Research Studies for the Labour Force Survey Sample Redesign

M.P. Singh, J. Gambino, N. Laniel, Statistics Canada
J. Gambino, Statistics Canada, Ottawa, Canada K1A 0T6

KEY WORDS: Survey design; Sampling; Household surveys.

1. Introduction

The Labour Force Survey (LFS) is designed to provide a monthly snapshot of the Canadian labour market. The LFS frame and systems are also used by most household surveys conducted by Statistics Canada. This results in greatly reduced costs for these surveys since development of sample selection systems and data collection programs is carried out in an integrated fashion.

The LFS has been redesigned following every decennial census. In the 1970s, major changes were made to the questionnaire, the use of computers was increased and the sample grew from about 35,000 to 55,000 households per month. The objective of the survey was to produce reliable labour force estimates at the national and provincial levels. Following the 1981 census, the LFS was redesigned so that reliable subprovincial estimates could also be published.

The scope of the current redesign is similar in breadth to the one in the 1970s. A major overhaul of computer hardware and software systems, as well as changes to data collection and to the questionnaire are planned. The purpose of this paper is to discuss these design changes. See Drew et al. (1991) for discussion of other aspects of the redesign.

A redesign provides an opportunity to update the sampling frame. The current LFS uses 1981 census geography and the corresponding 1981 census counts. As a result, selection probabilities for primary sampling units (PSUs) and clusters are based on obsolete counts. Another manifestation of out-of-datedness is the increasing occurrence of growth clusters, i.e., clusters that turn out to have many more dwellings when they enter the sample than the census counts indicated. This usually leads to sub-sampling, increased field costs and sometimes estimation problems because of the large sampling weights introduced by the sub-sampling.

In previous designs, the sample was allocated within each province to Economic Regions (ERs). In 1989, an increase of 16,500 households in the monthly sample was introduced to produce good estimates for a different set of sub-provincial regions, the Unemployment Insurance Regions (UIRs). Thus, for the current redesign, two sets of regions need to be considered. An effort will be made to deal with both sets of regions simultaneously.

In previous designs, primary sampling units were designed so that their sample would correspond to an LFS interviewer's monthly assignment. This was important until the last redesign since all interviewing in rural areas was conducted in person. Now a single procedure, namely, an initial personal interview followed by five telephone interviews, is used everywhere. Probably as a result of this, the one-to-one correspondence between PSU and assignment has largely broken down for the LFS.

In addition, the LFS frame and systems are used by an increasing number of surveys---two major new examples are the Survey of Labour and Income Dynamics (SLID) and the National Population Health Survey (Health). As a result, the usefulness of the current PSU concept has diminished.

The idea of creating a general household survey vehicle by making the LFS design and system flexible enough to handle most household surveys is not new; see Singh and Drew (1981). However, changes like those discussed in the two previous paragraphs may allow us to come closer to achieving the goal of a general-purpose design. As we will see following the next section, the major design changes being considered involve simplification of the design.

2. The Current LFS Design

In this section, we give a brief, simplified description of the current Labour Force Survey design. For a detailed description, see Singh et al. (1990).

The core sample, which was approximately 52,000 households at the time of the last redesign, was first allocated to Canada's ten provinces and then to Economic Regions within each province. Several sample size reductions have occurred since 1985 and

the current core sample consists of 42,300 households. The addition of the UI sample of 16,500 households brings the total monthly sample size to almost 59,000 households.

The current LFS has two major types of design: (i) rural areas and small urban centres follow a non-self-representing (NSR) design, and (ii) all other urban centres are self-representing (SR) in the sense that such urban centres always have households in the sample. Traditionally, an urban centre was made self-representing if it could support a sample of at least twenty dwellings per month.

