

Multiple Frame Surveys: a simplified and unified review

Multiple Frame Surveys: una rassegna semplificata e unificata

Fulvia Mecatti

Abstract Multiple Frame Surveys are tools for dealing with imperfect sampling frames, difficult to sample populations and other complex sampling situations. However they are challenging with respect to estimation of population parameters. Since they have been introduced in the sixties, several estimators have appeared in the literature, derived under different approaches. This lack of homogeneity as well as their complex structure tricky to implement, with no easy formulae available for variance estimation, can limit the use of a multiple frame design despite its practical appealing. In this paper a review is given at both the selection and the estimation stage of a Multiple Frame Survey, with the purpose of unifying and simplifying the main available methods.

Key words: Generalized Horvitz-Thompson estimation, Imperfect Sampling Frames, Indirect Sampling, Multiplicity.

1 Introduction

A basic assumption in conventional survey sampling is the availability of an up-to-date list recording all and only the individuals eligible for the survey, each identified by a label and assigned a positive probability to be selected in the sample. This *perfect* list wholly covering the target population is used as a frame for sample selection. Perfect frames are seldom the case in modern practice, under and over-coverage being usual examples of frame *imperfection* that can also occur jointly. Multiple frame surveys are tools for dealing with imperfect frames by avoiding costly and error-prone screenings aiming to recover an utilizable sampling frame out of an imperfect one.

Fulvia Mecatti

Department of Sociology & Social Research, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy e-mail: fulvia.mecatti@unimib.it

In a multiple frame survey a collection of two or more lists of units are used in combination for sample selection. Lists may be singularly partial and overlapping each other at an unknown extent. However their union, though possibly unknown, is believed to offer an adequate coverage of the target population.

Multiple Frame (MF) surveys can be convenient even if a complete population list did exist, a typical example being the availability of a total-coverage area-frame, usually expensive to sample for requiring local visits, supported by one (or more) available list-frame cheaper to sample. An efficient MF design combines the available frames, by under-sampling from the expensive complete area-frame and over-sampling from the cheaper partial list-frame. Indeed, this was the motivational-example when multiple frame surveys were first proposed: they can result in a considerable reduction of survey costs versus a conventional unique-frame survey, by achieving at least the same accuracy of the final estimates. In recent literature MF surveys are often suggested for improving population coverage and response rate beside controlling survey costs which may vary significantly among frames. Moreover, a multiple frame survey could offer greater precision and flexibility than a conventional unique-frame survey. For instance in surveying complex phenomena on complex populations, modular sampling frames can help to better capture differences and subgroups, as well as allowing for mixing different sampling designs and different data collection modes. See [4] for a comprehensive account of sampling designs for elusive, hidden and rare populations.

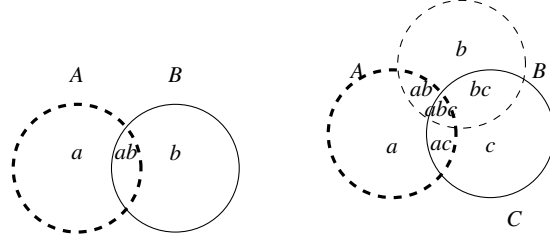
In contemporary applications the potential of using MF appears even more promising. For instance in a web survey both the population coverage and the selection bias might be controlled by using multiple web sites simultaneously for data collection, each targeting different segments of the study population. MF have been suggested for implementing multivariate stratified sampling design in multipurpose surveys [19] and the use of MF estimators has been proposed for cross-population comparisons in a *big (official) data* context [5].

On the other hand, MF estimation is challenging. It essentially means combining data from two or more frames with unknown overlapping, i.e dealing with increased inclusion probabilities in the final sample and with potential, un-identified data duplications, for units appearing in more than one frame. Since the seminal Hartley's papers in late sixties [3], several multiple frame estimators have appeared in the literature, derived under different approaches to estimation (Section 4). This lack of homogeneity can make it difficult for practical users to compare and select the best suited method for the problem at hands. In addition, as a direct consequence of the complexity of a multiple frame setup, available MF estimators have usually a complex structure tricky to implement, with no easy formulae available for variance estimation. This can limits the use of a MF design despite its practical appealing. In this paper a review of MF is given with the purpose of unifying and simplifying the main available methods, introducing a simplified notation (Section 2) and discussing both selection (Section 3) and estimation stage (Section 5).

