# INDIRECT SAMPLING IN CONTEXT OF MULTIPLE FRAMES

Manuela Maia
School of Economics & Management, Catholic University of Portugal
email: mmaia@porto.ucp.pt

## Abstract

Under-coverage is one of the most common problems of sampling frames. To reduce the impact of coverage error on survey estimates several frames can be combined in order to achieve a complete coverage of the target population. Multiple frame estimators have been developed to be used in the context of multiple frame surveys. Sampling frames may overlap which is the case when a single unit of the sampling frame is related with more than one element of the target population. Indirect sampling (Lavallée, 1995) is an alternative approach to classical sampling theory in dealing with the overlapping problem of sampling frames on survey estimates. In this paper a new class of estimators is presented which is the result from merging dual frames estimators with indirect sampling estimators in order to bring together in a single estimator the effect of several frames on survey estimates.

**Key Words:** Indirect Sampling, Generalized Weight Share Method, Multiple

Frame Surveys

## 1. Introduction

In any survey the random selection of the sample requires that a sampling frame is available. The sampling frame is used to identify the elements of the target population. The frames may be maps of areas in which elements can be found, among others. At their simplest, sampling frames consist of a list of population elements (Groves et al, 2007). There are populations for which lists are readily available, such as members of a professional organization, hospitals or schools. There are many populations, though, for which lists of individuals elements are not readily available. For example the adults living in a country, or the students attending school on a specific district.

When available, one central statistical concern for the survey researcher is how well the sampling frame actually covers the target population. A sampling frame is perfect when there is a one-to-one mapping of frame elements to target population elements. In practice, perfect frames seldom exist; there are always problems that disrupt the desired one-to-one mapping, namely: (a) under coverage, (b) duplication and (c) over coverage. Under coverage happens when some elements of the target population do not appear in the sampling frame; therefore such elements cannot appear in any sample drawn for the survey. Duplication happens when several frame units within a given frame are mapped onto the single elements in the target population, which makes the mapping not unique, not one-to-one. Over coverage occurs when multiple elements of different sampling frames are linked to the same single unit of the target population, i.e., a many-to-one mapping. There are also cases that combine the duplication and the over coverage problems in which multiple frame units map to one target population element.

Selecting a sample from a sampling frame that suffers from under coverage can cause coverage error on survey statistics. One of the strategies to reduce coverage error is to use

multiple frames. A principal frame that provide nearly complete coverage of the target population may be supplemented by a frame that provides better or unique coverage for the population elements absent or poorly covered in the principal frame. In most cases supplemental frames overlap with the principal frame requires estimation procedures to be adapted in order to correct probabilities of selection, which might to yield improved precision for survey estimates.

Selecting a sample from a sampling frame that suffers from either over coverage or duplication poses several difficulties to estimation, namely in what concerns sample weights computation.

## 2. Multiple frame estimators

The estimation under multiple frame designs was originally proposed by Hartley (1962) and others. They suggested that the union of the frames be used in estimation to obtain a more efficient estimator. They proposed that a dual frame design be examined as a set of no overlapping domains and results from each domain combined to obtain a target population estimate. By taking Q sampling frames - $A_1$, $A_2$, ..., $A_q$ - (that may overlap) to cover the target population $2^Q$-1 domains mutually exclusive can be defined. In the particular case of Q=2 three mutually exclusive domains can be defined: $D_1$, contains elements exclusively from frame 1, that is $D_1 = A_1 \cap \overline{A}_2$, $D_2$, contains the elements that belong simultaneously to both frames, that is, $D_2 = A_1 \cap A_2$, and $D_3$, contains elements exclusively from frame 2, that is $D_3 = \overline{A}_1 \cap A_2$. In this context, the dual frame estimator of the total population proposed by Hartley (1974) is based on the weighted average of the total estimates from the domains:

$$\hat{Y}(\theta) = \hat{Y}_{D_1}^{A_1} + \theta\, \hat{Y}_{D_2}^{A_1} + (1-\theta)\, \hat{Y}_{D_2}^{A_2} + \hat{Y}_{D_3}^{A_2} \qquad (1)$$

where θ (0≤θ≤1) is a parameter chosen to maximize $V[\hat{Y}(\theta)]$, $\hat{Y}_{D_1}^{A_1}$ is the total estimate from $D_1$, $\hat{Y}_{D_3}^{A_2}$ is the total estimate from $D_3$, $\hat{Y}_{D_2}^{A_2}$, the total estimate from $D_2$ using a sample from $A_2$ and $\hat{Y}_{D_2}^{A_1}$ is the total estimate of from $D_2$ using a sample from $A_1$.

