Indirect Sampling and the Problem of Links Identification

Pierre Lavallée[1] and Xiaojian Xu[2]
1 Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6, CANADA, plavall@statcan.ca
2 Department of Mathematics, Brock University, St. Catharines, Ontario, L2S 3A1, CANADA, xxu@brocku.ca

Abstract

Indirect Sampling involves two populations, U^A and U^B, that are linked by some relation, and for which we wish to produce estimates for U^B. Unfortunately, we may have a sampling frame for U^A only. In response, we can select a sample from U^A in order to produce estimations for U^B using the links that exist between the two populations. In order to calculate the weights for the units collected from U^B, we use the generalized weight share method (GWSM). Unfortunately, there are some applications for which the links between U^A and U^B can be difficult to identify. In these cases, the number of links tends to be under-estimated, which leads to over-estimating quantities such as totals. In order to correct this over-estimation, a number of solutions are proposed. In this paper, we present in details the problem of links identification, along with the different proposed solutions.

Keywords: Generalized Weight Share Method, Cluster Sampling, estimation, non-response.

1. Introduction

In order to select the probabilistic samples needed for social or economic surveys, it is useful to have access to sampling frames, that is, lists of units that are intended to represent the target populations. Unfortunately, it can happen that there is no direct access to a list of the desired collection units, but rather a list of units that are connected in some way to the list of collection units. Consider two populations U^A and U^B, that are linked by some relation, and for which we wish to produce an estimate for U^B. Unfortunately, we have a sampling frame for U^A only. In this case, we can conceivably select a sample s^A from U^A in order to produce an estimate for U^B using the existing relationship between the two populations. This is called *Indirect Sampling*.

Here, we are interested in target populations where the units are organized into clusters. Estimating the total (or mean) of a target clustered population U^B using a sample selected from another population U^A that is connected in some way to U^B can be a considerable challenge, especially if the links between the units from the two populations are not one-to-one. The problem is largely due to the difficulty of associating a selection probability, or an estimation weight, to the surveyed units from the target population. In order to resolve this estimation problem, the *Generalized Weight Share Method* (GWSM) has been developed.

2. The Generalized Weight Share Method

The GWSM enables us to obtain an estimation weight for each surveyed unit from the target population U^B. This estimation weight generally corresponds to a mean of the survey weights for the units from population U^A that are linked to the unit from U^B. Lavallée (1995) first described the GWSM in connection with the cross-sectional weighting of longitudinal surveys of households. The GWSM constitutes a generalization of the *weight share method* described by Ernst (1989).

Let us suppose that there are links between the units j of population U^A and the units k of the clusters i of population U^B. Each link is identified by an indicator variable $l_{j,ik}$, where $l_{j,ik}=1$ if there is a link between the unit $j \in U^A$ and the unit $ik \in U^B$, and 0 otherwise. Note that each cluster i from U^B must have at least one link with a unit j from U^A.

The survey process is as follows: we select a sample s^A containing m^A units from population U^A containing M^A units according to a given sampling design. Let $\pi_j^A > 0$ represents the selection probability of unit j. On the other hand, the target population U^B contains M^B units. This population is divided into N clusters, where cluster i contains M_i^B units. For each unit j selected in s^A, we identify the units ik from U^B that have a non-null relationship with j, i.e., $l_{j,ik}=1$. For each unit ik identified, we suppose that a list can be compiled of all M_i^B units of cluster i containing this unit. For each cluster i of U^B, let $L_{j,i} = \sum_{k=1}^{M_i^B} l_{j,ik}$. Also, let $\Omega^B = \{i \in U^B \mid \exists j \in s^A \text{ and } L_{j,i} > 0\}$ be the set of the n clusters identified by the units $j \in s^A$. We survey all the units k from the clusters $i \in \Omega^B$ for measuring a certain variable of interest y_{ik} and the number of links $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$ between the unit ik from the target population U^B and population U^A.