2 A Simplified Multiple Frame Notation

Hartley's pioneering papers dealt with the simple case of 2 frames, known as Dual Frame surveys (DF henceforth) and introduced the popular notation which combines capital letters A and B to denote overlapping frames and lowercase letters to denote non-overlapping domains a, ab and b generated by the two frames, as shown in the left panel of Figure 1. Several papers both improving Hartley's estimators and introducing new DF estimators followed, usually with a mere mention about the possibility to generalize them to any number of frames (MF henceforth). The DF notation indeed becomes quickly un-useful as the number of frames increases, as for 3 frames shown in the right panel of Figure 1.

Fig. 1 Typical 2-frames (DF) and 3-frames (MF) survey setup



Lohr and Rao first gave a MF notation in 2006 [9], allowing for showing estimators formulae in a closed form and for applying linearization or resampling for variance estimation. However, their notation might be complex to understand for relying on index sets and block matrices, and for not being related with the familiar DF notation. We reckon that an essential initial step for a simplified and unified account to MF surveys would be a simplified indexed notation, naturally extending the familiar alphabetical DF notation. This is depicted in Figure 2. First notice that since frames are in fact subsets (possibly improper) of the target population U , i.e. $A \cup B = U$, a natural indexed DF notation would denote frame A by subset U_1 and frame B by subset U_2 so that $\bigcup_{q=1}^Q U_q = U$ with $Q = 2$. Next notice that each frame is partitioned into a fixed number of domains (possibly empty): $A = a(A) \cup ab(A)$ and $B = b(B) \cup ab(B)$ with $ab(A) \equiv ab(B)$. Thus a natural indexed DF notation for domains would be $U_1 = U_{1(1)} \cup U_{2(1)}$ and $U_2 = U_{1(2)} \cup U_{2(2)}$.

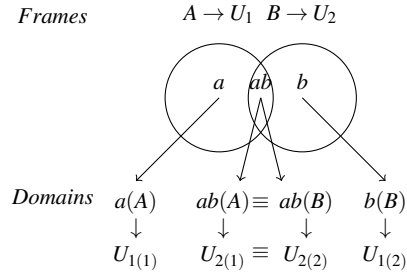


Fig. 2 From the alphabetical DF notation to an indexed MF notations

In the following we will refer to the general Q -frames case with $Q \geq 2$. Let $U_1 \cdots U_q \cdots U_Q$ be the collection of frames available for the survey; each frame U_q is partitioned into $D_q \leq 2^{Q-1}$ non empty domains $U_{d(q)}$, where $q = 1 \cdots Q$ and $d = 1 \cdots D_q$.

3 A Unified Theoretical Setup

In order to deal with *imperfect* frame, it is usually suggested to implement some form of *indirect* selection. When a complete frame perfectly representing the target population is not available, a working selection list easier to acquire might be used instead, based on a known linkage pattern between target and selection units. Popular examples are *Network* sampling [17], also known as *Multiplicity* or *Snowball* sampling as a link-tracing design for surveying rare or elusive populations, and *Indirect sampling* [6] suggested for social and economic surveys lacking of a sampling frame *directly* representing all the target population units. In Figure 3 a typical example is depicted. In a survey on small children a complete list of the target population is difficult to acquire; instead a list of households is available and can be used as an indirect sampling frame. Since not all the householders listed are also parents and parents may live in separate houses, selection units might be either linked to the target units according to a one-to-one pattern, or to a many-to-one pattern or not be linked at all.

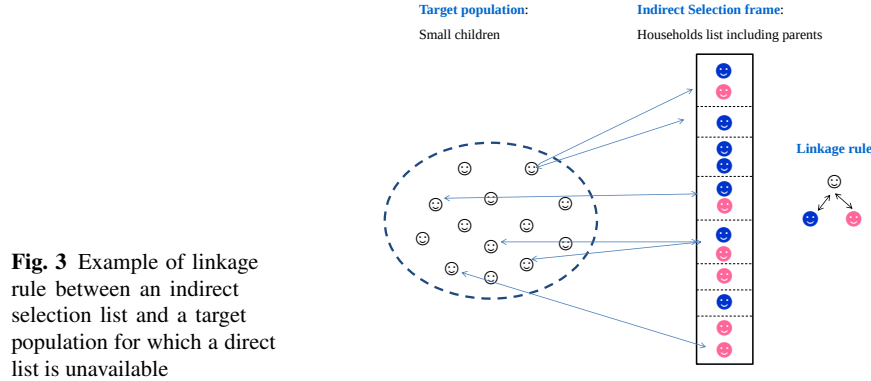
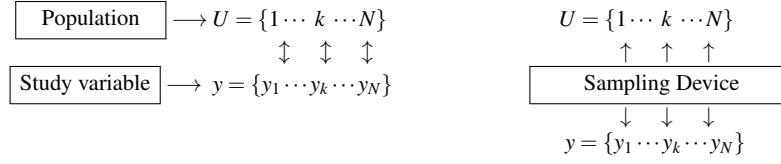


