methods of H, and FB can be classified in the category of separate frame approach in which sampling weights are adjusted separately for each of the samples  $s_A$  and  $s_B$  in light of the auxiliary information. This can be done by suitably defining optimal regression for complex designs along the lines of Rao (1994) - see also Section 4. Note that in computing expansion estimates for the

separate frame approach, the two estimates for the common domain need to be averaged.

In the class of methods belonging to the combined frame approach, the main limitation of B- and KA-methods lies in their requirement of the knowledge of inclusion probabilities under both designs for the units in  $c$  in each of the samples  $s_A$  and  $s_B$ , while that of the SR-method is its restriction to only two frames. The L- and FB\*-methods of the above class are, however, further restricted to simple random samples from two frames. Moreover, for the above class, an additional calibration stage is required for incorporating auxiliary information. In contrast, the separate frame approach is more flexible. It can be easily adapted to multiframe problems and does not necessarily require a separate stage for incorporating auxiliary information. The H- and FB- methods do, however, implicitly use a two-stage regression and not a single multiple regression on all the auxiliary variables. In the first stage, GR-estimates of domains are obtained from each sample using the frame-specific auxiliary information and in the second stage, optimal regression is performed to incorporate additional auxiliary predictors which are constructed from estimates of the overlapping domain - see section 4 for details. A better alternative would be to perform optimal regression on all the predictors in one stage. The resulting estimator will have good conditional properties (*cf.*, Valliant, 1993) in view of better "balancing" of the samples. Unfortunately, this is not practical because optimal regression may lead to instability of estimates in the presence of several auxiliary variables due to inadequate degrees of freedom available for estimating covariance matrices (see, *e.g.*, Rao, 1994).

In view of the above limitations of H- and FB-methods, a method using the separate frame approach is proposed which employs some of the ideas underlying FB- and SR-methods, provides asymptotically design consistent estimates under general conditions, and provides a GR-type alternative to the optimal regression (*i.e.*, optimal for simple random samples but expected to be robust for complex designs). The main differences between the proposed method for multi-frames and the usual GR method for single frames are that general predictors (in the form of difference of two estimates), and relative measures of the inverse effective sample size are allowed. The proposed method uses the modified regression (MR) methodology of Singh (1994, 1996) which was inspired by the contributions in survey statistics of Fuller (1975) and Särndal (1980) on regression methods, and of Rao and Scott (1981) on pseudo-

maximum likelihood estimation, and the contributions in classical statistics by Liang and Zeger (1986) on generalized estimating equations, and of Godambe and Thompson (1989) on optimal estimating functions. The MR-methodology, based on the idea of finite population (semiparametric) modelling with working covariance structure within the estimating function framework, encompasses the GR-methodology of Särndal (1980) (see also Fuller, 1975), based on the idea of superpopulation modelling within the model-assisted framework. Section 2 provides a heuristic motivation of the proposed method while Section 3 contains a detailed description. A comparison of the proposed method with alternative methods is given in Section 4. Empirical evaluation of various methods based on a Monte Carlo study is presented in Section 5. Finally, Section 6 contains concluding remarks.

## 2. HEURISTIC MOTIVATION

First, suppose, the two frames A and B overlap completely. In other words, we have two independent samples from a single frame. Thus, the problem reduces to the familiar problem of combining two (approximately) unbiased and independent estimates  $\hat{\theta}_{yA}^{GR}$  and  $\hat{\theta}_{yB}^{GR}$  of the same parameter  $\theta_y$ . Such problems often arise in practice with rotating panel surveys. The best linear combination is given by

$$\hat{\theta}_{y,comb} = (1-\gamma)\hat{\theta}_{yA}^{GR} + \gamma\hat{\theta}_{yB}^{GR} \quad (2.1a)$$

$$= \hat{\theta}_{yA}^{GR} - \gamma(\hat{\theta}_{yA}^{GR} - \hat{\theta}_{yB}^{GR}) \quad (2.1b)$$

$$= (\hat{\theta}_{yA}^{GR} + \hat{\theta}_{yB}^{GR})/2 - \gamma^*(\hat{\theta}_{yA}^{GR} - \hat{\theta}_{yB}^{GR}) \quad (2.1c)$$

where  $\gamma$  is chosen to minimize the variance and  $\gamma^* = \gamma - 1/2$ . Notice that equivalently,  $\hat{\theta}_{y,comb}$  can be obtained as an optimal regression estimator by regressing on the predictor  $\hat{\theta}_{yA}^{GR} - \hat{\theta}_{yB}^{GR}$ ; the predictor, being a difference of two unbiased estimators of the same parameter, is a parameter-free zero function, *i.e.*, a function with zero expectation. Thus,  $\hat{\theta}_{y,comb}$  is computed as a residual from an initial estimator (which may be taken as one of  $\hat{\theta}_{yA}^{GR}$ ,  $\hat{\theta}_{yB}^{GR}$ , or the average of the two) after regressing (or projecting) on the predictor zero function. If the two estimates  $\hat{\theta}_{yA}^{GR}$  and  $\hat{\theta}_{yB}^{GR}$  have same variance, then  $\gamma = 1/2$  and  $\hat{\theta}_{y,comb}$  will be simply  $(\hat{\theta}_{yA}^{GR} + \hat{\theta}_{yB}^{GR})/2$ .

