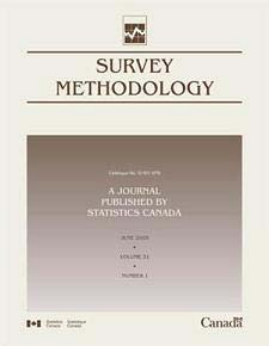
Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Weighting in rotating samples: The SILC survey in France

by Pascal Ardilly and Pierre Lavallée



December, 2007

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2007
Vol. 33, No. 2, pp. 131-137
Statistics Canada, Catalogue no. 12-001-X

Weighting in rotating samples: The SILC survey in France

Pascal Ardilly and Pierre Lavallée [1]

Abstract

The European Union's Statistics on Income and Living Conditions (SILC) survey was introduced in 2004 as a replacement for the European Panel. It produces annual statistics on income distribution, poverty and social exclusion. First conducted in France in May 2004, it is a longitudinal survey of all individuals over the age of 15 in 16,000 dwellings selected from the master sample and the new-housing sample frame. All respondents are tracked over time, even when they move to a different dwelling. The survey also has to produce cross-sectional estimates of good quality.

To limit the response burden, the sample design recommended by Eurostat is a rotation scheme consisting of four panels that remain in the sample for four years, with one panel replaced each year. France, however, decided to increase the panel duration to nine years. The rotating sample design meets the survey's longitudinal and cross-sectional requirements, but it presents some weighting challenges.

Following a review of the inference context of a longitudinal survey, the paper discusses the longitudinal and cross-sectional weighting, which are designed to produce approximately unbiased estimators.

Key Words: Longitudinal survey; Panel; Weight share method; Longitudinal weighting; Cross-sectional weighting.

1. Introduction

Statistics on Income and Living Conditions (SILC) is a European survey that produces data on the income and living conditions of persons living in regular households (persons living in communal households are excluded). It was introduced in 2004 as a replacement for the European Panel. While it is a European Union (EU) survey and therefore under Eurostat responsibility, it is conducted independently in each EU member state. Hence, the member states - France in this case - are free to adjust the sample design suggested by Eurostat to meet their national requirements. The data are also processed by the individual member states, as is usually the case for Eurostat surveys in the EU. This article deals only with the SILC survey conducted in France, but it may also be of interest to other EU member states.

SILC is a *longitudinal survey* conducted once a year in May. It focuses on individuals rather than households, and data are collected through personal interviews with every person in the sampled dwellings. SILC can be thought of as the European version of the Statistics Canada's Survey of Labour and Income Dynamics (SLID) (see Lavallée 1995, and Lévesque and Franklin 2000).

The SILC sample is rotating: each year, it is formed by combining nine panel subsamples selected under identical steady-state conditions, partly from the master sample and partly from the new-housing sample frame. The master sample and the new-housing sample frame are two dwelling frames constructed from the French census of population and the information and automated data processing system for dwelling and office space (SITADEL) respectively (see Ardilly 2006).

Each incoming panel includes all individuals living in the selected dwellings. Surveying all members of the households living in the selected dwellings makes it possible to produce both individual-level and household-level estimates and helps keep collection costs down by maximizing the number of individuals reached in each contact. On the other hand, some of the estimates are narrower in scope, applying only to the population aged 16 and over on December 31 of the survey year.

Each year, one subsample is rotated out and replaced with another subsample. In the survey's starting year, 2004, each subsample consisted of 1,780 dwellings (give or take a few units because of rounding). In the second and subsequent years (*i.e.*, from 2005 on), the size of the year's incoming subsample was 3,000 dwellings. Note that at the outset in 2004, the sample was 16,000 dwellings, divided into nine equal parts. One of those parts was surveyed only once (in 2004), another twice (2004 and 2005), a third three times (2004, 2005 and 2006), and so on. After the start-up phase, a given panel will be surveyed for nine consecutive years. During the start-up phase, which will end in 2012 with the departure of the ninth and last subsample from the 2004 selection, the subsamples will have been surveyed fewer than nine times.