SR design: In CMAs, there are two levels of stratification, using Census Tracts as stratification units. In other SR areas, Census Enumeration Areas (EAs) were used. The primary sampling unit is the cluster, which is typically a city block. The Rao-Hartley-Cochran (1962) random group method is used to select six (or a multiple of six) clusters per stratum. Each cluster then corresponds to one of six rotation groups, i.e., the dwellings selected in a cluster enter the sample at the same time and are replaced by new ones after six months. About four or five dwellings are selected systematically in each cluster.

An exception to the above design is the apartment frame which exists in the larger CMAs. Large apartment buildings were removed from the area frame at the time of the last design and their list is maintained separately. The apartment frame has worked well and it is unlikely that there will be major changes made to it.

NSR design: There are several variations on the NSR design; only the basic one will be described here. Strata were formed using EAs as units in the eight smallest provinces and using Census Subdivisions (CSDs) in Ontario and Quebec. Using EAs, PSUs were then formed within strata. In each stratum, a sample of two or three PSUs was selected using randomized probability-proportional-to-size (PPS) systematic sampling. Then a sample of six clusters was selected using randomized PPS systematic sampling in each selected PSU. A cluster is usually an EA; the exceptions occur when large EAs are split and when very small EAs are combined. Within each rural cluster, a systematic sample of about ten dwellings is selected, and within NSR urban clusters three are selected.

For estimation, the design weights are modified to agree with census-based demographic projections for several age-sex groups, as well as ER and CMA populations. This is accomplished using generalized least squares. The jackknife is used to estimate variances. For details and further references, see Singh et al. (1990).

3. Redesign Studies

In this section, we describe the design alternatives being considered and the studies for comparing these alternatives that have taken place or that are planned. The section also contains a brief discussion of estimation methods.

3.1 Time and Cost Study

Before describing the design alternatives, we will discuss the time and cost study which was conducted in 1992. Its results will be used as inputs when comparing different designs.

In this study, all 1000 LFS interviewers recorded travel and time information for the two interview weeks in October and November 1992. They noted all attempted telephone and personal contacts. For personal attempts, they recorded information that allows us to determine distances travelled between dwellings (where appropriate), between clusters in a PSU, between PSUs and between a PSU/cluster and home/office. The information also allows the separation of travel time and personal interviewing time. The data from the study are augmented by data from the weekly pay claims that all interviewers submit. These provide information on number of hours worked on various tasks, distance travelled and other expenses.

3.2 Stratification

The stratification algorithm is a modification of one due to Friedman and Rubin (1967). In the last redesign, 16 socioeconomic variables (income, education, employment by major industry group, four housing variables) were used. Minor changes to this list of variables are being studied. The robustness over time of the various stratification alternatives is also being studied by creating strata using data from one census and then evaluating the strata (with respect to homogeneity) using data from the next census.

Major changes to other aspects of stratification are being considered. In the current non-self-representing part of an Economic Region, strata are created using

either Enumeration Areas or Census Subdivisions as stratification units. Then primary sampling units are created and selected in each stratum. The alternative being studied would introduce another level of stratification while eliminating the current PSU stage (see the section below on the NSR design).

Several ways of achieving this have been compared and the most promising variation is the following: Create strata using Census Divisions (CDs) as units; a typical stratum would consist of one or two CDs. Then sub-stratify each stratum using EAs as units. Whether sub-strata need to be compact, i.e., whether the EAs forming a sub-stratum should not be allowed to be scattered throughout the parent stratum, is also being investigated. The stratification studies indicate that the efficiency of this approach to stratification is comparable to that of the current approach. There is also no loss in robustness over time. Finally, there are time and cost savings if EAs are used as units only within strata instead of within a whole ER.

3.3 Allocation

Since the last redesign, several decreases to the core sample plus the sample increase for Unemployment Insurance purposes have distorted the relative sample sizes among provinces. This redesign has provided an opportunity to review the provincial allocations and to propose changes.

An important factor to consider is the possibility that funding for the UI sample will stop at some point in the future. In that situation, only the core sample of 42,300 households will be available. It has been proposed that allocation be done in two phases: first, the core sample will be allocated among provinces and within provinces, and second, the UI-funded sample will be allocated to UI regions to meet their reliability requirement.