As mentioned in the introduction, it is preferable to

perform regression on all the predictors simultaneously, *i.e.*, a multiple regression on all the usual predictor zero functions denoted by  $\hat{\theta}_{x_A}^{HT} - \theta_{x_A}$  (used in  $\hat{\theta}_{y_A}^{GR}$ ),  $\hat{\theta}_{y_B}^{HT} - \theta_{y_B}$  (used in  $\hat{\theta}_{y_B}^{GR}$ ), and the additional predictor,  $\hat{\theta}_{y_A}^{GR} - \hat{\theta}_{y_B}^{GR}$  (or  $\hat{\theta}_{y_A}^{HT} - \hat{\theta}_{y_B}^{HT}$ ), due to overlapping frames. Here,  $x_A, x_B$  denote respectively the vector of auxiliary variables specific to frames  $A$  and  $B$  (which are identical for the present case) and  $HT$  signifies the Horvitz-Thompson estimator. Now, instead of optimal regression, the MR-estimator,  $\hat{\theta}_y^{MR}$ , uses a suboptimal regression under a working covariance matrix. This is done under a finite population (semiparametric) common mean model for the elementary estimates, consisting of four types of estimates of  $\theta_y$ : (i)  $\hat{\theta}_y^{HT} := (\hat{\theta}_{y_A}^{HT} + \hat{\theta}_{y_B}^{HT})/2$ , (ii)  $\hat{\theta}_y^{HT} + (\hat{\theta}_{x_A}^{HT} - \theta_{x_A})_y$ , (iii)  $\hat{\theta}_y^{HT} + (\hat{\theta}_{x_B}^{HT} - \theta_{x_B})_y$ , (iv)  $\hat{\theta}_y^{HT} + (\hat{\theta}_{y_A}^{HT} - \hat{\theta}_{y_B}^{HT})_y$ . In other words, each predictor zero function gives rise to a new estimator of  $\theta_y$  by adding it to  $\hat{\theta}_y^{HT}$ ; the predictor is of course assumed to be correlated with  $\hat{\theta}_y^{HT}$  in order to be useful. In the above model, elementary estimates can be visualized as working sufficient statistics for  $\theta_y$  as they represent a condensed form of raw survey data before being modelled. When the designs  $p_{s_A}$  and  $p_{s_B}$  are identical, MR can be defined in a manner similar to GR as follows; expressions for  $\hat{\theta}_{y_A}^{GR}$  and  $\hat{\theta}_{y_B}^{GR}$  are also given for comparison purposes.

$$\hat{\theta}_{y_A}^{GR} = \hat{\theta}_{y_A}^{HT} + (y_A' \Gamma_A X_A)(X_A' \Gamma_A X_A)^{-1} \times (\theta_{x_A} - X_A' \Gamma_A \mathbf{1}_A), \quad (2.2a)$$

$$\hat{\theta}_{y_B}^{GR} = \hat{\theta}_{y_B}^{HT} + (y_B' \Gamma_B X_B)(X_B' \Gamma_B X_B)^{-1} \times (\theta_{x_B} - X_B' \Gamma_B \mathbf{1}_B), \quad (2.2b)$$

$$\hat{\theta}_y^{MR} = [(\hat{\theta}_{y_A}^{HT} + \hat{\theta}_{y_B}^{HT}) + (y' \Gamma X)(X' \Gamma X)^{-1} \times (\theta_x - X' \Gamma \mathbf{1})]/2, \quad (2.2c)$$

where  $\hat{\theta}_{y_A}^{HT} = \sum_{k \in s_A} y_k h_{kA}$ ,  $\hat{\theta}_{y_B}^{HT} = \sum_{k \in s_B} y_k h_{kB}$ ,  $y_A = \text{vec}(y_k; k \in s_A)$ ,  $y_B = \text{vec}(y_k; k \in s_B)$ ,  $h_A = \text{vec}(h_k; k \in s_A)$ ,  $h_B = \text{vec}(h_k; k \in s_B)$ ,  $\Gamma_A = \text{diag}(h_A)$ ,  $\Gamma_B = \text{diag}(h_B)$ ,  $\theta_{x_A}$ ,  $\theta_{x_B}$  are auxiliary control totals,  $\mathbf{1}_A, \mathbf{1}_B, \mathbf{1}$  are vectors of 1s of dimensions  $n_A, n_B$ , and  $n (= n_A + n_B)$  respectively,  $y' = (y_A', y_B')$ ,  $\Gamma = \text{block diag}(\Gamma_A, \Gamma_B)$ ,  $\theta'_x = (\theta'_{x_A}, \theta'_{x_B}, 0)$  and

$$X: = \begin{pmatrix} X_A' \\ X_B' \end{pmatrix} = \begin{pmatrix} X_A & \mathbf{0} & y_A \\ \mathbf{0} & X_B & -y_B \end{pmatrix}. \quad (2.3)$$

