Proceedings of Statistics Canada's Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

CORRECTING FOR NON-RESPONSE IN INDIRECT SAMPLING

Pierre Lavallée[1]

ABSTRACT

In practice, we do not always have a list of the desired collection units. Instead we may have a list of different units that are somehow related to the collection units. We therefore have two related populations, U^A and U^B, and we want to produce an estimate for U^B. Unfortunately, we only have a sampling frame for U^A. One solution is to select a sample s^A from U^A and produce an estimate for U^B using the existing relationship between the two populations. This may be referred to as *indirect sampling*. To assign a selection probability, or an estimation weight, to the survey units, Lavallée (1995) developed the *generalized weight share method* (GWSM). The GWSM produces an estimation weight that basically constitutes an average of the sampling weights of the units in sample s^A.

With indirect sampling, there can be total non-response in sample s^A from U^A, or in the units selected to be surveyed in U^B, i.e., in the collection units. Since the units in population U^B are surveyed by cluster, there are two types of total non-response associated with U^B: cluster non-response and unit non-response. With indirect sampling, there is also relationship non-response. This type of non-response is associated with the situation where it is impossible to determine whether a given unit of U^B is related to a unit of U^A, which presents serious estimation problems in the application of the GWSM.

KEYWORDS: Generalized weight share method; non-response; cluster sampling

1. INTRODUCTION

To select samples for social or economic surveys, it is useful to have sampling frames, i.e., lists of units that are intended to represent the target populations. Unfortunately, we do not always have a list of the desired collection units but rather a list of different units that are somehow related to the collection units. Thus we have two related populations, U^A and U^B, and we want to produce an estimate for U^B. Yet we only have a sampling frame for U^A. One solution is to select a sample s^A from U^A and produce an estimate for U^B using the existing relationship between the two populations. This may be referred to as *indirect sampling*.

Estimating a total (or an average) for a target population U^B of clusters using a sample selected from population U^A, which is somehow related to the target population, can be a major challenge, especially if the relationships between the units of the two populations are not one to one. The problem is due chiefly to the difficulty of assigning a selection probability, or an estimation weight, to the units surveyed in the target population. The *generalized weight share method* (GWSM) was developed to solve this type of estimation problem.

The GWSM provides an estimation weight for each unit surveyed in target population U^B. This estimation weight basically constitutes an average of the sampling weights of the units in sample s^A. Lavallée (1995) first introduced the GWSM to deal with the problem of cross-sectional weighting of longitudinal household surveys. The GWSM is a generalization of the *weight share method* described by Ernst (1989).

---

1 Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 CANADA. E-mail: pierre.lavallee@statcan.ca.

2. DESCRIPTION

Following a particular sample design, we select a sample s^A containing m^A units from population U^A consisting of M^A units. Let π_j^A be the selection probability of unit *j*, where $\pi_j^A > 0$ for all units $j \in U^A$. In addition, target population U^B contains M^B units. This population is divided into *N* clusters, where cluster *i* contains M_i^B units. It is assumed that there is a relationship (or a correspondence) between units *j* of population U^A and units *k* of clusters *i* of population U^B. That relationship is identified by an indicator variable $l_{j,ik}$, where $l_{j,ik}=1$ if there is a relationship between unit $j \in U^A$ and unit $ik \in U^B$, and 0 otherwise. Note that there may be cases where there is no relationship between a unit *j* of population U^A and units *k* of clusters *i* of population U^B, i.e., $L_j^A = \sum_{i=1}^{N}\sum_{k=1}^{M_i^B} l_{j,ik} = 0$. On the other hand, there may be zero, one or more than one relationship for a unit *k* of cluster *i* of population U^B; that is, it is possible to have $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik} = 0$, $L_{ik}^B = 1$, or even $L_{ik}^B > 1$ for units $ik \in U^B$. However, for us to use the GWSM and for the GWSM to be unbiased, we must satisfy the following constraint: each cluster *i* of U^B must have at least one relationship with a unit *j* of U^A, i.e., $L_i^B = \sum_{k=1}^{M_i^B}\sum_{j=1}^{M^A} l_{j,ik} > 0$.

