



Catalogue no. 11-522-XIE

**Statistics Canada International Symposium
Series - Proceedings**

**Symposium 2004: Innovative
Methods for Surveying
Difficult-to-reach Populations**

2004



IMPROVING THE QUALITY OF ESTIMATES FOR A LOW-INCOME POPULATION: USE OF A DUAL FRAME IN THE SURVEY OF HOUSEHOLD SPENDING

Bruno Lapierre, Christian Nadeau, Johanne Tremblay and José Gaudet¹

SUMMARY

Statistics Canada's Survey of Household Spending is primarily intended to provide reliable provincial estimates of household spending. The multi-stage sample is selected from an area frame covering the entire population, and the data are collected by means of a personal interview. During the 2003 survey, an additional objective was to improve the quality of the estimates for a sub-population of low-income households representing approximately 2.5% of all households in the province of Quebec. To meet this objective, a supplementary sample was selected from a list of dwellings located in a limited number of geographic areas previously identified using auxiliary data. In this article we present the dual sample design used, along with different scenarios considered and some findings that led to the choices made.

KEYWORDS: Optimum Allocation; Rare Population; Stratification.

1. INTRODUCTION

In the context of developing an anti-poverty strategy, the *Ministère des finances du Québec* (MFQ) wants to redo a study conducted by the Ministère de la Main d'œuvre et de la Sécurité du Revenu that sought to establish minimum income thresholds for the province of Quebec (Fugère and Lanctôt, 1985). The model chosen for this study is largely based upon estimates of average expenditures for a sub-population of low-income households representing only 2.5% of all households in Quebec. To redo this study using more precise and up-to-date spending estimates than those currently available, the MFQ funded the addition of a supplementary sample to the Survey of Household Spending (SHS) of Statistics Canada (STC).

The regular SHS sample has a sample of 3,000 Quebec dwellings selected from an area frame using a multi-stage stratified design (Arsenault and Tremblay, 2001). According to this plan, the size of the supplementary sample needed to achieve the MFQ's objective is estimated at more than 4,000 dwellings. A sample design to better target the population of interest defined by the MFQ was devised so as to reduce the total sample size required.

This document presents the main features of the sample design devised to meet the needs of the MFQ, along with some findings that led to the choices made. It begins with a brief background and a description of the MFQ's objective. In Section 3, various concepts are discussed relating to the efficiency of a sample design when seeking to estimate means for a rare population. Various findings and analyses that led to the choice of a sample frame are presented in Section 4. Section 5 describes and justifies the choice of methods for stratifying the sample frame, allocating the sample and selecting dwellings. The article concludes with Section 6, which provides a brief description of the sample design chosen.

¹Bruno Lapierre, Statistics Canada, Main Building, Room 2500, Ottawa, Canada, K1A 0T6, Bruno.Lapierre@Statcan.ca; Christian Nadeau, Statistics Canada, R.H. Coats Building, 16th floor, Ottawa, Canada, K1A 0T6, Christian.Nadeau@Statcan.ca; Johanne Tremblay, Statistics Canada, R.H. Coats Building, 16 floor, Ottawa, Canada, K1A 0T6; José Gaudet, Statistics Canada, R.H. Coats Building, 11th floor, Ottawa, Canada, K1A 0T6.

2. BACKGROUND AND OBJECTIVE

The SHS is an annual survey designed to collect detailed data on household spending for the year preceding the collection period, using personal interviews that last an average of 110 minutes. The survey imposes a heavy burden on respondents, and the use of experienced interviewers helps to limit the impact of non-sampling errors on the quality of the data.

The MFQ's objective is to improve the accuracy of the estimates of ten expenditure variables² for households in the first income decile among one-earner households for which at least 50% of income comes from remuneration. According to data from the 2001 Census, these account for only 2.5% of Quebec households. The coefficients of variation (CV) of these estimates, obtained in SHS 2000 and SHS 2001, are shown in Table 2.1. According to the quality criteria established in the agreement between STC and the MFQ, the goal is to obtain CVs of less than 15% for the estimates of at least seven of the ten variables. As may be seen, the CVs obtained from the regular SHS sample are clearly greater than 15% for four of the ten variables and are substantially below the 15% threshold for only two variables.

