Survey Methodology, June 1995
Vol. 21, No. 1, pp. 25–32
Statistics Canada

Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method

PIERRE LAVALLÉE[1]

ABSTRACT

Statistical agencies are conducting increasing numbers of longitudinal surveys. Although the main output of these surveys consists of longitudinal data, most of them are also expected to produce reliable cross-sectional estimates. In surveys of individuals and households, population dynamics significantly changes household composition over time. For this reason, methods of cross-sectional estimation must be adapted to the longitudinal aspect of the sample. This paper discusses in a general context the Weight Share method, of which one application is to assign a basic weight to each individual in a household. The variance estimator associated with the Weight Share method is also presented. The weighting of a longitudinal sample is then discussed when a supplementary sample is selected to improve the cross-sectional representativeness of the sample. The paper presents as an application the Survey of Labour and Income Dynamics (SLID) introduced by Statistics Canada in 1994. This longitudinal survey covers individuals' work experience, changes in income and changes in family composition.

KEY WORDS: Weight share method; Longitudinal survey; Cross-sectional estimate; Supplementary sample.

1. INTRODUCTION

Longitudinal surveys, *i.e.* surveys that follow units over time, are steadily gaining importance within statistical agencies. Statistics Canada is currently developing three major longitudinal surveys of individuals: the National Population Health Survey, the National Longitudinal Survey of Children; and the Survey of Labour and Income Dynamics (SLID).

The primary objective of these surveys is to obtain longitudinal data. One of the uses of these data is to study the changes in variables over time (*e.g.*, longitudinal data may be used to analyze the chronic aspect of poverty). A secondary objective is the production of cross-sectional estimates, in other words estimates that represent the population at a given point in time. Although these estimates are far less important than the longitudinal data, to many users they are an essential aspect of the survey. Obtaining a representative cross-sectional view of the current population constitutes a means of measuring changing situations over time. The longitudinal aspect of the survey also improves the accuracy of the measurement of change.

This paper presents an extension of the Weight Share method presented by Ernst (1989). Although the method has been developed in the context of longitudinal household surveys, it is shown that the Weight Share method can be generalized to situations where a population of interest is sampled through the use of a frame which refers to a different population, but linked somehow to the first one. In the context of longitudinal surveys, the frame can be associated to the initial population, while the population of interest can be the population a few years later. The paper also provides a new proof of the unbiasedness of the Weight Share method together with the variance formula and variance estimator to be used with the method.

Using the Weight Share method, the question addressed in this paper is that of ensuring that the longitudinal sample can be used for cross-sectional estimation. The difficulty arises from the fact that, although the longitudinal sample remains constant, distribution of the population (individuals and households) changes over time. At the individual level, these changes are produced by such events as births and deaths, immigration and emigration, and moves within the country. Obviously, the birth or death of an individual also changes household composition; and such events as marriage, divorce, separation, departure of a child and cohabitation, are all factors that affect population distribution within the household. If we are to obtain accurate, unbiased cross-sectional estimates based on a longitudinal sample, we need an estimation method that takes these changes into account.

Our initial topic is the presentation of the Weight Share method in a general context. Secondly, we present the sample design for SLID. This is one of the major longitudinal surveys for which the production of cross-sectional estimates from a longitudinal sample is a significant problem. The survey itself is a typical longitudinal survey of individuals and households. Thirdly, we describe the use of a supplementary sample added to the initial longitudinal sample to improve the cross-sectional representativeness. Fourthly, we present the concept of basic weights, the equivalent, as it were, of sample weights. Finally, we describe the use of the Weight Share method to calculate basic weights for all individuals interviewed in SLID.

[1] Pierre Lavallée, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

2. THE WEIGHT SHARE METHOD IN A GENERAL CONTEXT

The Weight Share method is described in Ernst (1989) in the context of longitudinal household surveys. In the same context, Kalton and Brick (1995) discuss different weighting schemes, including the Weight Share method. Various implications of using the Weight Share method for longitudinal household surveys have been described by Gailly and Lavallée (1993).

We now present this method in a general context that can be applied to several sampling situations where the population of interest needs to be sampled through the use of a frame which refers to a different population, but is linked somehow to the first one. Note that this can be viewed as a form of Network Sampling (see Thompson 1992). For example, one can imagine the need to sample young children where the only available frame is a list of names of parents. The population of interest is really the children but we need to select a sample of parents from the frame in order to obtain the sample of children. Note that the children of a particular family can be sampled through either the father or the mother. Another example is one of business surveys where an incomplete frame of establishments is available. For each selected establishment from the frame, we wish to sample the entire set of establishments belonging to the same enterprise. The missing establishments from the frame are expected to be sampled via the establishments present on the frame.