For each unit *j* selected in s^A, we identify units *ik* of U^B that have a non-zero relationship with *j*, i.e., $l_{j,ik}=1$. If $L_j^A = 0$ for a unit *j* of s^A, there is simply no unit of U^B identified by that unit *j*, which affects the effectiveness of sample s^A but introduces no bias. For each unit *ik* identified, it is assumed that a list of the M_i^B units of cluster *i* containing that unit can be made. Each cluster *i* represents, then, by itself, a population U_i^B where $U^B = \bigcup_{i=1}^{N} U_i^B$. Let Ω^B be the set of *n* clusters identified by units $j \in s^A$.

An important constraint on the survey (or measurement) process is that **all** units belonging to the same cluster must be considered. In other words, if a unit is selected for the sample, then all units in the cluster containing the unit selected will be surveyed. This constraint arises frequently in surveys for two reasons: (i) for economy and (ii) because of the need to produce estimates for the clusters. Thus, all units *k* of clusters $i \in \Omega^B$ are surveyed where we are measuring variable y_{ik} and the number of relationships $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$ between each unit *ik* and population U^A.

For target population U^B, we want to estimate the total $Y^B = \sum_{i=1}^{N}\sum_{k=1}^{M_i^B} y_{ik}$. In applying the GWSM, we want to assign an estimation weight w_{ik} to each unit *k* of surveyed cluster *i*. To estimate the total Y^B for target population U^B, then, we can use the estimator

$$\hat{Y}^B = \sum_{i=1}^{n}\sum_{k=1}^{M_i^B} w_{ik} y_{ik} \qquad (2.1)$$

where *n* is the number of clusters surveyed and w_{ik} is the weight assigned to unit *k* of cluster *i*.

Steps in the GWSM:

Step 1: For each unit *k* of clusters *i* in Ω^B, we calculate the initial weight $w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} t_j / \pi_j^A$, where $t_j=1$ if $j \in s^A$, and 0 if not. Note that a unit *ik* that has no relationship with any unit *j* in U^A automatically has an initial weight of zero.

Step 2: For each unit *k* of clusters *i* in Ω^B, we obtain the total number of relationships $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$. This quantity is the number of relationships between the units of U^A and unit *k* of cluster *i* of population U^B. Hence, the quantity $L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$ is the total number of relationships in cluster *i*.

Step 3: We calculate the final weight $w_i = \sum_{k=1}^{M_i^B} w'_{ik} \big/ L_i^B$.

Step 4: Last, we set $w_{ik} = w_i$ for all $k \in U_i^B$.

Because the estimation weights produced by the GWSM are the same for all M_i^B units in each cluster *i*, estimator (2.1) can written in terms of clusters alone. Thus we have $\hat{Y}^B = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik} y_{ik} = \sum_{i=1}^{n} w_i \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^{n} w_i Y_i$, where $Y_i = \sum_{k=1}^{M_i^B} y_{ik}$.

3. BIAS AND VARIANCE

To compute the bias and variance of estimator $\hat{Y}^B$, we use the theorem described below.

Theorem: Duality of form of $\hat{Y}^B$ in relation to U^A and U^B

Let $z_{ik} = Y_i / L_i^B$ for all $k \in U_i^B$. We set $Z_j = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} z_{ik}$. Hence, the estimator $\hat{Y}^B$ given in (2.1) can also be written in the form

$$\hat{Y}^B = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j \, . \tag{3.1}$$

Estimator $\hat{Y}^B$ can thus be written as a function of either units *ik* of U^B or units *j* of U^A. We note that estimator $\hat{Y}^B$ is simply a Horvitz-Thompson estimator with variable of interest Z_j. This observation leads us to two important corollaries: (i) estimator $\hat{Y}^B$ is unbiased for the purposes of estimating Y^B, and (ii) the formula for the variance of estimator $\hat{Y}^B$ is that of a Horvitz-Thompson estimator with variable Z_j. This variance is given in Särndal, Swensson and Wretman (1992) and elsewhere. Proofs of the theorem and its corollaries can be found in Lavallée (2001).

4. NON-RESPONSE

In censuses and sample surveys alike, non-response is inevitable. It is the subject of countless publications. There are sizable, though hardly comprehensive, bibliographies in, for example, Droesbeke and Lavallée (1996), Hedges and Olkin (1983), and Bogeström, Larsson and Lyberg (1983). Since the topic is so broad, we will confine our remarks to *total non-response*, as opposed to *partial non-response*. As the GWSM generates estimation weights, we will focus on the adjustment of those weights for total non-response.