Note that the  $X$  and  $\Gamma$  matrices for MR are simply enlarged versions of the corresponding matrices for GR. If the two designs  $p_{s_A}$  and  $p_{s_B}$  are different, then one could use a relative measure of inverse effective sample size (*e.g.*, design effect as used by Skinner-Rao for a chosen variable) to give differential weights to the two matrices  $\Gamma_A, \Gamma_B$  used in the working covariance structure for MR. Denoting by  $\lambda_A, \lambda_B$  the relative measures of the inverse effective sample size for designs  $p_{s_A}, p_{s_B}$ , the working covariance matrix for the case of different designs is modified by replacing  $\Gamma$  by  $\Lambda \Gamma$  defined as

$$\Lambda \Gamma = \text{block diag}(\lambda_A \Gamma_A, \lambda_B \Gamma_B) \quad (2.4)$$

Clearly  $\lambda_A = \lambda_B = 1$  if the two designs are identical. Also notice that the final calibrated weights  $w' = (w'_A, w'_B)$ , through which  $\hat{\theta}_y^{MR}$  can be represented as an expansion estimator (*i.e.*,  $\hat{\theta}_y^{MR} = y'w/2$ ) except that estimates for the common domain are averaged, can be obtained as

$$w = h + \Lambda \Gamma X(X' \Lambda \Gamma X)^{-1} (\theta_x - X' \Gamma \mathbf{1}), \quad (2.5)$$

where  $h' = (h'_A, h'_B)$ . We remark that MR extends GR in that it allows for general predictors (such as difference of two estimates), and a relative measure of the inverse effective sample size. Also since MR is GR-type, it is optimal for simple random samples, and is expected to be robust for complex designs.

Now, in the realistic situation where the two frames overlap only partially, we get three domains  $a, b, c$ , and the additional predictors for MR are generated from two estimates for  $c$  corresponding to several study variables. This does not pose any new problem as the estimator  $\hat{\theta}_y^{MR}$  of (2.2c) can be easily modified by redefining the part of  $X$  corresponding to additional predictors, as shown in the next section.

### 3. MR-MULTIFRAME: The Proposed Method

For simplicity, we present the proposed method for two frames only. It is easily generalized to multiple frames. From the samples in the common domain  $c$ , we get additional predictors such as  $\hat{\theta}_{y_{cA}}^{HT} - \hat{\theta}_{y_{cB}}^{HT}$  corresponding to each selected  $y$ , and of course  $\hat{N}_{cA}^{HT} - \hat{N}_{cB}^{HT}$  for the counting variable. The usual predictors are  $\hat{\theta}_{x_A}^{HT} - \theta_{x_A}$  and  $\hat{\theta}_{x_B}^{HT} - \theta_{x_B}$  respectively for the two frames. In practice, one may also have some predictors for the combined frame which can be expressed as  $\hat{\theta}_{x_a}^{HT} + \hat{\theta}_{x_b}^{HT} + (\hat{\theta}_{x_{cA}}^{HT} + \hat{\theta}_{x_{cB}}^{HT})/2 - \theta_x$ . (If

there are three frames, say, then the divisor will be 3 for the additional component involving three estimates for the common domain.) The proposed method combines above pieces of information (available in the form of predictors) with the domain estimates  $\hat{\theta}_{ya}^{HT}$ ,  $\hat{\theta}_{yb}^{HT}$ ,  $\hat{\theta}_{yCA}^{HT}$ , and  $\hat{\theta}_{yCB}^{HT}$  to get  $\hat{\theta}_y^{MR}$ . Note that the proposed method also allows for using other correlated study variables ( $z$ , say) in estimating total for  $y$ . In practice, one could choose a set of key study variables for use in producing a set of final weights  $w$  which, in turn, can be used for all study variables. This gives rise to the multivariate nature of MR-multiframe.