Table 2.1: CVs of estimates of means for the population of interest, 2000 and 2001 SHS

	SHS 2000	SHS 2001
Food	8.4%	6.6%
Furnishings	24.3%	24.0%
Communications	10.7%	13.9%
Household operation	12.2%	14.5%
Clothing	13.9%	14.8%
Reading	21.5%	33.4%
Shelter	5.4%	4.4%
Recreation	23.4%	17.9%
Personal care	15.0%	14.7%
Transportation	31.8%	20.9%

We wanted to devise a sample design in order to select a supplementary sample that will serve to achieve the MFQ's objective. We considered it important to control the size of the supplementary sample in order to limit both the hiring of new interviewers and an extension of the collection period. We thus hope to minimize the risks of a reduction in the quality of the data collected and a delay in to the timetable for release. The agreement between the MFQ and STC stipulates that the size of the supplementary sample must be between 2,000 and 3,000 dwellings. It also states that preference should be given to those sample designs that meet the quality criteria established on the basis of a sample of 2,000 dwellings. The option of increasing the regular sample size of the SHS without making changes to the sampling frame was rejected at the outset, since the required supplementary sample size under this option was estimated at more than 4,000 dwellings.

3. EFFICIENCY OF A SAMPLE DESIGN FOR ESTIMATING A MEAN FOR A RARE POPULATION

To set up a sample design for efficiently estimating means for a rare population, we want to increase the sampling fraction where the rare population is most concentrated. Kalton (2001) discusses the case of allocating a sample between two strata when the goal is to estimate a mean for a rare population under the assumption that the mean and the variance of the variable of interest for the rare population are the same in the two strata. First, the prevalence is defined by $P = M/N$, where M and N represent respectively the size of the rare population and the size of the total

²The definitions of the expenditure variables used for the MFQ are defined in Nadeau et al. (2005) and may differ from the definitions used in the SHS.

population. Similarly, the prevalence in stratum h is defined by $P_h = M_h/N_h$ and the coverage of the rare population by stratum h is defined by $A_h = M_h/M$. When simple random sampling is used in each stratum, the optimal allocation is an allocation of the sample proportional to $N_h\sqrt{P_h}$. Kalton (2001) shows that in order for the optimal allocation to result in an appreciable reduction in the variance in comparison to an allocation proportional to size, the stratum with the highest prevalence ($h=1$) must be such that the values of P_1/P and A_1 are relatively high. In other words, in order for the efficiency gain to be appreciable, such a stratum must have a prevalence much greater than that observed in the population, while covering a substantial portion of the rare population. These principles are used in choosing the sampling frame, stratification methods and sample allocation presented in the following sections.

4. CHOICE OF SAMPLING FRAME

The first step in developing a sample design is to identify the sampling frame according to which the supplementary sample will be selected. The sampling frame options considered for increasing the size of the SHS sample in Quebec will be grouped into two major categories.

A first set of options is to select the supplementary sample based upon the area frame that is used to select the regular sample for the SHS. One particular approach often used at STC is to select households that have already participated in other surveys that select their sample from the same sampling frame. With this option, information already available on the respondents can be utilized, and thus the households to be included in the population of interest can be better targeted. However, in light of the already heavy response burden of the SHS, this option was eliminated at the outset. Another approach considered was to select the supplementary sample from the area frame but to use a sample design different from the one used for the regular sample. This approach was rejected because of the difficulty of assigning relevant and up-to-date auxiliary information to geographic units in the area frame that are small enough to effectively target the population of interest (Nadeau et al., 2005).

The other set of options consists of selecting the SHS sample from a dual frame, that is, to use the area frame only to select the regular SHS sample and to identify another frame covering only a part of the Quebec population for selecting the supplementary sample. This second sampling frame must be designed in such a way so as to respect the principles set out in the preceding section; it must feature both a prevalence and a coverage of the population of interest that are large enough for the efficiency gain to be appreciable.