Suppose that a sample s^A of m^A units is selected from a population U^A of M^A units using some sampling design. Let π_j^A be the selection probability of unit j. We assume $\pi_j^A > 0$ for all $j \in U^A$.

Let U^B be a population of M^B units. This population is divided into N clusters where cluster i contains M_i^B units. For example, in the context of social surveys, the clusters can be households and the units can be the persons within the households. For business surveys, the clusters can be enterprises and the units can be the establishments within the enterprises. From population U^B, we are interested in estimating the total $Y = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik}$ for some characteristic y.

An important constraint that is imposed in the measurement (or interviewing) process is to consider all units within the same cluster. That is, if a unit is selected in the sample, then every unit of the cluster containing the selected unit will be interviewed. This constraint is one which often arises in surveys for two reasons: cost reductions and the need for producing estimates on clusters. Referring back to the example of social surveys, there is normally a small marginal cost for interviewing all persons within the household. On the other hand, household estimates are often of interest with respect to poverty measures, for example.

We assume that there exists a *link* (or a correspondence) between each unit j of population U^A and at least one unit k of population U^B. Also, each cluster i of U^B has at least one link with a unit j of U^A. The link is identified through an indicator variable l_{jk} where $l_{jk} = 1$ if there is a link between unit $j \in U^A$ and unit $k \in U^B$ and 0 otherwise. All units of population U^A have at least one link with population U^B, *i.e.*, $L_j^A = \sum_{k \in U^B} l_{jk} \geq 1$ for all $j \in U^A$. However, there can be zero, one or more links for a unit k of population U^B, *i.e.*, it is possible to have $L_k^B = \sum_{j \in U^A} l_{jk} = 0$ or $L_k^B = \sum_{j \in U^A} l_{jk} > 1$ for some $k \in U^B$. This is illustrated in Figure 1.

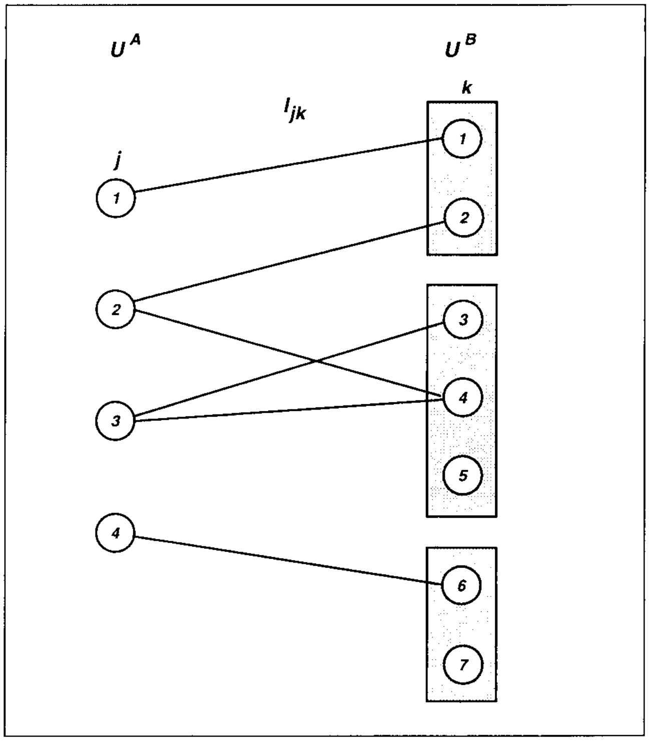

Figure 1. Links between units of populations U^A and U^B.

The estimation process presented now uses the sample s^A together with the links existing between U^A and U^B to obtain an estimation of the total Y belonging to population U^B. The links are in fact utilized as a bridge to go from population U^A to population U^B, and vice versa. Note that in practice, it might not be physically possible to directly select a sample s^B from U^B, as it has been described in the introductory examples.

To estimate the total Y, one can use the estimator

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

where n is the number of interviewed clusters and w_{ik} is the weight attached to unit k of cluster i. To obtain unbiased estimates, a possible set of weights could be the inverse of the selection probabilities of the units entering into the estimator $\hat{Y}$. For each unit k of cluster i having a link $l_{j,ik} = 1$ with a unit j in U^A, this is possible since we have $\pi_k^B = \pi_j^A$. However, not all units of U^B necessarily have a link to U^A. Moreover, even if a link exists, it is not guaranteed that the selection probability π_j^A is known when $j \notin s^A$; the sample design used to select s^A could be, for example, a multistage sample design where the ultimate selection probability of each unit j is only known at the end of the selection process. To assign a nonzero weight w_{ik} to each unit k of cluster i entering into $\hat{Y}$, the Weight Share method can be used.