The sampling procedure itself is the standard method of selecting units from the master sample and the new-dwelling sample frame (see Ardilly 2006). In this case, no

1. Pascal Ardilly, Sampling and statistical data division, INSEE, Paris, France. E-mail: pascal.ardilly@atih.sante.fr; Pierre Lavallée, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 Canada. E-mail: pierre.lavallee@statcan.ca.

category of individuals is overrepresented. The survey has a uniform sampling fraction - ignoring rounding - except for vacant rural dwellings and dwellings that were secondary residences in the 1999 census and became principal residences by survey date, which are traditionally under-represented.

Under the collection process, each subsample is considered a true panel of individuals. Panel members who move are tracked, and their files are sent to the appropriate regional branch of INSEE. More details on SILC's sample design are available in the November 17, 2003, issue of the *Official Journal of the European Union* and internal INSEE documents describing sampling practices in France.

Since SILC is a longitudinal survey with panels that overlap in time, weighting the sample presents a special problem. This paper provides a detailed picture of the two types of weighting used for SILC. We will begin by discussing some general principles related to SILC's sample design. Then we will examine longitudinal weighting, followed by cross-sectional weighting.

Note that we will not consider the topics of non-response correction and estimate adjustment. Those issues are dealt with in the same way as they are generally for any other longitudinal survey, such as the SLID (see Lavallée 1995, and Lévesque and Franklin 2000).

2. General principles

2.1 Two approaches: The longitudinal view and the cross-sectional view

Each year, we have a sample of fully panelized individuals, eight ninths of whom were interviewed at least once in previous years (barring non-response).

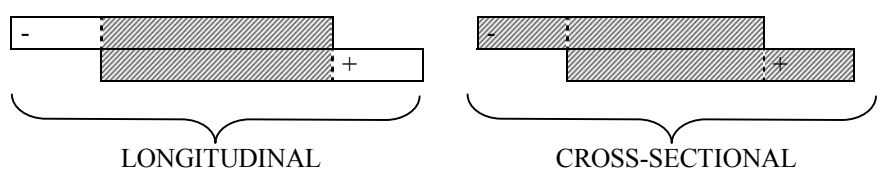
Two types of parameters may be of interest: annual totals Y_t (or their satellites), and changes in totals $\Delta_{t,t+1}$ between two years, consecutive or otherwise. For simplicity, we will confine ourselves to differences in totals between two consecutive years. When discussing changes, we have to be clear about the inference populations involved. We can look at the data in two different ways: either as populations that change over time - the cross-sectional approach - or as a fixed population - the longitudinal approach. If we let Ω_t be the entire in-scope population in year *t*, the annual total for year *t* is given by $Y_t = \sum_{i\in\Omega_t} Y_i^t$, where Y_i^t is a variable of interest measured for individual *i*. When we look at change, we may want to estimate the difference $\Delta^*_{t,t+1}$ between the total Y_{t+1} at $t+1$ over Ω_{t+1} and the total Y_t at *t* over Ω_t, that is, $\Delta^*_{t,t+1} = Y_{t+1} - Y_t$. This is a cross-sectional view. Alternatively, we may want to estimate the difference $\Delta_{t,t+1}$ between the totals for the units that are common to populations Ω_{t+1} and Ω_t, where the size difference between the two populations is due to their incoming units (births) and outgoing units (deaths). This is a longitudinal view. Let $\Omega_{t,t+1} = \Omega_t \cap \Omega_{t+1}$, the population that is common to *t* and $t+1$. Then $\Delta_{t,t+1}$ is defined as $\Delta_{t,t+1} = \sum_{i\in\Omega_{t,t+1}} (Y_i^{t+1} - Y_i^t)$.

The two approaches are illustrated in the diagrams below. The upper rectangle represents the entire population at time *t*, and the lower one represents the entire population at time $t+1$. The "minus" side represents deaths in the broad sense (persons who have died, emigrated, moved to a communal household, and so on), and the "plus" side represents births in the broad sense (newborns, new immigrants, persons who have become part of the survey population by passing an age threshold, and so on). The grey portion represents the inference population on each date.