To limit the cost and complexity of operations and to ensure that the addresses in the sampling frames considered for selection of the supplementary sample are of good quality, these frames all correspond to dwelling lists extracted from Statistics Canada's Address Register (AR) (Swain et al., 1992). The AR is kept up to date by STC in geographic areas with higher population densities. The version of the AR that is used is the one that results from the update that followed the 2001 Census; it covers 86% of the Quebec population.

The main options considered are described in sections 4.1 and 4.2, and then the accuracy of the estimates for samples drawn from these sampling frames is evaluated in section 4.3.

4.1 Post-censal approach

Post-censal surveys are usually identified and planned before the Census. The census questionnaire is used as the first stage of a two-stage sample design so as to better target the population of interest. Households or persons with certain characteristics form the sampling frame, and a sample is selected from it. The approach studied here differs from traditional post-censal surveys in the sense that the second-stage sample is made up of dwellings and not households, so as to avoid the tracing operations needed to locate households or household members who have moved since the 2001 Census. For the first time in the 2001 Census, the questionnaire mentioned the possibility of using the census data for purposes of sample selection for other STC surveys. Such a possibility was recently exploited in an STC survey (Duggan, Neusy and Bélanger, 2003) and can thus be considered for the selection of a sample for the SHS.

Our intention was to create a list of dwellings likely to be occupied by households in the population of interest when collecting data for the SHS 2003, using data on the households occupying those dwellings in the 2001 Census. To identify the characteristics of households so as to better identify dwellings that might be occupied by households in the population of interest three years later, tables were constructed cross-tabulating characteristics of households in year “0” and belonging or not belonging to the population of interest in year “3,” using data from the Survey of Labour Income Dynamics (SLID) (see Lavigne and Michaud, 1998 for further details on SLID).³ The results obtained using these tables led to a study of a list frame consisting of dwellings occupied by a household in the “expanded population of interest”⁴ in the 2001 Census for the post-censal approach. When the appropriate proportions observed in the cross-tabulations based on the SLID data are applied to the 2001 Census data, prevalence within this frame is calculated to be 8.3%, and the coverage of the population of interest to be 3.0% for reference year 2003.

4.2 Targeting of high-concentration geographic areas

Households in the population of interest have certain common characteristics, such as low income and renter status, that suggest that they will live in higher concentrations in particular geographic areas. We want to create a list frame made up of households located in geographic areas with a high concentration of households in the population of interest. Such geographic areas can be identified using data from the 2001 Census. These data contain information on household income, work and socio-demographic characteristics as well as geography. Income data in the Census refer to the year 2000 and are available for one household in five. The geographic units selected must be small enough to permit precise targeting of areas where members of the population of interest are concentrated. They must also be large enough to be able to attribute some reliability to the estimate of prevalence.

Three list frames with a high geographic concentration are considered here, and their characteristics are shown in Table 4.1. All three consist of dwellings located in dissemination areas (DAs) exhibiting the highest prevalences amongst those covered by the AR according to the 2001 Census data. A DA is a small, relatively stable geographic unit composed of one or more neighbouring blocks with a population of 400 to 700 persons (approximately 250 households on average). This is the smallest standardized geographic area for which all census data are disseminated. The province of Quebec is divided into 12,153 DAs.

Table 4.1 : Geographic concentration: Characteristics of list frames studied

List frame options with high geographic concentrations				
	Number of DAs in list frame	Prevalence in list frame	Coverage of population of interest	Coverage of total population
Geographic frame 1	1,083	9.5%	38%	10%
Geographic frame 2	2,125	7.3%	59%	20%
Geographic frame 3	5,060	4.4%	88%	50%

These frames differ from one another in terms of the proportion of the total population that they cover. The increase in the coverage of the population is obtained by adding DAs with increasingly low concentrations. The result is a reduction in overall prevalence in the frame and an increase in the coverage of the population of interest. It may also be seen that because of the way the frames are constructed, geographic frame 1 is included in geographic frame 2,

³SLID collects labour and income data on the same persons over a period of six consecutive years. It was therefore possible to analyse the characteristics of households over a three-year period. The data used here cover the period from 1997 to 2000.