In general, the Weight Share method allocates to each sampled unit a basic weight established from an average of weights calculated within each cluster i entering into $\hat{Y}$. An *initial weight* that corresponds to the inverse of the selection probability is first obtained for unit k of cluster i of $\hat{Y}$ having a link $l_{j,ik} = 1$ with a unit $j \in s^A$. An initial weight of zero is assigned to units not having a link. The *basic weight* is obtained by calculating the mean of the initial weights for the cluster. This weight is finally assigned to all units within the cluster. Note that the fact of allocating the same basic weight to all units has the considerable advantage of ensuring consistency of estimates for units and clusters.

Formally, each unit k of cluster i entering into $\hat{Y}$ is assigned an initial weight w'_{ik} as follows:

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}, \tag{2}$$

where $t_j = 1$ if $j \in s^A$ and 0 otherwise. Note that a unit k having no link with any unit j of U^A automatically has an initial weight of zero.

The basic weight w_i is given by

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}}, \tag{3}$$

where $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik}$. The quantity L_{ik} represents the number of links between the units of U^A and the unit k of cluster i of population U^B. The quantity $L_i = \sum_{k=1}^{M_i^B} L_{ik}^B$ then corresponds to the total number of links present in cluster i.

Finally, we assign $w_{ik} = w_i$ for all $k \in i$.

2.1 Unbiasedness of the Weight Share Method

We now show that the estimator $\hat{Y}$ with the Weight Share method is unbiased for Y. Starting with $\hat{Y} = \sum_{i=1}^{n} w_i \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^{n} w_i Y_i$, we replace the definition of w_i in $\hat{Y}$ to get

$$\hat{Y} = \sum_{i=1}^{n} Y_i \left[\frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}} \right] = \sum_{i=1}^{n} \frac{Y_i}{L_i} \sum_{k=1}^{M_i^B} w'_{ik}.$$

Letting $z_{ik} = Y_i / L_i$ for all $k \in i$, we then have

$$\hat{Y} = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w'_{ik} z_{ik}. \tag{4}$$

Let a single index k be used to identify the m^B units entering into $\hat{Y}(m^B = \sum_{i=1}^{n} M_i^B)$. By replacing w'_k by its definition (2), we obtain

$$\hat{Y} = \sum_{k=1}^{m^B} w'_k z_k$$

$$= \sum_{k=1}^{m^B} \left[\sum_{j=1}^{M^A} l_{jk} \frac{t_j}{\pi_j^A} \right] z_k.$$

Now since $t_j \neq 0$ only for the units k entering into $\hat{Y}$, we can extend the first summation to all units k in U^B. That is,

$$\hat{Y} = \sum_{k=1}^{M^B} \left[\sum_{j=1}^{M^A} l_{jk} \frac{t_j}{\pi_j^A} \right] z_k.$$

Rearranging $\hat{Y}$, we finally obtain

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{k=1}^{M^B} l_{jk} z_k$$

$$= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j. \tag{5}$$

Now, taking the expectation gives

$$E(\hat{Y}) = \sum_{j=1}^{M^A} \frac{E(t_j)}{\pi_j^A} Z_j$$

$$= \sum_{j=1}^{M^A} Z_j = Z$$

since $E(t_j) = \pi_j^A$.

It suffices now to show that $Z = Y$. First, we have

$$Z = \sum_{j=1}^{M^A} Z_j = \sum_{j=1}^{M^A} \sum_{k=1}^{M^B} l_{jk} z_k = \sum_{k=1}^{M^B} z_k \sum_{j=1}^{M^A} l_{jk}.$$

By rearranging these summations in terms of the N clusters of population U^B, we then obtain

$$Z = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} z_{ik} \sum_{j=1}^{M^A} l_{j,ik} = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} z_{ik} L_{ik}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} \frac{Y_i}{L_i} L_{ik} = \sum_{i=1}^{N} Y_i = Y.$$

The unbiasedness of the Weight Share method can also be proved using an approach similar to the one presented by Ernst (1989).

2.2 Variance Estimation

To obtain a variance formula for $\hat{Y}$, we start with equation (5). Since $\hat{Y}$ turns out to be nothing more than a Horvitz-Thompson estimator of Z (see Horvitz and Thompson 1952), the variance of $\hat{Y}$ is then directly given by

$$\operatorname{Var}(\hat{Y}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'} \qquad (6)$$

where $\pi_{jj'}^A$ is the joint probability of selecting units j and j' (see Särndal, Swensson and Wretman 1992 for the calculation of $\pi_{jj'}^A$ under various sampling designs).

In practice, equation (6) is simple to use. Initially, it suffices to calculate $z_k = Y_i / L_i$ for each unit $k \in i$. Then, we compute $Z_j = \sum_{k=1}^{M^B} l_{jk} z_k$. All that remains is to plug each Z_j into the variance equation of the Horvitz-Thompson estimator.