Alternatively, the population total Y may be represented by the following expression (Hartley 1962, 1974):

$$Y = \sum_{i\in\bigcup_q \mathbf{A}_q} \mathrm{y_i} = \sum_K \sum_{i\in\bigcup_q \mathbf{A}_q} \delta_i(K)\,\mathrm{y_i} \tag{2}$$

where $\delta_i(K)$ is an indicator of domain variable:

$$\delta_i(K) = \begin{cases} 1 & ,\mathrm{i}\in \mathrm{D_k} \\ 0 & ,\text{otherwise} \end{cases}$$

The sample selected from each frame is then used to produce an estimate for the total in each domain, which in turn, is combined to produce a single estimate for the population total. The estimator is given by

$$\hat{Y} = \sum_K \sum_{q\in \mathrm{K}} \sum_{i\in\bigcup_q \mathbf{A}_q} w_i^{(q)} \delta_i(K)\,\mathrm{y_i} \tag{3}$$

and it requires the weights $w_i^{(q)}$ to be computed.

In the literature there are two approaches to estimate these weights: the Domain Membership approach and the Multiplicity Unit approach. According to Domain Membership approach a partition of domains is defined in the frames, in such a way that it is always possible to correctly identify to which domain belongs each element of the sample. There are three types of estimators, depending on the fixed weights they use, in this class of estimators:

(a) The Optimal Estimator $w_{i,opt}^{(q)}$ - presents good theoretical properties - it has minimal variance (Hartley 1962, 1974; Lund 1968; Fuller and Burmeister 1972) - but, in operational terms, is very complex.

(b) The Single Based Estimator $w_{i,SF}^{(q)}$ - uses fixed weights guaranteeing unbiased estimates (Bankier 1986; Kalton and Anderson 1986; Skinner 1991; Skinner, Holmes and Holt 1994), however, are less efficient than the optimal estimator (Lohr and Rao 2000).

(c) The Pseudo Maximum Likelihood Estimator $w_{i,PML}^{(q)}$ - extends the applicability of the optimal estimator increasing its efficiency when compared with the single based estimator (Skinner and Rao 1996; Lohr and Rao 2000).

The Unit Multiplicity estimators are based on the concept of unit multiplicity which reflects the number of frames to which a sample element belongs (Mecatti 2007). This concept was first used by Casady and Sirken (1980). Under this approach, the population total can be written as:

$$\sum_q \sum_{\mathrm{i}\in \mathrm{A_q}} \mathrm{y_i} = \sum_{\mathrm{i}\in\bigcup_q \mathrm{A_q}} \mathrm{m_i\ y_i} = \sum_{\mathrm{q=1}}^{\mathrm{Q}} \sum_{\mathrm{i}\in \mathrm{A_q}} \mathrm{y_i\ m_i^{-1}} = \mathrm{Y} \tag{4}$$

and involves solely the frames to which the sample element belongs. The expression of the population total estimator is given by:

$$\hat{Y}_M = \sum_{q=1}^{Q} \sum_{i\in \mathbf{s}_\mathrm{q}} w_i^{(q)}\ \mathrm{y_i}\ m_\mathrm{i}^{-1} \tag{5}$$

where Q is the number of frames, $\mathrm{m_i} = \sum_\mathrm{q} \delta_\mathrm{i}^{(\mathrm{A_q})}$ and $\delta_\mathrm{i}^{(\mathrm{A_q})} = \begin{cases} 1 & \text{if} \quad \mathrm{i}\in \mathrm{A_q} \\ 0 & \text{if} \quad \mathrm{i}\notin \mathrm{A_q} \end{cases}$ is the number of frames in which each unit is include among the frames involved in the survey.

Mecatti (2007) provides argument to apply Unit Multiplicity estimators in surveys with more than two sampling frames based on overlapping.

