Survey Methodology, December 2006
Vol. 32, No. 2, pp. 165-176
Statistics Canada, Catalogue No. 12-001

Indirect Sampling: The Foundations of the Generalized Weight Share Method

Jean-Claude Deville and Pierre Lavallée [1]

Abstract

To select a survey sample, it happens that one does not have a frame containing the desired collection units, but rather another frame of units linked in a certain way to the list of collection units. It can then be considered to select a sample from the available frame in order to produce an estimate for the desired target population by using the links existing between the two. This can be designated by *Indirect Sampling*.

Estimation for the target population surveyed by Indirect Sampling can constitute a big challenge, in particular if the links between the units of the two are not one-to-one. The problem comes especially from the difficulty to associate a selection probability, or an estimation weight, to the surveyed units of the target population. In order to solve this type of estimation problem, the Generalized Weight Share Method (GWSM) has been developed by Lavallée (1995) and Lavallée (2002). The GWSM provides an estimation weight for every surveyed unit of the target population.

This paper first describes Indirect Sampling, which constitutes the foundations of the GWSM. Second, an overview of the GWSM is given where we formulate the GWSM in a theoretical framework using matrix notation. Third, we present some properties of the GWSM such as unbiasedness and transitivity. Fourth, we consider the special case where the links between the two populations are expressed by indicator variables. Fifth, some special typical linkages are studied to assess their impact on the GWSM. Finally, we consider the problem of optimality. We obtain optimal weights in a weak sense (for specific values of the variable of interest), and conditions for which these weights are also optimal in a strong sense and independent of the variable of interest.

Key Words: Indirect Sampling; Generalized Weight Share Method; Unbiasedness; Optimal Weights.

1. Introduction

To select the samples needed for social or economic surveys, it is useful to have sampling frames, *i.e.*, lists of units intended to provide a way to reach desired target populations. Unfortunately, it happens that one does not have a list containing the desired collection units, but rather another list of units linked in a certain way to the list of collection units. One can speak therefore of two populations U^A and U^B linked to each other, where one wants to produce an estimate for U^B. Unfortunately, a sampling frame is only available for U^A. It can then be considered to select a sample s^A from U^A in order to produce an estimate for U^B by using the correspondence existing between the two populations. This can be designated by *Indirect Sampling*.

Estimation for a target population U^B surveyed by Indirect Sampling can constitute a big challenge, in particular if the links between the units of the two populations are not one-to-one. The problem comes especially from the difficulty to associate a selection probability, or an estimation weight, to the surveyed units of the target population. In order to solve this type of estimation problem, the Generalized Weight Share Method (GWSM) has been developed by Lavallée (1995) and Lavallée (2002), and presented also in Lavallée and Caron (2001). The GWSM provides an estimation weight for every surveyed unit of the target population U^B. Basically, this estimation weight corresponds to a weighted average of the survey weights of the units of the sample s^A. The GWSM is an extension of the Weight Share Method described by Ernst (1989) in the context of longitudinal household surveys.

The purposes of this paper are to describe Indirect Sampling – the foundations underlying the GWSM – and to obtain optimal weights from the GWSM that provide unbiased estimates with minimum variance. First, we will describe Indirect Sampling together with the GWSM in a theoretical framework that will use, for instance, matrix notation. The use of matrix notation for the GWSM has previously been presented by Deville (1998). Second, we will use this theoretical framework to state some general properties associated with the GWSM that include unbiasedness and transitivity. Transitivity is to go from the population U^A to a target population U^C, through an intermediate population U^B. Third, we will show the correspondence between the matrix formulation and the one that has been described in Lavallée (1995), Lavallée (2002), and Lavallée and Caron (2001). Fourth, we will study the effect of various typical link matrices between U^A and U^B on the precision of the estimates obtained from the GWSM. Finally, we will assess the problem of optimality. We will obtain optimal weights in a weak sense (for specific values

1. Jean-Claude Deville, Laboratoire de Statistique d'Enquête (ENSAI/CREST), Campus de Ker Lann, rue Blaise Pascal, 35170 Bruz, FRANCE. E-mail: deville@ensai.fr; Pierre Lavallée, Statistics Canada, Ottawa, Ontario, K1A 0T6, CANADA. E-mail: pierre.lavallee@statcan.ca.

of the variable of interest), and conditions under which these weights are also optimal in a strong sense and independent of the variable of interest.

2. Indirect Sampling

As mentioned in the introduction, with Indirect Sampling, we select a sample s^A from a population U^A in order to produce an estimate for a target population U^B. For that, we use the correspondence existing between the two populations. For example, assume that we want to produce estimates for a population of children (collection units) while we only have a sampling frame of parents. The target population U^B is the one of the children, but we need to select a sample of parents before being able to interview the children. This is illustrated in Figure 1.

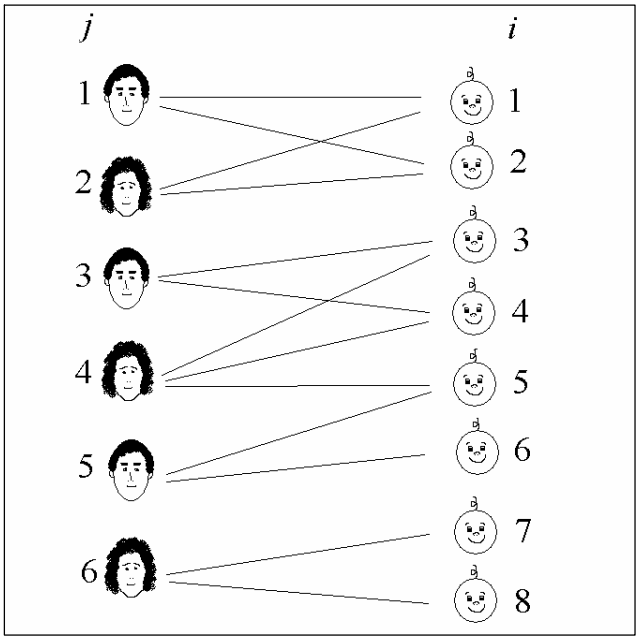

Figure 1. Population U^A of parents and population U^B of children with the links between the two.

Let the population U^A contain N^A units, where each unit is labeled by the letter j. Similarly, let the target population U^B contain N^B units, where each unit is labeled by the letter i. The correspondence between the two populations U^A and U^B can be represented by a *link matrix* $\mathbf{\Theta}_{AB} = [\theta_{ji}^{AB}]$ of size $N^A \times N^B$ where each element $\theta_{ji}^{AB} \geq 0$. That is, unit j of U^A is related to unit i of U^B provided that $\theta_{ji}^{AB} > 0$, otherwise the two units are not related to each other. For the above example, the link matrix is given by

$$\mathbf{\Theta}_{AB} = \begin{bmatrix} \theta_{11}^{AB} & \theta_{12}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{21}^{AB} & \theta_{22}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{33}^{AB} & \theta_{34}^{AB} & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{43}^{AB} & \theta_{44}^{AB} & \theta_{45}^{AB} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{55}^{AB} & \theta_{56}^{AB} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_{67}^{AB} & \theta_{68}^{AB} \end{bmatrix}.$$

Obtaining the link matrix *link matrix* $\mathbf{\Theta}_{AB} = [\theta_{ji}^{AB}]$ is a critical issue in Indirect Sampling. For the case where two units $j \in U^A$ and $i \in U^B$ are not linked, we simply set $\theta_{ji}^{AB} = 0$. When there is a link between two units j and i, the choice of $\theta_{ji}^{AB} > 0$ is important. As we will see, it influences the precision of the estimates issued from Indirect Sampling. Now, in several applications, the values of θ_{ji}^{AB} for the linked units are simply set to 1. Of course, the values of θ_{ji}^{AB} for the linked units can be chosen to be different from 1. Lavallée and Caron (2001) discussed the use of the linkage weights obtained from a record linkage process between U^A and U^B for assigning values to the θ_{ji}^{AB}. The linkage weights are proportional to the probability of two units $j \in U^A$ and $i \in U^B$ being linked. Since the choice of $\theta_{ji}^{AB} > 0$ for two linked units j and i can affect the precision of the estimates, it is natural to seek for those θ_{ji}^{AB} that will minimize the variance of the estimates. This optimization problem is considered in section 6 of the paper.

With Indirect Sampling, we select the sample s^A of n^A units from U^A using some sampling design. Let π_j^A be the selection probability of unit j. We assume $\pi_j^A > 0$ for all $j \in U^A$. For each unit j selected in s^A, we identify the units i of U^B that have a non-zero correspondence, *i.e.*, with $\theta_{ji}^{AB} > 0$. Let Ω^B be the set of the n^B units of U^B identified by the units $j \in s^A$, *i.e.*, $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A$ and $\theta_{ji}^{AB} > 0\}$. For each unit i of the set Ω^B, we measure a variable of interest y_i from the target population U^B. Let $\mathbf{Y} = \{y_1, \ldots, y_{N^B}\}'$ be the column vector of that variable of interest. In a practical view point, it is important to mention that although the sample size n^A is usually determined in advance, the number of units n^B is difficult to control because it depends on the selected sample s^A and the link matrix $\mathbf{\Theta}_{AB}$. As a consequence, it turns out to be difficult in general to establish a budget for measuring the variable of interest y_i. Fortunately, in most applications (*e.g.*, the parents-children case above), the number of links that start from a given unit j of s^A is somewhat predictable (for example, a parent typically has one, two, or three children), which helps to assess how many units i of U^B will finally be measured.