The variance $\operatorname{Var}(\hat{Y})$ may be unbiasedly estimated from the following equation:

$$\widehat{\operatorname{Var}}(\hat{Y}) = \sum_{j=1}^{m^A} \sum_{j'=1}^{m^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_{jj'}^A \pi_j^A \pi_{j'}^A} Z_j Z_{j'} . \qquad (7)$$

Another unbiased estimator of the variance $\operatorname{Var}(\hat{Y})$ may be developed in the form of Yates and Grundy (1953). Other variance estimators are available in the literature, such as jackknife variance estimators. A jackknife variance estimator in the context of the SLID sample design is discussed in Section 3.2.3. For further details, see Wolter (1985) and Särndal, Swensson and Wretman (1992).

3. APPLICATION TO SLID

In January 1994, SLID was launched by Statistics Canada. Its aim is to observe individual activity in the labour market over time, and changes in individual income and family circumstances. To repeat, the primary aim of SLID is to provide longitudinal data. However, cross-sectional estimates will also be produced. The target population of SLID is all persons, with no distinction as to age, who live in the provinces of Canada. For operational reasons, the Territories, institutions, Indian reserves and military camps are excluded (see Lavallée 1993).

3.1 Sample Design

3.1.1 Initial Sample

The SLID longitudinal sample was drawn in January 1993 from two groups rotating out of the Canadian Labour Force Survey (LFS), making the sample a sub-sample of the LFS. The longitudinal sample for SLID is made up of close to 15,000 households. A household is defined as any person or group of persons living in a dwelling. It may consist of one person living alone, a group of people who are not related but who share the same dwelling, or it may be a family.

LFS is a continuing survey designed to produce monthly estimates of employment, self-employment and unemployment. This survey uses a stratified multi-stage design which uses an area frame in which dwellings are the final sampling units. All the individuals who are members of households that occupy the selected dwellings make up the LFS sample. In other words, LFS draws a sample of dwellings and all individuals in the households that live in the selected dwellings are interviewed. A six-group rotation plan is used to construct the sample: every month, one group that has been in the sample for six months is rotated out. Each rotation group contains approximately 10,000 households, or approximately 20,000 individuals 16 years old or more. For further details on the LFS sample plan, see Singh *et al.* (1990).

For SLID, the longitudinal sample will not be updated following its selection in January 1993. However, to give the sample some cross-sectional representativeness, *initially-absent individuals* in the population (*i.e.*, individuals who were not part of the population in the year the longitudinal sample was selected) will need to be considered in the sample in January 1994 and later. Initially-absent individuals include *newborns* (births since January 1993) and *in-migrants*. Note that this addition to the sample will be cross-sectional in that only the longitudinal individuals will be permanently included in the sample.

Table 1 presents the terminology developed for SLID. After sample selection in January 93 (year 1), the population contains longitudinal individuals and initially-present individuals. In January 94 (year 2), the population contains longitudinal individuals, initially-present individuals and initially-absent individuals. Focusing on the households containing at least one longitudinal individual (*i.e.*, *longitudinal households*), initially-present and initially-absent individuals who join these households are referred to as *cohabitants*.

Table 1

SLID Terminology

Individuals:

Longitudinal individuals: Individuals selected at year 1 in the longitudinal sample.

Initially-absent individuals: Individuals who were not part of the population in the year the longitudinal sample was selected (year 1). It includes in-migrants and newborns.

Initially-present individuals: Individuals who were part of the population of year 1 but were not selected then.

Cohabitants: Initially-absent and initially-present individuals who join a longitudinal household.

In-migrants: Individuals who, in January of year 1, were outside the ten provinces of Canada and individuals in excluded areas (the Territories, institutions, Indian reserves and military barracks).

Newborns: Births since January of year 1.

Households:

Longitudinal households: Households containing at least one longitudinal individual.

SLID will follow individual and household characteristics over time. At the time of each wave of interviews, all the members of a longitudinal household will be interviewed. The composition of the longitudinal households will change over time, as the result of a birth or the arrival of an in-migrant in the household. A part of the selection of initially-absent individuals may be based on individuals who join longitudinal households.

3.1.2 Supplementary Sample

The restriction to initially-absent individuals who join longitudinal households will unfortunately exclude households made up of initially-absent individuals only (*e.g.*, in-migrant families). To offset this shortcoming, one possibility is to draw a *Supplementary Sample*. This sample could be one of dwellings drawn directly from the ongoing LFS at each wave of interviews. Supplementary questions would then be added to the LFS questionnaire to detect households that contain *at least one in-migrant*; the households selected would then be interviewed.