### 3.1 Description of MR-multiframe

To define  $\hat{\theta}_y^{MR}$ , all we need is to modify the matrix  $X$  in (2.5) appropriately to get  $w$  so that  $\hat{\theta}_y^{MR}$  can be represented as

$$\begin{aligned}\hat{\theta}_y^{MR} &= \hat{\theta}_{ya}^{MR} + \hat{\theta}_{yb}^{MR} + \hat{\theta}_{yc}^{MR} \\ &= y'_a w'_a + y'_b w'_b + (y'_{cA} w'_{cA} + y'_{cB} w'_{cB})/2\end{aligned}\quad (3.1)$$

where

$$y' = (y'_a, y'_{cA}, y'_{cB}, y'_b), \quad w' = (w'_a, w'_{cA}, w'_{cB}, w'_b). \quad (3.2)$$

Now, the matrix  $X$  is  $n \times q$  where  $n = n_A + n_B$ , and  $q$  is the total number of predictors. The number  $q$  is, in general, equal to  $q_1 + q_2 + q_3 + q_4$  where  $q_1$  is the number of frame A-specific predictors,  $q_2$  is the number of frame B-specific predictors (these are the usual predictors for GR),  $q_3$  is the number of predictors for the combined frame, and  $q_4$  is the number of predictors chosen for the common domain  $c$ . As in (2.3), the matrix  $X$  can be horizontally partitioned into a  $n_A \times q$  matrix  $X_A^*$  and a  $n_B \times q$  matrix  $X_B^*$ . These matrices are defined as follows in four parts:

- (i) the first  $q_1$  columns of the matrix  $X_A^*$  represent  $n_A$  observations on the usual predictors for frame A,
- (ii) the next  $q_2$  columns are zeros because they correspond to frame B predictors,
- (iii) the next  $q_3$  columns contain for each  $x$  for the combined frame either  $x_k$  or  $x_k/2$  depending on whether  $k$  is in  $a$  or  $c$ , and finally
- (iv) the last  $q_4$  columns contain for each chosen variable  $y$  for the common domain either  $+y_k$  or  $0$  depending on whether  $k$  is in  $c$  or not.

The matrix  $X_B^*$  is similarly defined, the main difference being in the last  $q_4$  columns which contain either  $-y_k$  or  $0$  depending on whether  $k$  is in  $c$  or not. This completes the description of the proposed method. Note that the control totals  $\theta_x$  corresponding to predictors  $\hat{\theta}_{yCA}^{HT} - \hat{\theta}_{yCB}^{HT}$  for the common domain  $c$  will be simply zeros. Moreover, for these selected  $y$ -variables, we will have  $\hat{\theta}_{yCA}^{MR} = \hat{\theta}_{yCB}^{MR} = \hat{\theta}_{yc}^{MR}$ .

### 3.2 Variance Estimation for MR-multiframe

It can be shown that  $\hat{\theta}_y^{MR}$  is the solution of the estimating equation  $G' \Gamma_g^{-1} g = 0$  where  $g$  is the  $(q+1)$ -vector  $(\hat{\theta}_{ya}^{HT} + \hat{\theta}_{yb}^{HT} + (\hat{\theta}_{yCA}^{HT} + \hat{\theta}_{yCB}^{HT})/2 - \theta_y, \mathbf{1}' \Gamma X - \theta_x)'$ ,  $\Gamma_g$  is the  $(q+1) \times (q+1)$  working covariance matrix of  $g$  with first row as  $(y' \Gamma y, y' \Gamma X)$  and the matrix of the last  $q$  rows as  $(X' \Gamma y, X' \Gamma X)$ , and  $G$  is the  $(q+1) \times 1$  vector  $(1, 0, \dots, 0)'$ . Then the estimated asymptotic variance,  $\hat{V}(\hat{\theta}_y^{MR})$ , of  $\hat{\theta}_y^{MR} - \theta_y$  has the sandwich form,

$$\hat{V}(\hat{\theta}_y^{MR}) = B' G' \Gamma_g^{-1} \hat{V}(g) \Gamma_g^{-1} G (B^{-1})', \quad (3.3)$$

where  $\hat{V}(g)$  is a consistent estimate of the true covariance matrix of  $g$ , and  $B = G' \Gamma_g^{-1} G$  which is a scalar in our case. Note that the vector  $g$  consists of HT-estimators for various parameters, and therefore, standard results in sampling can be used for estimating its covariance matrix.

### 3.3 Asymptotic Properties of MR-multiframe

We assume that under the asymptotic setup of Isaki and Fuller (1982) for a sequence of finite populations and samples, as  $n_A, n_B, N_A, N_B \rightarrow \infty$  such that  $n_A/n_B, N_A/N_B$  tend to positive constants, the HT estimators in vector  $g$  defined in Section 3.2 above are consistent estimates and follow the multivariate central limit theorem, *i.e.*,

$$n^{\frac{1}{2}} N^{-1} (g - \mathbf{0}) \rightarrow_d N_{q+1}(\mathbf{0}, V(g)n/N^2). \quad (3.4)$$

Suppose also that  $N^{-1} \Gamma_g$  converges in probability to a positive definite matrix where  $N = N_A + N_B$ . Now using the delta method, it follows that  $\hat{\theta}_y^{MR}$  is asymptotically design consistent. Moreover, it is asymptotically normal with mean  $\theta_y$  and variance given by (3.3), which can then be used for constructing confidence intervals.