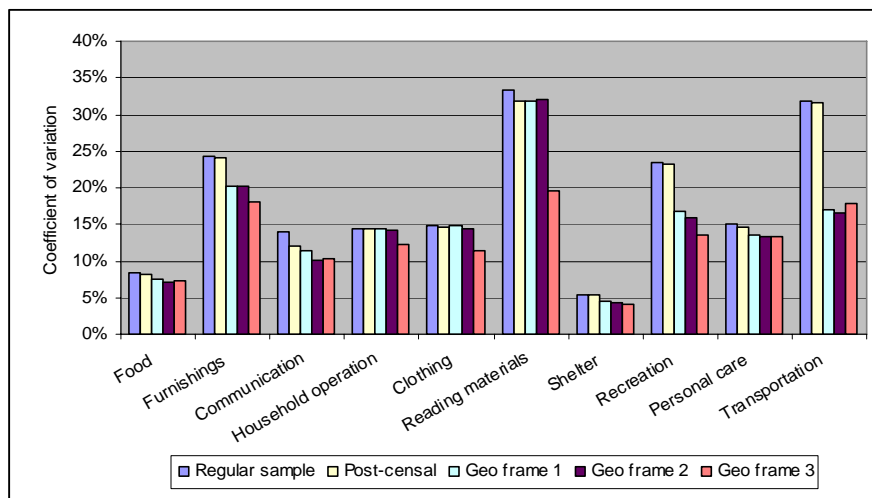
⁴As was seen in Section 2, the population of interest consists of the 10% of households with the lowest income in the subpopulation of one-earner households having at least 50% of the household income coming from remuneration. The term “expanded population of interest” means the 20% of households with the lowest income in the same subpopulation.

which in turn is included in geographic frame 3. To choose an efficient frame, it is therefore necessary to find the best compromise between prevalence and the coverage of the population of interest.

4.3 Evaluation of options and choice of list frame

To evaluate the four dual frame options identified, namely the post-censal approach and the three geographic frames, we calculate the CVs expected for the estimates of the means of the ten expenditure variables for the population of interest. Figure 4.1 shows the CVs calculated by combining the regular SHS sample and a supplementary sample of 2,000 dwellings. The CVs of the estimates obtained using only the regular SHS sample are also shown in this chart for purposes of comparison. The methodology used to calculate the expected CVs is described in the appendix. Additional details are provided in Nadeau et al. (2005).

Figure 4.1: Expected coefficients of variation for estimates of mean expenditures for the population of interest according to different list frame options



It is immediately clear that adding a sample of 2,000 dwellings using the post-censal option yields only a very slight improvement in CVs compared with the regular sample. Geographic frame 3, covering the largest proportion of the population of interest, yields better CVs for most of the variables. It is especially efficient for the variables with the highest CVs using the regular sample. Thus, for CVs, a good coverage of the population of interest by the frame appears to be more important than better prevalence.

As may be seen, the use of geographic frame 3 is the only option that satisfies the quality criteria set out in the agreement between STC and the MFQ, namely, CVs of less than 15% for the estimates of at least seven of the ten variables. The list frame formed by the dwellings in DAs that have the highest prevalences and cover 50% of all households in Quebec is therefore the one used to select the supplementary sample.

5. CHOICE OF METHODS OF STRATIFYING THE SAMPLING FRAME AND ALLOCATING AND SELECTING THE SAMPLE

The results presented thus far have led to the choice of a sampling frame for selecting the supplementary sample. The second stage consists in determining the methods of stratifying the frame, allocating the sample and selecting dwellings in such a way as to reduce the CVs for the estimates of means for the population of interest.