We assume that for any unit j of s^A, the correspondences for $i = 1, \ldots, N^B$ can be obtained. That is, we can identify all the links between the two populations by direct interview or by some administrative source for any sampled

unit j. Also, for any identified unit i of U^B, we assume that the links for $j = 1, ..., N^A$ can be obtained (as mentioned by Lavallée (2002), there are cases where this last constraint can be difficult to satisfy in practice. Referring to the example of parents and children, it might not be easy for a very young child, selected through his mother, to mention back his father, when the two parents are divorced. In order to simplify the discussion, such a problem of identification of links will be assumed to be negligible). Therefore, the values of the links need not to be known between the entire populations U^A and U^B. In fact, we need to know the links (and consequently the values of θ_{ji}^{AB}) only for the lines j of $\mathbf{\Theta}_{AB}$ where $j \in s^A$, and also for columns i of $\mathbf{\Theta}_{AB}$ where $i \in \Omega^B$.

Suppose that we are interested in estimating the total Y^B of the target population U^B where $Y^B = \sum_{i=1}^{N^B} y_i$. We can also write $Y^B = \mathbf{1}_B' \mathbf{Y}$ where $\mathbf{1}_B$ is the column vector of 1's of size N^B (note that we use for simplification the notation $\mathbf{1}_B$ instead of $\mathbf{1}_{N^B}$). Now let $\theta_{+i}^{AB} = \sum_{j=1}^{N^A} \theta_{ji}^{AB}$ and let $\tilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \theta_{+i}^{AB}$. We have $\mathbf{1}_A' \mathbf{\Theta}_{AB} = \{\theta_{+1}^{AB}, ..., \theta_{+N^B}^{AB}\}$. We then define the *standardized link matrix* $\tilde{\mathbf{\Theta}}_{AB} = \mathbf{\Theta}_{AB} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1}$, where $\text{diag}(\mathbf{v})$ is the square matrix obtained by putting the elements of the row-vector (or column-vector) $\mathbf{v}$ in the diagonal, and 0 elsewhere. Note that in order for the matrix $\tilde{\mathbf{\Theta}}_{AB}$ to be well defined, we must have $[\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1}$ to exist, which is the case if and only if $\theta_{+i}^{AB} > 0$ for all $i = 1, ..., N^B$. For the parents-children example, this means that every child must be linked to at least a parent.

Result 1:

The link matrix $\tilde{\mathbf{\Theta}}_{AB}$ is a standardized link matrix if and only if

$$\tilde{\mathbf{\Theta}}_{AB}' \mathbf{1}_A = \mathbf{1}_B. \tag{2.1}$$

The proof of Result 1 follows directly from the definition of a standadized link matrix. Using Result 1, we directly obtain Result 2 that can also be found in Deville (1998):

Result 2:

$$\begin{aligned} Y^B &= \mathbf{1}_B' \mathbf{Y} \\ &= \mathbf{1}_A' \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y} = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \frac{\theta_{ji}^{AB}}{\theta_{+i}^{AB}} y_i. \end{aligned} \tag{2.2}$$

Let us define the column vector $\mathbf{Z} = \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$ of size N^A. Considering each line of $\mathbf{Z}$, the variable $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$ is defined for each unit j of the population U^A and measured for each unit $j \in s^A$.

For estimating Y^B, we want to use the values of y_i measured from set Ω^B. For this, we will use an estimator of the form:

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i y_i \tag{2.3}$$

where w_i is the estimation weight of the unit i of Ω^B, with $w_i = 0$ for $i \notin \Omega^B$. Let $\mathbf{W}' = \{w_1, ..., w_{N^B}\}$. The estimator (2.3) can be rewritten as

$$\hat{Y}^B = \mathbf{W}'\mathbf{Y}. \tag{2.4}$$

Usually, to get an unbiased estimate of Y^B, one can simply use as the weight the inverse of the selection probability π_i^B of unit i. As mentioned by Lavallée (1995) and Lavallée (2002), with Indirect Sampling, this probability can however be difficult, or even impossible, to obtain. It is then proposed to use the GWSM, which is defined as follows.

Let $\boldsymbol{\pi}^A = \{\pi_1^A, ..., \pi_{N^A}^A\}'$ and let $\mathbf{\Pi}_A = \text{diag}(\boldsymbol{\pi}^A)$ be the diagonal matrix of size $N^A \times N^A$ containing the selection probabilities used for the selection of sample s^A. Accordingly, let $\mathbf{t}^A = \{t_1^A, ..., t_{N^A}^A\}'$ where $t_j^A = 1$ if $j \in s^A$, and 0 otherwise. Let $\mathbf{T}_A = \text{diag}(\mathbf{t}^A)$ be the diagonal matrix of size $N^A \times N^A$ containing the indicator variables t_j^A. Starting from $Y^B = \mathbf{1}_A' \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y} = \mathbf{1}_A' \mathbf{Z}$, we can directly form the following Horvitz-Thompson estimator in terms of the vector $\mathbf{Z}$:

$$\hat{Y}^B = \mathbf{1}_A' \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{Z}. \tag{2.5}$$

Using the fact that $\mathbf{Z} = \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$, we have $\hat{Y}^B = \mathbf{1}_A' \mathbf{T}_A \mathbf{\Pi}_A^{-1} \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$ and therefore we can define the column vector $\mathbf{W}$ of weights:

$$\mathbf{W} = \tilde{\mathbf{\Theta}}_{AB}' \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A. \tag{2.6}$$

The vector $\mathbf{W}$ is of size N^B and for each $i = 1, ..., N^B$, we have $w_i = \sum_{j=1}^{N^A} t_j^A \tilde{\theta}_{ji}^{AB} / \pi_j^A$. The weights w_i of that vector are said to be obtained from the GWSM, as described by Lavallée (2002).

3. Properties of the GWSM

3.1 Unbiasedness

As mentioned by Ernst (1989), to get an unbiased estimator, we only need to have $E(\mathbf{W}) = \mathbf{1}_B$. By construction, because the estimator (2.5) is a Horvitz-Thompson estimator, this condition is directly satisfied and therefore, the GWSM produces unbiased estimates.

From this discussion, we can in addition obtain the following result:

Result 3:

The vector of weights $\mathbf{W}$ given by (2.6) provides unbiased estimates if and only if the matrix $\tilde{\mathbf{\Theta}}_{AB}$ is a standardized link matrix.

Proof:

Starting from (2.6), we have

$$E(\mathbf{W}) = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{1}_A. \tag{3.1}$$

Using Result 1, we directly get $E(\mathbf{W}) = \mathbf{1}_B$ and therefore we have unbiased estimates. Now, assume that $E(\mathbf{W}) = \mathbf{1}_B$. From (3.1), we must have $\tilde{\mathbf{\Theta}}'_{AB}\mathbf{1}_A = \mathbf{1}_B$ and therefore, $\tilde{\mathbf{\Theta}}_{AB}$ is a standardized link matrix.

3.2 Variance

Because the estimator (2.5) is a Horvitz-Thompson estimator, we directly obtain the following result:

Result 4:

The variance of $\hat{Y}^B$ is given by

$$\begin{aligned}\mathrm{Var}(\hat{Y}^B) &= \mathbf{Z}'\mathbf{\Delta}_A\mathbf{Z} \\ &= \mathbf{Y}'\mathbf{\Delta}_B\mathbf{Y}\end{aligned} \tag{3.2}$$

where $\mathbf{\Delta}_A = [(\pi^A_{jj'} - \pi^A_j\pi^A_{j'})/\pi^A_j\pi^A_{j'}]_{N^A\times N^A}$ is a non-negative definite matrix of size $N^A \times N^A$ and where $\pi^A_{jj'}$ is the joint selection probability of units j and j' from U^A, and where $\Delta_B = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AB}$.

For a proof of the variance of the Horvitz-Thompson estimator, see Särndal, Swensson and Wretman (1992).

3.3 Transitivity

Let us suppose that we are interested in producing estimates for a target population U^C that can only be reached through the population U^B. We assume that the target population U^C contains N^C units, where each unit is labeled by the letter k. The correspondence between the two populations U^B and U^C can be represented by the link matrix $\mathbf{\Theta}_{BC} = [\theta^{BC}_{ik}]$ of size $N^B \times N^C$ where each element $\theta^{BC}_{ik} \geq 0$. That is, unit i of U^B is related to unit k of U^C provided that $\theta^{BC}_{ik} > 0$, otherwise the two units are not related to each other.

We can now use Indirect Sampling by *transitivity*. For this, we select a sample s^A from the population U^A and first identify the set Ω^B of U^B. From this set Ω^B, we then identify the units of U^C that are associated in order to form the set $\Omega^C = \{k \in U^C \mid \exists\, i \in \Omega^B \text{ and } \theta^{BC}_{ik} > 0\}$ of units to be measured from the target population U^C. An important question is to see if the GWSM, when applied in the context of Indirect Sampling by transitivity, is also transitive. That is, is applying the GWSM from U^A to U^B, and then from U^B to U^C, is equivalent to directly applying the GWSM from U^A to U^C?