## 4. ALTERNATIVE METHODS: REVIEW AND COMPARISON

### 4.1 Separate Frame Approach

As mentioned in the introduction, regression-based methods via domain totals, H and FB, like MR, fall in this category. The H-estimator is given by

$$\hat{\theta}_y^H = \hat{\theta}_{ya}^{GR} + \hat{\theta}_{yb}^{GR} + (1-\beta)\hat{\theta}_{yca}^{GR} + \beta\hat{\theta}_{ycb}^{GR}, \quad (4.1a)$$

or

$$\hat{\theta}_y^H = \left(\hat{\theta}_{ya}^{GR} + \hat{\theta}_{yb}^{GR} + \hat{\theta}_{yca}^{GR}\right) - \beta\left(\hat{\theta}_{yca}^{GR} - \hat{\theta}_{ycb}^{GR}\right) \quad (4.1b)$$

where  $\beta$  is chosen to get an optimal regression estimator.

The H-estimator first uses frame-specific auxiliary information in obtaining GR (this is implicit in the paper) for both samples  $s_A$  and  $s_B$ , and then uses only one additional predictor from  $c$  for the variable  $y$ . The FB-estimator uses another predictor from  $c$  corresponding to the counting variable. We thus have

$$\hat{\theta}_y^{FB} = \hat{\theta}_{ya}^{GR} + \hat{\theta}_{yb}^{GR} + (1-\beta_1)\hat{\theta}_{yca}^{GR} + \beta_1\hat{\theta}_{ycb}^{GR} + \beta_2(\hat{N}_{cb}^{GR} - \hat{N}_{ca}^{GR}) \quad (4.2a)$$

$$= (\hat{\theta}_{ya}^{GR} + \hat{\theta}_{yb}^{GR} + \hat{\theta}_{yca}^{GR}) - \beta_1(\hat{\theta}_{yca}^{GR} - \hat{\theta}_{ycb}^{GR}) - \beta_2(\hat{N}_{ca}^{GR} - \hat{N}_{cb}^{GR}), \quad (4.2b)$$

where  $\beta_1, \beta_2$  are chosen to get optimal regression. Clearly,  $\hat{\theta}_y^{FB}$  is superior to  $\hat{\theta}_y^H$  as it uses more information.

One can get a set of final weights for optimal regression estimators for complex designs if estimates of  $\beta$ -coefficients are obtained from the estimated covariance matrix under the assumption of with replacement selection of primary sampling units, as shown by Rao (1994). This property carries over to the present case of optimal regression for multiframe provided that a set of key study variables are chosen in advance for the additional predictors from  $c$ . This also implies that the above regression-based methods can be made multivariate in nature, *i.e.*, information about other correlated study variables ( $z$ ) can be used for estimating total for  $y$ . The only main limitation of these methods, as mentioned in the introduction, is that they, unlike MR, do not use all the auxiliary information simultaneously and thus the resulting sampling weights may not be properly balanced. On the other hand, optimal regression with all the

predictors simultaneously may lead to instability of estimates.

### 4.2 Combined Frame Approach

As mentioned in the introduction, regression-based methods via domain means, FB\*, L, and SR, and Horvitz-Thompson-based methods B, and KA fall in this category.

#### 4.2.1 Regression-based Methods

For the simple random sample case, the domain mean estimates  $\hat{\mu}_{ya}^{SRS}, \hat{\mu}_{yb}^{SRS}$ , and  $\hat{\mu}_{yc}^{SRS}$  (which are simply sample means  $\bar{y}_a, \bar{y}_b$ , and  $(n_{cA}\bar{y}_{cA} + n_{cB}\bar{y}_{cB})/n_c$  with  $n_c = n_{cA} + n_{cB}$ ), are unbiased and mutually uncorrelated, conditional on domain sample sizes  $n_a, n_b, n_{cA}$  and  $n_{cB}$ . Therefore, the predictors  $(\hat{\mu}_{yca}^{SRS} - \hat{\mu}_{ycb}^{SRS})$  and  $(\hat{N}_{cA}^{SRS} - \hat{N}_{cB}^{SRS})$ , which are simply  $(\bar{y}_{cA} - \bar{y}_{cB})$  and  $[(N_A/n_A)n_{cA} - (N_B/n_B)n_{cB}]$ , do not provide any extra information in estimating domain means. However, for estimating domain population counts  $N_a, N_b$ , and  $N_c$ , the predictor  $(\hat{N}_{cA}^{SRS} - \hat{N}_{cB}^{SRS})$  will be useful because it is correlated with  $\hat{N}_a^{SRS}, \hat{N}_b^{SRS}$ , and  $\hat{N}_c^{SRS}$ , while the other predictor  $(\hat{\mu}_{yca}^{SRS} - \hat{\mu}_{ycb}^{SRS})$  will not be useful. Now, noting that  $N_a = N_A - N_c$ ,  $N_b = N_B - N_c$ , for a regression-based estimate of  $\theta_y$ , it is enough to find the optimal regression estimator of  $N_c$  as

