Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

THE WEIGHT SHARE METHOD AND THE EUROPEAN SURVEY ON INCOME AND LIVING CONDITIONS

Pascal Ardilly and Pierre Lavallée[1]

ABSTRACT

The European Survey on Income and Living Conditions (SILC) will replace the European Panel beginning in 2004. This annual longitudinal survey will produce annual comparative statistics on income distribution, poverty and social exclusion. While the survey is longitudinal, it will also provide quality cross-sectional estimates. The sample design developed for the SILC by Eurostat is a rotational design based on four panels of four years each with the replacement of one panel per year. This sample design meets the survey's longitudinal and cross-sectional requirements. However, it poses weighting challenges. After a description of the survey, this article looks at the longitudinal and cross-sectional weightings where, in the case of the latter, the Weight Share Method is used.

KEYWORDS: Longitudinal survey; Weight Sharing Method; Weighting.

1. INTRODUCTION

The European Survey on Income and Living Conditions (SILC) is a major undertaking launched in 2000 by Eurostat with the encouragement of the Commission of the European Union. Replacing the European Community Household Panel that ended in late 2001, the purpose of SILC is to become, as of 2004 and for every country of the European Union (EU), a single statistical source on income and living conditions in order to measure social inclusion within the community. SILC is similar to the Survey of Labour and Income Dynamics (SLID) developed by Statistics Canada (see Lavigne and Michaud, 1998).

The population covered at the household level is all so-called "ordinary" households (that is, except for communities) living in the Member State's territory on the collection date. For individual data on income and living conditions, the population is restricted to individuals aged 16 and older present in the households in questions. Each year, every country of the EU must provide Eurostat with a microdata file with the willingness to desegregate the concepts from national accounting at the microeconomic level. These data will be used to calculate structural social indicators commonly defined by all EU countries, which will then be used to develop the annual report of the Commission in the areas of income distribution, poverty and exclusion.

The data will include a *cross-sectional* and a *longitudinal* aspect. The longitudinal aspect relates to physical individuals, who will be followed over time, particularly when they change dwelling.

Sampling and capture methods are left to the discretion of the Member States, which can select the methods most appropriate to their context, with the restriction that sampling be random and highly "representative" (the data may be collected by survey, census, micro simulations, etc.). There is no imposed link between the longitudinal and cross-sectional operations, but Eurostat is developing a system of cross-sectional surveys that relies on the longitudinal sample of individuals, using a rotating design that renews one-quarter of the sample each year (an individual panel is therefore questioned for at least four consecutive years). The master framework requires that the cross-sectional survey in all 15 Member States includes at least 80,000 respondent households (or 156,000 individuals aged 16 or older) and that the longitudinal survey, over two consecutive years, be designed to question at least 60,000 households (and 116,500 individuals aged 16 years or older). More information on SILC can be obtained by consulting Eurostat (2001) and Eurostat (2003), and the Commission européenne (2003 a/b) for the European rules governing this survey.

[1] Pascal Ardilly, INSEE, 18 Boul Adolphe-Pinard, 75675 Paris CEDEX 14, France, (pascal.ardilly@insee.fr), Pierre Lavallée, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6 (pierre.lavallee@statcan.ca)

2. GENERAL PRINCIPLES ASSOCIATED WITH SILC

2.1 Sampling units and units of observation

For SILC, dwellings are sampled from a frame of dwellings updated from new housing construction. In each selected dwelling, the physical individuals and households are observed, the latter constituting clusters of individuals. The longitudinal units are, in fact, the individuals. In other words, we follow the individuals over time.

The individuals for the sample are selected by *indirect sampling*. We begin by selecting a sample of dwellings. Each unit produces one or more households, from which the survey's final sample of individuals is selected. There are two options for selecting the individuals: (i) use of all individuals within the selected households, or (ii) restriction to one individual (called Kish individual) selected randomly within each household. Option (i) allows collection of individual information covering the entire household and thus makes it easier to produce statistics at the household level. It has the disadvantage of requiring the cooperation of all individuals in the household, which is more restricting (especially when someone is absent). This is less of a problem with option (ii).

Use of indirect sampling to obtain the sample of households and individuals requires careful management of the dwelling-household relationship. This can be a complex relationship if we adopt a methodology that authorizes a household to be surveyed across several dwellings.