One important constraint on the survey process (or measurement process) is to consider all units that belong to the same cluster. In other words, if a unit is selected from the sample, then all the units from the cluster containing the selected unit are

surveyed. This constraint often occurs in surveys for two reasons: (i) cost savings and (ii) the need to produce estimates at the cluster level.

Let us suppose that we wish to estimate the total $Y^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik}$ for the target population U^B. In applying the GWSM, we want to assign an estimation weight w_{ik} to each unit k from a surveyed cluster i.

Steps of the GWSM:

Step 1: For each unit k from the clusters i in Ω^B, we calculate the initial weight w'_{ik}, as $w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} t_j^A / \pi_j^A$ where $t_j^A = 1$ if $j \in s^A$, and 0 otherwise.

Step 2: For each unit k from the clusters i in Ω^B, we measure the total number of links $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$, as well as $L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$. *The quantity L_i^B corresponds to the total number of links from the cluster i.*

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

Step 4: We set $w_{ik} = w_i$ for all units $k \in$ cluster i.

To estimate the total Y^B of the target population U^B, we can use the estimator

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

where n is the number of surveyed clusters. Following steps 1 to 4, we obtain

$$w_{ik} = \sum_{j=1}^{M^A} \frac{t_j^A}{\pi_j^A} \frac{L_{j,i}}{L_i^B} \qquad (2)$$

and thus,

$$\hat{Y}^B = \sum_{j=1}^{M^A} \frac{t_j^A}{\pi_j^A} \sum_{i=1}^{N} Y_i \frac{L_{j,i}}{L_i^B} \qquad (3)$$

where $Y_i = \sum_{k=1}^{M_i^B} y_{ik}$. Lavallée (1995) showed that $\hat{Y}^B$ is unbiased for the estimation of Y^B.

3. Types of total non-response

In Indirect Sampling, total non-response may occur within the sample s^A selected from U^A, or within the units identified for surveying within U^B, i.e., the collection units. Non-response within s^A is a classic case of non-response. It can be handled like the case where a sample s^A is selected in order to produce estimates for U^A. Given that the survey of units within the population U^B occurs by clusters, we can distinguish two types of total non-response within U^B: *cluster non-response* and *unit non-response*. Cluster non-response is where none of the cluster units responded to the survey. This case frequently occurs in practice. Unit non-response is a total non-response where one or more cluster units, but not all, did not answer. Unit non-response occurs as frequently as does cluster non-response, but for different reasons (Lavallée, 2002).

With Indirect Sampling, we find another form of non-response, namely the *problem of links identification*. This type of non-response is associated with the situation where it is not possible to establish whether a unit *ik* from U^B is linked to a unit j from U^A. This problem has already been mentioned by Sirken and Nathan (1988) in the context of Network Sampling. More recently, Ardilly and Le Blanc (2001) addressed this problem while using the GWSM to weight a survey of homeless persons. The problem of identifying links is particularly problematic for the GWSM because it can create serious bias problems in the estimates.

To use the GWSM, only the linkages between the units j from s^A and the units from the clusters in Ω^B are necessary. In practice, often these links can be obtained by interviewing the units j selected from s^A, and the units k from the clusters i identified in U^B. In the presence of problems in links identification, it is not possible to establish whether a unit *ik* from U^B is linked to a unit j from U^A.

For example, let U^B be the target population of homeless persons, and let U^A represent the set of services (meals, bed, etc.) that are provided to these homeless persons. Using Indirect Sampling, we select a sample s^A of services from U^A, in order to estimate the population U^B of homeless persons. Now, for each service selected in s^A, we are able to identify the homeless person that used this service. However, the GWSM requires to know <u>all</u> services that the identified homeless person has received, and this is often difficult to get because these persons are usually difficult to interview. This is a problem of identification of links (see Ardilly and Le Blanc, 2001).