## 3. Indirect Sampling and the Generalized Weight Share Method

In Classical Sampling Theory the weight for each sampled element is related to the inverse of its selection probability. The Horvitz-Thompson estimator for the population total - $\hat{Y}^{HT}$ - resumes this principle:

$$\hat{\mathrm{Y}}^{\mathrm{HT}} = \sum_{\mathrm{k} \in \mathrm{S}} \frac{\mathrm{y}_{\mathrm{k}}}{\pi_{\mathrm{k}}} \tag{6}$$

where $\pi_k = P(k \in S)$ is the probability of the *k* element be selected in the samples. This theory assumes that the sampling frame is a perfect representation of the target population, i.e., a one-to-one mapping and is difficultly applied outside this condition.

Indirect Sampling was first proposed by Lavallée (1995) to deal with the problem of Cross-sectional weighting for longitudinal household surveys.

Indirect sampling assumes that a sampling frame $\mathrm{U}^{\mathrm{A}}$ with $\mathrm{M}^{\mathrm{A}}$ units is available to represent the target population $\mathrm{U}^{\mathrm{B}}$. $\mathrm{U}^{\mathrm{B}}$ contains $\mathrm{M}^{\mathrm{B}}$ elements, divided into N clusters, each one with $M_i^B$ elements. A sample $\mathrm{s}_{\mathrm{A}}$ with $\mathrm{m}_{\mathrm{A}}$ units is then selected from the frame $\mathrm{U}^{\mathrm{A}}$ in order to estimate some parameter of the target population $\mathrm{U}^{\mathrm{B}}$. The Generalized Weight Share Method (GWSM), developed by Lavellée (1995) in the context of indirect sampling, uses the links between the units j $\in$ $\mathrm{U}^{\mathrm{A}}$ and the elements *k* of the i$^{\mathrm{th}}$ cluster of $\mathrm{U}^{\mathrm{B}}$ to compute the weight for each element in the sample.

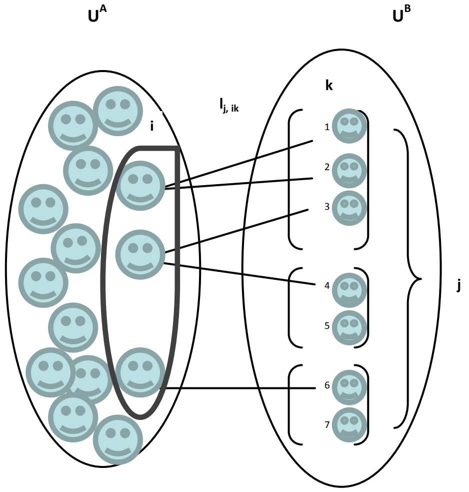

**Figure 1:** Example of links between sampling frame and the target population in Indirect Sampling

Under the GWSM the estimator for the population total is given by:

$$\hat{Y}^B = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik} y_{ik} \qquad (7)$$

where $w_{ik}$ is the weight attached to the element k of cluster *i*, defined by

$$w_i = \sum_{k=1}^{M_i^B} w'_{ik} \Big/ L_i^B \qquad (8)$$

where $w'_{ik}$ corresponds to the inverse of the selection probability of units *j* of $s^A$ that have non-zero link with unit *k* of cluster i of $\hat{Y}^B$.

The process to compute $w_{ik}$ can be resumed in four steps:

1) Compute the number of links $L_{ik}^B$ between the units' $j \in U^A$ and the element *k* of the i^th^ cluster of $U^B$ by

$$L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$$

where $l_{j,ik} = \begin{cases} 1 & \text{if a link between} \quad j \in U^A \text{ and } ik \in U^B \text{exists} \\ 0 & \text{otherwise} \end{cases}$

2) Obtain the total number of links in cluster i: $L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$ .

3) Compute an initial weight

$$w'_{ik} = \frac{\sum_{j=1}^{M^A} l_{j,ik} \; t_j}{\pi_j^A}$$

where $t_j = \begin{cases} 1 & \text{if} \quad j \in s_A \\ 0 & \text{if} \quad j \notin s_A \end{cases}$ and $\pi_j^A$ is the selection probability of unit $j \in s_A$ .

4) Compute the final weight $w_i$ to each sampled element $k \in U_i^B$.