2.2 Sampling strategies over time

There are three possible scenarios for building the SILC sample over time:

Scenario 1: Select a (pseudo) independent sample in each wave *t;*

Scenario 2: Select a sample identified at the initial time t_0 , and follow this sample over time, which constitutes a panel;

Scenario 3: Select a sample partially renewed in each wave *t*, which is called a *rotating sample*.

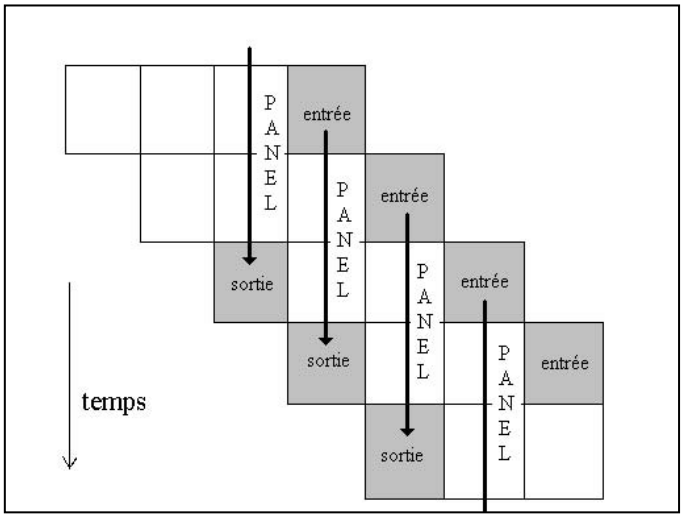
The sample must be designed based on the population of interest. The sample at date t_0 is selected from the longitudinal population. The cross-sectional population is the population at the current date: it is a changing population in the sense that its composition changes with inputs (births, immigrants) and outputs (deaths, emigrants). SILC is interested in both types of populations and the three possible scenarios are more or less adapted to the cross-sectional and longitudinal approaches. The following table describes the three scenarios in relation to the two possible approaches.

Sample type	Cross-sectional approach	Longitudinal approach
Independent	Natural	Possible but less effective
Panel	Impossible without a complementary sample	Natural
Rotating	Possible	Possible

Since the rotating sample adapts well to both the cross-sectional and longitudinal approaches, it is the scenario retained for SILC.

2.3 Principle, advantages and disadvantages of the rotating design

Figure 1. Selection of panels



The rotating sample operates from a juxtaposition of panels of physical individuals. In the case of SILC, each panel has a duration limited to four years (or four waves). Each year, one panel sample is added and one panel sample is removed. Each incoming panel is selected using the same sample design from the updated housing survey frame. The rotating process is presented in Figure 1.

The rotating design has the traditional advantages of a panel such as, for example, the constitution of a longitudinal sample in which evolution analyses can be made. On the other hand, this technique has the traditional disadvantage of panels of individuals, specifically, the tracing costs (follow-up over time). Using the rotating approach, the burden on surveyed individuals is limited (problem of attrition) by reducing their time in the sample to four years, with the disadvantage, however, of reducing the longitudinal use of the data. Finally, the rotating design – thanks to the addition each year of an incoming sample representing the updated population – allows for a natural adaptation of the sample to changes in the population, both in the longitudinal approach and the cross-sectional approach.

3. LONGITUDINAL WEIGHTING

In SILC, the unit of observation is exclusively the individual, although dwellings are identified. At date α of the selection of a given panel sub-sample u_α, the initial weights w_k^α are determined in order to represent the population noted as Ω_α.

Let $s_t = \bigcup_{\alpha=t-3}^{t} u_\alpha$ be the sample resulting from the union of the four panels present on date *t*. Since each panel u_α represents the population Ω_α defined on date α, sample s_t represents the population Ω_t defined at date *t*. This is possible because of the rotating nature of the sample design. In effect, as Figure 2 shows, births in wave *t* can be surveyed via u_t, and therefore via s_t.

