Income Statistics Division

75F0002MIE - 00004

Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics 1997 Reference year

Prepared by:
Isabelle Lévesque
Sarah Franklin

August 2000

Statistics Canada Statistique Canada

Canada

Data in many forms

Statistics Canada disseminates data in a variety of forms. In addition to publications, both standard and special tabulations are offered. Data are available on the Internet, compact disc, diskette, computer printouts, microfiche and microfilm, and magnetic tape. Maps and other geographic reference materials are available for some types of data. Direct online access to aggregated information is possible through CANSIM, Statistics Canada's machine-readable database and retrieval system.

How to obtain more information

Inquiries about this product and related statistics or services should be directed to: Client Services Income Statistics Division Canada, Ottawa, Ontario, K1A 0T6 ((613) 951-7355; (888) 297-7355; income@statcan.ca) or to the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

You can also visit our World Wide Web site: http://www.statcan.ca

Toll-free access is provided **for all users who reside outside the local dialing area** of any of the Regional Reference Centres.

National enquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Order-only line (Canada and United States)	**1 800 267-6677**

Ordering/Subscription information

All prices exclude sales tax

Catalogue no. 75F0002MIE-00004, is available on internet for CDN $0.00 per issue. Users can obtain single issues at http://www.statcan.ca/cgi-bin/downpub/freepub.cgi.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact your nearest Statistics Canada Regional Reference Centre.

Statistics Canada
Income Statistics Division

Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics 1997 Reference Year

Published by authority of the Minister responsible for Statistics Canada

August 2000

Catalogue no. 75F0002MIE - 00004
ISSN 0000-0000

Catalogue no. 75F0002MPE - 00004
ISSN 0000-0000

Frequency: Irr.

Ottawa

La version française de cette publication est disponible sur demande

---

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

ABSTRACT

The Survey of Labour and Income Dynamics (SLID), introduced in the 1993 reference year, is a longitudinal panel survey of individuals. The purpose of the survey is to measure changes in the economic well-being of individuals and the factors that influence these changes. SLID's sample is divided into two overlapping panels, each six years in length. Longitudinal surveys like SLID are complex due to the dynamic nature of the sample, which is in turn is due to the ever-changing composition of households and families over the years. For each reference year, SLID produces two sets of weights: one is representative of the initial population (the longitudinal weights) while the other is representative of the current population (the cross-sectional weights). For the production of the cross-sectional weights, SLID combines two independent samples and assigns a probability of selection to individuals who joined the sample after the panel was selected. The longitudinal weights, like the cross-sectional weights, are adjusted for nonresponse, influential values, post-stratification and confidentiality. The purpose of this document is to describe SLID's methodology for the longitudinal and cross-sectional weights, as well as to present problems that have been encountered and solutions that have been proposed. For the purpose of illustration, results for the 1997 reference year are used.

TABLE OF CONTENTS

1. Introduction

The Survey of Labour and Income Dynamics (SLID), introduced in the 1993 reference year, is a longitudinal panel survey of individuals. Its goal is to measure changes in the economic well-being of individuals and the factors that influence those changes, especially key factors in the areas of demographic and family characteristics and labour market activity. Its end product is a set of microdata files. For an overview of the survey, see Lavigne and Michaud (1998).

Originally, SLID was intended to provide longitudinal data to be used in producing longitudinal estimates and analyses. Its cross-sectional dimension has now become just as important as its longitudinal one, in part because the Survey of Consumer Finances (SCF), a cross-sectional survey, has been merged with SLID (Cotton et al., 1999). The two surveys were integrated for the 1998 reference year. For more information about SCF, see Statistics Canada, 1997.

Weighting a longitudinal survey can sometimes be a major methodological challenge, not only because of its longitudinal and cross-sectional dimensions but also because of the dynamic nature of the panel (which is in turn due to the ever-changing composition of families and households). Many weighting problems have arisen through the years, and there will be others in the future: the combining of overlapping panels, the integration of SCF, the change in software used to model nonresponse, the increase in the number of age-sex post-strata and the introduction of new post-strata (number of economic entities, income groups), changes in data processing, etc. All these considerations add to the complexity of SLID's longitudinal and cross-sectional weighting.

The main purpose of this document is to describe the various steps in the longitudinal and cross-sectional weighting processes for the 1997 reference year. Our second aim is to document certain concerns and questions about a few of the weighting steps. Our third objective is to inform the reader of future developments and major changes affecting weighting. The document consists of five sections. The survey's methodology is discussed in section 2. Section 3 presents the steps in the longitudinal weighting process, from determination of the initial weights to calculation of the final ones, while section 4 covers the various steps in the cross-sectional weighting process. Section 5 looks ahead to the future, and section 6 contains the document's conclusion.

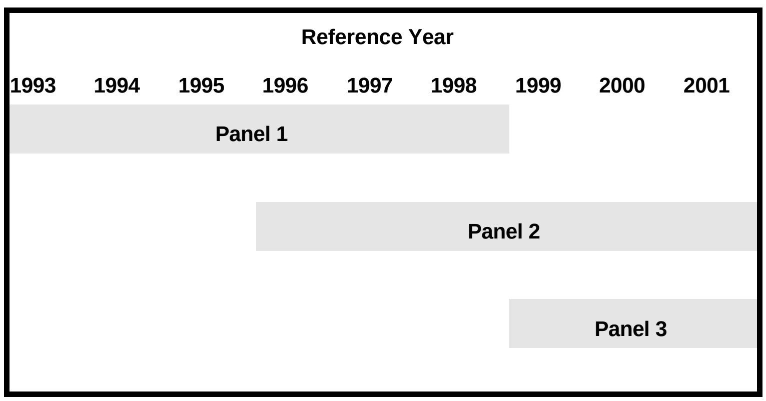
This document is partly based on previous reports on SLID's weighting (Grondin, 1996; Gagnon, 1997; Renaud, 1997). The latter documents cover the 1993 and 1994 reference years. As the longitudinal and cross-sectional weighting methodology has undergone a number of changes and innovations since then, an updated version was deemed necessary.

2. Survey Methodology

SLID sample is composed of two panels, each six years (or six waves) in length. Panel 1, selected in January 1993, covers the population living in the 10 Canadian provinces on December 31, 1992. Panel 2, chosen in January 1996, represents the provinces' population on December 31, 1995. In both cases, the population excludes persons living on Indian reserves, full-time members of the Canadian Armed Forces, and persons who had been living in an institution for more than six months. All references to the population of the 10 Canadian provinces in this document automatically imply all the above exclusions.

A new panel is introduced every three years to replace the oldest panel. The purpose of this rotation is to make the cross-sectional sample more representative (to reflect the new households added to the population in the preceding three years), to mitigate the effect of attrition, and to reduce response burden. Panel selection and overlap is illustrated in Figure 2.1.

Figure 2.1: Overlap of SLID Panels



Each SLID panel consists of a subsample of about 15,000 households (roughly 40,000 people) from the Labour Force Survey (LFS). The LFS sample is drawn from an area frame in accordance with a multistage sampling design. The LFS has six panels (rotation groups); each panel remains in the sample for six months, with one panel replaced each month. The last-stage sampling unit is the dwelling. Thus, all members of the households occupying the selected dwellings are included in the LFS sample. For further information about the LFS, see the reports on its methodology (Singh et al., 1990; Gambino et al., 1998).

The households selected for SLID belong to the panels that rotated out of the LFS sample in the first two months of SLID's first reference period (January and February 1993 for Panel 1, and January and February 1996 for Panel 2). SLID's initial sample is chosen from the LFS file for January. Of all the households that make up the outgoing LFS rotation groups in January and February, SLID selects only those which were LFS respondents in January. The final LFS interview serves as the introductory contact for SLID. A preliminary interview is conducted with each person selected by SLID. Basic information about work experience, level of education, and family and personal background is collected in the interview.

For Panel 1, the preliminary SLID interview was conducted at the same time as the final LFS interview. Because of budget constraints, households that responded in the preliminary interview had to be subsampled; the longitudinal sample for Panel 1 includes only members of those households. A subsample of interview nonespondents was also selected so that nonresponse bias could be studied. Data were collected for those households, but were not used for production. This subsample will not be discussed here.

For Panel 2, the preliminary interview was conducted at the same time as SLID's wave 1 interviews in order to reduce collection costs. The longitudinal sample for Panel 2 includes everyone who underwent a preliminary interview (respondents and nonrespondents). Figures 2.2 and 2.3 illustrate SLID's household selection process. More details on the selection of SLID panels are provided in subsection 3.1.

Figure 2.2: Selection of Households for SLID Longitudinal Sample - Panel 1

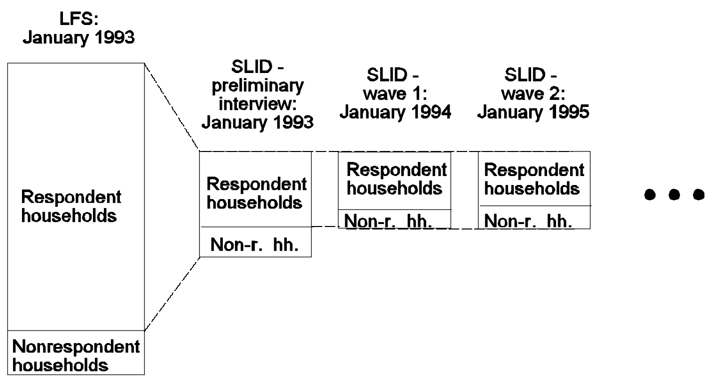
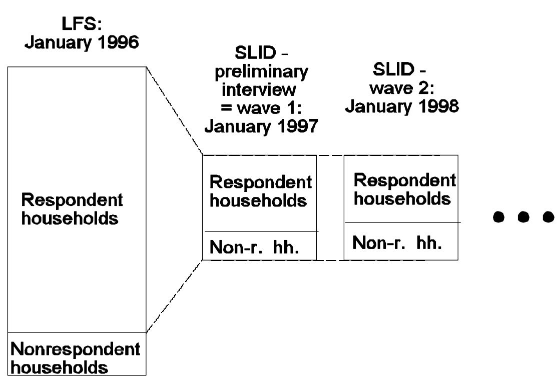


Figure 2.3: Selection of Households for SLID Longitudinal Sample - Panel 2



Each person selected for SLID is then interviewed twice a year for six years. Each year, information is collected about family and household composition, labour market activity and the previous year's income. The labour market questions are asked in January, and the income questions, in May. To reduce the response burden, respondents can avoid the May interview by preauthorizing Statistics Canada to access their Revenue Canada tax file.