4.1 Types of non-response

With indirect sampling, there can be total non-response in sample s^A from U^A, or in the units selected to be surveyed in U^B, i.e., in the collection units. Since the units in population U^B are surveyed by cluster, there are two types of total non-response associated with cluster sampling (direct or indirect): *cluster non-response* and *unit non-response*. Cluster non-response refers to situations where none of the units in the cluster responded to the survey. Unit non-response is a kind of total non-response in which one or more units in the cluster, but not all units, did not respond. With indirect sampling, there is also another form of non-response: *relationship non-response*. This type of non-response is associated with the situation where it is impossible to determine whether a unit *j* of U^A is related to a unit *ik* of U^B.

4.2 Response probabilities

The concept of response probability is very useful in adjusting estimates for total non-response. In a general context, let δ_k be an indicator variable that takes a value of 1 if unit *k* answers the survey questions, and 0 if not. It is generally assumed that this variable has a Bernoulli distribution with probability ϕ_k . In other words, it is assumed that each individual *k* in the survey population has a certain probability ϕ_k of responding to the survey, i.e., $P(\text{unit } k \text{ responds} \mid s) = P(\delta_k = 1 \mid s) = \phi_k$. In addition, for two units *k* and k', the indicator variables δ_k and $\delta_{k'}$ are deemed to be independent. This implies that the joint probability of response $\phi_{kk'}$ for these two units is given by $\phi_{kk'} = P(\delta_k = 1, \delta_{k'} = 1 \mid s) = P(\delta_k = 1 \mid s)P(\delta_{k'} = 1 \mid s) = \phi_k \phi_{k'}$. The independence between the indicator variables δ of two units *k* and k' follows from the assumption that the choice made by unit *k* to respond or not will have no bearing on the choice made by unit k'. Lastly, we have $E(\delta_k \mid s) = P(\delta_k = 1 \mid s) = \phi_k$ and $Var(\delta_k \mid s) = \phi_k(1 - \phi_k)$.

To estimate the response probabilities ϕ_k, we often use a model. A model frequently used in practice (Särndal, Swensson and Wretman, 1992) is the uniform model within *homogeneous response groups* (HRGs); that is,

$$\phi_{qk} = E(\delta_{qk} \mid s) = \beta_q \tag{4.1}$$

where q=1,…,Q, where Q is the number of HRGs and β_q is a fixed effect (or factor) to be estimated. The parameter β_q is in fact the expected probability of response in group *q*. With this model, we assume that all units in the same HRG have the same probability of response. The HRGs can be formed by a single factor or by a combination or two or more factors.

To estimate ϕ_{qk}, we can use the weighted maximum likelihood (or pseudo-maximum likelihood) method with weights set at $1/\pi_k$. The estimator derived from model (4.1) is given by the weighted response rate; that is,

$$\hat{\phi}_{qk} = R_q = \frac{\sum_{k=1}^{n_{r,q}} 1/\pi_k}{\sum_{k=1}^{n_q} 1/\pi_k} = \frac{\hat{N}_{r,q}}{\hat{N}_q} \tag{4.2}$$

where n_q is the number of units in the sample belonging to HRG *q*, and $n_{r,q}$ is the number of respondent units in this group.

4.3 Treatment of non-response in s^A

Non-response in sample s^A is a classic case of non-response. This type of non-response is covered in most books on sampling theory. In the theorem in section 3, we saw that the estimator $\hat{Y}^B$ produced by the GWSM can be written in the form of a Horvitz-Thompson estimator that is a function of units *j* of s^A. Hence, non-response in s^A is treated as we would treat non-response in the situation where we selected sample s^A to produce an estimate of a quantity related to population U^A.

It is assumed that a subset s_r^A of m_r^A units of sample s^A answered the survey questions. Let Ω_r^B be the set of n_r clusters identified by units $j \in s_r^A$. We survey all units *k* of clusters $i \in \Omega_r^B$, the variable of interest being *y*.