Let $p_k^\alpha = \Pr[k \in u_\alpha]$ be the probability that unit *k* is selected in panel u_α. The weight of unit *k* of this panel is given by $w_k^\alpha = 1/p_k^\alpha$. Similarly, for unit *k* of s_t, we have the longitudinal weight $w_{L,k}^t = 1/\pi_k^t$ where

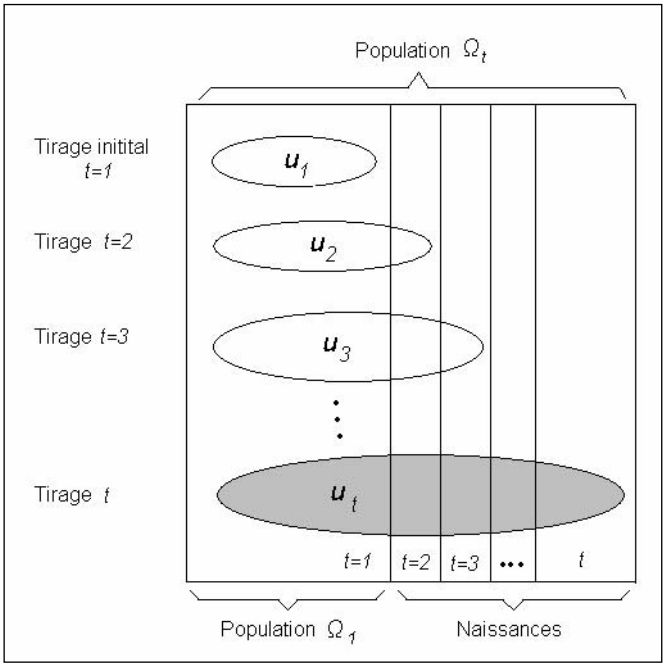
$\pi_k^t = \Pr[k \in s_t] = \Pr[k \in \bigcup_{\alpha=t-3}^{t} u_\alpha]$.

To conduct studies of the evolution of the parameters between two waves (let us use *t* and *t+1)*, we must have a sample of units present in both waves. This sample is referred to as *cylindered*. Between waves *t* and *t+1*, the cylindered sample is given by $s_{t,t+1} = \bigcup_{\alpha=t-2}^{t} u_\alpha$. In other words, to obtain the cylindered sample involving waves *t* and *t+1*, the panels selected at dates *t-2*, *t-1* and *t* must be used. The weight of unit *k* of the cylindered sample $s_{t,t+1}$ is then given — with a very slight approximation — by

$$w_{L,k}^{t,t+1} = \frac{1}{\pi_k^{t,t+1}} = \left(\sum_{\substack{\alpha=t-2 \\ k \in \Omega_\alpha}}^{t} 1/w_k^\alpha \right)^{-1}$$

since we have $\pi_k^{t,t+1} = \Pr[k \in s_{t,t+1}] \approx \sum_{\alpha=t-2}^{t} p_k^\alpha = \sum_{\alpha=t-2}^{t} 1/w_k^\alpha$. Note that if $k \notin \Omega_\alpha$ — which is the case with births after α — we have $p_k^\alpha = 0$. Using these initial unbiased weights, the final weights can be obtained after adjustments and corrections for the total non-response.

Figure 2. Longitudinal configuration



4. CROSS-SECTIONAL WEIGHTING

4.1 Population of extrapolation: problem of representativeness

From a cross-sectional standpoint, the extrapolation deals with the population of the current wave. More generally, it can be said that the population of interest is that defined at a given date *t*. The difference between the reference population Ω_{t_0} (from which the panel is selected) and the changing population Ω_t at a wave $t > t_0$ is due to

new-borns and to immigrants (in the broad sense, everyone who is not a new-born). To represent the latter within the sample, there needs to be a complementary sample to the panel sample. We note that there is also a difference between Ω_{t_0} and Ω_t due to deaths and emigrants, but it does not pose a problem because we simply ignore individuals who have disappeared without adjusting the weights of other individuals. Using a complementary sample to th po

i 7 (g h t i 7

t

produced from these individuals are unbiased. In other words, we are looking for cross-sectional weights $w_{TR,k}^t$ so that the estimator of the total

$$\hat{Y}^t = \sum_{k \in \tilde{s}_t} w_{TR,k}^t y_k^t$$

is unbiased, where $\tilde{s}_t$ is the sample of panel individuals (noted s_t) increased by cohabitants.