Fig. 3 Example of linkage rule between an indirect selection list and a target population for which a direct list is unavailable

Both indirect and multiple frame selection can be referred to a unique theoretical layout. Figure 4 gives an illustration of the conventional (unique) perfect frame setup versus the imperfect frame set up. When a perfect sampling frame is available, the conventional label-linkage between population units and the observable study values applies, i.e. there is a one-to-one linkage rule between the target population U and the study variable y as given by labels $k = 1 \cdots N$.

When imperfect frames occur, the direct one-to-one linkage rule is missed due to some sort of sampling device entering the picture for selection purposes. This sampling device may be any available list of units indirectly related to the target units, as in the indirect sampling example in Figure 3. When the sampling device is

**Fig. 4** Perfect frame setup versus imperfect frame setup

a collection of $Q \geq 2$ sampling frame, possibly partial and overlapping each other, we have a MF survey. The linkage pattern between the target population and the sampling device may be complicated to a different extent, for instance one-to-many or many-to-many as for the indirect sampling example in Figure 3. In a MF survey, it is one-to-one *within* each frame and many-to-one *between* frames. The linkage pattern consents the actual data collection and it is crucial at the estimation stage. Whatever complex, however, it is completely described by a *linkage matrix* whose entries define the linkage rule between target units (included in U) and selection units (included in the selection device). In a MF setup this is a $N \times Q$ matrix with population units on the rows and frames on the columns, whose entries are (non-random) Frame Membership Indicators $\mathbb{1}_{k \in U_q}$, i.e. non-random observable variable for every population $k \in \bigcup_{q=1}^Q U_q \equiv U$ and every frame U_q taking on value 1 if $k \in U_q$ and 0 otherwise (3)

Fig. 5 MF Linkage Matrix

$$\begin{matrix}
 & (U_1 & \dots & U_q & \dots & U_Q) \\
 \begin{matrix} 1 \\ \vdots \\ k \\ \vdots \\ N \end{matrix} & \left[\begin{array}{c} \\ \\ \mathbb{1}_{k \in U_q} = \begin{cases} 1 & \text{if } k \in U_q \\ 0 & \text{otherwise} \end{cases} \\ \\ \end{array} \right]
 \end{matrix}$$

Relevant definitions can be derived from the linkage matrix, such as domains (see for instance Figure 2) which are formed by the collection of units with equal rows in the linkage matrix. Frame sizes N_q are given by the sum of each column. The sum of each row gives the number of frames every population units belong into and is known as unit *multiplicity* $m_k = \sum_{q=1}^Q \mathbb{1}_{k \in U_q} \quad \forall k \in \bigcup_{q=1}^Q U_q \equiv U$

Multiplicity, first introduced in the contest of Network sampling [17], will be largely used in the following sections as the basic tool for a systematic treatment to MF estimation, which we will refer to as *multiplicity approach* [14].

4 MF Estimators: Toward a Simplified and Unified View

In a MF survey, data are generally collected by selecting independently a sample in each of the frames used in the survey and MF estimation essentially means combining such data. Figure 6 gives a hystorical glance of the main MF estimators appeared in the literature, derived under different theoretical approaches.

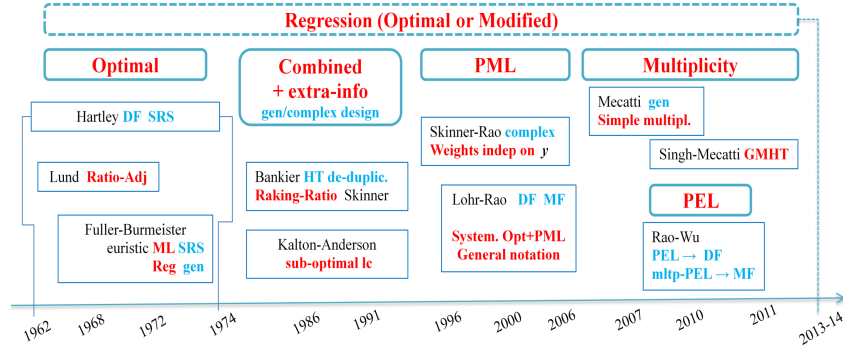


Fig. 6 Hystorical development of Multiple Frame estimators and estimation approaches

