# DESIGN AND ESTIMATION ISSUES FOR INCOME IN THE REDESIGN OF THE CANADIAN LABOUR FORCE SURVEY

E.J. Chen, J. Gambino, N. Laniel and J. Lindeyer, Statistics Canada
J. Gambino, Statistics Canada, Ottawa, Canada K1A 0T6

**Key Words: Stratification; Income representation; Complex surveys**

## 1. INTRODUCTION

The Canadian Labour Force Survey (LFS) is the largest ongoing household survey conducted by Statistics Canada. This monthly sample survey is a major source of data on labour market conditions. It collects and publishes monthly labour market indicators such as the unemployment rate as well as total employed and unemployed at various levels of aggregation.

The sampling scheme for the LFS is complex. For a detailed description of the sample design, see Singh et al. (1990). Essentially, the LFS has a stratified multi-stage area sample design. Each province in Canada is divided into sub-provincial regions, which are large areas of similar economic structure. The regions are divided into different types of areas, such as major urban areas and rural areas. Sampling strata are formed as groups of Census Enumeration Areas (EAs) or Census Tracts (which are sets of contiguous EAs) with similar socio-economic characteristics. The LFS sample is divided into six representative rotation groups within a stratum. Each month, one-sixth of the sampled households are replaced after a stay of six months. The sample size for the survey has fluctuated over time and is now about 59,000 households.

Increasingly, the LFS frame has been used by supplementary and special surveys to collect data for various needs. Supplementary surveys use additional questions asked of LFS respondents at the time of an LFS interview, whereas special surveys use samples of dwellings chosen from the LFS frame but not interviewed for the LFS itself. Examples of supplementary surveys and special surveys include the Survey of Consumer Finances (SCF) and the Survey of Family Expenditures (FAMEX), respectively. The SCF provides annual estimates of income and low income incidence for individuals and families. The FAMEX survey is conducted periodically to examine spending patterns in major cities and/or provinces in Canada.

The proper representation of households of different income levels is one of the major considerations in these surveys. However, households in certain income classes are not well-represented. In particular, as illustrated below, high income households are under-represented among survey respondents. This has an impact on the estimated proportion of individuals in various income classes, on the estimates of average income, and on the estimated variances of these estimates.

The under-representation can be attributed to sample frame problems, nonresponse, and reporting biases in the data. First, the current LFS frame is not the most efficient one for income estimates since stratification by income was not the primary goal; households are not separated by income level in the frame. Second, there is evidence that high income households are more likely to be unwilling or unable to provide income information. In addition, there are indications that some respondents have the tendency to under-report their income.

The LFS has been redesigned following every decennial census of population and is now in its fifth redesign since its inception in the 1940s. This redesign provides an opportunity to update the sample frame and to introduce other changes, such as computer assisted data collection and new questionnaire content. Singh, Gambino and Laniel (1993) outlined the studies undertaken for the current LFS sample redesign. It was in this context that a decision was made to use the redesign as an opportunity to deal with high income problems, both through design changes and through estimation.

Representation problems for low income households are not as evident as they are for the high end of the income spectrum. Nevertheless, because of the importance of the estimates produced by household surveys for, for example, establishing low income cut-offs (LICOs), the low end of the income scale was also dealt with in the redesign.

In this paper, the focus will be on high income problems, although we also describe the new low income strata. We first examine the under-representation problems in the current SCF. Then some preliminary results and problems with using income tax information as auxiliary data are discussed. Next, we describe the sample redesign aspects, namely, the formation of high income and low income strata in some Census Metropolitan Areas (CMAs).

## 2. PROBLEMS OF THE SCF SAMPLE DATA

The Survey of Consumer Finances (SCF) is conducted annually as a supplement to the April LFS. Its sample is two-thirds of the LFS sample, i.e., a sample size of about 40,000 households. The survey collects and publishes data on detailed estimates of income distribution and of low income incidence for individuals and families. In addition, the SCF income data are combined with data from several other household surveys to provide detailed information about the household population and its characteristics. For example, the SCF data are linked with the Household Facilities and Equipment Survey data to provide statistics on household facilities by income and other characteristics. In the following, we will use the SCF to illustrate the high income representation problem for surveys using the LFS frame or sample.