In applying the GWSM, we normally want to assign an estimation weight w_{ik} to each unit *k* of surveyed cluster i. To estimate the total Y^B for target population U^B, then, we can use estimator (2.1), which was constructed on the assumption that there is no non-response in sample s^A On the basis of the theorem in section 3, we can rewrite estimator (2.1) as (3.1), which is a function of units *j* of s^A. Since we have only subsample s_r^A of the respondent units, we have to use an estimator that has been corrected for non-response. To that end, we can use the following estimator:

$$\hat{Y}^{NRA,B} = \sum_{j=1}^{M^A} \frac{t_j \delta_j^A}{\pi_j^A \phi_j^A} Z_j = \sum_{j=1}^{m_r^A} \frac{Z_j}{\pi_j^A \phi_j^A} \tag{4.3}$$

where ϕ_j^A is the response probability of unit *j*. Indicator variable $\delta_j^A = 1$ if unit *j* of s^A responds, and 0 if not. Using the theorem in section 3, we can show that estimator (4.3) is unbiased.

To apply the GWSM with an adjustment for non-response in s^A, we need only replace $1/\pi_j^A$ with $1/\pi_j^A\phi_j^A$ in step 1 of the GWSM (see section 2). After applying the non-response-adjusted GWSM, we obtain estimation weight w_{ik}^{NRA}, which is used in the estimator

$$\hat{Y}^{NRA,B} = \sum_{i=1}^{n_r} \sum_{k=1}^{M_i^B} w_{ik}^{NRA} y_{ik} \tag{4.4}$$

In practice, estimator $\hat{Y}^{NRA,B}$ is useful only if the value of the response probabilities ϕ_j^A is known for all units *j* of s_r^A. Since the ϕ_j^A are probably unknown, we want to estimate them so that we can use one of the following forms:

$$\hat{\hat{Y}}^{NRA,B} = \sum_{j=1}^{m_r^A} \frac{Z_j}{\pi_j^A \hat{\phi}_j^A} \quad \text{or} \quad \hat{\hat{Y}}^{NRA,B} = \sum_{i=1}^{n_r} \sum_{k=1}^{M_i^B} \hat{w}_{ik}^{NRA} y_{ik} \tag{4.5}$$

where the weight $\hat{w}_{ik}^{NRA}$ is obtained by replacing ϕ_j^A with $\hat{\phi}_j^A$. To obtain $\hat{\phi}_j^A$, we can use model (4.1), which in this case takes the form $\phi_{qj}^A = \beta_q^A$. In view of the estimator based on this model, we use the response rate $R_q^A = (\sum_{j=1}^{m_{r,q}^A} 1/\pi_{qj}^A)/(\sum_{j=1}^{m_q^A} 1/\pi_{qj}^A)$. Thus we have

$$\hat{\hat{Y}}^{NRA,B} = \sum_{j=1}^{m_r^A} \frac{Z_j}{\pi_j^A \hat{\phi}_j^A} = \sum_{q=1}^{Q} \frac{\sum_{j=1}^{m_{r,q}^A} Z_{qj} / \pi_{qj}^A}{\sum_{j=1}^{m_{r,q}^A} 1/\pi_{qj}^A} \left(\sum_{j=1}^{m_q^A} 1/\pi_{qj}^A \right) \tag{4.6}$$

where m_q^A is the number of units in sample s^A belonging to HRG *q*, and $m_{r,q}^A$ is the number of respondent units in this group. If we look at estimator (4.6), we see that it is nothing more than a ratio estimator in two-phase sampling. Särndal, Swensson and Wretman (1992) present a proof that estimator (4.6) is asymptotically unbiased, under the conditions of model (4.1). The formula for the variance of estimator (4.6) is given in Lavallée (2001).

4.4 Treatment of cluster non-response

To address cluster non-response, we start with the situation where we select a sample s^A of m^A units, as in section 4.3. This time, however, we assume that all m^A units in the sample answered the survey questions. In carrying out the survey process, we attempt to survey all units *k* in clusters *i* of Ω^B. Unfortunately, for some entire clusters, we are unable to obtain any data. This is cluster non-response. We assume that there are no clusters in which only a non-null subset of units responded. Let Ω_r^B be the set of n_r respondent clusters.