First, consider using Indirect Sampling from U^A directly to the target population U^C. By going from the population U^A to U^B, and then to U^C, this can relate to having the link matrix $\mathbf{\Theta}_{AC} = [\theta^{AC}_{jk}]$ of size $N^A \times N^C$ defined as $\mathbf{\Theta}_{AC} = \mathbf{\Theta}_{AB}\mathbf{\Theta}_{BC}$. For each unit j selected in s^A, we identify the units k of U^C that have a non-zero correspondence, *i.e.*, with $\theta^{AC}_{jk} > 0$, to obtain the set $\vec{\Omega}^C = \{k \in U^C \mid \exists\, j \in s^A \text{ and } \theta^{AC}_{jk} > 0\}$. We measure the variable of interest y_k from the target population U^C. Applying the GWSM, we obtain from (2.6) the following weights:

$$\vec{\mathbf{W}}_C = \tilde{\mathbf{\Theta}}'_{AC}\mathbf{T}_A\mathbf{\Pi}^{-1}_A\mathbf{1}_A \tag{3.3}$$

where $\tilde{\mathbf{\Theta}}_{AC} = \mathbf{\Theta}_{AC}\,[\mathrm{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AC})]^{-1}$.

Let us now consider using Indirect Sampling in two steps. For each unit j selected in s^A, we identify the units i of U^B that have a non-zero correspondence, *i.e.*, with $\theta^{AB}_{ji} > 0$. As before, we have $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A \text{ and } \theta^{AB}_{ji} > 0\}$. For each unit i of the set Ω^B, we then identify the units k of U^C that have a non-zero correspondence, *i.e.*, with $\theta^{BC}_{ik} > 0$. We then have the set $\Omega^C = \{k \in U^C \mid \exists\, i \in \Omega^B \text{ and } \theta^{BC}_{ik} > 0\}$. From (2.6), we have the column vector $\mathbf{W}_B$ of weights associated to the units of population U^B:

$$\mathbf{W}_B = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{T}_A\mathbf{\Pi}^{-1}_A\mathbf{1}_A. \tag{3.4}$$

For each unit i of the set Ω^B, we then have a non-zero weight w^B_i. Now, the set Ω^B can be seen as a sample of units that are used in an Indirect Sampling process to identify the set Ω^C. By similarity with Indirect Sampling from the sample s^A to the target population U^B, applying the GWSM in the context of Indirect Sampling from the set Ω^B to the target population U^C produces the following weights:

$$\mathbf{W}_C = \tilde{\mathbf{\Theta}}'_{BC}\,\mathbf{T}_B\,\mathrm{diag}(\mathbf{W}_B)\mathbf{1}_B \tag{3.5}$$

where $\tilde{\mathbf{\Theta}}_{BC} = \mathbf{\Theta}_{BC}[\,\mathrm{diag}(\mathbf{1}'_B\mathbf{\Theta}_{BC})]^{-1}$ and $\mathbf{T}_B = \mathrm{diag}(\mathbf{t}_B)$ with $\mathbf{t}_B = (t^B_1, \ldots, t^B_{N^B})'$ and $t^B_i = 1$ if $i \in \Omega^B$, and 0 otherwise. Because the weights $w^B_i = 0$ for $i \notin \Omega^B$, we have $\mathbf{T}_B\,\mathrm{diag}(\mathbf{W}_B) = \mathrm{diag}(\mathbf{W}_B)$. Therefore, we obtain

$$\mathbf{W}_C = \tilde{\mathbf{\Theta}}'_{BC}\,\mathrm{diag}(\mathbf{W}_B)\mathbf{1}_B. \tag{3.6}$$

Replacing $\mathbf{W}_B$ by (3.4) in equation (3.6), we get

$$\begin{aligned}\mathbf{W}_C &= \tilde{\mathbf{\Theta}}'_{BC}\,\mathrm{diag}(\tilde{\mathbf{\Theta}}'_{AB}\,\mathbf{T}_A\,\mathbf{\Pi}^{-1}_A\mathbf{1}_A)\mathbf{1}_B \\ &= \tilde{\mathbf{\Theta}}'_{BC}\,\tilde{\mathbf{\Theta}}'_{AB}\,\mathbf{T}_A\,\mathbf{\Pi}^{-1}_A\mathbf{1}_A.\end{aligned} \tag{3.7}$$

Since $\tilde{\mathbf{\Theta}}'_{BC}\tilde{\mathbf{\Theta}}'_{AB}\mathbf{1}_A = \tilde{\mathbf{\Theta}}'_{BC}\mathbf{1}_B = \mathbf{1}_C$, from Result 1, the matrix $\tilde{\mathbf{\Theta}}_{AB}\tilde{\mathbf{\Theta}}_{BC}$ is a standardized link matrix. Because of this, the GWSM is therefore transitive, at least in some sense. That is, the weights $\mathbf{W}_C$ can be obtained in a single step by using the standardized link matrix $\tilde{\mathbf{\Theta}}_{AB}\tilde{\mathbf{\Theta}}_{BC}$ into the GWSM. Now, for the GWSM to be perfectly transitive, the weights $\mathbf{W}_C$ provided (3.7) would need to be exactly the same as the weights $\vec{\mathbf{W}}_C$ provided by (3.3). By comparing equations (3.3) and (3.7), we obtain the following result:

Result 5:

Applying the GWSM from U^A to U^B, and then from U^B to U^C, is transitive if and only if

$$\tilde{\boldsymbol{\Theta}}_{AC} = \tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}. \qquad (3.8)$$

Unfortunately, condition (3.8) does not hold in general. In fact, it is relatively easy to construct examples where $\tilde{\boldsymbol{\Theta}}_{AC} \neq \tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}$.

4. A Structural Property of the GWSM

In the present section, we stress the fact that with Indirect Sampling, the sampling process depends only on the links between the two populations U^A and U^B. The values of the θ_{ji}^{AB} themselves, apart from being zero or not, do not interfere in the sampling process. On the other hand, the values of the θ_{ji}^{AB} **do** have a role in the weights, and therefore the estimator, issued from the GWSM. We extend this idea in the following paragraphs.

Indirect Sampling associates to each sample s^A in U^A a sample Ω^B in U^B, namely $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A$ and $\theta_{ji}^{AB} > 0\}$. Thus, a function $f : s^A \to \Omega^B$ that maps the sample s^A to the sample Ω^B is uniquely determined by the set of couples (j, i) with $\theta_{ji}^{AB} > 0$. Let $l_{ji}^{AB} = 1$ if $\theta_{ji}^{AB} > 0$, and 0 otherwise. These are the elements of the incidence matrix of the graph linking U^A to U^B.

Suppose we are given a function ϕ from the set of subsets of U^A into the set of subsets of U^B. Like f, suppose that ϕ satisfies the "Union Property": $\phi(s_1^A \cup s_2^A) = \phi(s_1^A) \cup \phi(s_2^A)$, where s_1^A and s_2^A are two subsets of U^A.

Result 6:

The function ϕ is determined unequivocally by *a zero-one link matrix*.

Proof:

This can be shown as follows: Take $s_j^A = \{j\}$ for some unit j in U^A. Then, $\phi(s_j^A)$ is a set in U^B. Let $l_{ji}^{AB} = 1$ if unit i of U^B belongs to $\phi(s_j^A)$, and 0 otherwise. By the Union Property, $\phi(s^A) = \bigcup_{j \in s^A} \phi(s_j^A)$ and the set of l_{ji}^{AB} defines the *zero-one link matrix* $\mathbf{L}_{AB} = [l_{ji}^{AB}]$ of size $N^A \times N^B$, which precisely defines the function ϕ.

This provides us an equivalence relation between link matrices, associated with a deeper property. Let p^A be a sampling design on U^A (*i.e.*, a probability distribution on the set of subsets of U^A). The function f induces a sampling design on U^B by $p^B(\Omega^B) = \sum_{s^A : \Omega^B = f(s^A)} p^A(s^A)$. As the design is induced by f, it does not depend on the particular link matrix $\boldsymbol{\Theta}_{AB}$ defining the function, but is rather a characteristic of the equivalence class through the zero-one link matrix $\mathbf{L}_{AB}$. As a consequence, the Horvitz-Thompson estimator in U^B depends only on this class. It is therefore of some interest to choose in this class a matrix $\boldsymbol{\Theta}_{AB}$ having, in some sense, an optimal characteristic (see section 6).

5. Special Link matrices

As it can be seen from the previous sections, the link matrix $\boldsymbol{\Theta}_{AB}$ drives the form of the estimator (2.4) obtained from the GWSM. In this section, we present some special link matrices $\boldsymbol{\Theta}_{AB}$ that correspond to extreme cases. Although not all such cases are likely to be seen it practice, they illustrates the effect of the link matrix on the estimator (2.4).