It has been observed that high income individuals and households are under-represented in the SCF sample. Moreover, the number of high income respondents in the sample and consequently their weighted count fluctuates noticeably from year to year. Table 1 presents the estimated number of individuals at different high income levels for the income reference years 1989 to 1992 (see the end of this paper). The estimated number of individuals was obtained from the weighted sum of the SCF sample respondents. Here, data are presented for four provinces. It is clear that the year-to-year fluctuation increases as the income level increases. Note that since there were no respondents with income over $250,000 in the 1990 Quebec sample, the corresponding estimate is zero. Occasionally, the presence of extremely high income respondents distorts the variance estimates greatly since the variance is sensitive to extreme values. Typically, the problem arises when one of the six clusters selected in a stratum contains a household with a very high income, leading to a large inflation of the variance estimate.

To examine the degree of under-representation of high-income households in the SCF sample, we compare the SCF sample data with the number of income tax filers obtained from Revenue Canada taxation. Table 2 illustrates the degree of under-representation for certain high income levels in reference year 1990. Here, the four provinces in Table 1 are treated together. The number of sample respondents and its weighted estimate were again obtained from the SCF data whereas the number of income tax filers was obtained from income tax returns for the same reference year. The relative degree of under-representation is calculated as the difference between the SCF weighted estimate and the number of income tax filers over the number of income tax filers. We see that the under-representation problem is more pronounced as the income level increases. The differences are too big to be accounted for by conceptual differences between the survey and the tax files.

## 3. CURRENT REDESIGN CONSIDERATIONS

In the current redesign, two alternative approaches were identified to deal with income under-representation problems, namely, the estimation and sample redesign approaches.

### 3.1. Estimation Approach

The SCF uses a regression estimator in its weighting procedure. This regression method (Lemaître and Dufour, 1987) incorporates auxiliary information as control totals. These include population projection counts, such as counts by age-sex groups and by household size. The method adjusts the survey weights so that the final, adjusted weights respect the control totals (post-stratification is a special case of this).

One approach to the problem of high-income under-representation is to augment the current set of control totals by including a count of high income individuals. Such a count can be obtained from the income tax data after adjusting for conceptual and definitional differences between these two data sources. This additional control would adjust the survey weights such that the survey estimate obtained for the control variable would be equal to the known total.

Table 3 presents the preliminary results for the characteristic *individual average income* of using the number of high-income tax filers as an additional control. The results are shown for the reference year 1990 and the number of income tax filers in the provincial version of Table 2 (not shown here) was used as the additional control.

The impact on average income when using this additional control, compared to the current set of controls, is measured in the column of percentage relative change. Clearly, the impact depends on the degree of under-representation. Such an impact can be more or less significant depending on the income cut-off level used. With regard to the variance estimates and the coefficients of variation (C.V.), the impact is again related to the degree of under-representation. Generally, we obtained a better C.V. when the additional control was used. In this data set, the cut-off level of $100,000 was a reasonable choice.

However, there are several issues regarding the use of income tax information as an additional control total. First, the income tax information is not available at the time of SCF weighting and estimation. There is at least a two-year lag before the income tax information becomes available. Second, the high income cut-off level to be used to derive the control total needs to be specified since the impact will depend on the choice of the level. These two problems are discussed next.

Availability of the Control Total: With respect to the first problem, research is needed on projecting control totals at various income cut-off levels since the income tax information is not available at the time of estimation. One approach is to use the proportion of high income individuals to derive a control total since such a proportion can be derived from historical income tax information and adjusted to account for conceptual and definitional differences. The usefulness of this approach depends on the stability over time of the proportion.