5.1 Stratification of the sample frame

While the main objective is to improve the quality of estimates for the population of interest at the provincial level, the possibility of using the data at another level in the analysis was raised by the MFQ during the project. To do so, Quebec was divided into three strata: the Montreal census metropolitan area (CMA), the Quebec City CMA, and the rest of Quebec. By doing this, we wanted to obtain a minimal sample size in each of these regions. As will be seen further on, the chosen allocation scenario ultimately does not take such an objective into account.

A second stage of stratification is created in order to form strata that are homogeneous with respect to the means of the ten expenditure variables in the population of interest. An examination of the data from SHS surveys from 1997 to 2001 shows that mean expenditures on *Transportation* differ considerably according to geography. Therefore, the three first-stage strata were divided into twelve second-stage strata. For the Montreal and Quebec City CMAs, the census economic regions were used as a second stage of stratification. For the rest of Quebec, the other four CMAs each form a separate stratum; another stratum is formed by grouping together all the census agglomerations that are not CMAs, and another by grouping the portion of Quebec outside census agglomerations.

To benefit from the optimal allocation described in Section 3, a third stage of stratification is introduced. We want to divide each of the twelve second-stage strata into two parts so as to increase the efficiency gains obtained using this allocation method. To do so, we partition the DAs so as to form two strata; the first one is made up of DAs with the highest prevalences and the second one, of DAs with the lowest prevalences. Under the assumptions stated in Section 3, Kalton and Anderson (1986) show that the variance of the estimator of the mean is approximately

proportional to $\left(\frac{A_1^2}{M_1 \sqrt{P_1}} + \frac{A_2^2}{M_2 \sqrt{P_2}} \right) (N_1 \sqrt{P_1} + N_2 \sqrt{P_2})$. Using the information on the DAs as to the number of

households in the total population and the population of interest in the 2001 Census, each stratum is partitioned into two sets of DAs in order to minimize this quantity and improve the efficiency gain. We thus obtain twenty-four final strata. The Monte Carlo study described in Section 5.3 confirms that using the third stage of stratification yields a reduction in the variance compared with the two-stage stratification design.

5.2 Sample allocation

The supplementary sample is first allocated among the first-stage strata proportionally to the number of dwellings in each of the three strata. This option represents a particular case of a family of power allocation methods evaluated in order to arrive at a compromise satisfying both provincial and sub-provincial objectives. While the sub-provincial objective was abandoned along the way, it was advantageous, from an operational standpoint, to use one of the methods evaluated during the process. Also, given the very similar prevalences for the three first-stage strata, an optimal allocation would have been very similar to the one thus obtained.

The sample of each first-stage stratum is distributed between second-stage strata according to the optimal allocation method described in Section 3, where the size of the sample attributed to each stratum is proportional to $N_h \sqrt{P_h}$. Thus, a larger proportion of the sample is allocated to strata exhibiting the highest prevalences than would have been obtained using a allocation proportional to size. The optimal allocation is also used to distribute the sample allocated to the second-stage strata amongst third-stage strata, since these were constructed specifically in order for this type of allocation to be efficient.

5.3 Selection of dwellings

The two-stage sampling method is used to select dwellings within each stratum so as to avoid excessive dispersion of the sample. The primary sampling units (PSUs) correspond to the DAs and are selected with a probability proportional to their size, that is, to the number of dwellings in the list frame. A given number of dwellings are then selected within each selected DA, using systematic sampling. The dwellings are ranked in advance according to a sequence that assures some geographic proximity to consecutive dwellings in the file.

A Monte Carlo study was conducted to measure the efficiency of two options for selecting dwellings in the second stage. A first option was to select five dwellings in each of the DAs selected, which is the average yield of the PSUs in the sample design for the regular SHS sample. The second option was to select three dwellings per DA in the twenty strata for CMAs and five dwellings per DA in the other four strata. This option would make it possible to select more PSUs in CMAs, while limiting the dispersion of the sample in smaller urban areas and rural areas.