5.1 Identity Matrix

Assume that the link matrix $\boldsymbol{\Theta}_{AB}$ is given by the identity matrix $\mathbf{I}$. In practice, this means that the population U^A and the target population U^B have a one-to-one relationship. Of course, this implies that $N^A = N^B = N$ and that the identity matrix $\mathbf{I}$ is of size $N \times N$.

As a first result, we have $\tilde{\boldsymbol{\Theta}}_{AB} = \mathbf{I}$. As a consequence, the vector of weights (2.6) is given by $\mathbf{W}' = (t_1^A / \pi_1^A, ..., t_{N^A}^A / \pi_{N^A}^A)$ and we also have $\mathbf{Z} = \tilde{\boldsymbol{\Theta}}_{AB}\mathbf{Y} = \mathbf{Y}$. Therefore, the estimator $\hat{Y}^B$ given by (2.5) turns out to be nothing else than the Horvitz-Thompson estimator $\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \boldsymbol{\Pi}_A^{-1} \mathbf{Y}$.

5.2 One for All (Within Clusters)

Consider the case where the population U^B is divided into Γ clusters where each cluster γ is of size N_γ^B. Theses clusters are such that each cluster γ from U^B is associated to exactly one unit j of U^A. Because of this, we can use the letter γ for both the units j from U^A and the clusters from U^B. Note also that $\Gamma = N^A$.

This situation corresponds to a link matrix $\boldsymbol{\Theta}_{AB}$ being block diagonal where each submatrix contains only one line. Let the row vector $\mathbf{1}'_{B\gamma}$ be of size N_γ^B and containing only 1's. The link matrix $\boldsymbol{\Theta}_{AB}$ is then defined as

$$\boldsymbol{\Theta}_{AB} = \begin{bmatrix} \mathbf{1}'_{B1} & \mathbf{0} & \cdots & \mathbf{0} \\ & \ddots & & \ddots \\ \mathbf{0} & & \mathbf{1}'_{B\gamma} & \mathbf{0} \\ \vdots & \ddots & & \ddots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}'_{B\Gamma} \end{bmatrix}. \qquad (5.1)$$

We can also write $\boldsymbol{\Theta}_{AB} = \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})$. Using this, we have $\text{diag}(\mathbf{1}'_A \boldsymbol{\Theta}_{AB}) = \text{diag}(\mathbf{1}'_A \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})) = \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})$ and hence $\tilde{\boldsymbol{\Theta}}_{AB} = \boldsymbol{\Theta}_{AB}$. From equation (2.6), we obtain the column vector of weights $\mathbf{W}' = (t_1^A / \pi_1^A \mathbf{1}'_{B1}, ..., t_\Gamma^A / \pi_\Gamma^A \mathbf{1}'_{B\Gamma})$. As we can see, the elements of the column vector $\mathbf{W}$ have the values $t_\gamma^A / \pi_\gamma^A$ repeated within each cluster γ of U^B. From (2.4), we obtain

$$\hat{Y}^B = \sum_{\gamma=1}^{\Gamma} \frac{t_\gamma^A}{\pi_\gamma^A} Y_\gamma^B \tag{5.2}$$

where $Y_\gamma^B = \sum_{i=1}^{N_\gamma^B} y_i$.

5.3 All for One (Within Clusters)

Consider the case where the population U^A is divided into Γ clusters where each cluster γ is of size N_γ^A. Theses clusters are such that each cluster γ from U^A is associated to exactly one unit i of U^B. Because of this, we can use the letter γ for both the clusters from U^A and the units i from U^B. Note also that $\Gamma = N^B$.

This situation corresponds to a link matrix $\mathbf{\Theta}_{AB}$ being block diagonal where each submatrix contains only one column. Let the column vector $\mathbf{1}_{A\gamma}$ be of size N_γ^A and containing only 1's. The link matrix $\mathbf{\Theta}_{AB}$ is then defined as

$$\mathbf{\Theta}_{AB} = \begin{bmatrix} \mathbf{1}_{A1} & \mathbf{0} & \cdots & \mathbf{0} \\ & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{1}_{A\gamma} & & \mathbf{0} \\ \vdots & \ddots & \ddots & \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}_{A\Gamma} \end{bmatrix}. \tag{5.3}$$

We can also write $\mathbf{\Theta}_{AB} = \text{diag}(\{\mathbf{1}_{A1}, ..., \mathbf{1}_{A\Gamma}\})$. Using this, we have $\tilde{\mathbf{\Theta}}_{AB} = \text{diag}(\{1/N_1^A \mathbf{1}_{A1}, ..., 1/N_\Gamma^A \mathbf{1}_{A\Gamma}\})$. From equation (2.6), we obtain the column vector of weights $\mathbf{W}' = (1/N_1^A \sum_{j=1}^{N_1^A} t_j^A/\pi_j^A, ..., 1/N_\Gamma^A \sum_{j=1}^{N_\Gamma^A} t_j^A/\pi_j^A)$. Thus, the elements γ (or i) of the column vector $\mathbf{W}$ have the averaged values $\sum_{j=1}^{N_\gamma^A} t_j^A/\pi_j^A N_{\gamma,\gamma}^A \gamma = 1, ..., \Gamma$. From (2.4), we obtain $\hat{Y}^B = \sum_{\gamma=1}^{\Gamma} y_\gamma / N_\gamma^A \sum_{j=1}^{N_\gamma^A} t_j^A/\pi_j^A$.

5.4 Inefficient Sampling

Suppose that some rows of the link matrix $\mathbf{\Theta}_{AB}$ contain only zeros. This means that some units of the population U^A are not associated to any unit of the target population U^B. Then, if such units are selected in the sample s^A, this will lead to the identification of no unit from U^B. This can be seen as inefficient in a sampling point of view. In a more formal way, assume that each of the first N^{1A} rows of the link matrix $\mathbf{\Theta}_{AB}$ contains at least one $\theta_{ji} > 0$, and that they form the submatrix $\mathbf{\Theta}_1$. Assume that the other N^{0A} rows of $\mathbf{\Theta}_{AB}$ have $\theta_{ji} = 0$ for $i = 1, ..., N^B$. We therefore have

$$\mathbf{\Theta}_{AB} = \begin{bmatrix} \mathbf{\Theta}_1 \\ \mathbf{0} \end{bmatrix}.$$

As a first result, we obtain

$$\tilde{\mathbf{\Theta}}_{AB} = \begin{bmatrix} \mathbf{\Theta}_1 [\text{diag}(\mathbf{1}_{1A}' \mathbf{\Theta}_1)]^{-1} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{\Theta}}_1 \\ \mathbf{0} \end{bmatrix} \tag{5.4}$$

where $\mathbf{1}_{1A}$ is the column vector of 1's of size N^{1A}. From equation (2.6), we obtain the column vector of weights $\mathbf{W} = [\tilde{\mathbf{\Theta}}_1' \ \mathbf{0}'] \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A$. Let $\mathbf{\Pi}_{1A} = \text{diag}(\{\pi_1^A, ..., \pi_{N^{1A}}^A\})$ be the diagonal matrix of size $N^{1A} \times N^{1A}$ and accordingly, let $\mathbf{T}_{1A} = \text{diag}(\{t_1^A, ..., t_{N^{1A}}^A\})$ be the diagonal matrix of size $N^{1A} \times N^{1A}$. We then get

$$\begin{aligned} \mathbf{W} &= [\tilde{\mathbf{\Theta}}_1' \ \mathbf{0}'] \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A \\ &= \tilde{\mathbf{\Theta}}_1' \mathbf{T}_{1A} \mathbf{\Pi}_{1A}^{-1} \mathbf{1}_{1A}. \end{aligned} \tag{5.5}$$

As we can see from (5.5), the weights only depend on the probabilities of selection π_j^A of the units of U^A that have at least one $\theta_{ji} > 0$ for $i = 1, ..., N^B$. From (2.4), we finally obtain $\hat{Y}^B = \mathbf{1}_{1A}' \mathbf{T}_{1A} \mathbf{\Pi}_{1A}^{-1} \tilde{\mathbf{\Theta}}_1 \mathbf{Y}$.

5.5 Biased Estimator

Suppose that some columns of the link matrix $\mathbf{\Theta}_{AB}$ contain only zeros. This means that some units of the population U^B are not associated to any unit of the target population U^A. Recall that in order for the matrix $\tilde{\mathbf{\Theta}}_{AB}$ to be well defined, we must have $\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})^{-1}$ to exist. As we will see, the present case does not satisfy this condition. This results in a biased estimator for the total Y^B.

In a more formal way, assume that each of the first N^{1B} columns of the link matrix $\mathbf{\Theta}_{AB}$ contains at least one $\theta_{ji} > 0$, and let them form the submatrix $\mathbf{\Theta}_1$, different from the one of the previous section. Assume that the other N^{0B} columns of $\mathbf{\Theta}_{AB}$ have $\theta_{ji} = 0$ for $j = 1, ..., N^A$. We therefore have $\mathbf{\Theta}_{AB} = [\mathbf{\Theta}_1, \mathbf{0}]$.