The GWSM estimator can be re-written as (Lavallée 1995):

$$\hat{Y}_B = \sum_{i=1}^{n} w_{ik} \; y_i = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j$$

where $Z_j = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} \; z_{ik}$ and $z_{ik} = \frac{Y_i}{L_i^B}$

The application of the GWSM requires the matching between sampling frame and target population and needs to satisfy the follow constraint:

*There exists, at least, one link between the unit $j \in U^A$ and the elements k of $i^{th}$ cluster of $U^B$ i.e.* $L_j^A = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} \geq 1$ for every $j \in U^A$;

This constraint is essential to ensure de unbiasedness of the GWSM.

Lavallé (1995) proved that the GWSM estimator is unbiased and its variance is directly given by:

$$Var(\hat{Y}^B) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'} \tag{9}$$

where $\pi_{jj'}^A$ is the joint probability of selecting units $j$ and $j'$.

## 4. Combining Multiple Frame Estimators with Indirect Sampling

Both multiple frame designs and indirect sampling seek to improve estimation in surveys where a "perfect" sampling frame does not exist. Suppose for example a RDD survey is used to reach the general adult population of a country. RDD will, in principle, cover all adults living in households with fixed line telephone access but it fails to cover adults living in households without a fixed line telephone. A remedy to under coverage may be a supplementary frame of mobile phone numbers. Under such a dual frame design the two frames together will likely provide a complete (or nearly complete) coverage of the adult population, however an important statistical problem will raise researcher' concern: some adults of the target population may be reachable both by mobile phone and fixed line phone, which means there is a many -to- one mapping. Under these circumstances the estimation approach should merge the solutions coming from dual frame estimation and indirect sampling.

Our proposal is to put dual frame estimators – both the Domain Membership estimators and Unit Multiplicity estimators – to the context of indirect sampling and thus provide an estimation approach adequate for surveys where the sampling frame suffers from under coverage (and several frames are combined to reduce the coverage error).

### 4.1 The Domain Membership estimator

In the context of Indirect Sampling the Domain Membership estimator for the population total can be expressed by:

$$\hat{Y}_{DM} = \sum_{j \in A_1} \frac{z_j(\theta)}{\pi_j^{A_1}} y_j + \sum_{j \in A_2} \frac{x_j(\theta)}{\pi_j^{A_2}} y_j \tag{10}$$

where $z_j(\theta) = \begin{cases} 1 & \text{if} \quad j \in D_1 \\ \theta & \text{if} \quad j \in D_2 \end{cases}$ and $x_j(\theta) = \begin{cases} 1 & \text{if} \quad j \in D_3 \\ (1-\theta) & \text{if} \quad j \in D_2 \end{cases}$ are indicators of domain variables, $(0 \le \theta \le 1)$ and $\pi_j^{A_q}$ represents the selection probability of unit j from the $q^{th}$ frame.

### 4.2 The Unit Multiplicity Estimator

In the context of Indirect Sampling the Unit Multiplicity estimator for the population total can be expressed by:

$$\hat{Y}_M = \sum_{q=1}^{2}\sum_{j=1}^{m_{A_q}} \frac{1}{\pi_j^{A_q}} \sum_{i \in U^B} \frac{L_{ji,q}}{L_i^B} y_j \tag{11}$$

where $L_i^B$ represents the total number of links between the unit j $\in$ $A_q$, (q=1,2) and the element i from $U^B$ and $\pi_j^{A_q}$ represents the selection probability of unit j from $A_q$, (q=1,2). $L_{ji,q}$ is given by:

$$L_{ji,q} = \begin{cases} 1 & \text{if there is a link beteween j-th unit, from } A_q\text{, and the i unit from } U^B \\ 0 & \text{otherwise} \end{cases}$$

### 4.3 The Dual Frame estimator

The estimator proposed by Hartley (1974) (eq. 1) can, in the same way, be converted to Indirect Sampling context:

$$\hat{Y}_H = \sum_{j \in s_{A_1}} \frac{1}{\pi_j^{A_1}} \underbrace{\frac{N_{A_1}}{\hat{N}_{A_1}} \varphi_j^{A_1}}_{C_j} y_j + \sum_{j \in s_{A_2}} \frac{1}{\pi_j^{A_2}} \underbrace{\frac{N_{A_2}}{\hat{N}_{A_2}} \varphi_j^{A_2}}_{D_j} y_j \tag{15}$$

where $\varphi_j^{A_1} = \begin{cases} 1 & \text{if } \delta_j^{A_2} = 0 \\ \tilde{\theta}_{j1}^{A_1} & \text{if } \delta_j^{A_2} = 1 \end{cases}$ $\qquad \varphi_j^{A_2} = \begin{cases} 1 & \text{if } \delta_j^{A_1} = 0 \\ 1 - \tilde{\theta}_{j1}^{A_1} & \text{if } \delta_j^{A_1} = 1 \end{cases}$