MF surveys have been first introduced by Hartley, essentially focused on DF case ($Q = 2$) and simple random sampling (SRS) in each frames; he proposed a unbiased linear combination of customary Horvitz-Thompson (HT) domain estimators under an optimal approach, i.e. minimizing the estimator variance, successively improved by Lund [10] and by Fuller and Burmeister [2] who introduced an euristic maximum likelihood (ML) argument for SRS and a regression DF estimator for general sampling designs. Optimal estimators depend on unknown population variances and covariances that must be estimated from sample data. Beside being computationally demanding, this also leads to an *approximate* optimality in practice that motivated sub-optimal approaches. In the eighties an approach inspired by the stratified design, based on pulling together the Q frame samples into a unique sample was introduced. Bankier [1] focused on the induced inclusion probabilities system so that the usual HT estimator applies. His estimator has the advantage to be easily generalized to any number of frames but requires extra information at the estimation stage in order to identify and eliminate any duplication in the pulled sample. He also empirically showed that efficiency gains can be reached by performing a Raking Ratio adjustment to the known frame sizes N_q , later analytically studied by Skinner [18]. Bankier named this approach *single-frame*, which might be confusing in a MF setup as opposed to the conventional unique-frame setup. For this reasons the term *combined-frames* has been proposed [16] as opposed to the *separate frame* approach of the optimal estimators above. A second unbiased combined DF frame estimator, not requiring de-duplication of the final sample but based on the inclu-

sion probability from *all* frames, was introduced by Kalton and Anderson under SRS [4]. In the nineties Skinner and Rao proposed a Pseudo-Maximum Likelihood (PML) estimator for complex designs, which guarantees a unique set of weights not depending on the study variable in multipurpose surveys [20]. Then Lohr and Rao first tried a systematic treatment of the problem, including variance estimation via resampling, for DF estimation in 2000 [8] and for MF estimation in 2006 [9], by developing a first MF notation as already mentioned in section 2. Singh and Wu introduced a *Modified Regression* DF estimator able to catch the design effects when different sampling designs are used in each frame [16]. A *multiplicity* MF estimator has been introduced in 2007 [11] based on fixed weights, with the main advantages of *i*) being very simple to implement for any number of frames; *ii*) requiring limited frame membership information to be collected; and *iii*) offering simple closed formulae for unbiased variance estimation. A Pseudo-Empirical Likelihood (PEL) DF estimator and a multiplicity-based PEL estimator applying to MF were proposed in 2010 [13]. A unified and simplified approach to MF estimation based on multiplicity is given in Singh and Mecatti (2011). A Generalize Multiplicity-adjusted class of HT estimators is derived (GMHT), which *i*) applies to any number of frames; *ii*) allows for exact variance estimation derived straightforward from usual HT theory; and *iii*) has the potential of including all the main estimators previously derived under different methods (provided that not requiring de-duplications so that with the exception of Bankier's HT estimator). Current work is focusing on further generalizing GMHT class via regression, either optimal or sub-optimal, in a GMHT-reg class of MF estimators [12] [15].

5 Multiplicity-Adjusted Multiple Frame Estimation

In the following sub-sections a sketch of a unified and simplified view to MF estimation is given, as derived by the natural HT approach in survey sampling estimation, according to the multiplicity approach. We focus on the conventional MF estimation problem with the total $Y = \sum_{k=1}^N y_k$ of a quantitative survey characteristic y as the population parameter to be estimated. We consider the general MF case of $Q \geq 2$ independent frame samples with unknown overlapping and selected under possibly different designs. Then, three main issues have to be handled at the estimation stage: 1) data are independent with respect to the parent overlapping frames but not with respect to the disjoint domains into which each frame sample can be partitioned (see Figure 2); 2) units appearing in more than one frame have an increased probability to be selected in the final sample and, at the same time, to be selected more than once, even if this have not occurred in practice; and 3) information about frame membership and the linkage matrix (see Figure 3) must be available, usually collected from every sampled unit besides the study values y_k .