Let δ_i^B be an indicator variable that takes a value of 1 if cluster *i* answers the survey questions, and 0 if not. As in section 4.2, it is assumed that each cluster *i* in U^B has a certain probability Φ_i^B of responding to the survey, i.e., $P(\text{cluster } i \text{ responds} \mid \Omega^B) = P(\delta_i^B = 1 \mid \Omega^B) = \Phi_i^B$. In addition, for two clusters *i* and i', the indicator variables δ_i^B et $\delta_{i'}^B$ are deemed to be independent.

In applying the GWSM, we want to assign an estimation weight w_{ik}^{NRG} to each unit *k* of respondent cluster i. To estimate the total Y^B for target population U^B, then, we can use the estimator

$$\hat{Y}^{NRG,B} = \sum_{i=1}^{n_r}\sum_{k=1}^{M_i^B} w_{ik}^{NRG} y_{ik} = \sum_{i=1}^{n}\delta_i^B \sum_{k=1}^{M_i^B} w_{ik}^{NRG} y_{ik} \qquad (4.7)$$

To obtain the weight w_{ik}^{NRG} from the GWSM, we use the response probability Φ_i^B for each cluster $i \in \Omega_r^B$. Then we need only replace w_i with w_i / Φ_i^B in step 4 of the GWSM given in section 2.

In practice, we want to estimate the known probabilities Φ_i^B for all clusters *i* in Ω_r^B so that we can use the estimator

$$\hat{\hat{Y}}^{NRG,B} = \sum_{i=1}^{n_r}\sum_{k=1}^{M_i^B} \hat{w}_{ik}^{NRG} y_{ik} \qquad (4.8)$$

where $\hat{w}_{ik}^{NRG} = w_{ik} / \hat{\Phi}_i^B$. To obtain $\hat{\Phi}_i^B$, we can follow the example of estimator (4.2). In this case, model (4.1) takes the form $\Phi_{qi}^B = \beta_q^B$. If we use this model, we define $\hat{\Phi}_{qi}^B$ as follows:

$$\hat{\Phi}_{qi}^B = R_q^B = \frac{\sum_{i=1}^{n_{r,q}}\sum_{k=1}^{M_{qi}^B} w_{qik}}{\sum_{i=1}^{n_q}\sum_{k=1}^{M_{qi}^B} w_{qik}} = \frac{\hat{M}_{r,q}^B}{\hat{M}_q^B} \qquad (4.9)$$

where w_{qik} is the estimation weight provided by the GWSM (assuming non-response is zero) for units *k* in clusters *i* belonging to HRG *q*. With (4.9), the estimator $\hat{\hat{Y}}^{NRG,B}$ given by (4.8) becomes

$$\hat{\hat{Y}}^{NRG,B} = \sum_{i=1}^{n_r}\sum_{k=1}^{M_i^B} \frac{w_{ik}}{\hat{\Phi}_i^B} y_{ik} = \sum_{q=1}^{Q} \frac{\hat{M}_{r,q}^B}{\hat{M}_q^B} \sum_{i=1}^{n_q}\delta_{qi}^B \sum_{k=1}^{M_{qi}^B} w_{qik} y_{qik} \qquad (4.10)$$

We can look at estimator (4.10) as a ratio estimator in two-phase sampling. As in the case of non-response in s^A, we can show that estimator (4.10) is asymptotically unbiased under the conditions of model (4.1). The formula for the variance of estimator (4.10) is given in Lavallée (2001).

4.5 Treatment of unit non-response

To address unit non-response, we start with the situation where we select a sample s^A of m^A units. We assume that all m^A units in s^A answered the survey questions. For each unit *j* in s^A, we identify units *ik* of U^B that have $l_{j,ik} = 1$. In carrying out the survey process, we attempt to survey all units *k* in clusters *i* of Ω^B. Unfortunately, for some units in the identified clusters, we are unable to obtain any data. This is unit non-response. We assume here that we have a response for at least one unit in each cluster *i* in Ω^B. Let $s_{r,i}^B$ be the set of respondent units in identified cluster *i*, and let $M_{r,i}^B > 0$ be the size of that set.

