Using Income as an Auxiliary Variable to Improve the Design of Household Expenditure Surveys

Charles Mitchell and Christian Nadeau
Statistics Canada, Ottawa, Ontario, Canada K1A 0T6

Keywords:
stratification, sample allocation, simulation, spending

1. Introduction

Surveys that collect data on household finance, such as the Survey of Household Spending (SHS) and the Survey of Financial Security (SFS), both conducted by Statistics Canada, can possibly benefit from the availability of auxiliary information on household income at different stages of the survey. The use of administrative data on wages and salaries for the calibration of the Statistics Canada income-related surveys, for example, helped to improve the coherence among these surveys as well as their coherence with other data sources (Tremblay, 2005). As reported by Arsenault et al. (2001), it also helped to reduce the variance of the SHS estimates.

Despite the use of related auxiliary information at the estimation stage, it might also be worthwhile to use auxiliary information on household income at the design stage, namely for stratification and sample allocation purposes. These two surveys draw their samples (or part of their samples) from a multipurpose area frame, referred to as the Labour Force Survey (LFS) frame, through a stratified multistage design. In the last two LFS redesigns, efforts have been made to create high income strata to help in improving the design of such surveys.

In this paper, we are interested to study how the use of household income auxiliary information at the design stage could help to improve the estimation efficiency in household finance surveys. In particular, we investigate the stratification of primary sampling units based on the prevalence of households with higher income, assuming a design which is somewhat similar to that of the SHS. After a quick overview of the SHS sampling design, we describe the method used to create an efficient high income stratum and compare it to the actual high income strata. A simulation study is then presented to assess, in the context of a two-stage design, the impact of using such stratification on the variance of calibrated estimates for variables that present different types of relations with household income.

2. Background

The SHS is an annual survey that collects detailed household expenditure data for the previous calendar year from a sample of approximately 21,000 households. Its sample is selected from the LFS area frame through a stratified multistage design. At the primary stage, small geographic clusters (or grouping of clusters) are selected in each stratum for a total of about 3,000 clusters. In all 10 provinces, approximately 7 households per cluster are then selected at the final stage from the clusters selected at the previous sampling stage (Gambino et al., 1998).

The clusters of the LFS area frame were formed by grouping together contiguous geographical blocks and include around 200 households each. These clusters generally respect the limit of some sub-provincial regions that are used as the first level of stratification within each province for LFS dissemination purposes. A second and final level of stratification allows for the creation of approximately 1,000 strata by grouping clusters that present similar socio-economic characteristics within the first level strata. The stratification of the most recent LFS frame is based on 2001 Census of Population data.

Some special strata were created to meet the needs of different surveys. Because household spending habits are linked to household income, one type of special strata that is of particular interest for the SHS is the high income strata. These strata are made of clusters with a large prevalence of households with higher income. The high income strata were first introduced to the LFS frame stratification in the redesign of 1994 based on the 1991 Census. At the time, they only covered the 9 largest cities in Canada and were only present in 5 provinces. The strata grouped together enumeration areas in each city with the highest average household income, based on the 1991 Census (Chen et al., 1994). Chun (1995) discusses the expected benefits of the inclusion of such high income strata for aggregate income estimation.

The LFS frame was redesigned in 2004 based on the 2001 Census. Under this redesign, the definition of the

high income strata was modified. A household was arbitrarily qualified as a "high income household" if the annual household income was greater than $125,000. The prevalence of high income households in each cluster was then determined using income tax data. If the prevalence of high income households was greater than a certain level, the cluster was considered to be high income. The high income strata were then formed by grouping high income clusters in the largest cities of each province (Dochitoiu, 2004). The LFS frame now includes at least one high income stratum in the 32 largest cities in Canada and they appear in 9 of the 10 provinces. However, none are in rural areas or in smaller cities. There are now 50 high income strata in total.

Following the LFS redesign, the SHS sample design was modified for the 2005 survey to take into account the increase of the high income strata as well as the increase of their population coverage. In each province, the sample was allocated between each of the high income strata and the set of regular strata proportional to their household income total, as estimated from the 2001 Census. The sample that is allocated to the set of regular strata is then allocated between each stratum proportional to their population size. Some adjustments are made in other strata to reduce the costs associated with the survey, but these areas will not be considered in this paper. For more detailed information see Mitchell (2005).

For household income aggregate estimates, a quick evaluation based on 2001 Census data on household income showed that the current allocation approach could lead to a 14% reduction of variance as compared to the use of a proportional to population size allocation approach.

3. High Income Stratum Definition

One goal of this study is to assess the impact of a modification of the coverage of the high income strata and to include clusters outside of the largest cities. In order to do so, high income stratum alternatives with a different coverage of the population will be defined. Assuming a simple two-stage sampling design, the variance of the household income aggregate estimate will then be estimated for each alternative and compared.