2.2 Surveys repeated over time and potential strategies

The goal, of course, is to produce both longitudinal estimates and cross-sectional estimates. There are essentially three possible strategies:

1. An "independent" sampling each year. In fact, because we have a master sample and a new-housing sample frame, the panels are selected from the same localities each year, and as a result, the subsamples are not truly independent. This solution can be improved for estimating changes.
2. A fully panelized sampling, *i.e.*, initial selection of a sample that is surveyed each year. This scenario presents a response burden problem, since the SILC survey is to continue indefinitely. It is therefore unrealistic.
3. A rotational sample. This is the scenario that was chosen, because of its advantages in satisfying both longitudinal and cross-sectional goals.



The table below characterizes the three potential sample designs in terms of the two desired approaches.

Sample TYPE	CROSS-SECTIONAL approach	LONGITUDINAL approach
"Independent" each year	CUSTOMARY	POSSIBLE but less efficient
Panel	IMPOSSIBLE without a top-up sample	CUSTOMARY
Rotational	POSSIBLE	POSSIBLE

The rotation strategy has four major advantages:

i. It reduces the sampling error associated with measuring change (in principle, as do panels, though it is theoretically less efficient than a "pure" panel).

ii. It has a smaller response burden than a "pure" panel. Under the circumstances, since France has a nine-year panel, this argument must be used with restraint. It is more persuasive in the scenario recommended by Eurostat, which consists of an annual survey for four consecutive years.

iii. It takes into account very "naturally" how the population changes over time. This point will become clearer when we look at the coverage of new populations.

iv. It reduces observation errors (as do panels).

On the other hand, the strategy also has at least three weaknesses:

i. Participants have to be tracked over time, which results in tracing costs and non-response due to moves.

ii. The length of the individual series is limited to nine years, which is substantial, though not as informative as a pure panel.

iii. The longitudinal/cross-sectional weighting method is not straightforward.

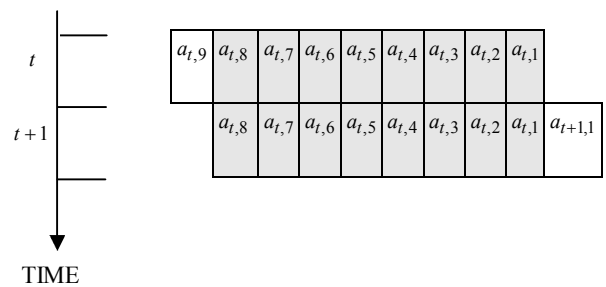
3. Longitudinal weighting

This type of weighting is inherently somewhat easier to understand than cross-sectional weighting because there is no need to take account of how the population changes over time (except for "deaths", units that leave the survey population over time, but they are not much of a technical problem). The idea behind longitudinal estimation, of course, is to make an inference based on a single population at an initial date.

Clearly, the rotational nature of the sample design is what makes weighting difficult, since between two consecutive years t and $t + 1$, we have to deal with eight different panels, each selected from a different population (a population made up of individuals who are, naturally, different from year to year). If we were dealing with just one panel, we would only need to use the sampling weights associated with the panel members who were still in-scope on date t, since those weights are calculated once and for all at the time of selection and can be used to make inferences about the initial population each year throughout the panel's life.

The essential difficulty is to represent population Ω_t on date t using eight panel subsamples selected on different dates and therefore from different populations. Intuitively, it makes sense that a given individual would have a probability of selection on date t that would depend on the number of panel subsamples for which he or she could be chosen. For this discussion, it is assumed that there is no non-response. This situation can be expressed formally by letting $a_{t,k} =$ a panel subsample to be surveyed in year t for the kth time, and $s_{t,t+1} = \bigcup_{k=1}^{8} a_{t,k}$.

Note that we can write $a_{t+1,k+1} = a_{t,k}$ $(\forall t, \forall k \neq 9)$ since we are obliged to use each (non-outgoing) panel subsample in its entirety year after year. This is pictured below.



The grey part represents $s_{t,t+1}$, which is the sample used in this longitudinal approach. It is from the individuals in $s_{t,t+1}$ that we obtain both Y_i^t and Y_i^{t+1}, *i.e.*, information about individual i on dates t and $t + 1$ respectively.