From this definition, we directly have

$$\begin{aligned} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1} &= [\text{diag}([\mathbf{1}_A' \mathbf{\Theta}_1, \mathbf{1}_A' \mathbf{0}])]^{-1} \\ &= \begin{bmatrix} \text{diag}(\mathbf{1}_A' \mathbf{\Theta}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^{-1}. \end{aligned} \tag{5.6}$$

Since this matrix is singular, $[\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1}$ does not exist. As a solution to this problem, it could be possible to use a *generalized inverse*. Recall that for a given square matrix $\mathbf{A}$, the matrix $\mathbf{A}^-$ is a generalized inverse of $\mathbf{A}$ provided that $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$ (Searle 1971). One possible generalized inverse of (5.6) is

$$[\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^- = \begin{bmatrix} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_1)]^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{5.7}$$

With this generalized inverse, we have the following standardized link matrix $\tilde{\mathbf{\Theta}}_- = \mathbf{\Theta}_{AB} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^- = [\tilde{\mathbf{\Theta}}_1, \mathbf{0}]$. Starting from equation (2.6), we can obtain the column vector $\mathbf{W}_-$ of weights:

$$\mathbf{W}_- = \begin{bmatrix} \tilde{\mathbf{\Theta}}_1' \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A \\ \mathbf{0}' \end{bmatrix}. \tag{5.8}$$

As we can see from (5.8), the weights are null for the units i of the target population U^B that $\mathbf{\Theta}_{AB}$ have $\theta_{ji} = 0$ for $j = 1, ..., N^A$. From (2.4) and using $\mathbf{W}_-$ instead of $\mathbf{W}$, we obtain $\hat{Y}_-^B = \mathbf{1}_A' \mathbf{T}_A \mathbf{\Pi}_A^{-1} \tilde{\mathbf{\Theta}}_1 \mathbf{Y}_1$ where $\mathbf{Y}_1 = \{y_1, ..., y_{N^{1B}}\}'$ is the subvector constructed from the N^{1B} first elements of $\mathbf{Y}$. Since in general $E(\hat{Y}_-^B) = \mathbf{1}_A' \tilde{\mathbf{\Theta}}_1 \mathbf{Y}_1 \neq \mathbf{1}_A' \mathbf{Y} = Y^B$, this estimator is biased for the total Y^B.

6. Optimality

Optimality is an important aspect of the GWSM. As it has been shown in Result 3, the estimator $\hat{Y}^B$ obtained by the GWSM will provide unbiased estimates provided that the matrix $\tilde{\mathbf{\Theta}}_{AB}$ is a standardized link matrix. Now, given that the variance (3.2) of this estimator depends on this matrix, there should be at least one matrix $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$ such that the variance of the estimator $\hat{Y}^B$ will be minimum. That is, for the θ_{ji}^{AB} that are greater than 0, we are interested in finding the values that these θ_{ji}^{AB} should have to obtain the most precise estimator $\hat{Y}^B$.

This optimality problem was first assessed by Kalton and Brick (1995). They obtained results based on the simplified situation where $N^A = 2$ and with s^A obtained through equal probability sampling. Their conclusions suggested the use of $\theta_{ji}^{AB,\text{opt}} = 1$ when $\theta_{ji}^{AB} > 0$, and $\theta_{ji}^{AB,\text{opt}} = 0$ when $\theta_{ji}^{AB} = 0$. Lavallée (2002) and Lavallée and Caron (2001) obtained results along the same lines by the use of simulations. In the present section, we present new results on the optimality of the GWSM.

6.1 Factorization

Factorization is the reverse problem of transitivity. It consists in finding a population U^G and standardized link matrices $\tilde{\mathbf{\Theta}}_{AG}$ and $\tilde{\mathbf{\Theta}}_{GB}$ such that $\tilde{\mathbf{\Theta}}_{AB} = \tilde{\mathbf{\Theta}}_{AG} \tilde{\mathbf{\Theta}}_{GB}$. This leads to an important simplification in searching for an optimal standardized link matrix $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$.

The population U^G can be taken as being one of clusters, the factorization being achieved in the context of "one for all (within clusters)" (from U^A to U^G) and "all for one (within clusters)" (from U^G to U^B), as presented in sections 5.2 and 5.3. This can be described in a very general way as follows. Consider a population U^G containing as many units as there are links starting from the units j of U^A. The population size N^G is then given by the number of θ_{ji}^{AB} of $\mathbf{\Theta}_{AB}$ that are greater than 0. Each unit g of U^G can be seen as the extremity of an "arrow" starting from some unit j of U^A. From this graph, there is only one link matrix $\mathbf{\Theta}_{AG}$ of size $N^A \times N^G$ keeping unbiasedness, namely $\mathbf{\Theta}_{AG} = [\theta_{jg}^{AG}]$ where $\theta_{jg}^{AG} = 1$ if there is a link (or an "arrow") leaving unit j of U^A to unit g from U^G, and $\theta_{jg}^{AG} = 0$ otherwise. Note that by construction, each unit g from U^G is linked to at most one unit j from U^A and therefore $\tilde{\mathbf{\Theta}}_{AG} = \mathbf{\Theta}_{AG}$. This corresponds to the "one to all within clusters" situation presented in section 5.2. Indirect Sampling from U^A to U^G is in fact standard Cluster Sampling and leading the GWSM to the usual Horvitz-Thompson estimator (see Lavallée 2002). For the parent-children example, the result of this factorization would be given by Figure 2.

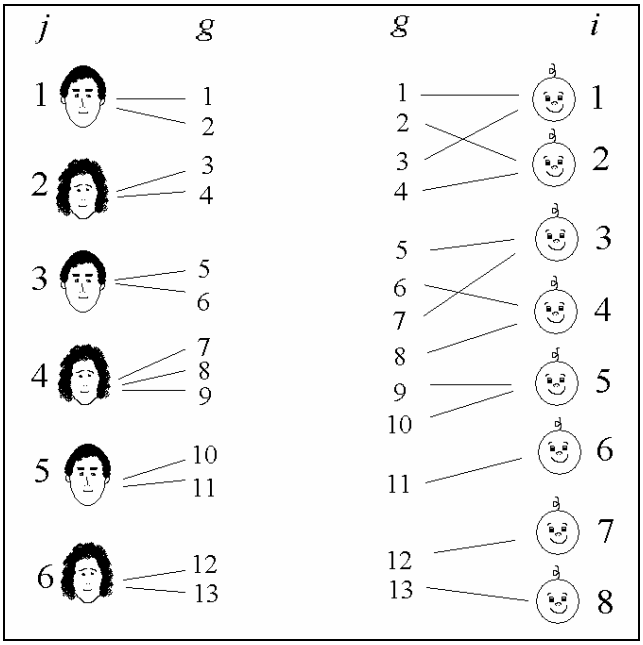

Figure 2. Result of the factorization of the parents-children populations.

Considering the graph from U^G to U^B, we can construct the link matrix $\mathbf{\Theta}_{GB}$ of size $N^G \times N^B$ as follows. Because of the definition of the population U^G, each unit g of U^G is linked to exactly one unit i of U^B. Note that Indirect Sampling in this context can be seen as sampling clusters (*i.e.*, the units i of U^B) from their elements (*i.e.*, the units g of U^G). It can also be seen as the "all to one within clusters" presented in section 5.3. Let $\tilde{\mathbf{\Theta}}_{GB} = \mathbf{\Theta}_{GB}[\text{diag}(\mathbf{1}_G' \mathbf{\Theta}_{GB})]^{-1}$ be the standardized link matrix obtained from $\mathbf{\Theta}_{GB}$. We have $\text{diag}(\mathbf{1}_G' \mathbf{\Theta}_{GB}) = \text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})$, and therefore $\tilde{\mathbf{\Theta}}_{GB} = \mathbf{\Theta}_{GB}[\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1}$.

Now,

$$\begin{aligned} \tilde{\mathbf{\Theta}}_{AG} \tilde{\mathbf{\Theta}}_{GB} &= \mathbf{\Theta}_{AG} \tilde{\mathbf{\Theta}}_{GB} \\ &= \mathbf{\Theta}_{AG} \mathbf{\Theta}_{GB} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1} \\ &= \mathbf{\Theta}_{AB} [\text{diag}(\mathbf{1}_A' \mathbf{\Theta}_{AB})]^{-1} \\ &= \tilde{\mathbf{\Theta}}_{AB}. \end{aligned} \tag{6.1}$$

Therefore, using this construction, the standardized link matrix $\tilde{\mathbf{\Theta}}_{AB}$ from U^A to U^B can always be factorized into the two matrices $\tilde{\mathbf{\Theta}}_{AG}$ and $\tilde{\mathbf{\Theta}}_{GB}$.

6.2 Strong Optimality: Statement of the Problem

As mentioned before, the optimality problem that we consider here is to minimize the variance (3.2) with respect to the standardized link matrix $\tilde{\mathbf{\Theta}}_{AB}$. Now, using the factorization presented in section 6.1, we have

$$\begin{aligned}\mathrm{Var}(\hat{Y}^B) &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{AB}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AB}\mathbf{Y} \\ &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\tilde{\mathbf{\Theta}}'_{AG}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AG}\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y} \\ &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\mathbf{\Delta}_G\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y}\end{aligned} \qquad (6.2)$$

where $\mathbf{\Delta}_G = \tilde{\mathbf{\Theta}}'_{AG}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AG}$.