3.1 Data Used

To conduct this study, the 2001 Census data that were collected from the 20% systematic sample of households that received a long form were used. The LFS frame stratum and cluster identifiers were added to these data. Apart from the geographic identifiers, this file also includes the census weight, household income, household wages and salaries and household investment income. The study is limited to the province of Ontario.

3.2 High Income Stratum Alternatives

In order to simplify the methodology, the high income strata used by the LFS frame were combined to form a unique high income stratum that is referred to as the "current high income stratum". The remaining strata in the province were left intact. This stratification is the baseline used by the study and is considered to be the "current stratification".

Using the data described in 3.1, it is possible to estimate the population coverage for a particular stratum (stratum h) as follows:

$$\hat{C}_h = \frac{\hat{N}_h}{\hat{N}} = \frac{\sum_{j=1}^{n_h} w_j^{Census}}{\sum_{h=1}^{H}\sum_{j=1}^{n_h} w_j^{Census}}$$

where $\hat{N}$ represents the estimated number of households in the province, $\hat{N}_h$ represents the estimated number of households in stratum h, n_h represents the number of households sampled in stratum h by the census, w_j^{Census} is the census weight and H represents the total number of strata in the province. Using this approach, the population coverage of the current high income stratum is estimated at 5.8%.

As for the current stratification, we arbitrarily qualify households as "high income households" if the annual household income is greater than $125,000. For each cluster i that is not already part of the current high income stratum, we can compute the proportion of high income households as follows:

$$\hat{P}_i^{125,000} = \frac{\hat{N}_i^{125,000}}{\hat{N}_i} = \frac{\sum_{j=1}^{n_i} w_j^{Census} \times I_j^{125,000}}{\sum_{j=1}^{n_i} w_j^{Census}}$$

where $\hat{N}_i$ and $\hat{N}_i^{125,000}$ respectively represent the estimated number of households and the estimated number of high income households in cluster *i*, n_i represents the number of sampled census households in cluster *i* and $I_j^{125,000}$ is a dummy variable indicating that household *j* is a high income household. This percentage is also referred to as the prevalence of high income households in cluster *i*.

In order to form the alternative high income strata, clusters that were not defined as high income are added to the current high income stratum in order of prevalence. The ones with the largest prevalence were added first. A coverage level was then targeted for each alternative stratification (10%, 15%, 17%, 20%, 22%, 24% and 40% in table 1). Once the high income stratum reaches a certain level of coverage, then the new high income stratum was considered complete and no additional clusters were added. This was repeated for different levels of coverage to obtain many high income stratum alternatives and thus, many stratification alternatives.

3.3 Sample Allocation

In order to be as similar as possible to the current SHS stratification, n_I = 614 clusters were allocated to the province of Ontario. For each stratification alternative, the 614 clusters were divided between the high income stratum and the remaining strata proportional to the household income aggregate in the two groups of strata. The number of clusters allocated in the high income stratum (*h*=1) and in the regular stratum (*h*=0) can be expressed as follows:

$$n_{h,I} = n_I \frac{\sum_{s_h} w_j^{Census} x_j}{\sum_s w_j^{Census} x_j}, \quad h = 0, 1$$

where $n_{h,I}$ is the number of clusters allocated in the high or regular strata and x_j is the household income for household *j* in the 20% census sample. Note that *s* comprises the whole 20% census sample, whereas s_h includes only the households in stratum *h*.

The $n_{0,I}$ clusters allocated to the regular strata are then allocated between these strata using the following methodology. Each stratum was initially allocated a minimum of one cluster, to ensure that the sample covered all strata. The remaining clusters were allocated to each stratum proportional to their population size. This could be expressed as follows:

$$n_{h,I} = 1 + (n_{0,I} - H + 1)\frac{\hat{N}_h}{\hat{N}} = 1 + (n_{0,I} - H + 1)\frac{\sum_{s_h} w_j^{Census}}{\sum_s w_j^{Census}}$$

$$h = 2, ..., H$$

The study is only considering one high income stratum, so no further allocation of clusters was required. For each cluster that was sampled, $n_{h,i}$=6 households were selected. Therefore, for every alternative stratification, 3,684 households were allocated.

3.4 Sampling Design

The sampling design considered was a stratified two stage SRS-SRS sample. For each targeted level of coverage, the sample allocation was computed and the variance of the household income aggregate estimate was produced. The variance was then compared to the variance of the current stratification.

At the second stage of sampling, we assume that $n_{h,i}$ households are sampled through SRS among the $N_{h,i}$ households in the clusters. In practice $N_{h,i}$ was estimated from the census sample and the sample of $n_{h,i}$ households is the results of an SRS selected from the systematic census sample.