When a panel is selected for SLID, every member of the households chosen, regardless of age, becomes part of the panel's longitudinal sample. These persons are considered part of the longitudinal sample for the panel's full six years, even if they move, die, are institutionalized or become full-time members of the Armed Forces living in military quarters. No one else can join the longitudinal sample for this panel. Thus, for each panel, the longitudinal sample is formed when the panel is selected and remains unchanged for the life of the panel. The sample's target population will stay the same throughout the six years that the panel is part of the survey.

For longitudinal purposes, the unit of interest is the individual; because of its constantly changing nature, a household can hardly be used as a tool for longitudinal analysis. For cross-sectional purposes, we are interested in both the individual and the household. Since SLID is concerned with the characteristics not only of longitudinal persons but also of households, everyone who lives with at least one longitudinal individual is interviewed. Hence, the cross-sectional sample for a given year is composed of all longitudinal persons in scope on December 31 of the corresponding reference year, and all persons living with them on that date. An individual is considered in scope on December 31 of a given reference year (cross-sectionally in scope) if on that date, he or she lives in one of the 10 Canadian provinces, does not live on an Indian reserve, has not been living in an institution for more than six months, and is not a full-time member of the Canadian Armed Forces living in military barracks. Every cross-sectionally in-scope longitudinal person will also be part of the cross-sectional sample. Interviewees who are not part of the longitudinal sample are called cohabitants. Adding cohabitants to the cross-sectional sample helps make it more representative (Latouche, Michaud, 1997).

To meet the above-mentioned requirements, longitudinal and cross-sectional weights are produced each year. Longitudinal weighting is designed to generate estimates that are representative of the population of Canada's 10 provinces at the time the longitudinal sample is selected, while cross-sectional weighting produces estimates that are representative of the provinces' population on December 31 of a given reference year. For example, the estimates based on Panel 1's longitudinal sample are representative of the provinces' population on December 31, 1992, and the estimates produced with Panel 2's longitudinal sample relate to the provinces' population on December 31, 1995. The estimates based on the cross-sectional sample for the 1997 reference year represent the provinces' population on December 31, 1997.

In January 1993, 39,745 persons were selected for Panel 1's longitudinal sample. In January 1996, 43,547 persons were chosen for Panel 2's longitudinal sample. The cross-sectional sample for the 1997 reference year consists of 81,090 individuals, 70,372 of whom are longitudinal persons.

Table 2.1 provides an overview of how the longitudinal samples have changed over time. Table 2.2 shows the composition of the cross-sectional samples of all reference years considered (1993 to 1997 inclusive).

Table 2.1: Composition of the Longitudinal Samples of Panels 1 and 2 (Number of Persons)

Reference year		1993	1994	1995	1996	1997
	Persons contacted at interview					
	In-scope on December 31 of the reference year[1]	39 456	36 241	34 336	33 159	31 802
	Living outside the 10 Canadian provinces	28	37	45	150	280
	Institutionalized	81	119	278	256	280
Panel 1	Deceased	180	408	657	908	1 134
	Duplicate, error	0	0	1	1	1
	Persons not contacted at interview[3]					
	Hard refusal, persons in households not traced[2]	0	2940	4428	5271	6 248
	Total	39 745	39 745	39 745	39 745	39 745
	Persons contacted at interview					
	In scope on December 31 of the reference year[1]	-----	-----	-----	41 767	38 366
	Living outside the 10 Canadian provinces	-----	-----	-----	126	270
	Institutionalized	-----	-----	-----	40	120
Panel 2	Deceased	-----	-----	-----	234	466
	Duplicate, error	-----	-----	-----	0	2
	Persons not contacted at interview[3]					
	Hard refusal, persons in households not traced	-----	-----	-----	1 380	4 323
	Total	-----	-----	-----	43 547	43 547

1 Still living in the 10 Canadian provinces, not deceased, not living in an institution for more than last six months.
2 Does not include refusals at SLID preliminary interview and the LFS interview.
3 Prior to the 1997 reference year, this category includes hard refusals and not-traced during collection for a given reference year. For the 1997 and subsequent reference years, all cases in this category relate to previous collection.

Table 2.2: Composition of Cross-sectional Samples, by Panel (Number of Persons)

Reference year	1993	1994	1995	1996	1997
Panel 1					
Longitudinal persons	39 456	36 241	34 336	33 159	31 802
Cohabitants	2 062	3 640	4 620	5 768	6 655
	41518	39881	38956	38927	38457
Panel 2					
Longitudinal persons	-----	-----	-----	41767	38366
Cohabitants	-----	-----	-----	2 351	4 011
	-----	-----	-----	44118	42377
Total	41 518	39 881	38 956	83 045	80 834

SLID's mandate is to collect longitudinal and cross-sectional data on labour-market and income characteristics and to make the data available to users inside and outside Statistics Canada. For the latter purpose, SLID produces two sets of microdata files: one anonymized to preserve confidentiality (external files) and one not anonymized (internal files). The internal files are available only to Statistics Canada users, as they contain weights that have not been anonymized (*internal weights*). External files, produced for users outside Statistics Canada, are anonymized through the addition of noise to all the weights (*external weights*).

Once the data for a given reference year have been processed, Statistics Canada users can access the relational database to perform their cross-sectional and longitudinal analyses. All information in the database is confidential since it has not been anonymized. Internal and external weights, as well as some other anonymized data, are also included in the database.

Internal cross-sectional and longitudinal files are also produced to make the data easier for Statistics Canada users to access. Releasing external files takes more work because they must be anonymized. All disclosure control procedures are reviewed and improved in order to ensure data confidentiality. For the 1997 reference year, the original dissemination strategy was to produce a set of public-use cross-sectional and longitudinal microdata files.

The cross-sectional files (which relate to the reference year in question) cover everyone (longitudinal persons and cohabitants) aged 16 and over in respondent households that are in scope on December 31 of the reference year concerned. The longitudinal files (which contain information for all reference years, from the first one to the most recent one) include all longitudinal persons regardless of status. The two sets of files have very similar contents. One feature of this dissemination strategy is that to maintain confidentiality, it prohibits the publication of microdata files that can be used to reconstruct households. Because of time constraints, the microdata for the 1997 reference year have not yet been released.

3. Longitudinal Weighting

This section describes the methodology used to weight the longitudinal persons in each wave of SLID. For illustrative purposes, we will use examples from wave 5 of Panel 1 and wave 2 of Panel 2. It is important to keep in mind that the longitudinal weights are intended to represent the population of Canada's ten provinces on December 31, 1992, for Panel 1 and on December 31, 1995, for Panel 2.

SLID produces two sets of longitudinal weights: a set of internal weights and a set of external weights. The external weights are for users outside Statistics Canada. Noise is added to these weights prevent the reconstruction of families and households. The two sets of longitudinal weights are produced separately for the two panels.

The longitudinal weighting process consists of the six steps discussed in the following subsections: determination of initial weights (3.1), classification of longitudinal individuals (3.2), nonresponse modelling and adjustment (3.3), adjustment for influential values (3.4), post-stratification (3.5) and the addition of noise (3.6). The resulting final longitudinal weights are described in subsection 3.7.

3.1 Determination of Initial Weights

For the first SLID sample (Panel 1), respondent households from the two outgoing LFS rotation groups in January and February 1993, a total of 20,486 households, were selected. Of these, 17,659 households responded at SLID's preliminary interview, and the remaining 2,827 were nonrespondent households. For budget reasons, the decision was made to reduce the initial SLID sample to about 15,000 households. To this end, the respondent households were subsampled using a Poisson process with a parameter of 0.84. The nonrespondent households were also subsampled using a Poisson process, but in this case the parameter was 0.06. This resulted in a reduced sample of 14,832 respondent households (39,745 persons) and 174 nonrespondent households (410 persons).

For budget reasons, only the respondent households (14,832) were selected for the initial sample in Panel 1. A study showed that this decision would have a relatively minor impact (Durning, 1994). Though excluded from the final sample, the 174 nonrespondent households were retained for analysis. These households form what is called a quality sample. This sample will not be discussed in the present document.

The sample for Panel 2 was selected in a much simpler way: all households that responded to the LFS in January 1996 and belonged to the outgoing LFS rotation groups for January and February 1996 were selected (16,472 households totalling 43,547 persons). The preliminary interviews for Panel 2 were conducted at the same time as the wave 1 interviews (that is, in January 1997).

The initial longitudinal weights are equal to the households' inverse selection probability Thus, all longitudinal members of a household (the one to which they belonged at the time the sample was selected) have the same initial weight.

Since SLID sample is drawn from the LFS sample, the initial weight will be as follows:

$$w_{initial,p_1} = w_{LFS} \times 3 \times 1.19$$
$$w_{initial,p_2} = w_{LFS} \times 3$$

where
$w_{initial,p_1}$ = initial longitudinal weight for Panel 1
$w_{initial,p_2}$ = initial longitudinal weight for Panel 2
w_{LFS} = LFS weight adjusted for nonresponse (LFS subweight)
3 = inverse selection probability of the LFS rotation groups (2 out of 6 groups were selected)
1.19 = inverse sampling ratio of the subsample of respondents at the preliminary interview for Panel 1 (selected using a Poisson process).

3.2 Classification of Longitudinal Individuals

For each wave, the first step in longitudinal weighting is to place the longitudinal individuals in one of the following categories: nonrespondents (in respondent or nonrespondent households), cross-sectionally in-scope respondents, and cross-sectionally out-of-scope persons. A household is defined as respondent if at least one household member responded to at least one of the two interviews (the labour interview or income interview). A household is considered nonrespondent if all household members are nonrespondents at both interviews. A person is deemed cross-sectionally out of scope if, on December 31 of the reference year, he or she was deceased, had been in an institution for more than six months, or was not living in one of Canada's ten provinces.

The classification of longitudinal individuals is necessary in order to determine which ones should be assigned a non-zero longitudinal weight. A longitudinal weight of zero is applied to individuals in nonrespondent households. All other individuals initially selected for the longitudinal sample have a non-zero longitudinal weight, even if they are cross-sectionally out of scope (since they are considered to be respondents). It should be noted that for weighting purposes, SLID defines nonrespondent persons in respondent households as respondent. This is to ensure weighting consistency within households (the data for nonrespondent persons in respondent households are imputed).

The classification of longitudinal individuals also determines whether an individual's weight should be adjusted for nonresponse and identifies those individuals who will be used in the nonresponse adjustment model. The purpose of this model is to find the value by which respondents' weights must be adjusted to make them representative of all respondents and nonrespondents.