For any standardized link matrix $\tilde{\mathbf{\Theta}}_{AB}$, the factorization presented in section 6.1 always produces the same first factor $\tilde{\mathbf{\Theta}}_{AG}$. Therefore, if we seek for some optimal matrix $\tilde{\mathbf{\Theta}}_{AB,\mathrm{opt}}$ that minimizes the variance (3.2), it is sufficient to optimize the second factor $\tilde{\mathbf{\Theta}}_{GB}$. We would also like the optimal matrix $\tilde{\mathbf{\Theta}}_{AB,\mathrm{opt}}$ to produce unbiased estimates.

Let U_i^G be the subpopulation of U^G containing the N_i^G links to the unit i of U^B. Note that the subpopulations U_i^G are disjoint. Thus, without loss of generality, we can order the links from U^A to U^B so that, for every i, the links to unit i in U^B are indexed consecutively. Now, let $\tilde{\tilde{\boldsymbol{\theta}}}_{GB,i}$ be the i^{th} column vector of the matrix $\tilde{\mathbf{\Theta}}_{GB}$, $i = 1, \ldots, N^B$. By construction, the vector $\tilde{\tilde{\boldsymbol{\theta}}}_{GB,i}$ contains non null elements only for the N_i^G links to the unit i of U^B. Hence, letting $\tilde{\boldsymbol{\theta}}_{GB,i}$ be a column vector of size N_i^G containing the non null elements of $\tilde{\tilde{\boldsymbol{\theta}}}_{GB,i}$, we have

$$\tilde{\tilde{\boldsymbol{\theta}}}_{GB,i} = \begin{bmatrix} \mathbf{0} \\ \tilde{\boldsymbol{\theta}}_{GB,i} \\ \mathbf{0} \end{bmatrix}.$$

Similarly, let $\dot{\mathbf{1}}_{G,i}$ be the column vector of size N^G containing 1's for N_i^G elements, and 0's elsewhere. Letting $\mathbf{1}_{G,i}$ be a column vector of size N_i^G containing 1's, we have

$$\dot{\mathbf{1}}_{G,i} = \begin{bmatrix} \mathbf{0} \\ \mathbf{1}_{G,i} \\ \mathbf{0} \end{bmatrix}.$$

Now, for the GWSM from U^G to U^B to be unbiased, we need to have $\tilde{\boldsymbol{\theta}}'_{GB,i}\mathbf{1}_{G,i} = 1$ for all i, or equivalently $\tilde{\tilde{\boldsymbol{\theta}}}'_{GB,i}\dot{\mathbf{1}}_{G,i} = 1$. All this together leads to the following optimization problem:

Find a matrix $\tilde{\mathbf{\Theta}}_{GB,\mathrm{opt}} = \{\tilde{\tilde{\boldsymbol{\theta}}}_{GB,\mathrm{opt},1}, \ldots, \tilde{\tilde{\boldsymbol{\theta}}}_{GB,\mathrm{opt},N^B}\}$ satisfying $\tilde{\tilde{\boldsymbol{\theta}}}'_{GB,\mathrm{opt},i}\dot{\mathbf{1}}_{G,i} = 1$ for all $i = 1, \ldots, N^B$, and minimizing the quadratic form $\mathrm{Var}(\hat{Y}^B) = \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\mathbf{\Delta}_G\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y}$.

This problem turns out to be nothing else than the minimization of a positive quadratic form under linear constraints. This is a relatively standard and simple problem to solve. It is well known that a solution always exists and is unique if the form (6.2) is positive definite, or if the null subspace of $\tilde{\mathbf{\Theta}}_{GB}$ is not included in the null-space of $\mathbf{\Delta}_G$.

The above optimization problem can be rewritten in a different form. Let $\mathbf{\Delta}_{G,ii'}$ be the submatrix of $\mathbf{\Delta}_G$ corresponding to the elements in positions g and g' if g has a link with unit i and g' has a link with unit i'. These matrices constitute a partition of $\mathbf{\Delta}_G$. Note that the matrices $\mathbf{\Delta}_{G,ii}$ are symmetric, positive definite, and $\mathbf{\Delta}'_{G,ii'} = \mathbf{\Delta}_{G,i'i}$. With these notations, the optimization problem can be written as:

Minimize

$$\sum_{i=1}^{N^B}\sum_{i'=1}^{N^B} y_i y_{i'} \tilde{\boldsymbol{\theta}}'_{GB,i}\mathbf{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,i'} \qquad (6.3)$$

under the constraints $\tilde{\boldsymbol{\theta}}'_{GB,i}\mathbf{1}_{G,i} = 1$ for all $i = 1, \ldots, N^B$.

Minimization is achieved for vectors $\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i}$ satisfying

$$y_i \sum_{i'=1}^{N^B} \mathbf{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i'} y_{i'} = \lambda_i \mathbf{1}_{G,i} \qquad (6.4)$$

for all $i = 1, \ldots, N^B$ and where λ_i are the Lagrange multipliers entering into the constrained minimization of (6.3). As we can see from (6.4), the optimal choice $\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i}$ (and therefore $\tilde{\mathbf{\Theta}}_{GB,\mathrm{opt}}$) will depend in general explicitly on the vector $\mathbf{Y}$, which is not useful in practice. Observe that the set of λ_i depends also of the variable $\mathbf{Y}$. This will appear more explicitly in section 6.3. This is the reason why we will seek, instead of a strong optimization, for a weaker form of optimality that will lead to the existence of an "optimal" solution $\tilde{\mathbf{\Theta}}_{GB,\mathrm{opt}}$ (and $\tilde{\mathbf{\Theta}}_{AB,\mathrm{opt}}$) not depending on $\mathbf{Y}$.

6.3 Weak Optimality

Equations (6.4) must be valid for any vector $\mathbf{Y}$. In particular, a necessary condition is to hold for a particular variable of interest, such as $y_i = 1$ for a unit i of U^B and $y_{i'} = 0$ for all other units i' of U^B ($i' \neq i$). This leads to the necessary conditions (one for each of those particular variables) $\mathbf{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i} = \lambda_i \mathbf{1}_{G,i}$. Assuming that $\mathbf{\Delta}_{G,ii}$ is invertible, we then have $\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i} = \lambda_i \mathbf{\Delta}_{G,ii}^{-1}\mathbf{1}_{G,i}$. It can be shown that this is also a sufficient condition. Now, because $\tilde{\boldsymbol{\theta}}'_{GB,\mathrm{opt},i}\mathbf{1}_{G,i} = 1$, we have $\lambda_i = 1/\mathbf{1}'_{G,i}\mathbf{\Delta}_{G,ii}^{-1}\mathbf{1}_{G,i}$. Therefore, a necessary and sufficient condition for equation (6.4) to be satisfied is when

$$\tilde{\boldsymbol{\theta}}_{GB,\mathrm{opt},i} = \frac{\mathbf{\Delta}_{G,ii}^{-1}\mathbf{1}_{G,i}}{\mathbf{1}'_{G,i}\mathbf{\Delta}_{G,ii}^{-1}\mathbf{1}_{G,i}}. \qquad (6.5)$$

This result corresponds to weak optimization in the following sense. The weight w_i given by (2.6) satisfies $E(w_i) = 1$ and moreover $E(w_i \mid i \in \Omega^B) = 1/\pi_i^B$ where π_i^B is the inclusion probability of unit i in Ω^B, which is generally difficult or even impossible to compute in practice. Now, note that the Horvitz-Thompson estimator is characterized by $\mathrm{Var}(w_i \mid i \in \Omega^B) = 0$. The weak optimization

obtained here consists in minimizing $\mathrm{Var}(w_i \mid i \in \Omega^B)$ over all possible standardized link matrices $\tilde{\boldsymbol{\Theta}}_{GB}$, or equivalently $\tilde{\boldsymbol{\Theta}}_{AB}$. This variance is strictly positive for the cases where unit i of U^B is in position to receive more than a unique weight for different sample s^A. Moreover, using (6.3), the multiplier λ_i appears to be the variance of the weight w_i and is, therefore, always strictly positive (except, a case that we exclude, when unit i is selected with a weight equal to one).