3.5 Estimator Used and Variance Estimation

The following estimator of the household income aggregate was assumed:

$$\hat{Y} = \sum_{h=1}^{H} \frac{N_{h,I}}{n_{h,I}} \sum_{i \in s_{h,I}} \frac{N_{h,i}}{n_{h,i}} \sum_{k \in s_{h,i}} y_k$$

where $N_{h,I}$ and $n_{h,I}$ represent the number of clusters and the number of sampled clusters in stratum *h* respectively, $N_{h,i}$ and $n_{h,i}$ represent the number of households and the number of households selected in cluster *i* of stratum *h* respectively, $s_{h,I}$ refers to the sample of clusters in stratum h while $s_{h,i}$ refers to the sample of households in cluster *i* of stratum *h*.

The variance of the household income aggregate estimate can be calculated for a stratified two stage SRS-SRS design as follows:

$$V_{2st}(\hat{Y}) = \sum_{h=1}^{H} {N_{h,I}}^2 \frac{1-f_{h,I}}{n_{h,I}} S^2_{tU_{h,I}} + \sum_{h=1}^{H} \frac{N_{h,I}}{n_{h,I}} \sum_{U_{h,I}} {N_{h,i}}^2 \frac{1-f_{h,i}}{n_{h,i}} S^2_{yU_{h,i}}$$

Note that in practice the variances $S^2_{tU_{h,I}}$ and $S^2_{yU_{h,i}}$ are estimated from the 20% census sample, as well as $N_{h,i}$ and $f_{h,i}$, respectively the number of households and the sampling fraction in cluster *i* of stratum *h*. $U_{h,I}$ and $U_{h,i}$ represent the set of clusters in stratum *h* and the set of households in cluster *i* of stratum *h* respectively.

3.6 Results

The results for different levels of coverage of high income households are shown in Table 1. The table demonstrates that the variance of total household income can be reduced by covering a larger part of the population and thus a larger proportion to the high income households with the high income stratum. The optimal stratification alternative is to have the high income strata covering 22% the population. This coverage level is highlighted in Table 1. It reduces the variance by 11.1% as compared to the current stratification. It is the optimal alternative stratification for the household income aggregate. This however, will vary from province to province.

The results presented in Table 1 also show that the efficiency of the high income strata is quite robust to population coverage variation. For high income stratum alternatives that cover between 15% and 30% of the population, the variance reduction only varies from 9.7% to 11.1%. For the remainder of the paper, we will consider the coverage level of 22% in the high income stratum to be the alternative stratification.

The alternative stratification has a prevalence of 28%, which is still not bad considering that the alternative high income stratum covers 22% of the province.

Table 1 - A Comparison of Alternative Stratifications for the SHS Sample in Ontario

High Income Stratum				Variance Reduction (%)
Population Coverage	High Income Hhld Coverage	Prevalence	% of Sample in High Income	
5.8%	22%	38%	12%	---
10%	36%	37%	20%	5.8%
15%	48%	33%	27%	9.7%
17%	53%	31%	29%	10.8%
20%	58%	29%	33%	11.0%
22%	61%	28%	36%	11.1%
24%	64%	27%	38%	10.8%
30%	72%	24%	45%	10.1%
40%	83%	21%	55%	7.4%

4. The Simulation

While the results in 3.6 have shown that a reduction in variance can be obtained for household income, some other scenarios will be considered. Calibration might reduce the impact of the alternative stratification. Furthermore, different variables that are not as well correlated with household income could show less desirable results. A simulation study was undertaken to investigate these possibilities.

4.1 Methodology

The census data that were used to look into the alternative stratification were also used for the simulation. Note that the same methodology and design that is described in section 3 was used at the sample selection stage of the simulation. The two stratification alternatives that were compared are the current stratification (population coverage of 5.8%) and the alternative stratification (population coverage of 22%). The differences between alternatives are the

number of clusters allocated and the coverage of the high income households in the high income stratum.

The census data does not include expenditure variables. For this reason, it was decided to use the variables household income, wages and investment income. Household income was desirable because it was the variable used to determine the alternative stratification. Wages was used because it shows a fairly strong relationship with household income and represents the most important source of income for the majority of Canadians. Investment income shows a more asymmetric distribution and its linear relationship with household income is not as strong. A few expenditure variables such as additions and renovations have comparable distributions and relationships with household income.

In addition to the estimator used in section 3, a calibrated estimator is used for the simulation. The calibration methodology used is described in Lemaître and Dufour (1987). Three calibration groups were used for this study, for a total of 11 control totals. They were:

age (0-17, 18+)
household size (1, 2, 3+)
wages (6 categories)

The controls on age and wages represent person level calibration constraints while the household size controls represent household level calibration constraints. Only province level controls were used. There were no sub-provincial controls.