$$\tilde{N}_c = \alpha \hat{N}_{cA}^{SRS} + (1-\alpha) \hat{N}_{cB}^{SRS} \quad (4.3)$$

and then substitute the component estimates in the nonlinear parametrization (1.1b). This was essentially done by Fuller-Burmeister to propose another estimator,  $\hat{\theta}_y^{FB*}$  for the simple random sample case, except that they used a maximum likelihood-type argument to estimate  $N_c$  via a binomial-type approximation to the joint distribution of  $\hat{N}_{cA}$  and  $\hat{N}_{cB}$ , as shown by Skinner (1991). Denoting this estimator of  $N_c$  by  $\hat{N}_c^{ML}$  (which is obtained as the smallest root of a quadratic equation, and turns out to be asymptotically equivalent to  $\tilde{N}_c$  under the normal approximation to binomial), we have

$$\hat{\theta}_y^{FB*} = (\hat{N}_A - \hat{N}_c^{ML}) \hat{\mu}_{ya}^{SRS} + (N_B - \hat{N}_c^{ML}) \hat{\mu}_{yb}^{SRS} + \hat{N}_c^{ML} \hat{\mu}_c^{SRS} \quad (4.4)$$

The L-estimator is also for sample random samples and uses  $\tilde{N}_c$  instead of  $\hat{N}_c^{ML}$  in estimating  $N_c$ , but does not use  $\tilde{N}_c$  to improve the other estimators  $\hat{N}_a^{SRS}$  and

$\hat{N}_b^{SRS}$ , resulting in a loss of efficiency compared to FB\*. The estimator  $\hat{\theta}_y^L$  is given by

$$\hat{\theta}_y^L = \hat{N}_a^{SRS} \hat{\mu}_{ya}^{SRS} + \hat{N}_b^{SRS} \hat{\mu}_{yb}^{SRS} + \tilde{N}_c \hat{\mu}_c^{SRS} \quad (4.5)$$

The SR-estimator makes an important generalization of the FB\*-estimator to complex designs using the idea of pseudo maximum likelihood (PML) estimation. It is given by

$$\hat{\theta}_y^{SR} = (N_A - \hat{N}_c^{PML}) \hat{\mu}_{ya}^{GR} + (N_B - \hat{N}_c^{PML}) \hat{\mu}_{yb}^{GR} + \hat{N}_c^{PML} \hat{\mu}_{yc}^{GR}, \quad (4.6)$$

where  $\hat{\mu}_c^{GR} = (\lambda_A^{-1} N_A^{-1} \hat{\theta}_{cA}^{GR} + \lambda_B^{-1} N_B^{-1} \hat{\theta}_{cB}^{GR}) / (\lambda_A^{-1} N_A^{-1} \hat{N}_{cA}^{GR} + \lambda_B^{-1} N_B^{-1} \hat{N}_{cB}^{GR})$ ,  $\lambda_A^{-1}, \lambda_B^{-1}$  being relative measures of effective sample sizes under the two designs  $p_{sA}$  and  $p_{sB}$ , and  $\hat{N}_c^{PML}$  is the smallest root of a quadratic equation similar to the one for FB\* but modified suitably for complex designs, see equation (5) of SR. Note that the term  $\lambda_A^{-1} N_A^{-1} \hat{\theta}_{cA}^{GR}$  in  $\hat{\mu}_c^{GR}$  corresponds to the term  $n_A N_A^{-1} \hat{\theta}_{cA}^{SRS}$  (or  $n_{cA} \hat{\mu}_{cA}^{SRS}$ ) in  $\hat{\mu}_c^{SRS}$ , and similarly for the term involving B. Notice also that  $\hat{\theta}_y^{SR}$  does not use all the predictors as the FB method for complex designs does, because it attempts to modify the FB\* method for simple random samples to the case of complex designs. Unlike MR, the main limitation of the above regression-based methods under the combined approach seems, as mentioned earlier, to be their restriction to only two frames. However, all the methods do have the desirable property of producing a set of final weights for the combined sample which can be used for any study variable. In the case of SR, this is achieved by choosing  $\lambda_A, \lambda_B$  to depend on a variable other than  $y$  such as the counting variable for the domain  $c$ . Further, these weights can be calibrated to satisfy auxiliary controls corresponding to the usual auxiliary information. However, information in the additional predictors from the common domain is no longer useful at this stage because these predictors become identically zero.