6.4 Strong Optimality Independent of Y

Weak optimality is a necessary condition for strong optimality independent of the vector $\mathbf{Y}$ of a variable of interest. It provides the necessary form of the vectors $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$ in (6.4). To get sufficient conditions for strong optimality independent of $\mathbf{Y}$, we go back to the equations (6.4). These equations need to be satisfied for all vectors $\mathbf{Y}$ and they must therefore be satisfied for a particular variable of interest such as $y_i = 1$ for a unit i of U^B, $y_{i'} = 1$ for another unit i' of U^B, and $y_{i''} = 0$ for all other units i'' of U^B $(i'' \neq i' \neq i)$. In that case, to satisfy equations (6.4), it is necessary to have the following relations for any i and i':

$$\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} + \boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} = \lambda_i^{ii'}\mathbf{1}_{G,i} \qquad (6.6)$$

$$\boldsymbol{\Delta}_{G,i'i'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} + \boldsymbol{\Delta}_{G,i'i}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \lambda_{i'}^{ii'}\mathbf{1}_{G,i'}.$$

As we must necessarily have weak optimality, we have $\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \lambda_i \mathbf{1}_{G,i}$. Considering the first line of (6.6), we then get

$$\begin{aligned}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} &= (\lambda_i^{ii'} - \lambda_i)\mathbf{1}_{G,i} \\ &= \Phi_{ii'}\mathbf{1}_{G,i}. \qquad (6.7)\end{aligned}$$

Multiplying both sides of (6.7) by $\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}$, we obtain

$$\begin{aligned}\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} &= \Phi_{ii'}\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\mathbf{1}_{G,i} \\ &= \Phi_{ii'}\end{aligned}$$

since $\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\mathbf{1}_{G,i} = 1$. Let $\boldsymbol{\Phi}$ be the matrix with elements $\Phi_{ii'}$ off the diagonal and $\Phi_{ii} = \lambda_i$ on the diagonal. Using again (6.2), it can be shown that the optimal variance (whenever it exists) has the expression $\mathbf{Y}'\boldsymbol{\Phi}\mathbf{Y}$.

Let us show that this set of conditions is also sufficient. Assume that (6.7) holds. Note that for $i = i'$, condition (6.7) is nothing else than (6.5) which gives the necessary values for the $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$. It is now straightforward to verify that (6.4) holds whatever the value of $\mathbf{Y}$ and that we have obtained the strong optimality. Now, the values of λ_i depend on $\mathbf{Y}$, as well as the variance $\mathrm{Var}(\hat{Y}^B)$, but we have that equations (6.4) always have the same solution (6.5) that does not depend on $\mathbf{Y}$. We therefore have the following result:

Result 7:

The conditions $\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} = \boldsymbol{\Phi}_{ii'}\mathbf{1}_{G,i}$ are necessary and sufficient for the existence of a standardized link matrix $\tilde{\boldsymbol{\Theta}}_{GB,\,\mathrm{opt}}$, or equivalently $\tilde{\boldsymbol{\Theta}}_{AB,\,\mathrm{opt}}$, that achieves strong optimality independent of the vector $\mathbf{Y}$ of the variable of interest. The values in the columns of this strong optimal matrix are given by (6.5), which are the vectors $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$ obtained from weak optimality.

It should be noted that since $\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \lambda_i \mathbf{1}_{G,i}$ (6.7) can be written in an equivalent way as

$$\Phi_{ii'}^{**}\,\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \boldsymbol{\Delta}_{G,ii}^{-1}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} \qquad (6.8a)$$

or

$$\Phi_{ii'}^{*}\,\mathbf{1}_{G,i} = \boldsymbol{\Delta}_{G,ii'}\boldsymbol{\Delta}_{G,i'i'}^{-1}\mathbf{1}_{G,i'} \qquad (6.8b)$$

where $\Phi_{ii'}^{**} = (\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'})(\mathbf{1}'_{G,i}\boldsymbol{\Delta}_{G,ii}^{-1}\mathbf{1}_{G,i})$ and $\Phi_{ii'}^{*} = (\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'})(\mathbf{1}'_{G,i'}\boldsymbol{\Delta}_{G,i'i'}^{-1}\mathbf{1}_{G,i'})$. In some situations, these can proved to be easier to use that the expresssion (6.7) stated in Result 7.

6.5 Two Examples

We now present two examples that illustrate the preceding theory on weak optimality and strong optimality independent of $\mathbf{Y}$.

Example 1: Poisson Sampling

Let us suppose that the sample s^A is selected using Bernoulli or Poisson Sampling. In that case, the $N^A \times N^A$ matrix $\boldsymbol{\Delta}_A$ is given by $\boldsymbol{\Delta}_A = \mathrm{diag}(1/\pi_j^A - 1)$. Considering the factorization of section 6.1, we have $\boldsymbol{\Delta}_G = \tilde{\boldsymbol{\Theta}}'_{AG}\boldsymbol{\Delta}_A\tilde{\boldsymbol{\Theta}}_{AG} = \tilde{\boldsymbol{\Theta}}'_{AG}[\mathrm{diag}(1/\pi_j^A - 1)]\tilde{\boldsymbol{\Theta}}_{AG} = [\mathrm{diag}((1/\pi_j^A - 1)\mathbf{1}_{A,jj})]$ where $\mathbf{1}_{A,jj}$ is a square matrix of size N_j^A, with N_j^A being the number of links (or "arrows") starting from unit j of U^A. From $\boldsymbol{\Delta}_G$, we extract the submatrices $\boldsymbol{\Delta}_{G,ii}$ that are, in the present case, diagonal. Each submatrix $\boldsymbol{\Delta}_{G,ii}$ is given by $\Delta_{G,ii} = \mathrm{diag}(1/\pi_g^A - 1)$, which is of size N_i^G. Note that each value $(1/\pi_g^A - 1)$ simply corresponds to a unit j of U^A that has previously been linked to the unit g of U^G, which is in turn linked to the unit i of U^B. Now, from (6.5), we directly obtain the optimal values $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$ that minimize $\mathrm{Var}(\hat{Y}^B)$, in the weak sense. These values are given by the vectors

$$\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i} = \left\{\frac{\pi_1^A}{(1-\pi_1^A)\tau_i^G}, \ldots, \frac{\pi_{N_i^G}^A}{(1-\pi_{N_i^G}^A)\tau_i^G}\right\}$$

where

$$\tau_i^G = \sum_{g=1}^{N_i^G} \pi_g^A /(1-\pi_g^A),\ i = 1,\ ...,\ N^B.$$

The $\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}$ are used to construct the vectors $\tilde{\tilde{\boldsymbol{\theta}}}'_{GB,\,\mathrm{opt},\,i}$, and then the matrix $\tilde{\boldsymbol{\Theta}}_{GB,\,\mathrm{opt}} = \{\tilde{\tilde{\boldsymbol{\theta}}}_{GB,\,\mathrm{opt},\,1},\ ...,\ \tilde{\tilde{\boldsymbol{\theta}}}_{GB,\,\mathrm{opt},\,N^B}\}$. Finally, after computing the optimal matrix $\tilde{\boldsymbol{\Theta}}_{AB,\,\mathrm{opt}} = \boldsymbol{\Theta}_{AG}\tilde{\boldsymbol{\Theta}}_{GB,\,\mathrm{opt}}$, we obtain the optimal weights $\mathbf{W}_{\mathrm{opt}}$ using (2.6).

It should be noted that if the inclusion probabilities π_j^A are equal, we get

$$\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i} = \left\{\frac{1}{N_i^G},\ ...,\ \frac{1}{N_i^G}\right\} = \frac{1}{N_i^G}\mathbf{1}_{GB,\,i},$$

where N_i^G is nothing else than the number of units of U^A linked to unit i of U^B. In other words, in the context of Bernoulli Sampling (*i.e.*, Poisson Sampling with equal probabilities), to minimize the variance $\mathrm{Var}(\hat{Y}^B)$, the choice of the values $\theta^{AB}_{\mathrm{opt},\,ji}$ should be given by 1 if there is a link between unit j of U^A and i of U^B, and 0 otherwise. This corresponds to the results obtained by Kalton and Brick (1995), Lavallée (2002), and Lavallée and Caron (2001).

Using Result 7, we now verify if conditions (6.7), (6.8a) or (6.8b) are satisfied for the optimal matrix $\tilde{\boldsymbol{\Theta}}_{AB,\,\mathrm{opt}}$ that we obtained through weak optimization. If it is the case, this matrix also provides strong optimality independent of the variable of interest y_i. First, we have

$$\Delta_{G,\,ii}^{-1} = \mathrm{diag}\left(\frac{\pi_g^A}{1-\pi_g^A}\right).$$

Also, each submatrix $\Delta_{G,\,ii'}$ of size $N_i^G \times N_{i'}^G$ has somewhat a diagonal structure, but "padded" with zeros. That is, a typical element of $\Delta_{G,ii'}$ is given by $(1/\pi_g^A - 1)$ on a part of the diagonal if both i and i' are linked to the same unit j of U^A (that is linked to unit g of U^G coming from the same j of U^A), and 0 otherwise. Because of this, if two units i and i' are not linked to the same units of U^A, then $\Delta_{G,\,ii'}$ is a matrix of zeros, and then the conditions (6.7), (6.8a) and (6.8b) are automatically satisfied. Referring to Figure 1, children $i = 2$ and $i' = 3$ of U^B are not related to the same parents j of U^A. If the selection of the parents is done using Poisson or Bernoulli Sampling, the 2×2 matrix $\boldsymbol{\Delta}_{G,\,23}$ will then contain only zeros, *i.e.*,

$$\boldsymbol{\Delta}_{G,23} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Because if this, the relations (6.7), (6.8a) or (6.8b) will be satisfied with $\Phi_{23} = 0$, expressing the fact that the weights of i and i' are not correlated.