Let $\delta_{(i)k}^B$ be an indicator variable that takes a value of 1 if cluster *i* answers the survey questions, and 0 if not. We assume that each unit *k* in clusters *i* of U^B has a probability $\phi_{(i)k}^B$ of responding to the survey, i.e., $P(\text{unit } k \in i \text{ responds} \mid \Omega^{\mathrm{B}}) = P(\delta_{(i)k}^B = 1 \mid \Omega^{\mathrm{B}}) = \phi_{(i)k}^B$. In addition, for two units k and k' of cluster *i*, or of two different clusters, the indicator variables $\delta_{(i)k}^B$ and $\delta_{(i)k'}^B$ are deemed to be independent.

In applying the GWSM, we want to assign an estimation weight w_{ik}^{NRU} to each respondent unit *k* of cluster *i* in Ω^B. To estimate the total Y^B for target population U^B, then, we can use the estimator

$$\hat{Y}^{NRU,B} = \sum_{i=1}^{n}\sum_{k=1}^{M_{r,i}^B} w_{ik}^{NRU} y_{ik} = \sum_{i=1}^{n}\sum_{k=1}^{M_i^B} \delta_{(i)k}^B w_{ik}^{NRU} y_{ik} \qquad (4.11)$$

Weight w_{ik}^{NRU} can be obtained by drawing a parallel with two-stage indirect sampling (Lavallée, 2001). In other words, we can look at the unit non-response process as the selection of a sample $s_{r,i}^{B}$ of $M_{r,i}^{B}$ units obtained from the M_i^B units in each cluster *i* of Ω^B. Thus, we obtain $w_{ik}^{NRU} = w_i / \phi_{(i)k}^{B}$ for all the $k \in s_{r,i}^{B}$ where *i*=1,…,*n* , and w_i is the estimation weight provided by the GWSM (assuming zero non-response).

In practice, we can estimate the probabilities $\phi_{(i)k}^{B}$ so that we can use the estimator

$$\hat{\hat{Y}}^{NRU,B} = \sum_{i=1}^{n} \sum_{k=1}^{M_{r,i}^{B}} \hat{w}_{ik}^{NRU} y_{ik} \tag{4.12}$$

where $\hat{w}_{ik}^{NRU} = w_i / \hat{\phi}_{(i)k}^{B}$. To obtain $\hat{\phi}_{(i)k}^{B}$, we can use an approach that involves considering each set $s_{r,i}^{B}$ of respondent units separately. Then the response probabilities are estimated within each cluster *i*. With this so-called individual approach, model (4.1) takes the form $\phi_{(qi)k}^{B} = \beta_{qi}^{B}$, and the HRGs are then defined within each cluster *i*. We define $\hat{\phi}_{(qi)k}^{B}$ as follows:

$$\hat{\phi}_{(qi)k}^{B} = R_{qi}^{B} = \frac{M_{r,qi}^{B}}{M_{qi}^{B}} \tag{4.13}$$

where $M_{qi}^{B} = M_i^B$ and $M_{r,qi}^{B} = M_{r,i}^{B}$ for $i \in q$. With (4.13), the estimator $\hat{\hat{Y}}^{NRU,B}$ given by (4.12) becomes

$$\hat{\hat{Y}}^{NRU,B} = \sum_{i=1}^{n} \sum_{k=1}^{M_{r,i}^{B}} \frac{w_i}{\hat{\phi}_{(i)k}^{B}} y_{ik} = \sum_{q=1}^{Q} \sum_{i=1}^{n_q} \frac{M_{qi}^{B}}{M_{r,qi}^{B}} \sum_{k=1}^{M_{r,qi}^{B}} w_{qi} y_{qik} \tag{4.14}$$

This estimator is nothing more than a ratio estimator within each PSU under two-stage sampling. Estimator $\hat{\hat{Y}}^{NRU,B}$ is asymptotically unbiased for estimating Y^B, under the conditions of model (4.1). The formula for the variance of estimator (4.14) is given in Lavallée (2001). In addition to the individual approach, Lavallée (2001) describes a so-called “aggregate” approach, which consists in considering the set $s_r^B = \bigcup_{i=1}^{n} s_{r,i}^{B}$ of respondent units as a whole.