Nonresponse adjustments are usually applied to the weights of all respondent units in a survey. In SLID, a nonresponse adjustment factor is applied to the weights of all persons considered to be respondent, except children in respondent households and cross-sectionally out-of-scope persons. A nonresponse adjustment factor is not used for cross-sectionally out-of-scope persons because the number of such persons is already overestimated (Franklin, 1999b). Children's weights are not adjusted for nonresponse because too little information is collected about children to properly model nonresponse. A nonresponse adjustment is made implicitly at post-stratification.

Table 3.1 shows the various response categories of the longitudinal individuals in wave 5 of Panel 1 and wave 2 of Panel 2. The *Nonresponse modelling* column indicates which categories of individuals are used in the nonresponse adjustment model. The last three columns show how the weights of individuals in the various categories will be processed.

Table 3.1: Classification of Longitudinal Individuals in Wave 5 of Panel 1 and Wave 2 of Panel 2

	Classification of individuals	Panel 1 wave 5	Panel 2 wave 2	Nonresponse modelling	W_{adjust} = 0	W_{adjust} = $W_{initial}$ (unad-justed)	W_{adjust}= $W_{initial}$ /R_{RHG} (adjusted)
CHILDREN	In a nonrespondent household	1 043	1 057		Ž		
age 0-15 (on Dec. 31, 1997)	In a respondent household	5 530	7 986			Ž	
	In a cross-sectionally out-of-scope household (deceased, institutionalized or outside the provinces)	39	65			Ž	
	Cross-sectionally out of scope (institutionalized, deceased or outside the provinces)	1 513	705			Ž	
ADULTS	Respondents	25 819	29 002	U			Ž
age 16 + (on Dec. 31, 1997)	Nonrespondents in a nonrespondent household	5 648	4 407	U	Ž		
	Duplicate, error	1	2		Ž		
	Nonrespondents in a respondent household (will be imputed)	151	323				Ž
	Quality sample[1]	411	-----		Ž		
	Total	40.155	43.547				

3.3 Nonresponse Modelling and Adjustment

Having generated the initial weights, we now have to compute a nonresponse adjustment factor for every individual deemed to be a respondent except for children in respondent households and cross-sectionally out-of-scope persons. As noted earlier, the two panels are processed separately. The nonresponse adjustment factor for an individual is defined as the inverse response rate of the response homogeneity group (RHG) to which the individual belongs. RHGs are composed of all individuals with similar nonresponse characteristics. If the RHGs are defined so that nonresponse is completely random within each group, then the nonresponse bias is negligible (Tambay et al., 1998). To find an appropriate adjustment value to make respondent weights representative of all respondents and nonrespondents, we use nonresponse modelling.

---

1 This category contains all longitudinal persons who are in the quality sample or in the same household as a member of the quality sample (there was one such person in the 1997 reference year).

The first step in nonresponse modelling is to determine which explanatory variables need to be in the model in order to correctly predict the dichotomous variable *response* (whose possible values are either *respondent* or *nonrespondent*). The simplest way of selecting these variables is to use information collected by the LFS, that is, prior to SLID's initial interviews. The advantage of this method is that the information is available for all longitudinal individuals, whether or not they are respondents in the current wave. However, some of the variables collected are subject to change (income, level of education, marital status, employment, presence of children, etc.) and may be outdated for the reference year of the wave being processed. Yet if we decided to use the most recent available information for each individual, we could end up with information from different reference periods (because of nonresponse in the wave being processed). In other words, for respondents, the information would be up-to-date, but for nonrespondents, older information would have to be used. Moreover, if the latest information were used for nonresponse modelling, the weight associated with the immediately preceding wave would have to be used as the base weight instead of SLID's initial weight. But that raises the problem of converted nonrespondents (persons who were respondents two waves ago, nonrespondents in the preceding wave and respondents in the current wave). Since converted nonrespondents have a longitudinal weight of zero in the wave preceding the current one, some way would have to be found to assign them a base weight. One way would be to use the weight computed two waves ago. In that case, it would be necessary to ensure that the method used to assign base weights from different periods was methodologically acceptable. The possibility of using the latest information for nonresponse modelling was also rejected because it was believed that the cumulative effect of all those nonresponse models would produce estimates that were too dependent on the nonresponse models.

SLID therefore elected to consider only known information gathered before the start of the first collection period. For Panel 1, this information was taken from the preliminary interview, since it was conducted at the time of the final LFS interview. For Panel 2, the information considered was collected at the final LFS interview.

The first step in nonresponse modelling (determining the explanatory variables) is not carried out until the first wave of a panel is processed, because the variables used in the nonresponse model must be determined at the outset. The variables selected for nonresponse modelling in the first wave are used in subsequent waves. To determine what explanatory variables should be used in the nonresponse modelling, categorical variables with $m>2$ possible values are recoded as m dichotomous variables. From all these variables, the ones most likely to affect response are selected subjectively.

Logistic regression is performed separately for each variable. On the basis of the results of these univariate regressions, the significant variables are selected. In all, 57 dichotomous variables were retained for Panel 1, and 43 for Panel 2. A complete list of the variables selected for each panel is provided in Appendix A. As mentioned earlier, the same variables are used for all waves of a panel; the only changes normally made are alterations in the possible values of some variables (e.g., age group and income bracket).

The second step is to generate the RHGs on the basis of the variables selected in the first step. Prior to the 1996 reference year, SLID created the RHGs by logistic regression: the SAS procedure LOGISTIC, with selection type STEPWISE, was used for this purpose. Since the 1996 reference year, SLID has been using segmentation modelling to generate the RHGs. The software package that applies this method is called KnowledgeSeeker (version 4.2.2). Segmentation modelling is preferred to logistic regression because it reduces nonresponse bias more effectively (Dufour et al., 1998).

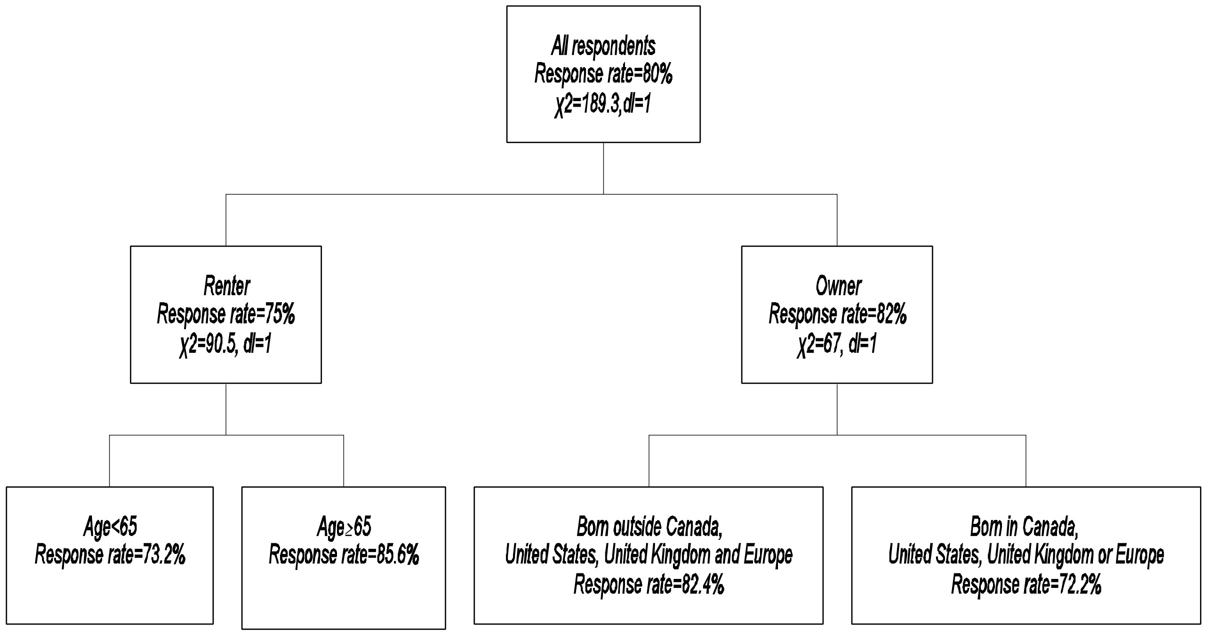
Segmentation modelling employs an iterative process to partition the data file and form the decision tree. The root node of the tree is simply the entire data set. The first iteration defines the tree's first split. The latter is created by determining which of the variables selected for modelling has the most significant effect on the *response* variable. For each variable, a statistical measure reflecting the extent to which the variable discriminates nonresponse is computed, and the one that has the most significant effect is selected. This is how the first split is formed, with a node at each end. For each node formed, another iteration of the procedure is carried out. This in turn produces more splits and more nodes. When the process reaches a node where no variable has a significant effect on response, there are no further splits, and the formation of the branch is complete. The iterative process ends when all branches of the tree are completed.

A number of methods can be used to find the variable that has the greatest influence on the response variable. SLID uses chi-square automatic interaction detection (CHAID), an algorithm that selects the variable with the highest Pearson chi-square statistic.

One of the reasons segmentation modelling is preferable to logistic regression is that it is more flexible. Logistic regression forces the model to be symmetric: once the significant variables are determined, the RHGs are formed by taking all possible combinations of the intersection of these variables. This process can create small cells or cells that are not significant, which can affect the reliability of the estimates. Segmentation modelling allows the model to be asymmetric: at each split in the decision tree, it determines which variable is most significant. This ensures that each RHG is significant, and it usually results in an asymmetric model.

For example, in Panel 1 of SLID, the most significant variable in discriminating between response and nonresponse is a variable that indicates whether the person is a homeowner. For homeowners, the most significant variable in accounting for nonresponse is one that indicates whether a person was born in the United States, the United Kingdom or Canada. For non-homeowners, the next most significant variable in explaining nonresponse is one that indicates whether a person is over age 65 (see Figure 3.1 below).

Figure 3.1: First Few Branches of the Decision Tree for Panel 1, 1997 Reference Year



Another advantage of segmentation modelling is that it can be performed with the very user-friendly KnowledgeSeeker, a software package that simplifies the task of creating a decision tree. However, segmentation modelling also has a few drawbacks. While creating a decision tree is easy, the next step in the process is tedious and slow. The definitions of all the RHGs have to be coded into SLID's longitudinal weighting program so that each individual's nonresponse adjustment factor can be computed. Knowledge Seeker does not provide a file defining each RHG and the associated response rate.