The simulation was carried out using data from the 2001 Census. It consisted of selecting 1,000 samples, then calculating the variance of the estimates for households in the population of interest, for the two sample designs under evaluation. Since the Census does not include expenditure variables other than *rental cost*, three income variables were also used to compare the variability of the estimates obtained according to the different designs. As noted in Section 5.1, a design limited to the first two stratification stages was also studied, so as to evaluate the efficiency of the third stage of stratification. Lastly, the variance according to a simple random sample design was calculated in order to be able to derive the design effect of each of the different options.

Contrary to expectations, the Monte Carlo study showed that selecting three dwellings for DAs in strata with CMAs (and thus increasing the number of PSUs selected) has little effect on the quality of the estimates of the variables studied for the population of interest. Even though the results for the two options are very similar, the selection of three dwellings per PSU was chosen, since selecting a greater number of PSUs generally results in a reduction in the variance.

This simulation indicates that the design effect that results when this design is used is less than the design effect observed when the regular sample is used. The impact of this efficiency gain in relation to the results presented in Section 4.3 is described in Nadeau et al. (2005).

The size of the supplementary sample was ultimately set at 2,211 dwellings, following a re-evaluation of the collection constraints once its geographic distribution was known. As a result of matching between this sample and the regular SHS sample, the risks of overlap between them were minimized.

6. CONCLUSION

The MFQ wants to improve the quality of the estimates of average expenditures from the SHS for households in the first income decile among one-earner households for which at least 50% of income comes from remuneration. An increase of 4,000 dwellings in the size of the regular SHS sample is considered necessary for achieving the MFQ's objective if no change is made to the SHS sample design. A sample design intended to better target the population of interest was developed in order to reduce the size of the supplementary sample needed to achieve this objective.

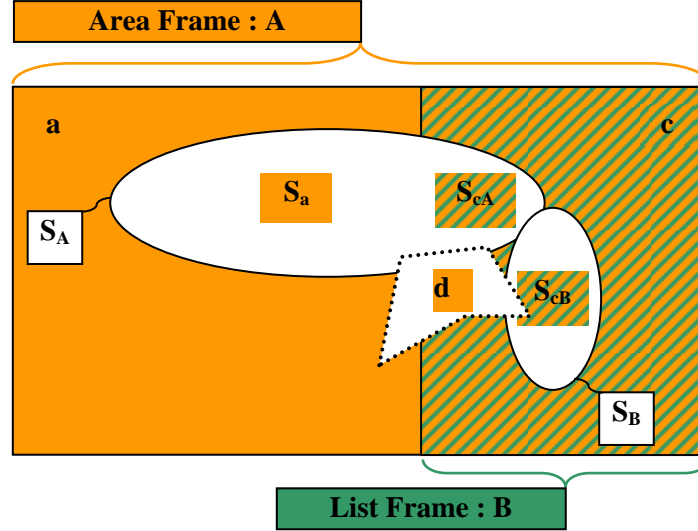
The MFQ's supplementary sample was selected from a list of dwellings according to a two-stage stratified sample design. The list frame used consists of dwellings entered in the AR in DAs with the highest prevalences in the 2001 Census, so as to cover 50% of Quebec households. It is stratified according to three stages, with the first two stages being geographic in nature and the third based upon the concentration of households belonging to the population of interest. The allocation of the sample between first-stage strata is proportional to the number of dwellings in these strata, whereas the allocation for subsequent stratification stages takes account of the number of dwellings and the concentration of household belonging to the population of interest, in such a way as to reduce the variance as much as possible. Within each stratum, a given number of DAs were selected with a probability proportional to the number of dwellings. In each of the DAs selected, systematic sampling was used to select three dwellings in the CMAs and five dwellings elsewhere. The size of the supplementary sample was determined attempting to satisfy the quality criteria established in the agreement between STC and the MFQ while minimizing the impact on the SHS; it was set at 2,211 dwellings.

The actual efficiency of the design chosen is currently being evaluated, with a view to presenting the results following the release of the 2003 SHS.

APPENDIX – METHODOLOGY FOR CALCULATING EXPECTED COEFFICIENTS OF VARIATION

We want to evaluate and compare the quality of the estimates of the mean for the domain of interest that are obtained by selecting an SHS sample from each of the four dual frames described in Section 4. Figure A.1 illustrates the use of these dual frames.