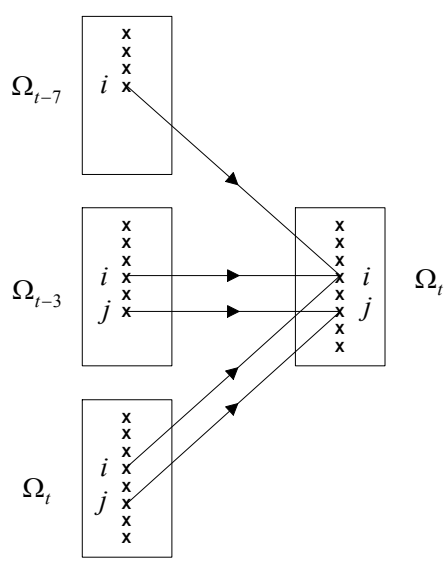
Suppose we have an individual i in Ω_t who is in-scope on date t. We denote as L_i the number of years in $\{t-7, t-6, ..., t-1, t\}$ during which individual i was in-scope and therefore had a chance of being selected as a member of an incoming panel. It is assumed here that each year, the sample frame covers the survey population exactly. We have $L_i \in \{1, 2, 3, \ldots, 8\}$. In addition, we denote as K_i the set of k indexes out of 1, 2, 3, …, 8 for which $i \in a_{t,k}$. These are the numbers of the panels of which individual i is a member on date t. For all i in $s_{t,t+1}$, K_i will be construed as a set containing at least one element. Most of the time, K_i will in fact have only one index, but in some cases, it will have two or even more indexes. That will be the case if i is selected for a panel, he/she moves and his/her new dwelling is chosen for another panel in a subsequent year. Note that our scenario excludes the possibility of selecting a

given dwelling twice, since dwellings from the master sample and the new-housing sample frame are not supposed to be surveyed more than once. This is just a practical convention, however, as the theory can easily accommodate a system in which dwellings can be selected multiple times.

If $i \in a_{t,k}$, let $W_i(t,k)$ be his/her "raw" sampling weight. In fact, it is the sampling weight of the dwelling in which *i* was living at the time he/she was chosen as a panel member, *i.e.*, at the time of the annual selection from Ω_{t-k+1}. This weighting system allows direct inference from subsample $a_{t,k}$ to the entire population Ω_{t-k+1}. In particular, $\sum_{i \in a_{t,k}} W_i(t,k)$ provides an unbiased estimate of the total number of in-scope individuals who are members of population Ω_{t-k+1}. For SILC in France, that total is roughly 60 million. The longitudinal weight assigned to each individual *i* in $s_{t,t+1}$ will therefore be as follows:

$$W_i^{t,t+1} = \frac{1}{L_i} \sum_{k \in K_i} W_i(t,k). \qquad (1)$$

This equation is derived from the application of the weight share method (see Lavallée 1995, and Lavallée 2002) in which the initial population (the population of sampling units) is defined as the union of the populations $\Omega_{t-7}, ..., \Omega_{t-1}, \Omega_t$ and the final population (the population of observation units) as Ω_t. This is illustrated in the diagram below; for greater clarity, only three of the initial subpopulations are shown. Clearly, the number of links is equal to L_i (in this case, *i* has exactly eight links, while *j* must have fewer than eight because it does not appear in the oldest sample frames). In practice, it is realistic to proceed as if $\Omega_{t-7} \subset \Omega_{t-6} \subset \ldots \subset \Omega_{t-1} \subset \Omega_t$. We can work with nested populations, since all individuals who leave the survey population in the time before *t* will not be part of $s_{t,t+1}$.



Equation (1) is the most general formula for the "raw" longitudinal weight. It can then be simplified for specific situations. For example, if we ignore the cases in which a panel member can be selected more than once, we have

$$W_i^{t,t+1} = \frac{W_i}{L_i} \qquad (2)$$

where W_i is the weight of *i* relative to the one panel subsample of which he/she was a member on date *t*. In France's case, because of the sample sizes involved, it seems quite appropriate to use that equation. If we assume that we are in the ideal position - though that seems simplistic in our circumstances - of having a population that does not change over time, we will have $L_i = 8$ for all *i*. The population changes a great deal in nine years, but with shorter panel lives, the ideal case might be an acceptable approximation. Moreover, if the panels are selected with equal probabilities, W_i will be equal to a constant *W* and we will have

$$W_i^{t,t+1} = \frac{W}{8}. \qquad (3)$$

Such a scenario is highly improbable in France's case. First, up to 2012, the sample will contain subsamples with very different raw weights. Second, the sampling process is likely to focus on generating a predetermined number of dwellings (as the total number of dwellings increases), and not a constant sampling fraction.