Another drawback of segmentation modelling is directly related to the decision tree itself. Once formed, a decision tree can quickly become cumbersome; beyond a certain number of splits, it becomes difficult to interpret. Decision trees are also rather unstable: the addition, removal or alteration of one variable or one individual can have a major impact on the tree. This makes it very difficult to compare several decision trees.

In each wave, SLID generates a nonresponse adjustment model separately for each panel. Each model is initially defined at the national level. Province-level splits are kept in the decision tree if they are significant. SLID uses a national model because it is easier and more convenient than creating a model for each province. According to a 1999 study, the difference between using a national model and using provincial models is negligible. For further details on that study, see Franklin (1999a).

In accordance with the sampling design, the normed initial weights are used to compute a weighted chi-square statistic. It should be noted that SLID does not distinguish between the various types of nonresponse (attrition, not traced, refusal, etc.). Also important is the fact that some constraints are applied to the model during segmentation modelling: each RHG must contain at least 30 people (weighted and unweighted) and have a weighted response rate of at least 50%. Kalton and Kasprzyk (1986) recommend these constraints because they prevent excessive variability in the weights and reduce loss of precision in the estimates.

Once the decision tree has been created, the RHGs are formed using the ends of the branches. The individuals used to model nonresponse (i.e., respondent adults and nonrespondent adults in nonrespondent households) and nonrespondent adults in respondent households are placed in the various RHGs. Note that even though nonrespondent adults in respondent households are deemed respondents, they are not used in nonresponse modelling because they did not actually respond to the survey.

The nonresponse adjustment is carried out within each RHG: the weights of the respondents in each RHG are divided by the RHG's weighted response rate (which is the same as multiplying the weights of the respondents by an adjustment factor equal to the inverse weighted response rate). For wave 5 of Panel 1, there were 282 RHGs, with an average weighted response rate of 84%. For wave 2 of Panel 2, there were 134 RHGs, with an average weighted response rate of 88%.

The nonresponse-adjusted weights are obtained by multiplying the initial weight of each respondent individual by the nonresponse adjustment factor (or dividing by the weighted response rate for the final RHG). The weights of nonrespondents are set to 0, and the initial weights of children and cross-sectionally out-of-scope individuals are left unadjusted.

In each wave, the nonresponse-adjusted longitudinal weights are derived as follows:

$$w_{adjust} = \begin{cases} 0, & \textit{for persons in non-respondent households} \\ w_{initial}, & \textit{for children or cross-sectionally out-of-scope persons} \\ \dfrac{w_{initial}}{R_{RHG}}, & \textit{for persons in respondent households} \end{cases}$$

where w_{adjust} = nonresponse-adjusted longitudinal weight
$w_{initial}$ = initial longitudinal weight (= $w_{initial,p_1}$ for Panel 1, $w_{initial,p_2}$ for Panel 2)
R_{RHG} = weighted response rate in the RHG

3.4 Adjustment for Influential Values

Once the weights have been adjusted to compensate for nonresponse, the next step is to identify influential values that have a major impact on cross-sectional income estimates and variance estimates. A value is considered influential if it has an excessive effect on the cross-sectional estimate of total provincial income. An adjustment factor between 0 and 1 is computed and applied to the weights of individuals identified as influential (and to the weights of the other members of the household) to reduce the influence their weights have on income estimates. There were no influential individuals in wave 5 of Panel 1, and only two influential individuals in wave 2 of Panel 2.

Only influential values affecting the cross-sectional estimates are identified in this way. The same procedure could be followed for influential values affecting longitudinal weights. This might result in adjustments that were different for longitudinal and cross-sectional weighting. However, tests for the 1995 reference year have shown that at that time, there would have been no difference in the individuals identified as influential and very little difference in the adjustment factors. Consequently, the decision was made to use the cross-sectional adjustments for longitudinal weighting as well. For further details on the identification of influential values and the calculation of the adjustment factor, see subsection 4.6.

The longitudinal weight adjusted for influential values is defined as follows:

$$w_{infl} = w_{adjust} \times \beta_{infl}$$

where w_{infl} = longitudinal weight after adjustment for influential values
w_{adjust} = nonresponse-adjusted longitudinal weight
β_{infl} = adjustment factor for influential values ($0 < \beta_{infl} \leq 1$).

3.5 Post-stratification

The purpose of post-stratification is to ensure that the sum of the weights within certain sample subgroups (post-strata) matches known control totals for the population of those post-strata for a given year. Post-stratification is carried out independently for the two panels. For each post-stratum, the sum of the weights prior to post-stratification is calculated, and the control total for the population is divided by that sum. Then the longitudinal weight following post-stratification is computed using the formula

$$w_{ps} = w_{infl} \times \frac{T_L}{\sum_L w_{infl}}$$

where w_{ps} = longitudinal weight after post-stratification
w_{infl} = longitudinal weight adjusted for nonresponse and influential values
T_L = control total for post-stratum L

For the longitudinal weighting of SLID, the post-strata consist of province-sex-age group cross classifications. The age groups are shown in Table 3.2.

Table 3.2: Age Groups Used for Post-stratification

Age groups	
Panel 1 (8 age groups)	**Panel 2 (11 age groups)**
0-15	0-6
	7-15
16-19	16-18
20-24	19-24
25-34	25-34
35-44	35-44
45-54	45-54
55-64	55-59
	60-64
65+	65-69
	70+

For the 1996 reference year (Panel 2), substantial changes were made in the age groups used to process data. This was done to streamline the integration of SCF into SLID (cross-sectionally) and to comply with certain standards adopted by a number of major surveys at Statistics Canada. The new age groups went into effect in the 1996 reference year and were used for Panel 2 from the outset since it was selected in 1996. For Panel 1, the old age groups continued to be used for post-stratification to ensure stability and comparability of the various sets of weights produced during the waves prior to 1996.

The province used for longitudinal post-stratification is the one in which the individual lived at the time the sample for a given panel was selected, and the age group to which an individual belongs is determined by his or her age at the time the panel was selected. In addition, Panel 1's longitudinal weights after post-stratification must match the December 31, 1992, control totals for the Canadian population, while Panel 2's weights must match the December 31, 1995, control totals. The control totals are based on population projections generated by the Demography Division. The projections are only available in the middle of each month (and thus represent the population of Canada's ten provinces, with certain exclusions, at the middle of the month). SLID uses the projections for January.

The average post-stratification adjustment applied to longitudinal weights adjusted for nonresponse and influential values is 1.11 for wave 5 of Panel 1 and 1.10 for wave 2 of Panel 2.

3.6 Addition of Noise

At the time the longitudinal and cross-sectional weights for the 1997 reference year were computed, two public-use microdata files – a longitudinal file and a cross-sectional file – were produced each year. (As of the 1998 reference year, this is no longer done.) To safeguard confidentiality, it is necessary to ensure that data in the two files cannot be used to reconstruct households. Despite the adjustments described in the preceding sections, it is possible for individuals from the same household to have the same weight, which can result in at least partial reconstruction of a household. To prevent this, a final step in the weighting process is carried out: noise is added to the weights on the public-use files. The magnitude of the noise added is determined from the distribution of differences between consecutive weights after they have been sorted in ascending order of magnitude.

There are several steps in the process of introducing noise into the post-stratified weights. First, all persons who are from the same original household and have the same weight, regardless of whether they still live together, are identified. Then they are grouped two by two. Where an odd number of persons in a household has the same weight, the weight of one person chosen at random is not altered. For each pair, a random value e is generated from a uniform distribution U(0,1) and added to the value a, determined as follows:

$$a . \frac{\max(w_{ps}) \& \min(w_{ps})}{n}$$

where n is the sample size and $\max(w_{ps})$ and $\min(w_{ps})$ are computed for the Canada level. The value $e+a$, which falls between a and $a+1$, is then added to the weight of the first person in the pair and subtracted from the weight of the second. Thus, noise is added to the weights, and identical weights in the same household are eliminated. The control totals for each post-stratum are respected since individuals from the same household who have the same weight before the addition of noise necessarily come from the same post-stratum and thus have the same adjustments for nonresponse, influential values and post-stratification.

The weight following noise addition, known as the external weight, is obtained as follows:

$$w_{noise} ' \begin{cases} w_{ps} \pm (e\%a), & \textit{for pairs of individuals in the same household whose weights are identical} \\ w_{ps}, & \textit{in all other cases} \end{cases}$$

where
- w_{noise} = longitudinal weight after noise addition
- w_{ps} = longitudinal weight after post-stratification
- e = random noise
- a = longitudinal noise

The value of a for the 1995 and 1996 reference years (1.25) was also used for the 1997 reference year (both panels). For more details, see Franklin and Lévesque (1999).

3.7 Description of the Final Longitudinal Weights

Two sets of longitudinal weights are produced: a set of internal weights and a set of external weights. The internal weights, whose production is complete at the end of subsection 3.5 above, are used in studies conducted within Statistics Canada; no noise has been added to them. Noise is only added to the external weights, which are intended for users outside Statistics Canada. These are the weights whose production is complete at the end of subsection 3.6.

Table 3.3 shows the median initial and final longitudinal weights by province for the 1997 reference year. Only individuals with non-zero final weights are included in the calculation of the medians. The weights considered here are internal weights (i.e., no noise added). The province is the individual's province of origin (i.e., the province in which the individual was living when the sample for the panel to which he belongs was selected).

Table 3.3: Median Internal Longitudinal Weights (Initial and Final), for Wave 5 of Panel 1 and Wave 2 of Panel 2

Province	Panel 1		Panel 2	
	Initial weights	Final weights	Initial weights	Final weights
Newfoundland	207	250	242	285
Prince Edward Island	138	164	95	107
Nova Scotia	237	305	281	322
New Brunswick	279	321	254	296
Quebec	565	699	566	676
Ontario	563	769	558	718
Manitoba	221	309	327	409
Saskatchewan	305	370	279	347
Alberta	783	870	693	967
British Columbia	661	848	811	1117

4. Cross-sectional Weighting

This section provides a detailed description of the steps required to create a set of weights representative of Canada's ten provinces on December 31 of a given reference year. All cross-sectionally in-scope longitudinal individuals and their cohabitants are included in the production of this set of weights. The methodology described here is the one used for the 1997 reference year.

SLID produces various sets of cross-sectional weights. These sets are composed of individual or integrated weights. While each member of a household may have a different individual weight, every member of that household has the same integrated weight. Integrated weights are suitable for household-level studies, but they can also be used for studies at the person level. Individual weights, on the other hand, can only be used at the person level. Some sets of weights are produced exclusively for use within Statistics Canada (internal weights), while others are produced for users outside the Agency (external weights). For the sake of confidentiality, noise is added to all external weights. In producing cross-sectional weights for the 1997 reference year, it was important to ensure that the public-use microdata files to be produced could not be used to reconstruct households (to safeguard confidentiality). For that reason, integrated weights are available only within Statistics Canada. For studies conducted outside the Agency, individual weights must be used.