$\frac{N_{A_1}}{\hat{N}_{A_1}}$ and $\frac{N_{A_2}}{\hat{N}_{A_2}}$ are the pos-stratified estimators of each sampling frame

$\tilde{\theta}_{j1}^{A_1}$ proportion of elements in frame $A_2$ that also belongs to frame $A_1$

$1 - \tilde{\theta}_{j1}^{A_1}$ - Proportion of elements in frame $A_1$ that also belongs to frame $A_2$

$\delta_j^{A_q}$ is an indicator of frame variable and $\pi_j^{A_q}$ represents the selection probability of unit j from $U^{Aq}$ with q=1,2.

From equation (1) is possible to obtain the classes of estimators above described. Considering that $C_j = z_j(\theta)$ and $D_j = x_j(\theta)$ the class of Domain Membership estimators can be deduced. Replacing $C_j$ and $D_j$ by the proportion of the links from the frames $A_1$ and $A_2$, respectively, i.e., $C_j = \sum_{i \in U^B} \frac{L_{ji,1}}{L_i^B}$ and $D_j = \sum_{i \in U^B} \frac{L_{ji,2}}{L_i^B}$ we obtain the class of Unit Multiplicity estimators.

## Acknowledgements

The authors would like to thank Prof. Pierre Lavallée from Statistics Canada for his helpful comments and contribution to this paper.

## References

Bankier, Michael D. (1986), Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys, *Journal of the American Statistical Association,* Vol. 81, pp.1074-1079.

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

Deville J. C., Lavallée, P. (2006), Indirect sampling: Foundations of Generalized Weight Share Method, *Survey Methodology*, Vol. 32, No. 2, pp. 165-176

Ernest, L. (1989), Weighting issues for longitudinal and family estimates, *Panel Surveys*, (Kasprzyk, D., Duncun, G., Kalton, G., Singh, M. P. Editors), John Wiley and Sons, New York, pp. 135-159.

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

Groves, R., Fowler Jr, F., Couper, M., Lepkowski, J., Singer, E. & Tourangeau, R. (2004) *Survey Methodology.* New York: John Wiley and Sons.

Hartley, H. O. (1974), Multiple Frame Surveys Methodology and Selected Applications. *Sankhyä*. C, 36, 99-118.

Hartley, H.O. (1962), Multiple Frame Surveys, *Proceedings of the American Statistical Association, Social Statistics Section*, pp. 99-118

Kalton G., e Anderson, D. W., (1986), Sampling Rare Populations, *Journal of the Royal Statistical Society, Series A,* vol. 149, nº 1, pp. 65-82

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

Lavallée, P.(1995) Cross–sectional weighting of longitudinal surveys of individuals and households using weight share method. *Survey Methodology*, Vol. 21, No. 1, pp. 25-32.

Lohr and Rao, J. N. K.(2000), Inference from Dual Frame Surveys, *Journal of the American Statistical Association,* Vol. 95, nº 449, pp.271-280.

Lund, Richard E. (1968), Estimators in Multiple Frame, *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 282-288.

Mecatti, F. (2007), A single frame multiplicity estimator for multiple frame surveys, *Survey Methodology*, Vol. 33, No. 2, pp. 151-157

Skinner C. J (1991), On the Efficiency of Ratio Estimation for Multiple Frame Surveys *Journal of the American Statistical Association,* Vol. 86, nº 415, pp.779-784.

Skinner C. J and Rao J. N. K (1996), Estimation on Dual Frame Surveys with Complex Designs, *Journal of the American Statistical Association,* Vol. 91, pp.349-356.

Skinner C. J., Holmes D. J.  e Holt d, (1994), Multiple Frame for Multivariate Stratification, *International Statistical Review*, Vo. 62, nº3, pp. 333-347

Thompson, S. K. (1992), *Sampling*, John Whiley and Sons, New York.