#### 4.2.2 Horvitz-Thompson-based Methods

If the duplicate sample units in  $s_A$  and  $s_B$  can be identified, then in the B-method, a HT-estimator based on distinct units is constructed for  $\theta_y$  by assigning inclusion probabilities  $\{\pi_k\}$  to units in the combined sample as follows: denoting by  $s_a, s_b, s_c$ , the three parts of the sample  $s$  in domains  $a, b, c$ , we have

$$\pi_k = \pi_{kA} \text{ if } k \in s_a; \pi_{kB} \text{ if } k \in s_b;$$

$$\pi_{kA} + \pi_{kB} - \pi_{kA} \pi_{kB} \text{ if } k \in s_c \quad (4.7a)$$

and

$$\hat{\theta}_y^B = \sum_{k \in s} y_k \pi_k^{-1}, \quad (4.7b)$$

where  $\pi_{kA}, \pi_{kB}$  are inclusion probabilities under designs  $p_{sA}, p_{sB}$  respectively, and  $s^*$  denotes the combined sample after discarding the duplicate units. Thus the B-estimator is unbiased by construction. It is assumed that inclusion probabilities  $\pi_{kA}$  and  $\pi_{kB}$  from both designs are available for all  $k$  in  $s_c$ . This assumption seems to be the main limitation of this method. However, for many situations involving simpler designs, this may not be a problem. Bankier also allows for a second stage to incorporate the usual auxiliary information by adjusting the basic weights  $\{\pi_k^{-1}\}$  of the combined sample via raking. However, as in the case of SR, information in the additional predictors cannot be used at this stage. The KA-estimator, on the other hand, ignores the possibility of duplicate sample units, and therefore, the product term  $\pi_{kA} \pi_{kB}$  in (4.7a) is dropped. Consequently, it is only an approximate HT-estimator under the assumption that both  $\pi_{kA}$  and  $\pi_{kB}$  are small for units in the common domain. The attractive feature of the KA-method is that it is simple and unbiased. However, like the B-method, KA also requires the knowledge of  $\pi_{kA}$  and  $\pi_{kB}$  for  $k$  in  $s_c$ .

## 5. EMPIRICAL RESULTS

Following Skinner and Rao (1996), we conducted a simulation study to compare empirical MSE (denoted as EMSE) of H, FB, KA, SR, GR and MR estimators. For MR, three versions were considered: MR(i) where  $\lambda_A/\lambda_B = 1$ , i.e., the effective sample sizes are assumed to be identical; MR(e) where  $\lambda_A/\lambda_B$  is estimated for each sample data as in SR; and MR(g) where  $\lambda_A/\lambda_B$  is given in advance using past data. A two-stage sampling with  $m_1$  sample clusters and  $m_2$  sample elements from each sampled cluster in frame A (i.e.,  $n_A = m_1 m_2$ ) and simple random sampling of  $n_B$  elements in frame B were employed.

The  $y$ -values for samples from frames A and B were generated according to a superpopulation model as in Skinner-Rao which approximates asymptotically a two stage cluster sampling for frame A and a simple random sampling for frame B for finite populations.

Thus, evaluation of estimators corresponds to the infinite population case. For generating samples from frame A, a beta distribution (with mean  $\gamma_a/(1-\gamma_b)$ ), where  $\gamma_a = N_a/N$ ,  $\gamma_b = N_b/N$  was used to choose the random proportion of units in the domain  $a$  for each cluster  $i$ , and then a binomial distribution given the observed proportion to define the number of units in  $a$ . Next the  $y$ -values using a nested error model (to account for clustering) were generated separately for domains  $a$  and  $c$ . The intracluster correlation ( $\rho$ ) was assumed to be common for each of the two domains while the across domain intracluster correlation was assumed to be different and taken as  $\rho\delta$ . For generating samples from frame B, binomial distribution (with mean  $(\gamma_b/(1-\gamma_a))$ ) was used to define number of units falling in the domain  $b$ , and then  $y$ -values were generated using separate common mean models for domains  $b$  and  $c$ .

We present results from a preliminary study where the population counts for the two frames serve as the only auxiliary control totals. For the MR(g) version, the relative inverse effective sample sizes were taken roughly as 3 to 1 based on some experimentation with estimated design effects for the  $y$ -variable. A total of 5000 replications were performed for each selected set of the design parameters. Table 1 shows empirical MSE when the parameters for the simulation experiment are set as  $(\gamma_a, \gamma_b) = (.1, 0), (.2, 0), (.1, .1), (.2, .1)$ ,  $\mu_a = 9, \mu_b = 11, \mu_c = 10$ ,  $\sigma^2 = 1, \rho = .1, \delta = .5$ ,  $m_{10} = m_2 = 15$ ,  $n_A = 225, n_B = 200$ .