5.1 Simple Multiplicity HT estimator

In a conventional unique-frame survey, the familiar Horvitz-Thompson estimator is computed by using data from a unique-sample s with design weights π_k^{-1} inverting the (first order) inclusion probability of unit k : $\hat{Y}_{HT} = \sum_{k \in s} y_k \pi_k^{-1}$. In a MF survey the target population is covered by $Q \geq 2$ frames so that the population total is given by $Y = \sum_{k \in \bigcup_{q=1}^Q U_q} y_k$. Since frames are assumed both partial and with unknown overlaps, it is convenient to alternatively express the target parameter from a sum over the unavailable union $\bigcup_{q=1}^Q U_q$ of size N (possibly unknown) to a sum over the Q overlapping frames with $\sum_{q=1}^Q N_q \geq N$. This is naturally accomplished by using unit multiplicity m_k , i.e. the number of frames in which every unit is included $Y = \sum_{q=1}^Q \sum_{k \in U_q} y_k m_k^{-1}$. A design-unbiased simple multiplicity estimator thus follows straightforward, also extending to MF the so called DF *average* estimator (see for instance [7], sec. 2.1)

$$\hat{Y}_{SM} = \sum_{q=1}^Q \sum_{k \in s_q} y_k m_k^{-1} \pi_k^{-1} = \sum_{q=1}^Q \sum_{k=1}^{N_q} y_k m_k^{-1} \pi_k^{-1} \mathbb{1}_{k \in s_q} \quad (1)$$

where $\mathbb{1}_{k \in s_q}$ denotes the sample membership indicator of unit k into each frame sample s_q , i.e. an observable random variable under the design-randomization. Expression (1) shows how the (inverse of) unit multiplicity would adjust the design weights with respect to the risk of data duplication in a MF survey. In a conventional unique-frame survey $m_k = 1$ for all $k = 1 \dots N$ and equation (1) reduces to the usual HT estimator. The case of Q non-overlapping frames is equivalent to a target population partitioned into Q strata. In this case too every population unit has unitary multiplicity and equation (1) is equivalent to the usual stratified HT estimator. As a consequence $\hat{Y}_{SM}$ naturally generalizes to MF surveys the conventional HT estimation [14].

Unit multiplicity is a natural choice, still it is not the unique. By defining a *general* multiplicity-adjustment coefficient, a generalized multiplicity-adjusted HT class of estimators for MF survey is also defined.

5.2 Generalized Multiplicity-adjusted HT estimators

In Singh and Mecatti [14] a generalized multiplicity-adjusted HT methodology for MF estimation is given. Let $\alpha_{k(q)} \neq 0$ be a general multiplicity-adjustment coefficient for every unit k in a given frame U_q with $\sum_q \alpha_{k(q)} = 1$. By substituting in equation (1) a class of design-unbiased estimators is given as *Generalized Multiplicity-adjusted Horvitz-Thompson* (GMHT) class of MF estimators

$$\hat{Y}_{GMHT} = \sum_{q=1}^Q \sum_{k=1}^{N_q} y_k \alpha_{k(q)} \pi_{k(q)}^{-1} \mathbb{1}_{k \in s_q} \quad (2)$$

where the coefficient $\alpha_{k(q)}$ ensures that y_k is counted once even if unit k is duplicated in more than one frame, i.e. $\sum_{q=1}^Q \alpha_{k(q)} = 1$. Different GMHT estimators follow by different choices for the multiplicity-adjustment α -coefficient in (2). The simple multiplicity-adjusted MF estimator as given in (1) is the simplest GMHT estimator with the basic choice $\alpha_{k(q)} = m_k$, unit multiplicity being constant whatever the frame(s) U_q listing unit k .

The GMHT class has the potential of encompassing the range of MF estimators available in the literature and briefly discussed in Section 4. Notice that, as defined in (2), the class includes all the known design-unbiased MF estimators. However, the unbiasedness requirement can be relaxed to embrace into the GMHT methodology all the MF estimators known so far, as it will be outlined in Section 6.

Singh and Mecatti [14] was primarily concerned with combined-frame estimators, requiring to adjust for multiplicity by a unit/frame-specific α -coefficient, so that equation (2) applies straightforward. Consider for instance the Kalton and Anderson estimator originally proposed for DF and SRS [4] :

$$\hat{Y}_{KA} = \hat{Y}_{a(A)} + \frac{\pi_A}{\pi_A + \pi_B} \hat{Y}_{ab(A)} + \frac{\pi_B}{\pi_A + \pi_B} \hat{Y}_{ab(B)} + \hat{Y}_{b(B)} \quad (3)$$

where $\pi_A = n_a/N_A$ and $\pi_B = n_b/N_B$. In other words Kalton and Anderson proposed to adjust data from the overlapping domain - where data duplication might occur - by giving higher weight to the estimator from the frame with higher inclusion probability. In Singh and Mecatti [14] a general MF form for estimator (3) is given for any number of frames $Q \geq 2$ and under a general sampling design in each frame

$$\hat{Y}_{KA} = \sum_{q=1}^Q \sum_{k \in U_q} y_k \mathbb{1}_{k \in s_q} \left(\sum_{d=1}^{D_q} \sum_{q'=1}^Q \pi_{k(q')} \mathbb{1}_{k \in U_{d(q')} \equiv U_{d(q)}} \right)^{-1} \quad (4)$$