Choice of Income Level Cut-off Value: With respect to the second problem, should we set the additional control at the $100,000 or $150,000 income level? The choice of cut-off value will depend on the variable of interest and the degree of under-representation of individuals or households with incomes exceeding the cut-off value. We can also use a fixed-proportion approach, for example the top one per cent of tax filers. Once such a cut-off or proportion is determined, it should be used consistently over time to minimize the disruption on the series of income estimates.

In addition to studying provincial-level income estimates for individuals, the impact on other characteristics, such as household average incomes requires study. The impact at different levels of aggregation is also important since small area estimation and domain estimation may be greatly affected by using the additional control.

## 3.2. Sample Redesign Considerations

In the old LFS design, sample strata were created using information from the 1981 census. In particular, large Census Metropolitan Areas (CMAs) in the current LFS design are divided into two separate frames containing regular private dwellings (the Area Frame) and apartment buildings (the Apartment Frame). In the Area Frame, a stratification algorithm was used to group Census Tracts with similar socio-economic characteristics. Clusters, which consist of groups of households in a city block or in a set of block faces, were then created and served as first stage sampling units. The LFS Apartment Frame exists in the seventeen largest CMAs in Canada. An apartment building in the frame must have at least thirty units and five floors of living quarters. The stratification of the apartment frame was done by size, i.e., the number of units in each building, and in some cities, by geography as well. With some exceptions, each apartment building is a cluster. LFS strata can be viewed as sets of clusters of households and must be sufficiently large to permit sample rotation and meet sample size requirements.

A two-stage sampling design is used in these two frames. In the Area Frame, a random sample of clusters is selected using the Rao-Hartley-Cochran (1962) random group method at the first stage. In the Apartment Frame, a random sample of apartment buildings is selected using the probability-proportional-to-size (PPS) systematic sampling method. In both frames, dwellings within clusters are randomly selected using a systematic sampling scheme.

After consultation with subject matter experts and investigation of the 1991 Census income data, it was decided to create high income and low income strata in large CMAs in Canada during the LFS sample redesign.

### 3.2.1 Creation of High Income Strata

The criteria used to create the high income strata are the following. The high income strata are to be created in large CMAs among the strata containing regular private dwellings. The EAs that rank in the highest 3% of the CMA in terms of their average household income, as reported in the 1991 Census of Population, are assigned to the high income stratum. It is desired that the average household income in the CMA's high income stratum be over $100,000. In addition, the stratum must be large enough to permit rotation and yield a sample of at least 24 dwellings. In some of the large cities, further sub-stratification is performed based on average income.

Table 4 presents the results for the creation of high income strata in the LFS sample redesign. The high income strata were created in nine major CMAs in Canada. There were 562 clusters that represented the highest household income areas in these cities with a cluster median income of $122,765. Each stratum will yield a sample of about 24 dwellings for the LFS. In certain strata, the sampling fraction was modified to yield this expected sample.

### 3.2.2 Creation of Low Income Strata

In addition to stratification by the number of apartment units and by geographical area in the current

apartment frame, it was decided to stratify the apartment buildings by income using income information from the 1991 Census of Population. Apartments are included in the low income stratum if they reported low average household income. It is also desired that the average household income of all the apartments in the city's low income stratum not exceed about $15,000. In the stratum, the dwelling count has to be large enough to yield a sample of at least 30 dwellings. In some large cities, sub-stratification is also performed to further group the apartment buildings by their average income.

Low income strata were created for 7 of the 9 high income cities (see Table 5). These strata group the lowest income apartments in these cities. The extent of stratification of the low income apartment frame varies from city to city. Because of its large population, income strata in Toronto are formed within geographical strata.