Recalling that the Supplementary Sample is used for the selection of households made up solely of initially-absent individuals (*i.e.*, in-migrants and newborns), restricting this sample to in-migrants only would not cause any representativeness problem. This is because it is highly unlikely that households containing only newborns would be found: each household normally contains at least one adult. The newborns are then already represented in the sample by the longitudinal households. Now, if the Supplementary Sample were to include newborns in addition to in-migrants, significant costs would be added to the survey. This is because the Supplementary Sample would include a complete household for each newborn selected in the Supplementary Sample, producing excessive sample growth and unnecessary costs since the newborns are already represented in the sample.

One other approach different from using the ongoing LFS could be to select the Supplementary Sample by revisiting the dwellings used for the selection of the initial sample. This method offers some practical advantages, one being the facility to go to known addresses. This approach however would bring the problem of new dwellings which were not there in January 1993. These dwellings would have a zero probability of being selected in the Supplementary Sample and a bias would therefore be introduced. This is one reason favouring the first approach, *i.e.*, detecting households that contain at least one in-migrant via the questionnaire of the ongoing LFS.

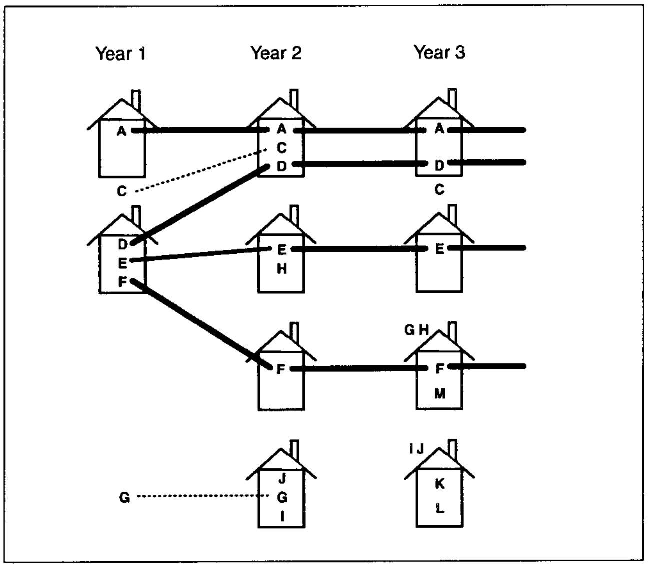

Figure 2. Selection of persons for SLID.

Figure 2 summarizes the longitudinal and cross-sectional selection of individuals. In Figure 2, the letters and houses represent individuals and households, respectively. Individuals A, D, E and F are longitudinal individuals whom we follow over time. Individual C is an initially-present individual, *i.e.*, an individual who was included in the population in year 1 but was not selected then. Initially-absent and initially-present individuals who join a longitudinal household are called cohabitants. In year 2, individual H represents an initially-absent individual who joins the sample as a cohabitant. The fourth house in year 2 represents a household selected for the Supplementary Sample of year 2 and in which individuals I and J are initially-absent individuals (with one of the two being necessarily an in-migrant since the Supplementary Sample is restricted to them). Individual G is an initially-present individual with the same status as C. In year 3, individuals C and H have left their longitudinal households and will therefore not be interviewed. Individuals I and J who were selected in the Supplementary Sample are now replaced with the individuals of the Supplementary Sample of year 3, *i.e.*, individuals K and L. Individual M is an initially-absent individual joining a longitudinal household as a cohabitant. It may finally be noted that, for cross-sectional purposes, a selected household may contain one or more longitudinal individuals, initially-present individuals and initially-absent individuals (newborns and in-migrants).

3.2 Basic Weighting

3.2.1 General Considerations

To produce cross-sectional estimates, the longitudinal sample augmented with initially-absent individuals and initially-present individuals must be weighted. The first step is to obtain a *basic weight* for each individual in each interviewed household. The basic weight is the weight prior to adjustment or post-stratification. It is, so to speak, the equivalent of the sample weight. Note that the basic weights are useful solely for cross-sectional estimation.

The basic weights are obtained from the selection probabilities. As described above, in January 1993 (year 1), we select for SLID a sample $s^{(1)}$ of $m^{(1)}$ individuals from a population $U^{(1)}$ of $M^{(1)}$ individuals. The sample is selected through dwellings which contain households. In other words, the $m^{(1)}$ individuals are obtained by selecting $n^{(1)}$ households from $N^{(1)}$, each household I being selected with probability $\pi_I^{(1)} > 0, I = 1, \ldots, N^{(1)}$. Let $M_I^{(1)}$ be the size of household I so that $M^{(1)} = \sum_{I=1}^{N^{(1)}} M_I^{(1)}$. Also let $\pi_j^{(1)}$ be the selection probability of individual j. This selection probability is retained throughout all waves of the survey.