Thus $\hat{Y}_{KA}$ is a GMHT estimator with the following choice for the multiplicity-adjustment α -coefficient $\alpha_{k(q)}^{KA} = \pi_{k(q)} \left(\sum_{d=1}^{D_q} \sum_{q'=1}^Q \pi_{k(q')} \mathbb{1}_{k \in U_{d(q')} \equiv U_{d(q)}} \right)^{-1}$.

5.3 GMHT Regression Form

Most of DF and MF estimators can be computed under a separate-frame approach such as *i)* first consider separately data from each frame-sample s_q post-classified into domain-samples $s_{d(q)}$, $d = 1 \cdots D_q$; *ii)* compute estimate of domain totals; and *iii)* finally aggregate all domain estimates to produce the overall estimate of the population total. A design-unbiased separate-frame estimator can be expressed as GMHT by considering a domain-specific multiplicity-adjustment α -coefficient

$$\alpha_{k(q)} = \sum_{d=1}^{D_q} \alpha_{d(q)} \mathbb{1}_{k \in U_{d(q)}} \quad \text{with} \quad \sum_{q=1}^Q \sum_{d=1}^{D_q} \alpha_{d(q)} \mathbb{1}_{k \in U_{d(q)}} = 1 \quad (5)$$

Expression (5) applies for instance to the original Hartley's estimator, proposed as a design-unbiased optimal estimator for DF under SRS in both frame [3]

$$\hat{Y}_H = \hat{Y}_{a(A)} + \alpha \hat{Y}_{ab(A)} + (1 - \alpha) \hat{Y}_{ab(B)} + \hat{Y}_{b(B)} \quad (6)$$

with α optimal, namely chosen to minimize the actual estimator variance $V(\hat{Y}_H)$

$$\alpha^H = \frac{V(\hat{Y}_{ab(B)}) + \text{Cov}(\hat{Y}_{b(B)}, \hat{Y}_{ab(B)}) - \text{Cov}(\hat{Y}_{a(A)}, \hat{Y}_{ab(A)})}{V(\hat{Y}_{ab(A)}) + V(\hat{Y}_{ab(B)})} \quad (7)$$

Estimator (6), under the same argument used above for the Kalton and Anderson estimator, can be generalized to a MF survey with any number $Q \geq 2$ of frames under a general sampling design in each frame

$$\hat{Y}_H = \sum_{q=1}^Q \sum_{k \in U_q} y_k \mathbb{1}_{k \in s_q} \pi_{k(q)}^{-1} \sum_{d=1}^{D_q} \alpha_{d(q)}^H \mathbb{1}_{k \in U_{d(q)}} \quad (8)$$

Hence Hartley's MF estimator is in fact GMHT, with multiplicity-adjustment α -coefficient as given by (5) and optimal $\alpha_{d(q)}^H$ minimizing $V(\hat{Y}_H)$ under the unitary sum constraint. No easy formulae in closed form generalizing to MF the DF solution (7) are available so far for optimal $\alpha_{d(q)}^H$. In Lohr and Rao [9] a method based on solving a linearized system is provided. A simpler solution in a closed form can be given (MecattiSingh2014) based on expressing $\hat{Y}_H$ under a regression representation, as already suggested for DF surveys [16].

First notice that each and all units included in the same domain share the same multiplicity, so that unit multiplicity m_k for all $k \in U_{d(q)}$ also defines the multiplicity $m_{d(q)}$ of domain $U_{d(q)} \ni k$. Therefore $m_{d(q)}$ estimators are available for the same domain total, one from each frame-sample intersecting into that domain. Secondly, by starting from the simple multiplicity-adjusted estimator (1) with initial basic multiplicity-adjustment $\alpha_{k(q)} = m_k$, excluding domains of unitary multiplicity which in fact do not need any multiplicity adjustment, a regression form would follow from the convex combination of the $m_{d(q)}$ independent HT estimators available for the same domain total which would act as predictors. Thus equation (8) can be rewritten as