If two units i and i' are linked to the same unit j of U^A, then, using (6.7), the column vector $\boldsymbol{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'}$ contains the scalar $(\tau_{i'}^G)^{-1} = [\sum_{g=1}^{N_{i'}^G} \pi_g^A /(1-\pi_g^A)]^{-1}$ for its first N_i^B components, and 0 for the remaining $N_{i'}^B - N_i^B$ ones (assuming $N_{i'}^B \geq N_i^B$). Because the quantity $\boldsymbol{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'}$ must be equal to $\Phi_{ii'}\mathbf{1}_{G,\,i}$ to satisfy (6.7), it must contain only the value $\Phi_{ii'}$. Since $\Phi_{ii'} = \tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\boldsymbol{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'}$, this will occur only if the vector $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = [1]$, which means that there is only one link to unit i of U^B. As we can see, this is not a condition that will be satisfied in general and therefore, it can be said that in the case of Poisson Sampling, strong optimality independent from $\mathbf{Y}$ will not occur in general.

As a conclusion, we might say that with Poisson or Bernoulli Sampling, the conditions (6.7), (6.8a) or (6.8b) will be satisfied in practice only when the units of U^A are linked to a single unit of U^B, as in the case of sampling households using a frame of individuals. In the other cases, the optimal matrix $\tilde{\boldsymbol{\Theta}}_{AB,\,\mathrm{opt}}$ obtained through weak optimality will not likely lead to strong optimization independent of $\mathbf{Y}$.

Example 2: Simple Random Sampling

Let us suppose that the sample s^A is selected using Simple Random Sampling. In that case, the $N^A \times N^A$ matrix $\boldsymbol{\Delta}_A$ is given by

$$\boldsymbol{\Delta}_A = \frac{N^A}{n^A}\frac{(N^A - n^A)}{(N^A - 1)}\left[\mathbf{I}_A - \frac{\mathbf{1}_A\mathbf{1}'_A}{N^A}\right].$$

Considering the factorization of section 6.1, we have

$$\begin{aligned}\boldsymbol{\Delta}_G &= \tilde{\boldsymbol{\Theta}}'_{AG}\boldsymbol{\Delta}_A\tilde{\boldsymbol{\Theta}}_{AG} \\ &= \frac{N^A}{n^A}\frac{(N^A - n^A)}{(N^A - 1)} \times \tilde{\boldsymbol{\Theta}}'_{AG}\left[\mathbf{I}_A - \frac{\mathbf{1}_A\mathbf{1}'_A}{N^A}\right]\tilde{\boldsymbol{\Theta}}_{AG} \\ &= \frac{N^A}{n^A}\frac{(N^A - n^A)}{(N^A - 1)} \times \left[\mathrm{diag}(\mathbf{1}_{A,jj}) - \frac{\mathbf{1}_G\mathbf{1}'_G}{N^A}\right] \qquad (6.9)\end{aligned}$$

where $\mathbf{1}_{A,jj}$ is a square matrix of size N_j^A, with N_j^A being the number of links (or "arrows") starting from unit j of U^A. From $\boldsymbol{\Delta}_G$, we extract the submatrices $\boldsymbol{\Delta}_{G,\,ii}$. Each submatrix $\boldsymbol{\Delta}_{G,\,ii}$ is given by

$$\boldsymbol{\Delta}_{G,\,ii} = \frac{N^A}{n^A}\frac{(N^A - n^A)}{(N^A - 1)} \times \left[\mathbf{I}_{G,\,i} - \frac{\mathbf{1}_{G,\,i}\mathbf{1}'_{G,\,i}}{N^A}\right],$$

which is of size N_i^G. Then, using a matrix result that can be found, amongst others, in Jazwinski (1970), we get

$$\boldsymbol{\Delta}_{G,ii}^{-1} = \frac{(N^A - 1)}{(N^A - n^A)}\frac{n^A}{N^A} \times \left[\mathbf{I}_{G,\,i} + \frac{1}{(N^A - N_i^G)}\mathbf{1}_{G,\,i}\mathbf{1}'_{G,\,i}\right].$$

Now, from (6.5), we directly obtain the optimal values

$$\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \frac{1}{N_i^G}\mathbf{1}_{G,\,i}$$

that minimize $\text{Var}(\hat{Y}^B)$, in the weak sense, $i = 1, \ldots, N^B$. These values are used to construct the vectors $\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}$, and then the matrix $\tilde{\boldsymbol{\Theta}}_{GB,\text{opt}} = \{\tilde{\boldsymbol{\theta}}_{GB,\text{opt},1}, \ldots, \tilde{\boldsymbol{\theta}}_{GB,\text{opt},N^B}\}$. Finally, after computing the optimal matrix $\tilde{\boldsymbol{\Theta}}_{AB,\text{opt}} = \boldsymbol{\Theta}_{AG}\tilde{\boldsymbol{\Theta}}_{GB,\text{opt}}$, we obtain the optimal weights $\mathbf{W}_{\text{opt}}$ using (2.6).

Again, this result is an important one because it goes directly in the direction of the results of Kalton and Brick (1995), Lavallée (2002), and Lavallée and Caron (2001). That is, with Simple Random Sampling, the optimal choice of $\theta^{AB}_{\text{opt},ji}$ should be 1 if there is a link between unit j of U^A and i of U^B, and 0 otherwise.

Using Result 7, we now verify if the conditions (6.7), (6.8a) or (6.8b) for strong optimality independent of y_i are satisfied for the optimal matrix $\tilde{\boldsymbol{\Theta}}_{AB,\text{opt}}$ that we obtain through weak optimization. First, each submatrix $\Delta_{G,ii'}$ of size $N_i^G \times N_{i'}^G$ is given by

$$\boldsymbol{\Delta}_{G,ii'} = \frac{N^A}{n^A}\frac{(N^A - n^A)}{(N^A - 1)} \times \left[\mathbf{H}_{G,ii'} - \frac{\mathbf{1}_{G,i}\mathbf{1}'_{G,i'}}{N^A}\right]$$

where $\mathbf{H}_{G,ii'}$ is a $N_i^G \times N_{i'}^G$ diagonal matrix of ones, "padded" with zeros. Exactly on the same pattern as in example 1, a typical element of $\mathbf{H}_{G,ii'}$ is given by 1 if both i and i' are linked to the same unit j of U^A (that is linked to unit g of U^G), and 0 otherwise. Therefore, we can easily see in which cases the conditions (6.7), (6.8a) or (6.8b) can be satisfied. In fact, because all components of $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i}$ are equal, $\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'}$ is a vector proportional to the sum of the lines of $\boldsymbol{\Delta}_{G,ii'}$, *i.e.*, the sum of the lines of

$$\left[\mathbf{H}_{G,ii'} - \frac{\mathbf{1}_{G,i}\mathbf{1}'_{G,i'}}{N^A}\right].$$

But (6.7) says that this vector must have the same components. This is possible if and only if the matrix $\mathbf{H}_{G,ii'}$ contains only zeros, or if it is of dimension 1×1, which occurs when both i and i' are each linked to only one element of U^A. Therefore, as for Poisson Sampling, strong optimality independent of $\mathbf{Y}$ does not occur in general for Simple Random Sampling.

7. Conclusion

In the present paper, we discussed the use of Indirect Sampling together with the method developed to obtained estimation weights: the Generalized Weight Share Method (GWSM). We then showed the following properties of the GWSM: unbiasedness, the variance computation and transitivity. We presented after a section on the use of the GWSM when the links between the populations U^A and U^B are expressed by ones and zeros, *i.e.*, there is a link or there is not. The section after was devoted to results that are obtained with different forms of link matrices. Finally, we assessed the problem of optimality, *i.e.*, the choice of optimal values to express the links between U^A and U^B in order to minimize the variance of the estimates issued from the GWSM. We have distinguished two kind of optimization: weak and strong optimization.

Weak optimization consists in finding the values of the links to be used in order to minimize, for each unit, the variance of the weights provided by the GWSM. The solution is always uniquely defined, easy to compute and to implement in practice. Weak optimization is also a necessary condition for strong optimization. Strong optimization consists in finding the values of the links in order to minimize the variance of estimation for the total of any variable of interest y. It does not exist for all sampling designs and type of links between the populations U^A and U^B. It also depends on somewhat complicated relations.

We recommend the use of weak optimization because of its flows naturally and the fact that it is very easy to use. Moreover, if our estimation problem can be as well optimized in the strong sense, we will have achieved it through weak optimization, even if it was not demonstrated!

Acknowledgements

The authors would like to thank all the people that showed an interest in Indirect Sampling, and especially in the GWSM. They motivated the writing of this paper that goes beyond what was made previously on this subject.

References

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, Inc. 135-159.

Deville, J.C., (1998). Comment attraper une population en se servant d'une autre. *Insee méthodes, n°84-85-86, Actes des Journées de methodologie statistique des 17-18 mars 1998*, 63-82.

Jazminski, A.H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press.

Kalton, G., and Brick, J.M. (1995). Weighting Schemes for Household Panel Surveys. *Survey Methodology*, 21, 1, 33-44.

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 1, 25-32.

Lavallée, P. (2002). *Le Sondage Indirect, ou la Méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Brussels.

Lavallée, P., and Caron, P. (2001). Estimation using the generalised weight share method: The case of record linkage. *Survey Methodology*, 27, 2, 155-169.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Searle, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.