Note that equation (3) is intuitive. Ultimately, everything proceeds "as if" any individual in the longitudinal sample $s_{t,t+1}$ had a selection probability eight times the selection probability of each panel subsample $s_{t,t+1}$ that is part of.

The foregoing applies to the survey in its steady state and must be adapted slightly during the start-up phase, *i.e.*, until 2012. The first longitudinal operation is performed on the combined 2004-2005 data, to estimate the changes between 2004 and 2005 with the 2004 reference population (from which the "deaths" are removed in 2005). In this case, we need only to divide all the weights W_i of the eight subsamples $a_{2004,1}$ to $a_{2004,8}$ by 8; in other words, $L_i = 8$ for all *i*. In 2006, when we look at the 2005-2006 changes, the denominator L_i may take only two values. In the first scenario, panel member *i* was in the sample frame used in 2004 (and hence could have been selected in 2004) and so $L_i = 8$. This is due to the fact that everything proceeds as if, in 2004, the seven selection processes for panels $a_{2005,2}$ to $a_{2005,8}$ had been carried out under exactly the same conditions. In the second scenario, individual *i* was not in the 2004 sample frame - but is in the 2005 frame and is necessarily in $a_{2005,1}$ - and $L_i = 1$. For the 2006-2007 changes, L_i can be equal to 1, 2 or 8, and so on. We will not

have the set of all possible values of L_i in {1, 2, 3, …, 8} until we measure the 2011-2012 changes.

Once we reach this point in the longitudinal weighting process, we can calculate the longitudinal weights $W_i^{t,t+1}$ and then derive the estimator of the difference $\Delta_{t,t+1}$ using

$$\hat{\Delta}_{t,t+1} = \sum_{s_{t,t+1}} W_i^{t,t+1}.(Y_i^{t+1} - Y_i^t). \quad (4)$$

Logically, the weights $W_i^{t,t+1}$ are used only to estimate change. They are of no value for point estimates because the inference population has little meaning on a particular date. Note that up to this point, the $W_i^{t,t+1}$ have not been corrected for non-response or adjusted in any other way. In practice, equation (4) will be subject to adjustments in the case of the SILC survey.

Estimation of the difference $\Delta^*_{t,t+1} = Y_{t+1} - Y_t$ is a cross-sectional matter and therefore involves the weighting process described in the next section.

4. Cross-sectional weighting

The aim is to make an inference about the total in-scope population Ω_t on the current date *t*. The essential difficulty lies in the fact that in theory, a given (panelized) subsample provides adequate coverage of the population only in the year in which it was selected. After that year, the panel subsample no longer represents the new population of "births", the units that become in-scope. That is the case for newborns, immigrants, individuals who reach specific age thresholds, homeless people who start living in a regular dwelling, people who leave communal dwellings, and so on. While in practice we might consider this coverage defect acceptable for a period of time, it very quickly becomes a serious problem (that is true each year for most panel subsamples), and a top-up sample must be obtained in some fashion. It is worth noting that the problem of population change over time is highly dissymmetrical, since the subpopulation that disappears from year to year (the "deaths") presents no particular difficulties for weighting.

In the SILC survey, the top-up sample is obtained as follows. We survey all individuals in the household of each panel member interviewed in the longitudinal tracking process. Thus, every household surveyed in the cross-sectional process is made up of two types of people: panel members and cohabitants (people who are surveyed but are not panel members). This method covers a large portion of the "births" (in the broad sense) in the population. However, it does not cover households consisting entirely of "births", such as households of new immigrants. "Birth" status is usually determined by asking the birth date of newborns and the landing date of immigrants. Moreover, in practice, the weakness in births coverage is generally regarded as very minor because it is partially corrected with adjustments.