The Weight Share Method is described by Ernst (1989) and by Deville (1998). It consists of calculating the sum of the survey weights linked to the panel individuals (selected at t_0) in each cluster *i* in Ω_t (and living at *t*) and dividing that sum by the total number of individuals (panel individuals and cohabitants) in cluster *i* present in Ω_{t_0} (and living at *t*). This gives

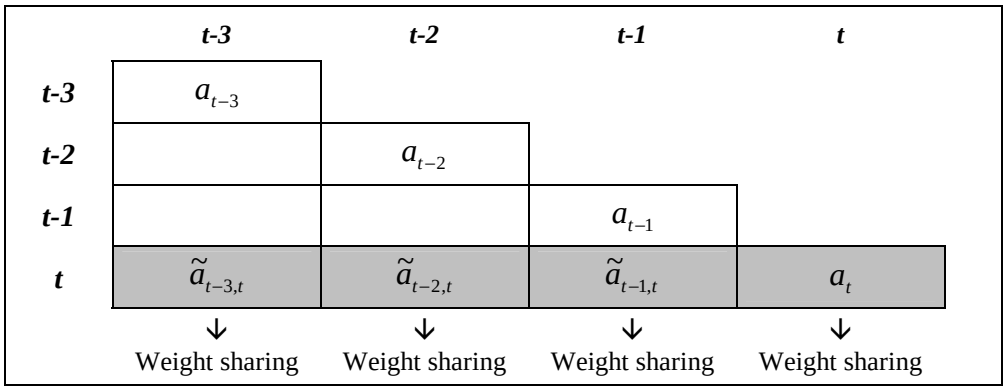
$$w_i^t = \sum_{\substack{k \in i \\ k \in \Omega_{t_0} \\ k \in s_t}} w_{ik}^0 \quad / \quad \sum_{\substack{k \in i \\ k \in \Omega_{t_0}}} 1$$

The next step, for a given household *i*, is to assign this weight w_i^t to all individuals *k* of household *i*; this produces $w_{TR,k}^t = w_i^t$ for $k \in i$. Lavallée (2002) generalized the Weight Share Method to any two populations Ω_{t_0} and Ω_t — with possibly different units — where the links between the two populations are not necessarily one-to-one. Note that households with no link to Ω_{t_0} are not represented by $\tilde{s}_t$ (and thus the problem posed by households composed only of immigrants, see 4.3).

4.3 Application to the rotating approach

The Weight Share Method is applied within each sub-sample (or panel) u_α, as illustrated in Figure 4. A global approach can also be used, which consists of considering all of the panels at the same time. This approach provides a more rigorous but less intuitive solution than the one proposed in this paper. It requires using the generalized version of the Weight Share Method presented in Lavallée (2002). The global approach is described by Merkouris (1999), among others, as part of the SLID.

Figure 4. Application of the weight sharing method



For each sub-sample u_α, the population of inference is the complete population in wave *t* of the survey or Ω_t. Each household considered at *t*, regardless of its composition, has at least one link with population Ω_α at selection date α unless it is constituted solely of persons who immigrated from wave $\alpha+1$ at *t*. Regardless, each household surveyed at *t* will necessarily contain at least one panel individual (coming from one of u_α, $\alpha = t-3,...,t$), but it may also contain cohabitants that may or may not also be found in Ω_α, $\alpha = t-3,...,t$. The essential complication in the weighting process arises from the fact that the probability of surveying a household composed solely of

immigrants depends on when they arrived in the area: the more recent their arrival, the less likely they will be included in panels u_α.

For the survey conducted in wave *t*, we use $\Omega_{\alpha,t}^{immig}$ as the population of immigrants at date α present in a household consisting of only immigrants who arrived on or after α $(t-3 \le \alpha \le t)$. Note that an immigrant individual *k* in α but present in a household that is not constituted solely of immigrants (referred to as an "integrated" immigrant) does not require any special treatment. In effect, he is treated in the "standard" way because he can be surveyed through a panel individual selected in one of the u_α. Thus, we see that population Ω_t consists of individuals already present in the area at *t*-3, immigrants integrated at *t*, and immigrants of $\Omega_{t-2,t}^{immig}$, $\Omega_{t-1,t}^{immig}$ and $\Omega_{t,t}^{immig}$.