$$\hat{Y}_H = \hat{Y}_{SM} - \sum_{d=1}^D \left[\sum_{q=1}^{m_d} \sum_{q' > q=1}^{m_d} \beta_{d(qq')} \left(\hat{Y}_{d(q)}^{HT} - \hat{Y}_{d(q')}^{HT} \right) \right] \quad (9)$$

where $\hat{Y}_{d(q)}^{HT} = \sum_{k \in U_q} y_k \pi_{k(q)}^{-1} \mathbb{1}_{k \in s_{d(q)}}$ is the HT estimator of the total of domain $U_{d(q)}$ computed with data from frame-sample s_q classified into domain-samples $s_{d(q)} = s_q \cap U_{d(q)}$, $d = 1 \dots D_q$. This yields to an explicit closed expression for the vector $\underline{\beta}$ of optimal regression coefficients $\beta_{d(qq')}$ in (9)

$$\underline{\beta} = V_{\varphi\varphi}^{-1} \text{Cov}(\underline{\varphi}, \hat{Y}_{SM}) \quad (10)$$

where $\underline{\varphi}$ is the vector of available predictors $\varphi_{d(qq')} = \hat{Y}_{d(q)}^{HT} - \hat{Y}_{d(q')}^{HT}$, $V_{\varphi\varphi}$ defines the Var-Cov matrix of $\underline{\varphi}$ and $Cov(\underline{\varphi}, \hat{Y}_{SM})$ is the vector of covariances between every predictor in $\underline{\varphi}$ and $\hat{Y}_{SM}$ (see [16] for details). Finally, some little algebra over the equivalent expressions (8) and (9) gives the following closed solution for the optimal α -coefficient of Hartley's, estimator

$$\alpha_{d(q)}^H = \begin{cases} 1 & \text{if } m_{d(q)} = 1 \\ m_{d(q)}^{-1} - \sum_{q' \neq q=1}^{m_{d(q)}} \beta_{d(qq')} (-)^{q' > q} \mathbb{1}_{U_{d(q)} \equiv U_{d(q')}} & \text{if } m_{d(q)} \geq 2 \end{cases} \quad (11)$$

where β -coefficients can be computed by standard software for linear regression analysis. Moreover equation (11) generalizes the DF solution given in (7) and the unitary sum constraint is readily proved.

5.4 Variance Estimation

Unbiased multiplicity-adjusted estimators included in the GMHT class (2) are in fact linear combinations of Q independent HT estimators - one from each frame-sample s_q . Thus the estimator variance as well as an unbiased variance estimator follow directly from HT-type estimation [14]. In other words the simplification induced by the multiplicity approach allows for standard variance estimation in closed form, while resampling and/or asymptotics have been usually suggested in the literature. The exact variance of any GMHT estimator in the well known HT form is given by

$$V(\hat{Y}_{GMHT}) = \sum_{q=1}^Q \left[\sum_{k \in U_q} y_k^2 \alpha_{k(q)}^2 \frac{1 - \pi_{k(q)}}{\pi_{k(q)}} + \sum_{k \neq k' \in U_q} \frac{y_k \alpha_{k(q)} y_{k'} \alpha_{k'(q)}}{\pi_{k(q)} \pi_{k'(q)}} (\pi_{k(q)k'(q)} - \pi_{k(q)} \pi_{k'(q)}) \right] \quad (12)$$

Moreover an unbiased variance estimator then follows straightforward in closed form by substituting in (12) standard estimators of variance and covariances between HT domain estimators.

6 Generalizing the multiplicity approach: research perspectives

The GMHT class of unbiased MF estimators as defined in (2) can be further generalized for also including approximately or asymptotically unbiased estimators shown in Figure 6. Let $\delta_{k(q)}$ be an observable random variable generalizing the sample membership indicator $\mathbb{1}_{k \in s_q}$ in expression (2) and let $w_{k(q)}$ define general design-weights. An alternative more general definition of GMHT class is then

$$\hat{Y}_{GMHT} = \sum_{q=1}^Q \sum_{k \in U_q} y_k \alpha_{k(q)} w_{k(q)} \delta_{k(q)} \quad (13)$$

which includes definition (2) with $\delta_{k(q)} = \mathbb{1}_{k \in s_q}$ and $w_{k(q)} = E(\mathbb{1}_{k \in s_q})^{-1} = \pi_{k(q)}^{-1}$. Moreover any GMHT estimator included in (13) will be exactly unbiased if $w_{k(q)} = E(\delta_{k(q)})^{-1}$ or else approximately unbiased if $w_{k(q)} \simeq E(\delta_{k(q)})^{-1}$. The α -coefficient in (13) still represents the multiplicity-adjustment with the unitary constraint, either frame-specific as in (13) or domain-specific as in (5). Internal consistency (see for instance [7] section 2) is granted by all positive α -coefficients.