3.2.3 Cluster Formation and Sample Design in the Income Strata

In this sample redesign, clusters in major cities (excluding the apartment frame) were created by using a Computer Assisted Districting Program (CADP), that had earlier been used to form EAs for the 1991 census. This automated procedure used 1991 Census Tracts, EAs and block faces as input to produce clusters, with the block faces serving as the basic building blocks. The clusters in most cities were designed to have 150 to 250 dwellings. In the three largest cities, i.e., Toronto, Montreal and Vancouver, the cluster size was designed to fall in the range of 200 to 300 dwellings. As in the current design, clusters in the apartment frame are apartment buildings.

The formation of clusters and the size requirements in the income strata are the same as in other strata in the same city. The same two-stage sample design method as in non-income strata is used. In other words, the sample selection of the clusters and the selection of the dwellings within the clusters are the same as in other strata in the LFS. Such a design will have minimal impact on the LFS and LFS frame based surveys with respect to operations and estimation.

3.3 Benefits of the Income Strata

Household surveys that are based on the LFS sample frame can now select the sample with different size requirements in the income strata. For example, they can specify a higher sample size in the high income strata. In addition, we are now able to monitor the income contribution and response rates of these income strata. If nonresponse is found to be an important contributor to the low representation in the sample, special measures can be implemented to deal with this problem. With the creation of income strata, we are also better able to assess the degree of under-reporting of incomes.

## 4. CONCLUDING REMARKS

The LFS redesign sample will be phased in from October 1994 to March 1995. The new income strata in the LFS are expected to improve the representation of incomes and lead to more reliable estimates. There is a study underway to examine the impact of the new design and the efficiency gains in the survey estimates.

Middle and low income individuals and families have received considerable public and statistical attention compared to those with high incomes. However, the high income earners can have a large impact on estimates of average income and the study of high income earners encompasses many important social and economic policy issues; see Murphy et al. (1993). Such analyses require accurate and reliable data. The estimation and sample design features discussed in this paper should enhance data quality and improve the representation of high income households.

## REFERENCES

LEMAITRE, G. and DUFOUR, J. (1987). An Integrated Method for Weighting Persons and Families, *Survey Methodology*, 13, 2, 199-207.

MURPHY, B., FINNIE, R. and WOLFSON, M. (1993). A Profile of High Income Ontarians. A paper prepared for The Ontario Fair Tax Commission.

RAO, J.N.K., HARTLEY, H.O. and COCHRAN, W.G. (1962). A Simple Procedure for Unequal Probability Sampling without Replacement, *Journal of the Royal Statistical Society, B*, 24, 482-491.

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.

SINGH, M.P., GAMBINO, J.G. and LANIEL, N. (1993). Research Studies for the Labour Force Survey Sample Redesign. Proceedings of Survey Research Methods Section, American Statistical Association.

TABLE 1. The Estimated Number of Individuals at Different High Income Levels, SCF 1989 - 1992.

| Province | Income Level | 1989 | 1990 | 1991 | 1992 |
|---|---|---|---|---|---|
| QUEBEC | $100,000+ | 31,382 | 27,112 | 46,560 | 23,776 |
| | $150,000+ | 8,081 | 6,222 | 13,606 | 5,343 |
| | $250,000+ | 1,478 | 0 | 795 | 230 |
| ONTARIO | $100,000+ | 60,740 | 79,085 | 76,187 | 80,714 |
| | $150,000+ | 21,237 | 18,418 | 23,266 | 23,124 |
| | $250,000+ | 8,540 | 2,979 | 10,940 | 6,231 |
| MANITOBA | $100,000+ | 4,213 | 2,243 | 3,212 | 7,006 |
| | $150,000+ | 979 | 487 | 1,600 | 1,019 |
| | $250,000+ | 328 | 395 | 148 | 59 |
| BRITISH COLUMBIA | $100,000+ | 13,390 | 33,151 | 13,364 | 21,015 |
| | $150,000+ | 3,654 | 9,026 | 3,195 | 6,604 |
| | $250,000+ | 796 | 1,370 | 1,729 | 1,811 |