Let $\tilde{u}_{\alpha,t}$ be the cross-sectional sample in wave *t* linked to u_α. This sample contains individuals from panel u_α and cohabitants. By applying the Weight Share Method within each cross-sectional sample $\tilde{u}_{\alpha,t}$, we get the weights $w_{TR,k}^{\alpha,t}$ for each individual *k* surveyed in wave *t*. These weights are then used to estimate the totals in wave *t*. It is assumed that there is zero probability of encountering cases of households at *t* that contain panel individuals from different samples $\tilde{u}_{\alpha,t}$. In practice, such a situation is not desirable because an individual from a first sample could extend his time in the survey by being part of another sample from a subsequent wave.

Let Y^t be the real total of y_k^t in Ω_t, and $Y_{\alpha,t}^{immig}$ be the real total of y_k^t in $\Omega_{\alpha,t}^{immig}$. By using the units of $\tilde{u}_{t-3,t}$ from the panel selected at date $\alpha = t-3$, it is clear that the quantity $\sum_{k\in\tilde{u}_{t-3,t}} w_{TR,k}^{t-3,t}\, y_k^t$ estimates without bias the total $Y^t - Y_{t-2,t}^{immig} - Y_{t-1,t}^{immig} - Y_{t,t}^{immig}$. Similarly, it is possible to verify that $\sum_{k\in\tilde{u}_{t-2,t}} w_{TR,k}^{t-2,t}\, y_k^t$ estimates $Y^t - Y_{t-1,t}^{immig} - Y_{t,t}^{immig}$, that $\sum_{k\in\tilde{u}_{t-1,t}} w_{TR,k}^{t-1,t}\, y_k^t$ estimates $Y^t - Y_{t,t}^{immig}$, and that $\sum_{k\in u_t} w_{TR,k}^{t,t}\, y_k^t$ estimates Y^t. By combining these estimators judiciously, we can obtain an unbiased estimator of the total Y^t that uses the data from all individuals surveyed in wave *t*, or in other words, all of the cross-sectional samples $\tilde{u}_{\alpha,t}$, $\alpha = t-3,\ldots,t$. To this end, we identify two scenarios.

Scenario 1:

This scenario is simple and rigorous but relies on a simplification that generates a bias: the individuals of $\Omega_{\alpha,t}^{immig}$, $\alpha = t-2, t-1, t$ before Ω_t are "ignored". In other words, the population of immigrants in wave α (for $\alpha = t-2, t-1, t$) present at *t* in a household consisting only of immigrants who arrived on or after α is assumed to be negligible. In this case, $\sum_{k\in\tilde{u}_t} w_{TR,k}^{t}\, y_k^t / 4$ estimates Y^t "almost" without bias, where $\tilde{u}_t = \tilde{u}_{t-3,t} \cup \tilde{u}_{t-2,t} \cup \tilde{u}_{t-1,t} \cup u_t$ and where $w_{TR,k}^{t} = w_{TR,k}^{\alpha,t}$ for $k \in \tilde{u}_{\alpha,t}$. Thus, the weight of each individual *k* of the cross-sectional sample $\tilde{u}_t$ is given by $w_{TR,k}^{t} / 4$.

Scenario 2 :

This scenario is more complicated but also more rigorous. It is clear that $\left(\sum_{u_t\cap\Omega_{t,t}^{immig}} w_{TR,k}^{t,t}\, y_k^t\right)$ estimates $Y_{t,t}^{immig}$, because it is an estimation of the domain $\Omega_{t,t}^{immig}$. Similarly, the quantity $\left(\frac{1}{2}\sum_{u_t\cap\Omega_{t-1,t}^{immig}} w_{TR,k}^{t,t}\, y_k^t + \frac{1}{2}\sum_{\tilde{u}_{t-1,t}\cap\Omega_{t-1,t}^{immig}} w_{TR,k}^{t-1,t}\, y_k^t\right)$ estimates unbiasedly $Y_{t-1,t}^{immig}$, and