GMHT class as defined in (13) is likely to encompass also all the approximately unbiased MF estimators appeared in the literature, a regression representation as discussed in subsection 5.3 appearing as a suitable, unifying argument [15]. The same form also offer a flexible weighting approach, suitable for including available extra information. Future research will consider calibration to known auxiliary totals of GMHT-reg estimators.

Finally note that a large amount of simulation results have appeared in the literature about DF and MF surveys. Although empirical evidence supplies some recommendations and suggestions for subsets of compared MF estimators, still the lack of homogeneity of the estimation approach can make it difficult to compare and choose. The GMHT-reg class would provide a simplified framework for comparisons as well as for the derivation of new estimators according to both theoretical requests and specific application conditions.

References

1. Bankier, M.D: Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys. J. Am. Stat. Ass. **81**, 1074–1079 (1986)
2. Fuller, W.A. and Burmeister, L.F.: Estimators of samples selected from two overlapping frames. ASA Proceedings Soc. Stat Sec., 245–249 (1972)
3. Hartley H.O.: Multiple frame surveys. ASA Proceedings Soc. Stat Sec., 203–206 (1962); Multiple Frame Methodology and Selected Applications. Sankhya, **C 36**, 99–118 (1974)
4. Kalton G. and Anderson D.W.: Sampling rare populations. J. Royal Statistical Society, **A 149**, 65–82 (1986)
5. Kwok, P.K. and Shin, H.C. and Murphy, W. and O’Muircheartaigh, C. and Debello, A. and Carris, K.: Cross-community Comparison and Multi-frame Weighting in REACH U.S. JSM Proceedings Joint Statistical Meeting - Survey Research Methods Sec., 2655–2662 (2009)

6. Lavalle, P.: Indirect Sampling. Springer Series in Statistics (2007)
7. Lohr S.L.: Multiple Frame Survey in Sample Survey: Design, Methods and Applications. In: Pfeffermann D. and Rao C.R. (eds.) Handbook of Statistics 29 A, pp. 71-88 North-Holland (2009)
8. Lohr, S. and Rao, J. N. K.: Inference from Dual Frame Surveys. J. Am. Stat. Ass. **95**, 271–280 (2000)
9. Lohr, S. and Rao, J. N. K.: Estimation in Multiple Frame Surveys. J. Am. Stat. Ass. **101**, 1019–1030 (2006)
10. Lund, R.E.: Estimators in multiple frame surveys. ASA Proceedings Soc. Stat. Sec., 282–288 (1972)
11. Mecatti, F.: A Single Frame Multiplicity Estimator for Multiple Frame Surveys. Survey Meth., **33**, 151–158 (2007)
12. Mecatti F. and Singh A.C.: Estimation in Multiple Frame Surveys: A Simplified and Unified Review using the Multiplicity Approach. J. French Statist. Soc., To Appear (2014)
13. Rao, J. N. K. and Wu, C.: Pseudo Empirical Likelihood Inference for Multiple Frame Surveys. J. Am. Stat. Ass. **105**, 1494–1503 (2010)
14. Singh, A.C. and Mecatti, F.: Generalized Multiplicity-adjusted Horvitz-Thompson type Estimation as a Unified Approach to Multiple Frame Surveys. J. Official Stat., **27**, 633–650 (2011)
15. Singh A.C. and Mecatti F.: Generalized Multiplicity-Adjusted Multiple Frame Estimation via Regression. Proceedings STC Symposium (2013)
16. Singh, A.C. and Wu, S.: An Extension of Generalized Regression Estimator to Dual Frame Surveys. JSM Proceedings Joint Statistical Meeting - Survey Research Methods Sec., 3911–3918 (2003)
17. Sirken, M. G.: Network Sample Surveys of Rare and Elusive Population: a Historical review. Proceedings STC Symposium (2004)
18. Skinner, C. J.: On the Efficiency of Raking Ratio Estimation for Multiple Frame Survey. J. Am. Stat. Ass. **86**, 779–784 (1991)
19. Skinner C. J., Holmes D. J., Holt D.: Multiple Frame Sampling for Multivariate Stratification. International Statistical Review **62**, 333–347 (1994)
20. Skinner, C. J. and Rao, J.N.K.: Estimation in dual frame surveys with complex designs. J. Am. Stat. Ass. **91**, 349–356 (1996)