Figure A.1: Dual frame consisting of an area frame covering the total population and a list frame covering only part of the population



Area frame **A**, from which is drawn the regular sample S_A , covers the entire Quebec population. A list frame **B**, which covers only a part of the Quebec population, is used to draw supplementary sample S_B . The area frame may be divided into two parts, one not covered by the list frame ($a=A \cap B^c$) and the other covered by the list frame ($c=A \cap B$), and sample S_A is divided in similar fashion to obtain $S_a=S_A \cap a$ and $S_{cA}=S_A \cap c$. Note that since $B \subseteq A$, then $c=B$ and $S_{cB}=S_B$. The domain defined by the MFQ's population of interest will be denoted by **d**.

The notation used for the estimators and population parameters is such that the first index is used to designate either the population or the domain. In the case of the estimators, a second index designates the sampling frame used, while an exponent is used to indicate the type of estimator where necessary ("r" for a regression estimator and "π" for a Horvitz-Thompson estimator). In general, \hat{Y} represents an estimator of the mean of a population or a domain \bar{Y} , \hat{Y} represents an estimator of a total, while \hat{N} represents an estimator of N , the size of a population or domain, as the case may be.

First assume the use of the estimator $\hat{Y}_{d,AB} = \hat{Y}_{d,AB} / \hat{N}_{d,AB}$ where $\hat{Y}_{d,AB} = \hat{Y}_{d,A}^r - \alpha(\hat{Y}_{d \cap c, A}^r - \hat{Y}_{d \cap c, B}^\pi)$, $\hat{N}_{d,AB} = \hat{N}_{d,A}^r - \alpha(\hat{N}_{d \cap c, A}^r - \hat{N}_{d \cap c, B}^\pi)$, and where $0 < \alpha < 1$. Using the Taylor linearization, we obtain the following approximate variance formula:

$$V(\hat{Y}_{d,AB}) \approx \frac{1}{N_d^2} \left[V(\hat{Y}_{d,AB}) + \bar{Y}_d^2 V(\hat{N}_{d,AB}) - 2\bar{Y}_d C(\hat{Y}_{d,AB}, \hat{N}_{d,AB}) \right] \quad (A.1)$$

$$\text{where } V(\hat{Y}_{d,AB}) = V(\hat{Y}_{d,A}^r) + \alpha^2 (V(\hat{Y}_{d \cap c, A}^r) + V(\hat{Y}_{d \cap c, B}^\pi)) - 2\alpha C(\hat{Y}_{d,A}^r, \hat{Y}_{d \cap c, A}^r), \quad (A.2)$$

$$\text{and } V(\hat{N}_{d,AB}) = V(\hat{N}_{d,A}^r) + \alpha^2 (V(\hat{N}_{d\cap c,A}^r) + V(\hat{N}_{d\cap c,B}^\pi)) - 2\alpha C(\hat{N}_{d,A}^r, \hat{N}_{d\cap c,A}^r). \quad (\text{A.3})$$

Since domains \mathbf{c} defined by the list frames considered are not identifiable using SHS data, we define, for each of the frames, a proxy domain \mathbf{c}' , identifiable using SHS data, for which the coverage of the population of interest and the incidence are similar to those observed for domain \mathbf{c} ($N_{d\cap c}/N_d \approx N_{d\cap c'}/N_d$ et $N_{d\cap c}/N_c \approx N_{d\cap c'}/N_{c'}$). Assume that $S_{d\cap c}^2 = S_{d\cap c'}^2 = S_d^2$, $\bar{Y}_{d\cap c} = \bar{Y}_{d\cap c'} = \bar{Y}_d$ and that the design effects are such that $\text{deff}(\hat{Y}_{d\cap c,A}^r) = \text{deff}(\hat{Y}_{d\cap c',A}^r)$, $\text{deff}(\hat{Y}_{d\cap c,B}^\pi) = \text{deff}(\hat{Y}_{d\cap c',A}^\pi)$, $\text{deff}(\hat{N}_{d\cap c,A}^r) = \text{deff}(\hat{N}_{d\cap c',A}^r)$ and that $\text{deff}(\hat{N}_{d\cap c,B}^\pi) = \text{deff}(\hat{N}_{d\cap c',A}^\pi)$.