In all, four sets of cross-sectional weights are produced. The first set consists of *integrated weights*, which, as noted above, are internal weights. The second set is composed of *usual external individual weights*. The third set is made up of *internal labour weights*. The latter are individual weights similar to usual individual weights, except that the concept of respondent used in producing them is somewhat different. Respondent identification is based on the labour interview alone rather than both the labour and income interviews. For example, a household that is nonrespondent at the labour interview (i.e., all members of the household are nonrespondents at that interview) but respondent at the income interview (because the members preauthorized access to their tax data) will be included in the usual individual weighting but not in the labour weighting. Using labour weights rather than integrated weights or usual individual weights provides a better estimate of labour characteristics. The fourth set of weights consists of *external labour weights*.

The cross-sectional weighting process involves the following eight steps: determination of individuals eligible for cross-sectional weighting (subsection 4.1), nonresponse adjustment (4.2), application of panel allocation factors (4.3), weight share (4.4), interprovincial migration adjustment (4.5), adjustment for influential values (4.6), post-stratification (4.7) and addition of noise (4.8). The final cross-sectional weights are described in subsection 4.9.

4.1 Determination of Individuals who are Eligible for Cross-sectional Weighting

The cross-sectional sample for a given year is representative of the population of Canada's ten provinces on December 31 of the reference year (except for households composed entirely of immigrants). It consists of longitudinal individuals who are still members of the population of Canada's ten provinces as of December 31 of the reference year, plus their cohabitants. These individuals are considered to be *cross-sectionally eligible*.

Before we can determine which individuals have a non-zero cross-sectional weight, we must first identify those who are cross-sectionally eligible (in scope). All longitudinal individuals who are no longer members of the target population on December 31 of the reference year (e.g., persons who have moved to a place outside the ten Canadian provinces, deceased persons, and persons living in an institution) are assigned a cross-sectional weight of zero, even if they are respondent (or deemed respondent). Of the longitudinal individuals who belong to the target population on December 31 of the reference year, only those who are members of respondent households have non-zero cross-sectional weights. For the purposes of producing usual individual weights and integrated weights, a household is defined as respondent if at least one member responded in at least one of the two interviews (the labour interview or the income interview). Information about nonrespondent persons in respondent households will be imputed. For the purposes of labour weights, a household is respondent if at least one member responded in the labour interview.

4.2 Nonresponse Adjustment

As noted above, the cross-sectional sample for a given year is representative of the population of Canada's ten provinces on December 31 of the reference year. Like any other survey, however, SLID is subject to nonresponse (in this case, we are referring to total nonresponse), which affects the representativeness of the sample. To mitigate the effects of nonresponse bias, we adjust the weights of respondents to compensate for nonrespondents.

From a cross-sectional standpoint, we have too little information about nonrespondent cohabitants to adjust for nonresponse at that level. For many cohabitants, the only available variables are province, age and sex, which in any case will be used in the post-stratification step. To circumvent this problem, two methods of cross-sectional nonresponse adjustment have been proposed (Latouche, Michaud, 1997). One solution is first to make a nonresponse adjustment for longitudinal individuals only and then to use the weight share method to assign a weight to cohabitants. The weights of cohabitants are thus indirectly adjusted for nonresponse. The other method is to adjust for nonresponse at the household level using the initial dwelling strata and rotation groups of the LFS sample design. However, household-level adjustment tends to underestimate the response rates slightly.

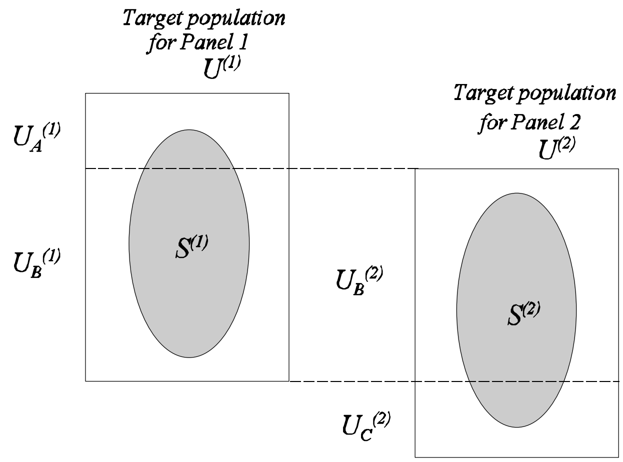
SLID adopted the first method described above. Consequently, the cross-sectional nonresponse adjustment step is not related to cross-sectional weighting but to longitudinal weighting. Longitudinal nonresponse modelling and adjustment were described in subsection 3.3 above. As indicated there, for longitudinal weighting each panel is treated separately. The cross-sectional weight adjusted for longitudinal nonresponse is therefore equal to the weight w_{adjust} derived in subsection 3.3. We denote the cross-sectional weight adjusted for longitudinal nonresponse as $w_{1997,adjust}$. This weight is used as the initial weight for the production of all derived cross-sectional weights (usual individual weight, integrated weight and labour weight).

4.3 Application of Panel Allocation Factors

The 1996 reference year was the first year for which SLID had overlapping panels. In theory, the two panels could be used independently of one another to produce cross-sectional estimates, but in practice, it is better to combine the two cross-sectional samples to produce the cross-sectional estimates. This doubles the sample size, which in turn reduces the variability of the estimates and permits the use of more control totals in the post-stratification step.

Since the two longitudinal samples are representative of the population of Canada's ten provinces at different reference periods (December 31, 1992, for Panel 1, and December 31, 1995, for Panel 2), a way has to be found to combine the two samples so that they both represent the target population for a given reference year. The problem that arises is the time lapse between them. Figure 4.1 shows how the population of the ten provinces changed between the selection of Panel 1 and the selection of Panel 2.

Figure 4.1: Change in the Population of Canada's Ten Provinces Between the Selection Dates of the Two Panels



Sample $S^{(1)}$ was selected from the population of Canada's ten provinces $U^{(1)}$ in January 1993. (This was the longitudinal sample for Panel 1.) Sample $S^{(2)}$ was selected from the population of Canada's ten provinces $U^{(2)}$ in January 1996 to form the longitudinal sample of Panel 2. Most individuals present in the January 1993 population are also present in the January 1996 population ($U_B^{(1)}=U_B^{(2)}$). The differences between populations $U^{(1)}$ and $U^{(2)}$ are due to departures from the target population between January 1993 and January 1996 (deaths, persons placed in institutions, persons who moved to places outside the provinces, etc.) and to additions to the population after January 1993 (immigrants, births). Part of the Panel 1 sample is composed of persons who left the population between January 1993 and January 1996 ($U_A^{(1)}$); since those individuals were no longer members of the Canadian population when the sample for Panel 2 was selected, they had no chance of being selected. Conversely, part of the Panel 2 sample is made up of persons who were in the population of Canada's ten provinces when that sample was selected but were not in the population when Panel 1 was selected ($U_C^{(2)}$); those persons had no chance of being selected for Panel 1.

To develop a method of combining the information from the cross-sectional samples of the two panels, we need only consider the two panels at the time Panel 2 was introduced. The population of interest is $U^{(2)}$, the target population at the time Panel 2 was introduced. The longitudinal samples of interest are the original samples selected, with the Panel 1 sample having been reduced by population departures and nonrespondents. SLID uses Merkouris's method (1999) to match overlapping samples (in this case, the samples for Panels 1 and 2). If Y is a population parameter, we can use the following combined estimator (which assigns a relative importance to each panel based on a certain factor p):

$$\hat{Y} = p\hat{Y}_1 + (1-p)\hat{Y}_2$$

where $\hat{Y}$ is the cross-sectional estimate for a variable of interest, $\hat{Y}_i$ is the estimate produced for that variable for Panel i, and p is the panel allocation factor. This estimator can be regarded as a weighted sum of the estimators of the two panels, where the weights are the panel allocation factors. The optimum value of the panel allocation factor was set so as to minimize the variance of $\hat{Y}$. The panel allocation factor is defined as follows:

$$p = \left(\frac{n_1}{d_1}\right)\left(\frac{n_1}{d_1} + \frac{n_2}{d_2}\right)^{-1} = \frac{n_1}{n_1 + n_2\left(\frac{d_1}{d_2}\right)} \qquad (4.1)$$

where d_i denotes the design effect for Panel i and n_i represents the number of Panel i longitudinal persons aged 16 and over who were respondent at time t (where t is the reference year concerned, which in this case is 1997). The design effect is defined as the ratio of the variance of the subweighted estimate (i.e. without benchmarking to population control totals) to the variance associated with a simple random sample of the same size. For SLID, the number of longitudinal persons aged 16 and over who were respondent at time t is the same as the number of cross-sectionally in-scope longitudinal persons aged 16 and over who are in a respondent household (since all members of respondent households are deemed respondent (see subsection 3.2)). Even though cohabitants are used to produce the cross-sectional estimates, only longitudinal persons respondent at time t are considered in computing the panel allocation factors. Merkouris (1999) and Lavallée (1994) suggest combining the two panels before incorporating cohabitants by the weight share method. Leaving cohabitants out of the computation of factor p helps to reduce the importance of the older panel and thus to minimize the potential effects of attrition bias.

To determine the optimum factor p, we need to estimate the design effects and the longitudinal sample sizes. However, no design effect was available for Panel 2 at the time of data processing. Since SLID is a subsample of the LFS, it was decided to use LFS data to derive the design effects needed to compute the p factors. Moreover, since we are interested in the ratio of two design effects, the LFS is a perfectly legitimate source for the required effects. The LFS's six rotation groups are used to determine the design effects. The effects are calculated for the *number of persons aged 15 and over* characteristic at the provincial level (which will require the calculation of ten panel allocation factors, one per province). For Panels 1 and 2 combined, the LFS files for January 1993 and January 1996 were used to compute the design effects. The required sample sizes were computed from the longitudinal samples of the two panels for the reference year concerned. As noted above, the sample sizes are the number of longitudinal persons aged 16 and over who were respondent at time t. They are obtained by counting all cross-sectionally in-scope persons aged 16 and over in respondent households.

Once the design effects and the sample sizes are known, it is easy to compute the panel allocation factors using equation 4.1. The factors must be calculated for each year involved since the composition of the longitudinal samples changes each year. For practical reasons, it was decided to compute the design effects at the inception of each new panel and to use those design effects in calculating the panel allocation factors for the three overlap years (Latouche et al., 2000). The panel allocation factors for the 1997 reference year are presented in Table 4.1 below.