Suppose that, based on interviewing, we know the links $l_{j,ik}$ for all the units j from s^A. However, for certain units from the clusters i in U^B identified by

s^A, we do not know all the links $l_{j,ik}$ leading back to U^A. In other words, we know $l_{j,ik}$ for $j \in s^A$, but we do not know all the $l_{j,ik}$ for $j \in \Omega^{A|B}$ where $\Omega^{A|B} = \{j \in U^A \mid \exists i \in \Omega^B \text{ and } L_{j,i} > 0\}$. The set $\Omega^{A|B}$ contains the units j from U^A that have a link to the clusters in Ω^B that were identified at the start by the sample s^A.

One direct consequence of not knowing all the links that lead to U^A is that it makes difficult to establish the values of L_i^B, which are essential to the GWSM (see equation (3)). Let L_i^{B*} be the total number of established links between cluster i and the population U^A. Here, $L_i^{B*} \le L_i^B$, which produces an over-estimation of Y^B because

$$\hat{Y}^{B*} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} Y_i \frac{L_{j,i}}{L_i^{B*}} \quad \geq \quad \hat{Y}^B \tag{4}$$

There are a number of conceivable solutions to correct this problem. They are presented in the following sections.

4. Some possible solutions to the problem of links identification

4.1 Record linkage

If we have access to two files A and B containing U^A and U^B, respectively, we can try to obtain all the links between these two populations. One way to obtain the values for $l_{j,ik}$ is to perform a *record linkage*. For more details on record linkage, the reader is directed to Fellegi and Sunter (1969), and to Lavallée and Caron (2001).

If obtaining the values $l_{j,ik}$ reveals to be too difficult because, for example, of the size of the files A and B, one can restrict the record linkage to the units k from the clusters i in Ω^B and the population U^A.

One can also use record linkage to try evaluating $L_{j,i}$ between the clusters i in Ω^B and the population U^A. As we can see from (3), it is sufficient to obtain the quantities $L_{j,i}$, rather than the individual links $l_{j,ik}$, for using the estimator $\hat{Y}^B$.

4.2 Modelling

It is possible to estimate the probabilities $\phi_{j,ik}$ of a link between the units j and ik by using a logistic-type model with vectors $\mathbf{x}_j^A$ and $\mathbf{x}_{ik}^B$ of auxiliary variables. Let $\Delta^{A|B}$ be the set of units from $\Omega^{A|B}$ for which the links have been identified. We can see that s^A is a subset of $\Delta^{A|B}$, which is itself a subset of $\Omega^{A|B}$. With the estimated probabilities $\hat{\phi}_{j,ik}$, we can produce estimations $\hat{l}_{j,ik} = \hat{\phi}_{j,ik}$ for the units j contained in the set $\Omega^{A|B} \setminus \Delta^{A|B}$. This can be seen as imputing links $l_{j,ik}$ for these units (see Ardilly and Le Blanc, 2001). Note that it is important to make maximum use of every constraints and information that would be associated with the values for $l_{j,ik}$ during modelling.

With $\hat{l}_{j,ik}$, we obtain

$$\hat{L}_{ik}^B = \sum_{j \in \Delta^{A|B}} l_{j,ik} + \sum_{j \in \Omega^{A|B} \setminus \Delta^{A|B}} \hat{l}_{j,ik} \tag{5}$$

and then

$$\hat{L}_i^B = \sum_{k=1}^{M_i^B} \hat{L}_{ik}^B \tag{6}$$

It is also possible to concentrate on L_i^B as a whole, without reference to the population U^A. The approach is then to use a log-linear model of the type $\log(L_i^B) = \boldsymbol{\beta}' \mathbf{x}_i^B$. Once again, it is important to make maximum use of every constraints and information that would be associated with the values for L_i^B during modelling.

4.3 Estimating $\theta_{j,i} = L_{j,i} / L_i^B$

Another way of solving the problem of links identification is to concentrate on the quantity $\theta_{j,i} = L_{j,i} / L_i^B$, rather than on the number of links $L_{j,i}$. In order to make the estimator (3) unbiased, we need only to ensure that $\sum_{j=1}^{M^A} \theta_{j,i} = 1$ (see Ernst, 1989, as well as Lavallée and Deville, 2002). Thus, it is not necessary to know every values of $L_{j,i}$ for $i \in \Omega^{A|B}$, but simply the values $\theta_{j,i}$ for $i \in s^A$, making sure that $\sum_{j=1}^{M^A} \theta_{j,i} = 1$.