For a given subsequent wave (which may be defined as year 2), the population U contains the $M^{(1)}$ individuals present at year 1, plus some $M^{(2)}$ initially-absent individuals (*i.e.*, initially absent from the population at year 1). The population of initially-absent individuals is indicated by $U^{(2)}$. Hence, the population $U = U^{(1)} \cup U^{(2)}$ contains $M = M^{(1)} + M^{(2)}$ individuals. Letting $U^{*(2)}$ be the population of $M^{*(2)}$ in-migrants of year 2, we have $U^{*(2)} \subseteq U^{(2)}$ and also $M^{*(2)} \leq M^{(2)}$. In our notation, the asterisk (*) is used to specify that the newborns have been excluded. The individuals of year 2 are contained in N households where household i is of size M_i, $i = 1, \ldots, N$.

For cross-sectional representativeness, some in-migrants are selected from the Supplementary Sample. At year 2, we then select a sample $s^{*(2)}$ of $m^{*(2)}$ individuals from the population $U^{*(2)}$ of $M^{*(2)}$ in-migrants. The Supplementary Sample is selected through households, *i.e.*, the $m^{*(2)}$ individuals are obtained by selecting $n^{*(2)}$ households. Let $\pi_j^{*(2)}$ be the selection probability of the in-migrant j. We assume $\pi_j^{*(2)} > 0$ for $j = 1, \ldots, M^{*(2)}$.

One implication of selecting in-migrants through households is that other individuals (such as newborns, initially-present individuals or longitudinal individuals) can be brought in by the Supplementary Sample by living in the same household as the selected in-migrants. However, since the selection units of the Supplementary Sample are restricted to the in-migrants, these other individuals are not properly selected, say, in the Supplementary Sample, even if they will be interviewed. The selection probabilities of these individuals are in fact not well defined.

The remaining in-migrants selected for cross-sectional representativeness are those individuals who join longitudinal households, who are then considered as cohabitants. As with the newborns and initially-present individuals of the previous paragraph, the addition of cohabitants to longitudinal households brings individuals with non-well defined selection probabilities.

The individuals with non-well defined selection probabilities have entered the survey process in a ''non-legitimate'' way. They complicate the determination of the basic weights, as their selection probability is not well defined. In order to override this difficulty, the Weight Share method is proposed.

3.2.2 Basic Weight Calculation

The Weight Share method described in Section 2 is now applied to the SLID sample, including the Supplementary Sample. The population U^A is here represented by the union of the two distinct populations $U^{(1)}$ and $U^{*(2)}$, *i.e.*, $U^A = U^* = U^{(1)} + U^{*(2)}$. The sample s^A of $m = m^{(1)} + m^{*(2)}$ individuals corresponds to the union of the two distinct samples $s^{(1)}$ and $s^{*(2)}$. The population U^B is represented by $U = U^{(1)} + U^{(2)}$. The population $U^A = U^*$ excludes the newborns while the population $U^B = U$ includes them. The clusters of population U^B simply correspond to the N households of year 2, and hence $M_i^B = M_i$.

One possible linkage between population U^A and U^B can be established by the same individuals in populations U^A and U^B. That is, $l_{jk} = 1$ if individual j in population U^A corresponds to individual k in population U^B, and $l_{jk} = 0$ otherwise. For each individual k not being a newborn, we then have $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik} = 1$. On the other hand, for each newborn k, we have $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik} = 0$ since they are excluded from U^A. We now have $L_i = \sum_{k=1}^{M_i^B} L_{ik} = M_i^*$ where M_i^* is the size of household i excluding the newborns.

Note that this last linkage is only one among several other possibilities. One other possible linkage would be to extend the linkage of the previous paragraph to all other persons within the household. That is, assign $l_{jk} = 1$ for all individuals k (of U^B) belonging to the same household i where individual j (of U^A) now belongs in U^B, and 0 otherwise. In other words, $l_{jk} = 1$ if individuals j and k belongs to household i. For each individual k in household i, we then have $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik} = M_i^*$. We also get $L_i = \sum_{k=1}^{M_i^B} L_{ik} = \sum_{k=1}^{M_i^B} M_i^* = M_i^B M_i^*$. One can show that this linkage produces the same basic weighting as the one from the previous paragraph. Because the first linkage corresponds to a more natural way to link the individuals (*i.e.*, by linking only the same individuals between U^A and U^B), we will adopt the linkage proposed in the previous paragraph.

By considering the definition (2) of the initial weight w'_{ik} of individual k in household i, we obtain

$$w'_{ik} = \frac{t_{ik}^{(1)}}{\pi_{ik}^{(1)}} + \frac{t_{ik}^{*(2)}}{\pi_{ik}^{*(2)}}, \tag{8}$$

where $t_{ik}^{(1)} = 1$ if individual k is part of $s^{(1)}$ and 0 otherwise, $t_{ik}^{*(2)} = 1$ if individual k is part of $s^{*(2)}$ and 0 otherwise. This can be written more explicitly as

$$w'_{ik} = \begin{cases} 1/\pi_{ik}^{(1)} & \text{for} \quad k \in s^{(1)} \\ 1/\pi_{ik}^{*(2)} & \text{for} \quad k \in s^{*(2)} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Note that the first line of (9) corresponds to the longitudinal individuals. The second line corresponds to the in-migrants selected through the Supplementary Sample. The third line represents altogether newborns, cohabitants (if the household is a longitudinal household not part of the Supplementary Sample) and/or initially-present individuals (if the household is part of the Supplementary Sample).