Table 4.1: Panel Allocation Factors for the 1997 Reference Year

Province	d_1 / d_2	n_1	n_2	p (Panel 1)	$1 \& p$ (Panel 2)
Newfoundland	1.71	1693	1306	0.4312	0.5688
Prince Edward Island	1.92	568	883	0.25095	0.74905
Nova Scotia	4.02	1788	1986	0.18298	0.81702
New Brunswick	1.94	1752	1819	0.33176	0.66824
Quebec	2.75	4908	5788	0.23568	0.76432
Ontario	2.87	6893	8901	0.21249	0.78751
Manitoba	1.36	1776	2111	0.38219	0.61781
Saskatchewan	2.21	1915	1872	0.31642	0.68358
Alberta	2.87	2378	2113	0.28168	0.71832
British Columbia	2.76	2238	2509	0.24425	0.75575

d_i = design effect for Panel i
n_i = number of Panel i respondent longitudinal persons aged 16 and over for the 1997 reference year
p, 1 - p = panel allocation factors

The table shows that the factors favour Panel 2 by a ratio of 3 to 1. In the most recent postcensal redesign of the LFS (completed in late 1994), its sample design was improved; as a result, the ratio d_1 / d_2 is high, which gives Panel 2 greater importance. When Panels 2 and 3 of SLID are combined, the ratio of the design effects will be closer to 1 since both panels will have been selected from the same LFS sample design. The panel allocation factors will be closer to 0.5, and the two panels will have similar importance.

The *p* factors are applied to all Panel 1 longitudinal persons who were in a respondent household at time t (in this case, December 31, 1997). The *1-p* factors should be applied to all Panel 2 longitudinal persons who were in a respondent household at time t, except persons who were outside the target population when Panel 1 was introduced (January 1993) and in the target population when Panel 2 was introduced (January 1996). For practical reasons, only newborns and immigrants who were not present when Panel 1 was introduced but were present when Panel 2 was introduced are exempted from application of the *1-p* factors. Those excluded will retain their weights as adjusted for longitudinal nonresponse. Accordingly, the cross-sectional weight adjusted for the combination of the two panels is as follows:

$$w_{1997,p} \,' \begin{cases} p\ w_{1997,adjust}, & \text{for Panel 1 persons} \\ (1\&p)\ w_{1997,adjust}, & \text{for Panel 2 persons not exempted from application of the panel allocation factors} \\ w_{1997,adjust}, & \text{for Panel 2 persons exempted from application of the panel allocation factors} \end{cases}$$

where $w_{1997,p}$ = cross-sectional weight adjusted for the combination of the two panels
$w_{1997,adjust}$ = cross-sectional weight adjusted for longitudinal nonresponse
p = panel allocation factor

This weight is identical for the usual individual and integrated cross-sectional weights. For the labour cross-sectional weight, the weight adjusted for the combination of the panels is computed in the same way, but the definition of respondent household is different (subsection 4.1). Note that the panel allocation factors are applied to the weights after the nonresponse adjustment step and before the weight share step, so that they will be included in the weights of cohabitants.

4.4 Weight Share

The weight share step is essential and unique to cross-sectional weighting. It is necessitated by the longitudinal nature of SLID. In a longitudinal survey, the sample evolves over time. New people move in with the selected individuals, and children are born into some of the sample households. In short, persons who were not originally selected become part of the sample. Since we do not have any probability of selection for these new persons (cohabitants), we have to find a way to assign them a weight. This is done by the weight share method. A brief explanation of this method is presented below. For further details, the reader should refer to the article by Lavallée (1995).

From the weight share step onward, no distinction is made between the two panels, and all individuals are treated as part of same sample, though the concepts of longitudinal individual and cohabitant remain. One might think that since we choose to perform the weight share at the household level, there would be no difference between treating the two cross-sectional samples separately or together. It is important to keep in mind that in a survey with multiple overlapping panels, some households may contain persons from Panel 1 and persons from Panel 2. When the longitudinal sample for Panel 2 was selected, any household containing at least one longitudinal person from Panel 1 was excluded in order to avoid surveying the same household for another six years. Because of the survey's dynamic, longitudinal nature, however, over a six-year period households may break up and new households may form. As a result, there is a possibility that the survey will include households with members from the two panels. If the cross-sectional samples are considered separately, two distinct sets of weights would be produced for such households, and this would cause all sorts of problems. For this reason, it is better to consider the two panels together.

It must first be determined at what level in the sample design hierarchy the weight share is to be performed: province, stratum, primary sampling unit or final sampling unit. SLID elected to perform the weight share at the level of the final sampling unit, the household. Once that decision is made, it is necessary to determine the status of each person in the cross-sectional sample: longitudinal individual, initially present cohabitant or initially absent cohabitant.

A cohabitant is initially present if he or she was a member of the target population when Panel 2 was selected (January 1996). Conversely, a cohabitant is initially absent if he or she was not a member of the target population when Panel 2 was selected (January 1996). Now these concepts – initially absent and initially present cohabitant – must reflect the fact that there are two overlapping panels in the survey and that these panels must be considered together at the time of Panel 2's selection (January 1996). It is important to note that all cohabitants have a longitudinal weight of zero since they are not part of the longitudinal sample. They may, however, have a non-zero cross-sectional weight, assigned by the weight share method.

In the weight share step, two different weights are produced for each individual: an individual weight (which may be different for each member of a given household) and an integrated weight (which is the same for all members of a given household).

In the case of the weight share for individual weights, the following steps are required. First, the households to which cohabitants have moved since the longitudinal sample was selected must be identified. For households composed entirely of longitudinal persons, the weight share method is not needed. For all members of such households, the shared cross-sectional weight is identical to the adjusted cross-sectional weight after the panel allocation factors are applied.

Second, for households with new cohabitants, the nonresponse-adjusted cross-sectional weights $w_{1997,p}$ of all longitudinal household members must be added together. Then the longitudinal individuals and the initially present cohabitants must be counted.

For households with at least one cohabitant initially present, the shared cross-sectional weight of all members is the sum of the cross-sectional weights $w_{1997,p}$ of all longitudinal members divided by the total number of longitudinal individuals and initially present cohabitants in the household.

For households with all cohabitants initially absent, the shared cross-sectional weight of longitudinal individuals and cohabitants is calculated in a different way. For longitudinal members of such households, the shared cross-sectional weight is equal to the adjusted cross-sectional weight following application of the panel allocation factors. For cohabitants, the shared cross-sectional weight is equal to the sum of the weights $w_{1997,p}$ of all longitudinal members divided by the total number of longitudinal members. Thus, the longitudinal individuals will not be assigned an average weight, and the nonresponse adjustment will be preserved. Only initially absent individuals will be assigned the average weight of the household's longitudinal members.

The shared individual cross-sectional weight (usual individual weight or labour weight) is defined as follows:

$$w_{1997,shared} = \begin{cases} w_{1997,p}, & \textit{for persons in households composed entirely of longitudinal persons} \\ & \textit{and for longitudinal members of households in which all cohabitants are IA} \\ \dfrac{\sum_h w_{1997,p}}{n_{L,h} + n_{IP,h}}, & \textit{for persons in households with at least one cohabitant IP} \\ & \textit{and for cohabitants in households where all cohabitants are IA} \end{cases}$$

where
$w_{1997,shared}$ = shared cross-sectional weight
$w_{1997,p}$ = cross-sectional weight adjusted for the combination of the two panels
IP = initially present
IA = initially absent
$n_{L,h}$ = number of longitudinal persons in household *h*
$n_{IP,h}$ = number of cohabitants initially present in household *h*

In the case of the weight share for integrated weights, the process is somewhat different. Every individual with a non-zero individual cross-sectional weight (usual individual weight only) also has a non-zero integrated cross-sectional weight. This is necessary because the weight prior to post-stratification must be identical for all members of a household so that an integrated weight can be produced (Lemaître, Dufour, 1987). The difference between integrated weights and individual weights derives from the fact that weight share has to be performed for all households, whether they contain cohabitants or not. Consequently, for all households the shared weight is computed in the same way for all individuals when the weight produced is integrated. The integrated shared cross-sectional weight is therefore defined as follows:

$$w_{1997,shared} = \frac{\sum_h w_{1997,p}}{n_{L,h} + n_{IP,h}}$$

where
- $w_{1997,shared}$ = shared cross-sectional weight
- $w_{1997,p}$ = adjusted cross-sectional weight after combination of the two panels
- $n_{L,h}$ = number of longitudinal persons in household *h*
- $n_{IP,h}$ = number of cohabitants initially present in household *h*

For the 1997 reference year, there were 80,834 persons in the cross-sectional sample. This total includes respondent and nonrespondent persons (for nonrespondent households, the last known household composition is used). It understates the actual number of persons in the cross-sectional sample because there is no information about the current composition of nonrespondent households. Of the 80,834 persons, 79,218 have non-zero usual individual and integrated cross-sectional weights, and 76,421 have non-zero labour cross-sectional weights.

4.5 Adjustment for Interprovincial Migration

Like the weight share step, the interprovincial migration adjustment step is directly related to the longitudinal nature of SLID. Over time, some people in the sample move from one province to another. Since the probability of selection can vary widely with geographical location, there is a substantial difference at the national level between the largest weights and the smallest ones. As a result, if a person with a fairly low selection probability (and therefore a fairly large weight) moves to a location where a given individual's selection probability is much higher (and that individual's weight is much smaller), problems can arise. In particular, certain cases of interprovincial migration can distort the results of analytical studies.

A simple example will help explain the problem. Suppose a heart surgeon moves from Toronto to Charlottetown. This would cause several difficulties. First, the surgeon's weight of about 400 would be much larger than any of the weights in Prince Edward Island, which average close to 65. The surgeon alone would immediately represent approximately 3.33% of the province's population, or about 5,000 people out of 150,000, which is completely unacceptable. Furthermore, from an analytic standpoint, he would artificially inflate the population's total income because of his weight of 5,000 and his salary.

A number of options are available to alleviate the problem of outlier weights due to interprovincial migration. Such weights could be replaced with the average weight, the median weight, a quantile of the weight distribution of in the province of destination, or some other value based on that distribution. The alternatives are arbitrary, and any of them could be used. What is needed is a compromise between assigning the migrant a weight similar to the average weight in the province of destination and excessively altering the migrant's original weight (based on selection probability). We must first identify those persons who have migrated to other provinces and whose weights are causing an imbalance in the estimates for their provinces of destination, and then adjust the weights.