It is important to note that $\theta_{j,i}$ can be defined generally, without reference to the links $L_{j,i}$. As mentioned by Lavallée (2002), some $\widetilde{\theta}_{j,i}$ can be defined arbitrarily by keeping $\sum_{j=1}^{M^A}\widetilde{\theta}_{j,i}=1$, which means that we can also use the unbiased estimator:

$$\hat{Y}^B=\sum_{j=1}^{M^A}\frac{t_j^A}{\pi_j^A}\sum_{i=1}^{N}\widetilde{\theta}_{j,i}Y_i \qquad (7)$$

As an example, this approach was used by Bankier (1983) to produce statistics from tax data (see Lavallée, 2002).

4.4 Calibration

Calibration has been generalized by Deville and Särndal (1992). The technique involves adjusting the survey weights in such a way that the estimations are calibrated to known totals. As a result, calibration can help to correct the over-estimation of the estimator $\hat{Y}^{B*}$ given by (4). This type of correction depends on the availability of auxiliary variables $\mathbf{x}_{ik}^B$ correlated with the variable of interest y_{ik}, and of control totals $\mathbf{X}^B$. Here, calibration can be performed either before the GWSM is applied, or after (Lavallée, 2002).

5. Proportional adjustments

Another possible approach for solving the problem of links identification is to directly estimate L_i^B by *proportional adjustment*.

Let $\Omega_i^{A|B}=\{j\in U^A \mid i\in\Omega^B \text{ and } L_{j,i}>0\}$ and let $M_i^{A|B}$ be the number of units j in $\Omega_i^{A|B}$. The set $\Omega_i^{A|B}$ contains the units from U^A that have a link to the cluster i in Ω^B. Note that, in general, we have $\Omega_i^{A|B}\cap\Omega_{i'}^{A|B}\neq\emptyset$, and thus, $\sum_{i=1}^{N}M_i^{A|B}\geq M^A$. Since for a given cluster i in Ω^B, $L_{j,i}$ are non-nulls only for the units j of the set $\Omega_i^{A|B}$, we directly obtain $L_i^B=\sum_{j=1}^{M^A}L_{j,i}=\sum_{j=1}^{M_i^{A|B}}L_{j,i}$. Finally, we have that $M_i^{A|B}\leq L_i^B$.

The set $\Omega_i^{A|B}$ contains the units j from U^A that have a link to the cluster i, whether they are in the sample s^A or not. Let us define $s_i^{A|B}=\{j\in s^A \mid i\in\Omega^B \text{ et } L_{j,i}>0\}$ and let $m_i^{A|B}$ be the number of units j in $s_i^{A|B}$. The set $s_i^{A|B}$ contains the units from s^A that are linked to the cluster i. We can see $s_i^{A|B}$ as a "sample" of $\Omega_i^{A|B}$. Let the "selection probability" be $\pi_{j|i}^{A|B}=P(j\in s_i^{A|B}\mid j\in\Omega_i^{A|B})$. It should be noted that $\pi_{j|i}^{A|B}$ is a function of π_j^A. Accordingly, we can define the following estimator for L_i^B:

$$\hat{L}_i^B=\sum_{j=1}^{m_i^{A|B}}\frac{t_{j|i}^{A|B}}{\pi_{j|i}^{A|B}}L_{j,i} \qquad (8)$$

where $t_{j|i}^{A|B}=1$ if $j\in s_i^{A|B}$, and 0 otherwise. It can be shown that $E(\hat{L}_i^B)=\sum_{j=1}^{M_i^{A|B}}L_{j,i}=L_i^B$ and thus, the estimator (3) used with (8) is asymptotically unbiased for the estimation of Y^B.