The main technique used to produce cross-sectional weights is the weight share method (Lavallée 2002). As noted previously, in year *t* we have nine panel subsamples $a_{t,k} (1 \le k \le 9)$. We will describe two different ways of using the weight share method. The information that must be collected in the questionnaire is the same for both methods.

4.1 Method 1

The more rigorous approach involves linking all nine subsamples $a_{t,k}$ to the cross-sectional sample for year *t*, which we will denote $\tilde{u}_t$ (Merkouris 2001). In other words, the sample $\tilde{u}_t$ is the same as $s_{t,t} = \bigcup_{k=1}^{9} a_{t,k}$. First, we must determine the links associated with this approach. When a panel member in one of the nine subsamples $a_{t,k}$ is selected, he/she points to himself/herself as a member of the cross-sectional sample at *t* (similar to what is shown in the diagram in 3.1). Under these conditions, when the survey is in steady-state mode, the cross-sectional weight $W_i^{t(1)}$ of an individual *i* in $\tilde{u}_t$ is calculated as shown below. The household of which *i* is a member is denoted *m*. We have

$$W_i^{t(1)} = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in \Omega_{t-k+1}}} 1} \quad (5)$$

where $W_j(t,k)$ is the sampling weight from sample $a_{t,k}$.

This expression shows that all members of the same household ultimately have the same weight. In the numerator, we have the sum of all the "raw" weights (the sampling weights) of the household's panel members. It is understood that a panel member generally appears in only one subsample, but that there may be cases in which a panel member is selected two or more times over a period of nine consecutive years (usually because he/she has moved). Note that dwellings selected from the master sample and the new-housing sample frame are not supposed to be selected again and therefore, in the case of SILC, the probability that an individual who has not moved will appear in two different panels is zero.

As in the longitudinal case (see 3.1), weighting can be carried out only if the data management system is capable of linking each panel member in $\tilde{u}_t$ to all panel samples $a_{t,k}$ in which he/she is included. In the denominator, for each of the nine years $t - 8$ to t considered, we count the household members (both panel members and cohabitants) who are in the sample frame from which the incoming panel subsample

for the year in question is selected. This calculation clearly requires the information provided by the questionnaire.

There are two advantages to this approach: it is completely general, and it produces unbiased cross-sectional weights directly because every cross-sectional household is necessarily linked to one of the nine subsamples involved. The fact that there is an incoming subsample each year ensures the completeness of the cross-sectional population Ω_t; that is, in more technical terms, it ensures that there is at least one link for each household considered at *t*. This is a useful property of rotational sampling, as discussed in section 2.2. On the other hand, the weighting formula has a disadvantage, which is its (relative) complexity both in theoretical terms and for computer programming purposes.

In the start-up phase (up to and including 2011), the formula must be adjusted. The numerator remains the same, but the denominator covers all individuals who could be sampled in 2004 (the survey's first year) and subsequent years. In 2004, weighting is trivial since there is no weight share, but in 2005, we have

$$W_i^{t(1)} = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\left(\sum_{\substack{j \in m \\ j \in \Omega_{2005}}} 1 \right) + 8 \cdot \left(\sum_{\substack{j \in m \\ j \in \Omega_{2004}}} 1 \right)}. \tag{6}$$

In 2006, the formula will be

$$W_i^{t(1)} = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\left(\sum_{\substack{j \in m \\ j \in \Omega_{2006}}} 1 \right) + \left(\sum_{\substack{j \in m \\ j \in \Omega_{2005}}} 1 \right) + 7 \cdot \left(\sum_{\substack{j \in m \\ j \in \Omega_{2004}}} 1 \right)}. \tag{7}$$

4.2 Method 2

It is possible to take an alternative approach to cross-sectional weighting one that leads to a "slightly" simpler equation and is easier to program, but one that presents a difficulty that was not present in the previous method and may make the final weights somewhat less precise. The idea is to use one subsample at a time rather than all of them at once. We take one of the nine subsamples $a_{t,k}$ and the sample of households to which it leads. We then apply the weight share, which when the survey is in steady-state mode yields an individual weight equal to

$$\tilde{W}_i(t,k) = \frac{\sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\sum_{\substack{j \in m \\ j \in \Omega_{t-k+1}}} 1} \tag{8}$$

for any individual *i* in household *m*. It is very easy to verify that if $k = 1$ (which is the case for the incoming subsample), $\tilde{W}_i(t,1)$ is the sampling weight of household *m*.