$\left(\frac{1}{3}\sum_{u_t\cap\Omega_{t-2,t}^{immig}} w_{TR,k}^{t,t}\ y_k^t + \frac{1}{3}\sum_{\tilde{u}_{t-1,t}\cap\Omega_{t-2,t}^{immig}} w_{TR,k}^{t-1,t}\ y_k^t + \frac{1}{3}\sum_{\tilde{u}_{t-2,t}\cap\Omega_{t-2,t}^{immig}} w_{TR,k}^{t-2,t}\ y_k^t\right)$ estimates $Y_{t-2,t}^{immig}$. From these estimators, using a few calculations without particular difficulty, it is possible to obtain the weight of unit *k* from $\tilde{u}_t$, as: $w_{TR,k}^t$ if *k* is in $\Omega_{t,t}^{immig}$; $w_{TR,k}^t/2$ if *k* is in $\Omega_{t-1,t}^{immig}$; $w_{TR,k}^t/3$ if *k* is in $\Omega_{t-2,t}^{immig}$; $w_{TR,k}^t/4$ in all other cases; where $\tilde{u}_t = \tilde{u}_{t-3,t} \cup \tilde{u}_{t-2,t} \cup \tilde{u}_{t-1,t} \cup u_t$ and where $w_{TR,k}^t = w_{TR,k}^{\alpha,t}$ for $k \in \tilde{u}_{\alpha,t}$.

It is important to note that the second scenario requires determining the eventual attachment of an individual to populations $\Omega_{\alpha,t}^{immig}$. Thus, an individual question needs to be included dealing with the first year of the individual's presence in a possibly sample dwelling. As with the longitudinal weighting, adjustments and corrections for non-response will be applied to the weights $w_{TR,k}^t$.

5. CONCLUSION

We have seen that SILC is a Europe-wide annual longitudinal survey. To be able to produce both longitudinal and cross-sectional statistics, Eurostat plans to use a rotating design based on the joint use of four panels for a period of four years each with the replacement of one panel per year. This sample design has significant benefits, but complicates longitudinal and cross-sectional weighting.

The longitudinal weighting is accomplished by combining the panels to obtain a cylindered sample. The number of panels used depends on the length of the period over which changes are to be measured. For example, for a two-year period, we use the data from three panels.

Th addition of cohabitants to the sample is a judicious method that makes it possible to ensure cross-sectional "representativeness" from a panel, but makes the cross-sectional weighting more complex. The Weight Share Method is used to be able to attach a cross-sectional weight to each surveyed individual (longitudinal individual or cohabitant).

This paper deals only with the basic SILC weighting, that is, that which results from sampling. When SILC is operational, one of the first steps will be to adjust the weights to take into account non-response. The weights will then be adjusted through calibration.

REFERENCES

Commission européenne (2003a), "Règlement de codécision du Parlement européen et du Conseil européen (16 juin 2003)", *Journal Officiel de la Commission Européenne*, 3 juillet 2003.

Commission européenne (2003b), "Règlements d'application de la Commission européenne", *Journal Officiel de la Commission Européenne*, 17 novembre 2003.

Deville, J.-C. (1998), "Les enquêtes par panel : en quoi diffèrent-elles des autres enquêtes? suivi de: Comment attraper une population en se servant d'une autre ", *INSEE Méthodes*, No. 84-85-86, pp. 63-82.

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

Eurostat (2001), "Sampling", article présenté au Groupe de travail sur les Statistiques sur le revenu et les conditions de vie, EU-SILC 51/01, 24-25 septembre 2001.

Eurostat (2003), "First Ideas on Weighting", article présenté au Groupe de travail sur les Statistiques sur le revenu et les conditions de vie, EU-SILC 123/03, 10-11 juin 2003.

Lavallée, P. (2002), *Le sondage indirect, ou la Méthode généralisée du partage des poids*, Belgique : Éditions de l'Université de Bruxelles, France : Éditions Ellipses.

Lavigne, M., Michaud, S. (1998), "Aspects généraux de l'Enquête sur la dynamique du travail et du revenu", document de travail de l'Enquête sur la dynamique du travail et du revenu, no. 98-05, Statistique Canada, mars 1998.

Merkouris, T. (1999), "Cross-Sectional Estimation in Multi-Panel Household Surveys", document de travail de la Direction de la méthodologie, no. HSMD – 99 – 004E, Statistique Canada, octobre 1999.