4.6 Treatment of relationship non-response

To address the treatment of relationship non-response, we start with the situation where we select a sample s^A of m^A units. In carrying out the survey process, for each unit *j* in s^A, we identify units *ik* of U^B that have $l_{j,ik} = 1$. We assume that we can identify **all** relationships $l_{j,ik}$ associated with each unit *j* de s^A. For each identified un*it ik*, we assume that we can make a list of the M_i^B units of cluster *i* containing that unit. We survey all units *k* in clusters i of Ω^B Although we can measure the variable of interest *y* for **all** M_i^B units in each cluster *i* of Ω^B, for some units *k*, relationship non-response prevents us from determining whether there is a relationship between those units *k* and a unit *j* of U^A. In other words, for **some** units *k* of a cluster $i \in \Omega^B$, it is impossible to determine whether $l_{j,ik} = 1$ or $l_{j,ik} = 0$.

Let $L_{r,i}^{B}$ be the total number of relationships found between cluster *i* and population U^A. Note that $L_{r,i}^{B} \le L_i^B$. In addition, because we are assuming that we can identify **all** relationships $l_{j,ik}$ associated with each unit *j* of s^A, we have $L_{r,i}^{B} > 0$ for all clusters *i* in Ω^B. By using only the total number $L_{r,i}^{B}$ of relationships found, we overestimate the total Y^B. However, in contrast to the other types of non-response discussed above, there seems to be no “miracle” solution to the problem of relationship non-response. One possible approach is to try to model the quantity L_i^B using auxiliary variables. With the resulting estimate $\hat{L}_i^B$, we can construct the estimator

$$\hat{\hat{Y}}^{NRL,B} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} Y_i \frac{L_{j,i}}{\hat{L}_i^B} \quad (4.15)$$

where $L_{j,i} = \sum_{k=1}^{M_i^B} l_{j,ik}$. Estimator (4.15) follows from the theorem discussed in section 3. We can show that estimator $\hat{\hat{Y}}^{NRL,B}$ is asymptotically unbiased. The unbiased nature of this estimator depends, however, on the unbiased nature of the estimator of L_i^B. In practice, it is not easy to obtain an unbiased estimator of L_i^B. Ardilly and le Blanc (2000) encountered a relationship non-response problem when they used the GWSM to weight a survey of homeless persons. They suggested making an assumption of regularity to impute some relationships $l_{j,ik}$ to 1, which is actually the same as modelling the quantity L_i^B.

Another possible solution to the overestimation problem is benchmarking. While it offers an attractive solution to the relationship non-response problem, it depends on the availability of auxiliary variables $\mathbf{x}_{ik}^B$ correlated with the variable of interest y_{ik}, which is not always the case in practice. The best solution is still to measure L_i^B exactly or, failing that, to obtain an estimate $\hat{L}_i^B$ that is as close as possible to L_i^B.

REFERNCES

Ardilly, P., le Blanc, P. (2000). "*Comment pondérer une enquête auprès des personnes sans domicile?*", paper presented at Deuxième Colloque Francophone sur les Sondages, Université Libre de Bruxelles, Belgique, June 2000.

Bogeström, B., Larsson, M., Lyberg, L. (1983). "Bibliography on Nonresponse and Related Topics", in *Incomplete Data in Sample Surveys* (Madow, W.G., Olkin, I., Rubin, D.B., Eds), Vol. 2, Academic Press, New York, pp. 479-567.

Droesbeke, J.-J., Lavallée, P. (1996). "La non-réponse dans les enquêtes", *Methodologica,* No. 4, pp. 1-39.

Ernst, L. (1989). "Weighting issues for longitudinal household and family estimates", in *Panel Surveys* (Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P., Ed), John Wiley and Sons, New York, pp. 135-159.

Hedges, L.V., Olkin, I. (1983). "Selected Annotated Bibliography", in *Incomplete Data in Sample Surveys* (Madow, W.G., Olkin, I., Rubin, D.B., Ed), Vol. 2, Academic Press, New York, pp. 417-478.

Lavallée, P. (1995) : Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households using the Weight Share Method, *Survey Methodology* vol. 21, p.25-32.

Lavallée, P. (2001). "La Méthode généralisée du partage des poids (ou le Sondage indirect)", Ph’d thesispresented at l’Université Libre de Bruxelles, Belgique, April 2000.

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