The problem with this approach lies in the existence (*a priori*) on date *t* of individuals who cannot be surveyed because they belong to households that cannot be "reached" through the sampling $a_{t,k}$ (as long as $k \geq 2$), *i.e.*, individuals whose probability of being surveyed at *t* is zero. This impediment did not exist in the previous method because taking all the subsamples into account at once ensured that on date *t*, every household had a non-zero probability of being selected, at least through $a_{t,1}$. This illustrates once again one of the key advantages of rotational sampling, which is that it covers the entire population each year. In our approach, it is clear that if we consider $a_{t,k}$ for $k \geq 2$, the population of households consisting exclusively of "immigrants" (in the broad sense) between $t - k + 1$ and *t* is not covered. To formalize the situation and produce the final cross-sectional weight, we will use $\Omega_{\alpha,t}^{\text{immig}}$ to denote the population of "immigrants" (in the broad sense) on date *t* in households that consist only of immigrants who can be sampled after year α, with $t - 8 \leq \alpha \leq t - 1$. To be more precise, we should say "who can be sampled on or after a date that is strictly subsequent to the collection date in year α."

On date *t*, the entire population Ω_t is partitioned into nine components: the eight subpopulations $\Omega_{\alpha,t}^{\text{immig}}$, with α ranging from $t - 8$ to $t - 1$, and the subpopulation consisting of individuals who either were already surveyable at $t - 8$ or became surveyable on a date subsequent to $t - 8$ (*i.e.*, who immigrated after $t - 8$) but at *t* are members of a household containing at least one person who is surveyable at $t - 8$. We consider that if the household at *t* contains at least one person who is surveyable at $t - 8$, that will be the case on any date between $t - 8$ and $t - 1$. This ignores situations in which an individual who is in-scope on a given date becomes out-of-scope for a time (as a result of emigration, for example) and then becomes in-scope again.

Next, we use $\tilde{u}_{t,k}$ to denote the cross-sectional sample at *t* from panel $a_{t,k}$, which leads to $\bigcup_{k=1}^{9} \tilde{u}_{t,k} = \tilde{u}_t$. Let $Y_{\alpha,t}^{\text{immig}}$ be the total of the Y_i^t defined on $\Omega_{\alpha,t}^{\text{immig}}$. Following the weight share performed for all $k = 2, ..., 9$, we have

$$E\left(\sum_{j \in \tilde{u}_{t,k}} \tilde{W}_j(t,k) \cdot Y_j^t \right) = \sum_{\Omega_t} Y_j^t - \sum_{\alpha=t-k+1}^{t-1} Y_{\alpha,t}^{\text{immig}}$$

$$= Y_t - \sum_{\alpha=t-k+1}^{t-1} Y_{\alpha,t}^{\text{immig}} \tag{9}$$

and

$$E\left(\sum_{j\in\tilde{u}_{t,1}} \tilde{W}_j(t,1)\cdot Y_j^t\right)=\sum_{\Omega_t} Y_j^t = Y_t \qquad (10)$$

since $\tilde{u}_{t,1} = a_{t,1}$.

If we were using shorter-duration panels, we might be able to ignore the $Y_{\alpha,t}^{\text{immig}}$ and take the actual total over Ω_t. In that case, the "raw" final cross-sectional weight of any individual i would be $\tilde{W}_i(t,k)/9$ if i is from $a_{t,k}$, which would yield the final estimator

$$\hat{Y}_t = \frac{1}{9}\sum_{k=1}^{9}\sum_{i\in\tilde{u}_{t,k}} \tilde{W}_i(t,k)\cdot Y_i^t$$