As was noted in the introduction, this is the first LFS redesign that will have to contend with two different sets of regions. One question that had to be addressed was whether the core sample should be allocated with Economic Regions in mind, as in previous designs. Because the core sample is now relatively small, it will not be possible to produce reliable estimates for all ERs. The UI portion of the sample will alleviate this problem to some extent, but as was noted above, there is no guarantee that this sample will always be there. Thus there is a change in emphasis toward provincial and UIR estimates, with the quality of ER estimates becoming a byproduct of provincial and UIR allocations.

Several allocations of the core sample to provinces have been compared. These include proportional to population, proportional to square root of population, Neyman and equal-CV allocations, as well as mixtures of these allocations (e.g., a compromise between Neyman and equal-CV allocations). Since the populations of the provinces differ greatly, proportional and Neyman allocations are not acceptable. Conversely, equal-CV allocation gives far too much sample to the smallest provinces and increases national CVs significantly. For illustration, the table below compares the current allocation to the square root allocation (based on number of households). The provinces are listed from smallest to largest in population (persons).

Province	Current Allocation	Square Root Allocation
Prince Edward I.	1420	1050
Newfoundland	2230	2010
New Brunswick	3100	2390
Nova Scotia	3110	2780
Saskatchewan	4530	3240
Manitoba	3280	3150
Alberta	5200	4660
British Columbia	4450	5520
Quebec	6470	8250
Ontario	8520	9260
Canada	42,310	42,310

Other compromise allocations are similar: they tend to increase the proportion of the sample allocated to the three largest provinces. In the end, it was decided to keep the current core sample for six of the ten provinces. For the pairs Manitoba-Saskatchewan and Alberta-British Columbia, the core sample was adjusted to yield the same coefficient of variation (CV) for *unemployed* in each pair (e.g., Manitoba and Saskatchewan would have the same CV based on the core sample).

Once the provincial sample sizes are determined, sub-provincial allocations can take place. Several studies

were conducted to compare different allocation methods. The current allocation was compared to proportional and optimal allocations. In optimal allocation, the relative cost of data collection in SR and NSR areas was taken into account. Allocations assuming fixed target CVs were compared to those assuming fixed sample size. Allocation parameters that were varied included design effects, unemployment rates and participation (in the labour force) rates.

It was decided to allocate the core sample within province using allocation proportional to the size of each intersection of UIRs and ERs. This is consistent with the desire to ensure that the core sample can be used to produce reliable provincial and national estimates should the UI portion of the sample be eliminated in the future.

3.4 The SR Design

About sixty per cent of the current LFS sample falls in the self-representing parts of Canada, with the CMAs being the largest component. An important consideration in choosing between alternatives will be the ease with which population counts can be updated between redesigns to deal with the problems due to obsolete counts that were discussed in the introduction.

There are four major alternatives being considered for the SR design, namely, the current cluster-based design, a design based on the Address Register, one based on postal codes and one using Enumeration Areas as clusters (instead of blocks).

The Current Approach. In areas where information on blocks was available, the current approach used clusters consisting of city blocks or blockfaces as PSUs. In other SR areas, EAs were used as PSUs. A sample of PSUs was selected, followed by selection of dwellings. Generally, this approach has performed well. The exception to this is the out-of-datedness of the dwelling counts mentioned earlier. Should this approach be used again in the next design, a major change may be in the way clusters are formed: their creation using the computer-assisted districting program (CADP) that was used to delineate EAs for the 1991 census is being investigated. It has also been proposed that clusters be made bigger and that the sample size per cluster be increased, but by a smaller ratio (e.g., triple the cluster size and double the sample size).