One of the difficulties in using the estimator (8) involves calculating the probabilities $\pi_{j|i}^{A|B}$. If the π_j^A are relatively homogenous, then we can use $\hat{\pi}_{j|i}^{A|B}=m_i^{A|B}/M_i^{A|B}=f_i^{A|B}$. This approach allows us to focus not on the links themselves and the quantities $L_{j,i}$ and L_i^B, but solely on the units j from U^A that are involved in the survey of the units i in Ω^B.

Unfortunately, $M_i^{A|B}$ is often unavailable because of the error in observing the links, which makes it difficult to use the estimator (8). In this case, we can try to estimate L_i^B by *global proportional adjustment*. For this adjustment, the variations between the clusters i from U^B are ignored. Thus, s^A is considered to be a "sample" of $\Omega^{A|B}$ and the "selection probability" is defined as $\pi_j^{A|B}=P(j\in s^A\mid j\in\Omega^{A|B})$. Note that $\pi_j^{A|B}$ is a function of π_j^A. Accordingly, in order to estimate L_i^B, we can use

$$\hat{\hat{L}}_i^B=\sum_{j=1}^{m_i^{A|B}}\frac{t_j^{A|B}}{\pi_j^{A|B}}L_{j,i} \qquad (9)$$

As for the estimator (8), one of the difficulties in using the estimator (9) is obtaining $\pi_j^{A|B}$. Here, we can try to use the approximation $\hat{\pi}_j^{A|B} = m^A / M^{A|B} = f^{A|B}$. In practice, $M^{A|B}$ (or $f^{A|B}$) may be easier to obtain than $M_i^{A|B}$ (or $f_i^{A|B}$).

It is advisable to make maximum use of every constraints and information that would help to calculate the values of $M_i^{A|B}$ or $M^{A|B}$. For example, in the context of longitudinal surveys of individuals within households, one can rely on the fact that the household composition is often relatively stable through time. Note that in this context, U^A is the population of individuals at the starting wave, and U^B is the target population of individuals within households at a later wave. Letting the clusters *i* correspond to the households, and assuming that the household composition is relatively stable through time, we can then assume that $M_i^{A|B} \approx M_i^B$. For further details, see Xu and Lavallée (2006).

References

Ardilly, P., Le Blanc, P. (2001) Comment pondérer une enquête auprès des personnes sans domicile?, *Enquêtes, modèles et applications*, Dunod, Paris, 417-436.

Bankier, M. (1983) *Evaluation of the Partnership Correction for TRA's T1 Weights*, Internal note to Statistics Canada, January 13, 1983.

Deville, J.-C., Särndal, C.-E. (1992) Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, Vol. 87, No. 418, June 1992, 376-382.

Ernst, L. (1989) Weighting Issues for Longitudinal Household and Family Estimates, in *Panel Surveys* (Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P., Editors), John Wiley and Sons, New York, 135-159.

Fellegi, I.P., Sunter, A. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, Vol. 64, 1183-1210.

Lavallée, P. (1995) Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method, *Survey Methodology,* Vol. 21, No. 1, 25-32.

Lavallée, P. (2002) *Le sondage indirect, ou la méthode généralisée du partage des poids*, Éditions de l'Université de Bruxelles and Éditions Ellipse.

Lavallée, P., Caron, P. (2001) Estimation Using the Generalised Weight Share Method : The Case of Record Linkage, *Survey Methodology*, Vol. 27, No. 2, 155-170.

Lavallée, P., Deville, J.-C. (2002) Theoretical Foundations of the Generalised Weight Share Method, *Proceedings of the International Conference on Recent Advances in Survey Sampling,* Laboratory for Research in Statistics and Probability (Technical Report No. 386), Ottawa, 127-136.

Sirken, M.G., Nathan, G. (1988) Hybrid Network Sampling, Survey Research Section of the *Proceedings of the American Statistical Association*, 459-461.

Xu, X., Lavallée, P. (2006) *Treatment of Link Nonresponse in Indirect Sampling*, Paper presented at the annual conference of the Statistical Society of Canada, London, Ontario, June 2006.