The method adopted by SLID uses the 95th percentile of the distribution of weights. This is a quick, simple method that requires little or no manual intervention. First it is necessary to identify every person who has moved to another province and whose weight is larger than the largest weight of the province's current population. The weights of such persons are then set to the 95th percentile of the province's weight distribution. Next, a random number between 0.00 and 10.00 is subtracted from that percentile to ensure that the weights remain nearly unique (two people can have the same individual weight, but such cases are very rare).

Interprovincial migration adjustments are made independently for individual and integrated weights and are recalculated each year. Integrated weights are adjusted at the household level to keep the weights in a household equal. The cross-sectional weight adjusted for interprovincial migration is defined as follows:

$$w_{1997,mig} = w_{1997,shared} \times a_{mig}$$

where $w_{1997,mig}$ = cross-sectional weight adjusted for interprovincial migration
$w_{1997,shared}$ = shared cross-sectional weight
a_{mig} = interprovincial migration adjustment factor

For the 1997 reference year, an interprovincial migration adjustment was required for the usual individual cross-sectional weights of 36 persons, for the integrated cross-sectional weights of 15 households (43 persons), and for the labour cross-sectional weights of 34 persons.

4.6 Adjustment for Influential Values

Like any survey that attempts to estimate income distribution, SLID is very likely to face the problem of influential values (such distributions have very long tails). Influential values have a significant impact on provincial estimates of total and average income and their variance estimates. The impact is even greater on estimates for smaller domains such as age group, sex or family size. Hence, methods must be developed to identify the influential values and then adjust them so as to reduce the effect they have on estimates.

The process of identifying influential values and calculating adjustment factors applies only to individual weights (usual individual weights or labour weights), while the adjustment factors produced are used on both individual and integrated weights.

SCF devised a method of detecting and processing influential values (Tremblay, 1998). Since it was known that SCF would be incorporated into SLID for the 1999 collection period, the methodology takes into account the fact that the successor survey is longitudinal.

Accordingly, SLID adopted this approach to deal with the problem of influential incomes. The first step is to identify the influential items, that is, observations that contribute too heavily to the weighted estimate of total personal income for the province. Such observations usually occur in upper income brackets. An individual's income (in this case, total income less capital gains) is regarded as influential if the contribution of the individual's weighted income to the provincial weighted estimate exceeds a certain threshold. This threshold is equal to the lower limit of the interval computed by the quartile method. It is defined as MED + k (Q3 - MED), where MED and Q3 are the median and the third quartile, respectively, of the weighted contribution of individuals, and k is a fixed parameter for each province, determined following a review of the highest percentages of contribution to the provincial estimate. (The value of k is 75 for Quebec and Ontario and 25 for all other provinces.)

Having identified the influential values, we must now compute the adjustment factors for their weights. To do so, we determine whether incomes over $99,999 are overrepresented in the sample compared with the distribution of taxfiler incomes from Revenue Canada's T1 file. If so, the weight is adjusted so that the distribution of the weighted sample matches that of the tax data (in order to keep weight adjustments to a minimum, this adjustment is not made in cases where the difference between the two distributions is small). The adjustment is made with an adjustment factor whose value is between 0 and 1. A separate adjustment factor is computed for each individual with an influential income, as described below. Note that the calculations are done at the provincial level.

> [Translation]
> *If a province has more than one influential person, the adjustment factor is computed first for the influential person with the highest income (unweighted). Suppose the person is in income class c. The adjustment will ensure that the estimate of the number of persons in class c or any higher income class is equal to the number of persons in the same income classes based on tax data, subject to certain constraints. First, only the weights of influential persons will be changed by this adjustment. In addition, if the adjustment increases the weight, i.e., the adjustment factor is greater than 1, the adjustment will not be applied and the initial weight will be retained. If the tax file shows that there is no one in the income classes, the adjustment will lower the weight to 1. Similarly, the weight will be reduced to 1 if the estimated number of persons in these classes based on non-influential persons only is greater than the number of persons according to the tax file.* (Tremblay, 1998).

Calculation of the adjustment factor carries over from one individual to the next. When the adjustment factor has been computed for the influential individual with the highest income, that individual's adjusted weight is considered in computing the adjustment factor for the next individual.

Although the adjustment factor is calculated at the person level, it has to be applied to the weights of the influential individual and all other members of the same household, for the duration of panel even if the person is no longer considered influential in waves following the one in which he or she was identified. The longitudinal nature of the survey makes this procedure necessary in order to maintain consistency over time, even with respect to cross-sectional weighting. For households in which no individuals are identified as having extreme income values, the adjustment factor is 1 for all household members.

To detect influential values, the method employs the weight obtained after post-stratification (to be discussed in the next subsection) since the latter is based on income distribution.

Hence, temporary post-stratification must be carried out before the adjustment for influential values so that the temporary post-stratified weights can be used to identify and process the influential values. Once the adjustment factors have been computed, post-stratification must be redone; the factors are applied to the weights that we had before the temporary post-stratification (that is, to the cross-sectional weights adjusted for interprovincial migration $w_{1997,mig}$).

The cross-sectional weight adjusted for influential values is computed as follows:

$$w_{1997,infl} = w_{1997,mig} \times \beta_{infl}$$

where $w_{1997,infl}$ = cross-sectional weight adjusted for influential values
$w_{1997,mig}$ = cross-sectional weight adjusted for interprovincial migration
β_{infl} = adjustment factor for influential values

For the 1997 reference year, two individuals were identified as having influential incomes. In addition, the weights of seven other persons identified as influential in previous waves had to be adjusted again. When other members of the households are counted, the adjustment was applied to a total of 29 people for the 1997 reference year. The adjustment factors ranged from 0.39 to 0.93, and most of them were in the vicinity of 0.50.

4.7 Post-stratification

As noted earlier, the purpose of post-stratification is to ensure that the sum of the weights in certain sample subgroups (post-strata) match the known control totals for the population of the post-strata for a given year.

For individual cross-sectional weights, post-stratification is carried out in the same way as in the longitudinal weighting process (subsection 3.5). The weight after post-stratification is equal to the weight before post-stratification multiplied by the control total for the population of the post-stratum to which the individual belongs, divided by the sum of the post-stratum's weights before post-stratification.

The method for producing the integrated weight is slightly more complex. Aside from matching the control totals for each post-stratum, post-stratification must also ensure that all individuals in the same household have the same final weight. For further information on how this is done, see the article by Lemaître and Dufour (1987). Let $w_{1997,ps}$ be the cross-sectional weight after post-stratification.

Like the post-strata for longitudinal weighting, the post-strata for cross-sectional weighting are province-sex-age group cross classifications. Province here refers to province of residence on December 31 of the reference year. Similarly, the age group to which an individual belongs is determined by his or her age on December 31 of the reference year. The age groups are those presented earlier in Table 3.2.

For the 1997 reference year, the usual individual weights adjusted for influential values were multiplied by an average of 1.13 to calibrate SLID sample. The integrated cross-sectional weights were also multiplied by an average of 1.13 during post-stratification. The average adjustment made to labour cross-sectional weights to calibrate the sample was 1.18.

4.8 Addition of Noise

As in the case of longitudinal weights, noise is added to the cross-sectional weights to preserve confidentiality. Note that noise addition applies only to individual cross-sectional weights and not to integrated weights. The reason for this is that releasing integrated weights in public-use microdata files would conflict with the aim of preventing reconstruction of households with the microdata. Consequently, only individual weights are included in the public-use files and therefore require the addition of noise.

The procedure for adding noise to individual cross-sectional weights is the same as for longitudinal weights (see subsection 3.6). The random noise e, with a value between 0 and 1, and the cross-sectional noise a, both of which are added to the individual cross-sectional weights, are calculated in the same way as they are in longitudinal weighting. The noise-added individual cross-sectional weights are obtained as follows:

$$w_{noise} = \begin{cases} w_{ps} \pm (e \times a), & \textit{for pairs of individuals in the same household whose weights are identical} \\ w_{ps}, & \textit{in all other cases} \end{cases}$$

where
w_{noise} = cross-sectional weight after noise addition
w_{ps} = cross-sectional weight after post-stratification
e = random noise
a = cross-sectional noise

The value of a for the 1995 and 1996 reference years (1) was also used for the usual individual weights in the 1997 reference year. The value of a for the labour weights was 1.25.

4.9 Description of the Final Cross-sectional Weights

Cross-sectional weighting produces four different sets of cross-sectional weights: integrated weights, usual external individual weights, internal labour weights and external labour weights. The integrated weights and internal labour weights are final at the end of the post-stratification step (subsection 4.7), while the usual external individual weights and the external labour weights are final after the noise addition step (subsection 4.8). Table 4.2 presents the medians of the initial cross-sectional weights and of selected final cross-sectional weights for the 1997 reference year. Only individuals with non-zero final weights are used in computing these medians. The province indicated in the table refers to the province of residence on December 31, 1997.

Table 4.2: Median Initial and Final Cross-sectional Weights, 1997 Reference Year

Province	Initial weight	Final weights		
		Integrated weight	Usual external individual weight	Internal labour weight
Newfoundland	133	127	129	133
Prince Edward Island	71	71	75	77
Nova Scotia	105	113	113	117
New Brunswick	150	150	155	159
Quebec	270	309	309	318
Ontario	253	301	302	314
Manitoba	148	167	171	180
Saskatchewan	161	163	165	170
Alberta	261	303	302	316
British Columbia	305	421	411	426

5. Forthcoming Changes and Future Developments

To improve the quality and reliability of the estimates, SLID's longitudinal and cross-sectional weighting processes are revised regularly. In particular, the revisions take into account any changes in the survey's objectives, processing and dissemination.

One such change relates to SLID's microdata dissemination strategy. Before SCF was incorporated into SLID (the merger came after the 1997 reference year), the dissemination strategy consisted in the annual production of a set of public-use longitudinal and cross-sectional microdata files that cannot be used to reconstruct households. Now that SCF has been integrated with SLID, producing cross-sectional files that will satisfy the needs of SCF users has become a priority. To that end, the content of the public-use cross-sectional files must be redefined and must include identifiers that can be used to reconstruct households and families.

The dissemination of such cross-sectional files raises doubts about the dissemination of public-use longitudinal files. The risk of disclosure associated with longitudinal files (i.e., that they could be used to reconstruct households and families) would probably be too high, and it would probably be difficult to produce longitudinal files that would have a low potential for matching with cross-sectional files. In view of these priorities, SLID has decided to focus on the dissemination of cross-sectional microdata and to explore other options for making longitudinal data available to users outside Statistics Canada. Two possibilities are remote access to data and the research data centre. For more information about the disclosure control strategy, see the article by Nadeau, Gagnon and Latouche (1999).