Address Register Design. The Address Register (AR) was created for the 1991 census using administrative data from various sources. It exists in all cities with a population greater than 50,000. In such cities, the AR can be used as a list frame. Since any list of addresses quickly becomes out-of-date, an AR list frame would have to be supplemented with a procedure for covering new dwellings. The AR can also be used in a cluster-based design as a source of addresses, i.e., it can provide a list of addresses for a selected cluster. Another possible use of the AR, if it is revised on a regular basis between censuses, is as a source of dwelling counts in some of the other designs discussed in this subsection.

During 1992, two AR tests were carried out. In the first test, the current cluster listing method was compared to (i) using the AR to create a list of addresses for the cluster, (ii) using the AR to create a (pre-)list that the interviewer then updates in the field, and (iii) using the AR to create a (post-)list that is used for verification and correction after the interviewer has obtained a list in the field. The test results show that there is little difference in coverage among the four methods.

The second Address Register test looked at the feasibility of using a growth frame to supplement an AR list frame. The test was conducted in seven CMAs across Canada, with 30 EAs selected per CMA. Prior to selection, the EAs in a CMA were stratified into a group where high growth was expected (as determined by an administrative source) and a group with little growth expected. The test discovered much more growth than anticipated, even in the so-called low-growth stratum. Perhaps a better administrative source would have provided a better stratification.

Undercoverage of the AR ranged from about three percent in Victoria to almost 15 per cent in Winnipeg. Overcoverage ranged from 1.6 per cent to 2.8 per cent. The test confirms the need for a method to supplement an AR-based list frame.

The test also showed that use of the Address Register in the rural fringe of cities would be difficult. Problems with addresses (e.g., they may not be unique) can make file matching inaccurate and make interviewers' job in the field difficult. The AR would be most useful in the more urbanized parts of cities.

Postal Code Based Design. Canada uses a six digit postal code system. Unlike clusters and EAs, postal codes are not geographical units---they do not cover land areas (they form a network). A first stage sample

of postal codes can be selected, and a second stage sample of dwellings can then be selected within postal code.

To be useful, an accurate count of the number of dwellings in each postal code is needed. However, such counts are not available. One possibility that will be investigated is to obtain up-to-date counts by using administrative sources such as telephone billing files or income tax files. These are some of the same sources that were used to create the Address Register.

EA Based Design. In this option, the current block/blockface-based clusters are replaced by Enumeration Areas as primary sampling units. Two advantages of this are first, that EAs are already defined, with census counts and maps already available, and second, that the relative impact of high growth will be smaller on EAs since they are larger units than the current clusters. Another possible advantage is that administrative sources can be used to obtain updated counts by EA. One disadvantage is that EA definitions change with every census, although modern geographical methods make conversion between vintages more accurate than in the past.

Enumeration areas vary considerably in size. Large EAs take longer to list and maintain. To decrease the range in EA size, one option is to split the largest EAs and then use the parts as first stage units for sampling. To minimize the amount of actual splitting required, a procedure based on exact conceptual splits, followed by sample selection and actual splitting of selected EAs has been devised. If an actual split turns out to be very different from the conceptual one, then an sample update procedure can be used to re-select a sample.

3.5 The NSR Design

Although the non-self-representing areas cover only about one third of Canada's population, the LFS design is most complex there. This paper has presented only a simplified account of the design (in fact, there are four variations). One of the goals of the current redesign is to simplify the actual design---partly because it now seems possible to do so and partly because it will lead to greater flexibility, e.g., to deal with sample size decreases and increases and to select samples for other household surveys.

Elimination of Current PSU Stage. The most significant change being considered for NSR areas is the elimination of the current PSU stage, where a PSU comprises several Enumeration Areas. In this alternative, EAs would be sampled directly, i.e., they would become the primary sampling units. When this change is taken together with the proposed change in stratification, the NSR portion of the sample will be selected using a design that has most of the advantages of the current design as well as less clustering and increased flexibility.