The basic weight w_i of household i is obtained from

$$w_i = \frac{\sum_{k=1}^{M_i} w'_{ik}}{\sum_{k=1}^{M_i} L_{ik}} = \frac{1}{M_i^*} \sum_{k=1}^{M_i} w'_{ik}, \tag{10}$$

and finally $w_{ik} = w_i$ for $k \in i$.

Using the basic weights obtained from the Weight Share method, one can estimate the total $Y = \sum_{i=1}^{N} \sum_{k=1}^{M_i} y_{ik}$ of the characteristic y measured at year 2. The estimator used is the one given by equation (1). Using the definitions of the initial weights and the basic weights, $\hat{Y}$ can be rewritten as

$$\hat{Y} = \sum_{k=1}^{m^{(1)}} \frac{z_k^*}{\pi_k^{(1)}} + \sum_{k=1}^{m^{*(2)}} \frac{z_k^*}{\pi_k^{*(2)}}$$

$$= \hat{Z}^{*(1)} + \hat{Z}^{*(2)}, \tag{11}$$

where $z_k^* = \bar{Y}_i^*$ for $k \in i$ with $\bar{Y}_i^* = \sum_{k=1}^{M_i} y_{ik}/M_i^*$. Thus, estimator (11) is simply the sum of two Horvitz-Thompson estimators related to $s^{(1)}$ and $s^{*(2)}$. As shown in Section 2, this estimator is unbiased for Y.

3.2.3 Variance Estimation

The variance formula for $\hat{Y}$ is provided by equation (6). However, assuming that the two samples $s^{(1)}$ and $s^{*(2)}$ are selected independently, we have $\operatorname{Var}(\hat{Y}) = \operatorname{Var}(\hat{Z}^{*(1)}) + \operatorname{Var}(\hat{Z}^{*(2)})$, where each term has the form of equation (6). For SLID, this assumption of independance holds if the selection of the Supplementary Sample is done through LFS.

Considering $\hat{Z}^{*(1)}$, we can re-index the individuals to reflect the fact that the $m^{(1)}$ individuals were selected at year 1 through $n^{(1)}$ households. This gives

$$\hat{Z}^{*(1)} = \sum_{k=1}^{m^{(1)}} \frac{z_k^*}{\pi_k^{(1)}} = \sum_{I=1}^{n^{(1)}} \sum_{j=1}^{M_I^{(1)}} \frac{z_{Ij}^*}{\pi_{Ij}^{(1)}}$$

$$= \sum_{I=1}^{n^{(1)}} \frac{1}{\pi_I^{(1)}} \sum_{j=1}^{M_I^{(1)}} z_{Ij}^* = \sum_{I=1}^{n^{(1)}} \frac{Z_I^{*(1)}}{\pi_I^{(1)}}, \tag{12}$$

since, by selecting complete households $\pi_{Ij}^{(1)} = \pi_I^{(1)}$ for $j \in I$. The variance $\operatorname{Var}(\hat{Z}^{*(1)})$ is then directly obtained as

$$\operatorname{Var}(\hat{Z}^{*(1)}) = \sum_{I=1}^{N^{(1)}} \sum_{I'=1}^{N^{(1)}} \frac{(\pi_{II'}^{(1)} - \pi_I^{(1)} \pi_{I'}^{(1)})}{\pi_I^{(1)} \pi_{I'}^{(1)}} Z_I^{*(1)} Z_{I'}^{*(1)}. \tag{13}$$

Considering $\hat{Z}^{*(2)}$, the individuals can also be re-indexed for consistency with $\hat{Z}^{*(1)}$, although this modification has no effect on the form of $\hat{Z}^{*(2)}$. Following the same steps used for $\operatorname{Var}(\hat{Z}^{*(1)})$, $\operatorname{Var}(\hat{Z}^{*(2)})$ is obtained as

$$\operatorname{Var}(\hat{Z}^{*(2)}) = \sum_{I=1}^{N^{*(2)}} \sum_{I'=1}^{N^{*(2)}} \frac{(\pi_{II'}^{*(2)} - \pi_I^{*(2)} \pi_{I'}^{*(2)})}{\pi_I^{*(2)} \pi_{I'}^{*(2)}} Z_I^{*(2)} Z_{I'}^{*(2)}, \tag{14}$$

where $N^{*(2)}$ is the number of households of year 2 containing at least one in-migrant and $Z_I^{*(2)} = \sum_{j=1}^{M_I^{*(2)}} z_{Ij}^*$. The quantity $M_I^{*(2)}$ represents the number of in-migrants present in household I.