Since the dissemination of a public-use longitudinal microdata file looks very unlikely, it no longer seems necessary to produce external longitudinal weights. In addition, since the new dissemination strategy allows for the construction of households and families from cross-sectional data, external individual cross-sectional weights are no longer needed. For the moment, the two sets of weights are still being produced, but the possibility of discontinuing them will be discussed in the near future.

Another change in the survey will be the introduction of labour data imputation primarily for persons who are nonrespondent in the labour interview but respondent in the income interview. Before SCF was merged with SLID, records in the public-use files were allowed to have missing labour characteristics. Now that the two surveys have been combined, we wish to publish complete files with no missing data. Consequently, the labour weights are of questionable value. Since imputation will be used in cases of nonresponse in the labour interview, it seems pointless to continue producing labour weights that compensate for nonresponse by including only labour interview respondents in the weighting process. Labour weights are still being produced, but more serious consideration will be given to discontinuing them once all the labour data imputation procedures are in place.

There are also plans to change the calibration of the cross-sectional sample. SLID is currently revising its calibration strategy in order to improve the estimates and harmonize concepts across the various household surveys. Up to now, SLID has used province-sex-age group control totals to calibrate its cross-sectional sample. Additional sets of control totals are planned: number of economic families of size 1, 2 and 3 or more by province, and income level (low, middle, high) by province.

The introduction of these new control totals raises questions about the definition of the sets of control totals used so far, especially the age-group totals. It has also led to a change in the method of calibrating the sample: post-stratification is to be replaced by a raking ratio method. SLID's methodology team is currently reviewing its strategy for smoothly incorporating these new totals into the existing calibration method. The impact that this change will have on the weights will also be studied thoroughly.

In the area of longitudinal weighting, plans call for the production of weights that will reflect both SLID panels. The challenge will be to define the longitudinal population for more than one panel. At present, longitudinal weighting treats the two panels separately, producing weights for each panel independently. Computing longitudinal weights with the respondents of both panels would increase the precision of the estimates for the purposes of longitudinal studies (since the sample size would be doubled). The method used to produce such weights would be based on the method used to combine SLID's cross-sectional samples.

6. Conclusion

SLID's goal – to provide longitudinal and cross-sectional data representative of the population of Canada's ten provinces for various characteristics related to labour and income dynamics – is a formidable challenge. The required longitudinal and cross-sectional weights are produced by the weighting processes described in detail in this document. A number of steps must be carried out to produce those weights: determination of initial weights, nonresponse adjustment, weight share (for cross-sectional weights only), interprovincial migration adjustment (for cross-sectional weights only), adjustment for influential values, post-stratification and addition of noise. These steps have been improved and optimized over the years, but as processing and dissemination strategies evolve, further refinements are needed, in particular to maintain data quality and reliability. These refinements include the introduction of new post-strata for cross-sectional sample calibration, the production of longitudinal weights for the two panels combined, and a change in the dissemination strategy that raises questions about the usefulness of certain sets of longitudinal and cross-sectional weights.

Acknowledgements

The authors wish to thank Christian Nadeau for all the support and valuable advice he provided throughout the preparation of this document. They are also grateful to Michel Latouche, Guy Laflamme and Martin St-Pierre for their constructive comments.

References

Cotton, C., Bishop, K., Giles, P., Hewer, P. and Saint-Pierre, Y. (1999). *A Comparison of the Results of the Survey of Labour and Income Dynamics (SLID) and the Survey of Consumer Finances (SCF) 1993-1997 : Update*, Statistics Canada publication NE75F0002MIE - 99007.

Dufour, J., Gagnon, F., Morin, Y., Renaud, M. and Särndal, C.E. (1998). *Measuring the Impact of Alternative Weighting Schemes for Longitudinal Data*, Proceedings of the Survey Research Methods Section, American Statistical Association.

Durning, A. (1994). *Weighting of the SLID Preliminary File*, internal document, Statistics Canada.

Franklin, S. (1999a). *Modelling SLID's Nonresponse at the Provincial Level*, internal document, Statistics Canada.

Franklin, S. (1999b). *Study of the Impact of Modifying SLID's Nonresponse Model*, internal document, Statistics Canada.

Franklin, S. and Lévesque, I. (1999). *L'ajout de bruit aux poids post-stratifiés*, internal document, Statistics Canada.

Gagnon, F. (1997). *Pondération longitudinale - Deuxième vague de l'EDTR*, internal document, Statistics Canada.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J. (1998). *Methodology ot the Canadian Labour Force Survey*, Statistics Canada, Catalogue 71-526-XPB.

Grondin, C. (1996). *Pondération longitudinale - Première vague de l'EDTR*, internal document, Statistics Canada.

Kalton, G. and Kasprzyk, D. (1986). *The Treatment of Missing Survey Data,* Survey Methodology, vol. 12, no 1, pp. 1-16.

Knowledge SEEKER IV, Version 4.2.2, Copyright 1996, Angoss International Limited.

Latouche, M., Dufour, J. and Merkouris, T. (2000). *SLID Cross-Sectional Weighting: Combining Two or More Panels*, internal document, Statistics Canada.

Latouche, M. and Michaud, S. (1997). *Concerns pertaining to weighting of longitudinal surveys*, Proceedings of the Section on Government Statistics and Section on Social Statistics, American Statistical Association.

Lavallée, P. (1994). *Ajout du second panel à l'EDTR: sélection et pondération*, internal document, Statistics Canada.

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

Lavigne, M. and Michaud, S. (1998). *General Aspects of the Survey of Labour Income Dynamics*, Income and Labour Dynamics Working Paper, Statistics Canada, 75F0002M No. 98-05.

Lemaître, G. and Dufour, J. (1987). *An Integrated Method for Weightint Persons and Families*, Survey Methodology, vol. 13, no 2, pp. 119-207.

Merkouris, T. (1999). *Cross-sectional Estimation in Multiple-Panel Household Surveys*, Methodology Branch Working Paper, Statistics Canada, HSMD-99-004E.

Nadeau, C., Gagnon, É. and Latouche, M. (1999). *Disclosure Control Strategy for the Release of Microdata in the Canadian Survey of Labour and Income Dynamics*, Proceedings of the Survey Research Methods Section, American Statistical Association.

Renaud, M. (1997). *Pondération transversale - Deuxième vague de l'EDTR*, internal document, Statistics Canada.

Singh, M.P., Drew, J.D., Gambino, J.G. and Mayda, F. (1990). *Methodology ot the Canadian Labour Force Survey*, Statistics Canada, Catalogue 71-526.

Statistique Canada (1997). *Income Distributions by size in Canada*, Statistics Canada publication NE13-207-XPB, pp. 52-59.

Tambay, J. L., Schiopu-Kratina, I., Mayda, J., Stukel, D. and Nadon, S. (1998). *Treatment of Nonresponse in Cycle Two of the National Population Health Survey*, Survey Methodology, vol. 24, no 2, pp. 147-156.

Tremblay, J. (1998). *Détection des observations influentes pour l'enquête sur les finances des consommateurs (EFC) et l'enquête sur la dynamique du travail et du revenu (EDTR)*, internal document, Statistics Canada.

Appendix A: List of Variables Used in the Nonresponse Adjustment Model

The following variables were used to model nonresponse for the first panel, wave 5:

Aborig	- Aboriginal person
Canada	- born in Canada
US	- born in US
UK	- born in UK
Europe	- born in Europe
Autpays	- born elsewhere (not in Canada, US, UK or Europe)
EOBR	- ethnic origin British
EOCAN	- ethnic origin Canadian
EOEUR	- ethnic origin European
EOFR	- ethnic origin French
IMSTAT	- immigration status
VISMIN	- visible minority
OWNER	- owns home
RURAL	- lives in a rural district (based on EA)
REMOTE	- lives in remote district (based on EA)
URBAIN	- lives in urban district (based on EA)
EMPLOYE	- employed
NEMPLOYE	- unemployed
NLF	- not in labour force
MAR_UL	- married or common law
SEP_DIV	- separated or divorced
VEUF	- widowed
CELIB	- bachelor (never married)
PROX	- preliminary interview was via a proxy respondent
REV0	- indicates if household income was missing in the preliminary interview. This variable was added in 1997 since it was realized that the nonresponse rate for this variable is high: 29%. In previous years, missing was in the same category as <$10,000 (REV1).
REV1	- household income <=$10,000
REV2	- household income > $10,000 and <=$25,000
REV3	- household income > $25,000 and <=$50,000
REV4	- household income > $50,000
PRESENF	- children present at the time of the preliminary interview
P10 – P59	- one variable for each province
HOMME	- sex variable
AGE_1	- age >=16 and <19
AGE_2	- age >=19 and <25
AGE_3	- age >=25 and <35
AGE_4	- age >=35 and <45
AGE_5	- age >=45 and <55
AGE_6	- age >=55 and <65
AGE_7	- age >=65
Educ1 – Educ9	- 9 levels of education

The following variables were used to model nonresponse for the second Panel:

Educ1	- 0-8 years of schooling
Educ2	- 9-13 years of schooling (i.e. secondary)
Educ3	- some post-secondary or a post secondary certificate, diploma
Educ4	- post secondary college or university (incl. PhD)
RENTER	- 1=owns home, 0=rents home
EMPLOYE	- employed
NON_EMPL	- unemployed
NLF	- not in labour force
MARIE	- married or common law
SEP_DIV	- separated or divorced
VEUF	- widowed
CELIB	- bachelor (never married)
P10 – P59	- one variable for each province
SEXE	- sex variable
AGE_1	- age >=16 and <19
AGE_2	- age >=19 and <25
AGE_3	- age >=25 and <35
AGE_4	- age >=35 and <45
AGE_5	- age >=45 and <55
AGE_6	- age >=55 and <65
AGE_7	- age >=65
HHSZ1-HHSZ5	- number of people in the household (HHSZ5=>5)
FAMTYPE1	- unattached individual
FAMTYPE2	- husband-wife or single parent families without young children
FAMTYPE3	- husband-wife or single parent families with children aged 0-17
FAMTYPE4	- other types of families
STUDENTS	- 1=student (full-time or part time), 0=not a student
CWORK1	- class of worker is a public employee
CWORK2	- class of worker is a private employee
CWORK3	- not a public or private employee, specifically: private, self-employed with or without employees or private, working in a family business without pay.