TABLE 2. Comparison of SCF Sample Data and Income Tax Return Data in Four Provinces, 1990

| Income Level | No. Sample Respondents | Weighted Estimate | No. Income Tax Filers | Under Representation |
|---|---|---|---|---|
| $100,000+ | 302 | 141,591 | 197,162 | -28.2% |
| $150,000+ | 76 | 34,153 | 83,430 | -59.1% |
| $250,000+ | 12 | 4744 | 29,388 | -83.9% |

TABLE 3. Effect of Using Income Tax Filer Count as an Additional Control at Different High Income Levels, 1990 (Individual Average Income)

| Prov | Income Level | Average Income | Rel. Change | Std. Error | C.V. |
|---|---|---|---|---|---|
| QUE. | Current Control | $21,765 | --- | $270 | 1.24 |
| | $100,000+ | $22,089 | +1.5% | $236 | 1.07 |
| | $150,000+ | $22,071 | +1.4% | $257 | 1.16 |
| | $250,000+ | $21,765 | +0.0% | $270 | 1.24 |
| ONT. | Current Control | $25,246 | --- | $262 | 1.04 |
| | $100,000+ | $25,898 | +2.6% | $244 | 0.94 |
| | $150,000+ | $26,370 | +4.5% | $444 | 1.68 |
| | $250,000+ | $26,644 | +5.5% | $799 | 3.00 |
| MAN. | Current Control | $20,781 | --- | $327 | 1.57 |
| | $100,000+ | $21,249 | +2.3% | $297 | 1.40 |
| | $150,000+ | $21,176 | +1.9% | $339 | 1.60 |
| | $250,000+ | $20,855 | +0.4% | $351 | 1.68 |
| B.C. | Current Control | $24,906 | --- | $364 | 1.46 |
| | $100,000+ | $24,884 | -0.1% | $416 | 1.67 |
| | $150,000+ | $25,253 | +1.4% | $313 | 1.24 |
| | $250,000+ | $25,529 | +2.5% | $510 | 2.00 |

TABLE 4. Stratification Results on the High Income Strata in the LFS Redesign

| CMA | No. Dwg. | No. Str. | No. Clus. | Med. Income | Ave. Income |
|---|---|---|---|---|---|
| Montreal | 15,237 | 3 | 83 | $121,881 | $132,818 |
| Ottawa | 6,558 | 2 | 39 | $111,729 | $116,973 |
| Toronto | 35,433 | 4 | 185 | $144,387 | $156,477 |
| Hamilton | 6,584 | 1 | 34 | $101,875 | $107,130 |
| London | 4,036 | 1 | 21 | $108,009 | $108,604 |
| Winnipeg | 7,543 | 2 | 42 | $96,763 | $100,264 |
| Calgary | 7,501 | 1 | 41 | $123,066 | $131,543 |
| Edmonton | 5,835 | 1 | 28 | $111,334 | $118,600 |
| Vancouver | 16,483 | 3 | 89 | $119,777 | $122,739 |
| Total | 105,210 | 18 | 562 | $122,765 | $132,217 |

TABLE 5. Stratification Results on the Low Income Strata in the LFS Redesign

| CMA | No. Dwg. | No. Str. | No. Clus. | Med. Income | Ave. Income |
|---|---|---|---|---|---|
| Montreal | 21,932 | 2 | 267 | $14,500 | $14,630 |
| Ottawa | 8,256 | 1 | 58 | $15,849 | $14,857 |
| Toronto | 39,580 | 3 | 209 | $14,427 | $13,969 |
| Winnipeg | 10,425 | 3 | 107 | $13,732 | $14,270 |
| Edmonton | 4,513 | 1 | 50 | $17,201 | $14,860 |
| Calgary | 5,146 | 1 | 39 | $14,035 | $15,251 |
| Vancouver | 8,545 | 1 | 85 | $13,824 | $14,601 |
| Total | 98,387 | 12 | 815 | $14,557 | $14,386 |