It is seen from the evaluation results that the three methods FB, SR, and MR perform quite similarly. The two simple methods GR (under the separate frame approach) and KA (under the combined frame approach) perform well when  $\gamma_a = \gamma_b$ , but poorly when  $\gamma_a \neq \gamma_b$ . They also show a marked deterioration for the nested case, *i.e.*, when  $\gamma_b = 0$ . This may disappear if the extra information about  $N_c$ , which is now known, is used to poststratify GR and KA.

It is planned to enhance this study to include more control totals and also to allow predictors from the common domain based on other correlated study variables. This might produce some significant differences in the performance of the three main estimators.

## 6. CONCLUDING REMARKS

The existing methods for the dual frame estimation problem can be classified under separate or combined frame approaches depending upon the nature of final calibrated weights. The methods of H and FB fall in the former category while those of L, FB\*, B, KA and SR in the latter category. A method, termed MR-multiframe, under the separate frame approach, was proposed. It is analogous to the familiar GR-estimator for single frames except that it allows for general predictors and incorporates a relative measure of the inverse effective sample size (such as the design effect) for each sample. The proposed method is expected to provide a robust alternative to the methods based on optimal regression such as those of H and FB. The MR-method, being in the category of separate frame approach, can be easily adapted to problems with multi-frames as well as multivariate auxiliary information, *i.e.*, information about other correlated variables ( $z$ ) can also be added in estimating total for a given variable  $y$ . Such flexibility is not shared by methods using the combined frame approach. Note that the auxiliary information can be either standard (*i.e.*, with known population totals which may be frame specific such as known  $N_A, N_B$ , or not such as known  $N$ ) or nonstandard (such as the difference between estimators of the same parameter). Finally, we remark that although the MR method allows for several predictors, there is a need in practice to choose a set of good predictors in the interest of efficiency.

**Table 1: EMSE ( x 100) of Estimators**

$\gamma_a$	$\gamma_b$	SEPARATE						COMBINED	
		H	FB	GR	MR(i)	MR(e)	MR(g)	KA	SR
.1	0	.88	.40	1.45	.42	.38	.37	1.44	.38
.2	0	1.00	.44	3.11	.48	.41	.39	3.45	.41
.1	.1	3.26	3.32	2.92	3.26	3.24	3.17	2.93	2.99
.2	.1	3.43	3.34	4.03	3.91	3.26	3.10	4.03	3.21

Note: MR(i): MR with identical design effects  
 MR(e): MR with estimated design effects  
 MR-(g): MR with given design effects

## ACKNOWLEDGEMENT

We are grateful to Jon Rao and Mike Bankier for helpful comments. The first author's research was partially supported by a grant from Natural Sciences and Engineering Research Council of Canada held at Carleton University, Ottawa.

## 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.
- Fuller, W.A. (1975). "Regression analysis for sample surveys," *Sankhyā, Series C*, 37, 117-132.
- 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.
- Godambe, V.P., and Thompson, M.E. (1989). "An extension of quasi-likelihood estimation (with discussion)," *Journal of Statistical Planning and Inference*, 22, 137-172.
- 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.
- Isaki, C.T., and Fuller, W.A. (1982). "Survey design under the regression superpopulation 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.
- Liang, K.-Y., and Zeger, S.L. (1986). "Longitudinal data analysis using generalized linear models", *Biometrika*, 73, 13-22.
- Lund, R.E. (1968). "Estimators in multiple frame surveys," in *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.
- Rao, J.N.K. (1994). "Estimating totals and distribution functions using auxiliary information at the estimation stage", *Journal of Official Statistics*, 10, 153-165.
- Rao, J.N.K., and Scott, A.J. (1981). "The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two-way tables", *Journal of the American Statistical Association*, 76, 221-230.
- Särndal, C.-E. (1980). "On  $\pi$ -inverse weighting versus best linear unbiased weighting in probability sampling", *Biometrika*, 67, 639-650.
- Singh, A.C. (1994). "Sampling design-based estimating functions for finite population means", Invited paper, Abstracts of the Annual meeting of the Statistical Society of Canada, Banff, Alberta, 48.
- Singh, A.C. (1996). "Modified regression for combining information in survey sampling with applications", Invited paper, *Proceedings of the Survey Research Methods Section, American Statistical Association* (to appear).
- 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.
- Valliant, R. (1993). "Poststratification and conditional variance estimation", *Journal of the American Statistical Association*, 88, 89-96.