$$= \frac{1}{9}\sum_{i\in\tilde{u}_t} \tilde{W}_i(t,k)\cdot Y_i^t. \qquad (11)$$

However, since the panels used in France have long lives, we will probably not be able to ignore the $Y_{\alpha,t}^{\text{immig}}$ (an analysis of the collection files will provide the answer), which will mean having to compute specific weights for the individuals in $\Omega_{\alpha,t}^{\text{immig}}$. In those circumstances, we check that any individual i in $\Omega_{\alpha,t}^{\text{immig}}$ who ends up in the cross-sectional sample $\tilde{u}_t$ will have a raw cross-sectional weigh $W_i^{t(2)}$ equal to the weight share value $\tilde{W}_i(t,k)$ divided by $t-\alpha$ (and therefore $1 \le t-\alpha \le 8$). Any individual in Ω_t who does not belong to any of the $\Omega_{\alpha,t}^{\text{immig}}$ (*i.e.*, the vast majority of individuals) will have a final weight of $\tilde{W}_i(t,k)/9$. Note that if i is in $\Omega_{\alpha,t}^{\text{immig}}$, he/she can be surveyed only through $a_{t,1}, a_{t,2}, \ldots, a_{t,t-\alpha}$. Thus we have

$$W_i^{t(2)} = \begin{cases} \tilde{W}_i(t,k)/(t-\alpha) & \text{if } i \in \Omega_{\alpha,t}^{\text{immig}} \\ \tilde{W}_i(t,k)/9 & \text{otherwise} \end{cases} \qquad (12)$$

In the start-up phase, the weighting process has to be adjusted. In 2005, the final cross-sectional weight of individuals in $\Omega_{2004,2005}^{\text{immig}}$ will come directly from the selection of the dwelling from $a_{2005,1}$ (they can only be reached through this incoming panel). In contrast, all other individuals can be surveyed "normally" in the nine panels $a_{2005,k}$ $(1 \le k \le 9)$, so that their weights as calculated by the weight share method will all be divided by 9. In 2006, the weights of the individuals in $\Omega_{2005,2006}^{\text{immig}}$ will be equal to the weight of the dwelling in which they live, a weight that directly reflects the sampling from $a_{2006,1}$; the weights of the individuals in $\Omega_{2004,2006}^{\text{immig}}$ will be the weights from the weight share divided by 2; and the weights of all other individuals will be the weights from the weight share divided by 9.

This procedure can be carried out for one subsample after another and does not have to take account of what happens in other subsamples. If an individual is surveyed at t through two (or more) different subsamples $a_{t,k}$, we carry out the full procedure for each of the two (or more) subsamples. This could occur, for example, in the case of a household composed of two panel members from two different subsamples $a_{t,k}$ who married each other and before their marriage were each tracked separately as one-person households. In that scenario, each individual would be "formally" surveyed twice, once as a panel member and once as a cohabitant.

Finally, to estimate the difference $\Delta_{t,t+1}^* = Y_{t+1} - Y_t$, we can use the weights $W_i^{t(1)}$ from method 1 and calculate

$$\hat{\Delta}_{t,t+1}^* = \sum_{i\in\tilde{u}_{t+1}} W_i^{t+1(1)} Y_i^{t+1} - \sum_{i\in\tilde{u}_t} W_i^{t(1)} Y_i^t. \qquad (13)$$

Alternatively, we can use the weights $W_i^{t(2)}$ from method 2. In that case, the estimator of the difference $\Delta_{t,t+1}^*$ will be given by

$$\hat{\hat{\Delta}}_{t,t+1}^* = \sum_{i\in\tilde{u}_{t+1}} W_i^{t+1(2)} Y_i^{t+1} - \sum_{i\in\tilde{u}_t} W_i^{t(2)} Y_i^t. \qquad (14)$$

References

Ardilly, P. (2006). *Les techniques de sondage*, 2nd edition. Editions Technip, Paris.

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

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

Lévesque, I., and Franklin, S. (2000). Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics, 1997 Reference year. Research document on income, Statistics Canada, Catalogue No. 75F0002MIE-00004, june 2000.

Merkouris, T. (2001). Cross-sectional estimation in multiple-panel household surveys. *Survey Methodology*, 27, 171-181.

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). *Méthodologie de l'Enquête sur la population active*. Statistics Canada, Catalogue 71-526.