Finally, $\operatorname{Var}(\hat{Y})$ is simply given by

$$\operatorname{Var}(\hat{Y}) = \sum_{I=1}^{N^{(1)}} \sum_{I'=1}^{N^{(1)}} \frac{(\pi_{II'}^{(1)} - \pi_I^{(1)}\pi_{I'}^{(1)})}{\pi_I^{(1)}\pi_{I'}^{(1)}} Z_I^{*\,(1)} Z_{I'}^{*\,(1)}$$

$$+ \sum_{I=1}^{N^{*\,(2)}} \sum_{I'=1}^{N^{*\,(2)}} \frac{(\pi_{II'}^{*\,(2)} - \pi_I^{*\,(2)}\pi_{I'}^{*\,(2)})}{\pi_I^{*\,(2)}\pi_{I'}^{*\,(2)}} Z_I^{*\,(2)} Z_{I'}^{*\,(2)}. \tag{15}$$

The variance (15) may be unbiasedly estimated using the following equation:

$$\widehat{\operatorname{Var}}(\hat{Y}^*) = \sum_{I=1}^{n^{(1)}} \sum_{I'=1}^{n^{(1)}} \frac{(\pi_{II'}^{(1)} - \pi_I^{(1)}\pi_{I'}^{(1)})}{\pi_{II'}^{(1)}\pi_I^{(1)}\pi_{I'}^{(1)}} Z_I^{*\,(1)} Z_{I'}^{*\,(1)}$$

$$+ \sum_{I=1}^{n^{*\,(2)}} \sum_{I'=1}^{n^{*\,(2)}} \frac{(\pi_{II'}^{*\,(2)} - \pi_I^{*\,(2)}\pi^{*\,(2)})}{\pi_{II'}^{*\,(2)}\pi_I^{*\,(2)}\pi_{I'}^{*\,(2)}} Z_I^{*\,(2)} Z_{I'}^{*\,(2)}. \tag{16}$$

As SLID is in fact a sub-sample from LFS, the jackknife variance estimator developed for LFS (see Singh *et al.* 1990) may also be used, with minor modifications. In general, the jackknife method works as follows: the sample first is divided into random groups (or replicates, according to the LFS terminology). Then, each random group r is removed in turn from the sample and a new estimate $\hat{Y}_{(r)}$ of the total Y is computed. The different estimates $\hat{Y}_{(r)}$ are finally compared to the original estimate $\hat{Y}$ to obtain an estimate of the variance $\operatorname{Var}(\hat{Y})$. For further details on the jackknife method in general, see Särndal, Swensson and Wretman (1992).

Recall that LFS is based on a stratified multi-stage design which uses an area frame. Within each first-stage stratum h, the random groups (or replicates) correspond basically to the primary sampling units (PSUs). To compute the jackknife variance estimate for the estimation of the total Y, the following formula can be used:

$$\widehat{\operatorname{Var}}_J(\hat{Y}) = \sum_h \frac{(R_h - 1)}{R_h} \sum_{r \in h} (\hat{Y}_{(hr)} - \hat{Y})^2, \tag{17}$$

where R_h is the number of replicates in stratum h and $\hat{Y}_{(hr)}$ is the estimate of Y obtained after replicate r in stratum h is removed. For LFS, both $\hat{Y}$ and $\hat{Y}_{(hr)}$ are poststratified based on the integrated approach of Lemaître and Dufour (1987). The plan is to use the same poststratification approach for SLID but this discussion is out of the scope of the present paper.

ACKNOWLEDGEMENTS

The author would like to thank Carl Särndal, M.P. Singh, the Associate Editor and the referees for very useful comments which helped in improving the clarity of the paper. We would like also to gratefully acknowledge Jean-Claude Deville for suggesting the idea of an extension of the Weight Share method to a general context.

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 and Sons, 135-159.

GAILLY, B., and LAVALLÉE, P. (1993). Insérer des nouveaux membres dans un panel longitudinal de ménages et d'individus: simulations. CEPS/Instead, Document PSELL No. 54, Luxembourg, mai 1993.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

KALTON, G., and BRICK, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.

LAVALLÉE, P. (1993). Sample representativity for the Survey of Labour and Income Dynamics. Statistics Canada, Research Paper of the Survey of Labour and Income Dynamics, Catalogue No. 93-19, December 1993.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of The Canadian Labour Force Survey*. Statistics Canada, Catalogue No. 71-526.

THOMPSON, S.K. (1992). *Sampling*. New York: John Wiley and Sons.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15, 235-261.