The LFS already has some experience with this design---it was introduced in Prince Edward Island (for other reasons) in the last redesign. This design was compared to the current one by studying costs and variances:

- Interviewing costs were compared for the two alternatives. Since the current PSUs were designed to be compact, travelling among EAs within a given PSU will cost less than travelling among the same number of EAs without the compactness constraint. However, because only one-sixth of interviews are conducted in person, this study showed relatively small differences in overall cost between the two alternatives.

- The design alternatives were compared with respect to variance. Seventeen Economic Regions from across Canada were selected for this study. Census data were used to look at variances for several characteristics.

Based on the results of the study and the desire to simplify the design, it was decided to eliminate the current PSU stage with only a few exceptions, mainly is remote areas.

4. Estimation

The LFS modifies the design weights of sampled individuals so that monthly census-based demographic projections for several categories are respected, i.e., the sum of the revised weights of all individuals in a category equals the projected total for that category. The categories consist of Economic Regions, Census Metropolitan Areas and several age-sex groups. In the past, the reweighting was done via poststratification, then using raking-ratio estimation, and now using generalized least squares (see Singh et al (1990)).

One consequence of using least squares is that some individuals may receive negative weights. The incidence of negative weights increases as the number

of adjustment categories increases and as the sample size decreases. As a result, although the LFS itself obtains negative weights only rarely, they occur more frequently in other household surveys since they usually have both smaller samples and more adjustment categories than the LFS.

The LFS weighting system currently deals with negative weights in an ad hoc way: if negative weights occur in a province, then that province goes through a second pass of the weighting program; any remaining negative weights are set to one. Several alternatives to this ad hoc procedure are being investigated. These include methods based on the work of Huang and Fuller (1978) and Deville and Sarndal (1992) as well as recent work by Singh and Mohl (1993, ongoing work at Statistics Canada). These alternatives can deal not only with negative weights but also with weights that increase by a large factor.

A major concern with the current variance estimation system is that it is poorly documented and difficult to maintain. After comparing various altenatives, it was decided to incorporate LFS variance requirements into Statistics Canada's Generalized Estimation System (GES).

A study of composite estimation is planned. It will compare the traditional approach to the multivariate one described by Singh et al. (1992). In addition, it will investigate the use of composite estimates as additional control totals, i.e., the composite estimates will play the same role as the demographic projections described earlier in this section. Both AK-composite and multivariate composite estimators can be used in this way.

High Income Earners. Household surveys that collect data on income are sometimes faced with a sampled household which contributes significantly to estimates of average income at the CMA and province levels. Recent work has confirmed that, in fact, high income households are under-represented among respondents. Two alternative ways of dealing with this problem are being investigated.

In the estimation approach, one option is to use tax data to determine the proportion of households or individuals in the population with incomes exceeding a cut-off value and use this to derive an additional "demographic projection" in the weighting program. This amounts to a poststratification by income category. A drawback of this approach is that it ignores the possibility that the low incidence of high income households in the sample may be due to nonresponse.

A design-based solution being considered would create high-income strata in some Census Metropolitan areas. by placing the EAs with the highest average incomes in a CMA (as reported in the census) into a special stratum. An advantage of this approach is that, if nonresponse is an important contributor to the low representation of high incomes in the sample, then special measures can be implemented in the high-income strata to deal with this problem.

References

Deville, J.C. and Sarndal, C.E. (1992). Calibration estimation in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Drew, D., Gambino, J., Akyeampong, E. and Williams, B. (1991). Plans for the 1991 redesign of the Canadian Labour Force Survey. Proceedings of the Survey Research Methods Section, American Statistical Association.

Friedman, H.P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

Huang, E.T. and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. Proceedings of the Social Statistics Section, American Statistical Association, 300-305.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, B*, 24, 482-490.

Singh, A.C., Gambino, J.G. and Chen, E.J. (1992). Multivariate composite estimation for rotating panel surveys. Statistics Canada, unpublished.

Singh, M.P. and Drew, J.D. (1981). Redesigning continuous surveys in a changing environment. *Survey Methodology*, 7, 44-73.

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