The calibration method used is the same as the one used for the SHS weighting. However, the SHS uses more provincial controls and also uses a few sub-provincial controls. The simulation study uses less controls for simplicity.

The variables of interest are only available for 20% of the data from the census. Therefore, this 20% sample was considered as the population for the simulation. At each iteration, a two stage SRS-SRS sample was pulled from the population. 614 clusters were selected at the first stage and 6 households were selected per cluster at the second stage. The total number of iterations used was 2500.

After the 2500 iterations were completed for each method, the Monte Carlo mean squared error (MSE) and relative bias (RB) were calculated as follows:

$$MSE = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_r - Y)^2$$

$$RB = \frac{1}{R}\sum_{r=1}^{R}\frac{(\hat{Y}_r - Y)}{Y}$$

In the formulas above, $\hat{Y}_r$ is the estimate obtained for replicate R of the simulation. The value for Y is considered to be the "true value" obtained from the whole population.

The mean squared error (MSE) will be approximately equal to the variance, with a sufficient number of iterations. The relative bias should tend to zero if the estimate is unbiased.

4.2 Results

The results from the simulation are shown in Table 2. It shows that the gain from using the alternative stratification is similar for all of the variables. The improvement in variance without calibration is slightly less for investment income (17.8%) as compared to household income (20.9%) and wages (20.2%). Calibration lessens the gain in variance for all of the variables but especially household income (20.9% to 12.5%) and wages (20.2% to 16.9%). The difference between calibrated and uncalibrated estimates is less important for investment income (17.8% to 17.1%).

Table 2 - The Percentage Improvement in MSE from the Current Stratification

Variable of Interest	Without calibration	With calibration
Household Income	20.9%	12.5%
Wages	20.2%	16.9%
Investment Income	17.8%	17.1%

Note that the results in Table 2 without calibration are different than the results shown in Table 1. Part of the difference can be attributed to the fact that the simulation uses the 20% sample from the census data as the population. The results in Table 1 use all of Ontario.

The relative bias of all estimates without calibration was approximately zero. There was a small amount of bias under calibration, but this is to be expected. Furthermore, the results were repeated with additional iterations and they were shown to be stable.

5. Conclusion

The results from the first part of this study have shown that the coverage of high income households is an important factor to consider when developing high income strata. An alternative stratification was created for the province of Ontario by increasing the coverage of the high income strata for the LFS survey frame. The alternative stratification would reduce the overall variance of total household income. It was also observed that the coverage of the high income strata can vary substantially without much impact on their efficiency.

A simulation study was then conducted in the second part of this paper in order to compare the current stratification to the alternative stratification and to observe the impact of calibration. The results from the study showed that an increased coverage of high income households can improve the efficiency of estimates that are less correlated to household income. While calibration reduced the impact of the results, there was still a gain from using the alternative stratification.

Some caution should be taken in interpreting the results from this study. For simplification purposes, the SRS-SRS design used in this paper is different from the SHS design. In addition, the methods used in this paper do not correspond to those that would be used for the stratification of the LFS frame. In theory, at each coverage level, a re-stratification would be performed. This could impact the results. Finally, the LFS frame was designed using tax data whereas this study used census data. The two sets of auxiliary information would lead to different stratifications.

Acknowledgements

The authors would like to thank David Haziza and Gordon Kuromi for their help with the simulation part of this study and John Lindeyer, Edward Chen, Ron Carpenter and Victor Estevao for their valuable advice.

References

Aresenault, S., Gaudet, J., Nadeau, C. and Tremblay, J. (2001). Introduction of a New Calibration Strategy for the Survey of Household Spending, *2001 Proceedings of the Annual Meeting of the American Statistical Association*, on CD-ROM.

Chen, E.J., Gambino, J., Laniel, N. and Lindeyer, J. (1994). Design and Estimation Issues for Income in the Redesign of the Canadian Labour Force Survey, *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Chun, B. (1995). Efficiency of Income Estimates Using Income Stratification Variable. Household Survey Methods Division, Working Paper, Statistics Canada.

Dochitoiu, C. (2004). Forming Special Strata for the LFS. Internal Document, Statistics Canada.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B., Lindeyer, J. (1998) Methodology of the Canadian Labour Force Survey, 1995-2004. Statistics Canada, Catalogue Number 71-526-XPB.

Lemaître, G. and Dufour, J. (1987). An Integrated Method for Weighting Persons and Families, Survey Methodology, 13, 2, 1999-207.

Mitchell, C. (2005). The Provincial Sampling Design for the 2005 Survey of Household Spending. Internal Document, Statistics Canada.

Tremblay, J. (2005). Apperçu de la stratégie de calage harmonisée des statistiques du revenu de Statistique Canada, *Recueil 2005 de la Section des méthodes d'enquête*, sur CD-ROM, Société Statistique du Canada.