Under these assumptions, and using formulas for the variance of Horvitz-Thompson estimators of a size and a total for a domain according to a simple random design, we obtain:

$$V(\hat{Y}_{d\cap c,A}^r) = V(\hat{Y}_{d\cap c',A}^r) \frac{(N_{d\cap c} - 1)S_d^2 + N_{d\cap c}(1 - N_{d\cap c}/N_A)\bar{Y}_d}{(N_{d\cap c'} - 1)S_d^2 + N_{d\cap c'}(1 - N_{d\cap c'}/N_A)\bar{Y}_d}, \quad (\text{A.4})$$

$$V(\hat{Y}_{d\cap c,B}^\pi) = V(\hat{Y}_{d\cap c',A}^\pi) \left(\frac{N_B^2}{N_A^2} \right) \left(\frac{1 - n_B/N_B}{1 - n_A/N_A} \right) \left(\frac{N_A - 1}{N_B - 1} \right) \left(\frac{(N_{d\cap c} - 1)S_d^2 + N_{d\cap c}(1 - N_{d\cap c}/N_B)\bar{Y}_d}{(N_{d\cap c'} - 1)S_d^2 + N_{d\cap c'}(1 - N_{d\cap c'}/N_A)\bar{Y}_d} \right), \quad (\text{A.5})$$

$$V(\hat{N}_{d\cap c,A}^r) = V(\hat{N}_{d\cap c',A}^r) \left(\frac{N_{d\cap c}}{N_{d\cap c'}} \right) \left(\frac{N_A - N_{d\cap c}}{N_A - N_{d\cap c'}} \right) \text{ and} \quad (\text{A.6})$$

$$V(\hat{N}_{d\cap c,B}^\pi) = V(\hat{N}_{d\cap c',A}^\pi) \left(\frac{N_B - n_B}{N_A - n_A} \right) \left(\frac{N_A - 1}{N_B - 1} \right) \left(\frac{n_A}{n_B} \right) \left(\frac{N_{d\cap c}}{N_{d\cap c'}} \right) \left(\frac{N_B - N_{d\cap c}}{N_A - N_{d\cap c'}} \right). \quad (\text{A.7})$$

Also let

$$C(\hat{Y}_{d,A}^r, \hat{Y}_{d\cap c,A}^r) = C(\hat{Y}_{d,A}^r, \hat{Y}_{d\cap c',A}^r) \sqrt{\frac{(N_{d\cap c} - 1)S_d^2 + N_{d\cap c}(1 - N_{d\cap c}/N_A)\bar{Y}_d}{(N_{d\cap c'} - 1)S_d^2 + N_{d\cap c'}(1 - N_{d\cap c'}/N_A)\bar{Y}_d}}, \quad (\text{A.8})$$

$$C(\hat{N}_{d,A}^r, \hat{N}_{d\cap c,A}^r) = C(\hat{N}_{d,A}^r, \hat{N}_{d\cap c',A}^r) \sqrt{\left(\frac{N_{d\cap c}}{N_{d\cap c'}} \right) \left(\frac{N_A - N_{d\cap c}}{N_A - N_{d\cap c'}} \right)} \text{ and} \quad (\text{A.9})$$

$$C(\hat{Y}_{d,AB}^r, \hat{N}_{d,AB}^r) = C(\hat{Y}_{d,A}^r, \hat{N}_{d,A}^r) \sqrt{\left(\frac{V(\hat{Y}_{d,AB}^r)}{V(\hat{Y}_{d,A}^r)} \right) \left(\frac{V(\hat{N}_{d,AB}^r)}{V(\hat{N}_{d,A}^r)} \right)}. \quad (\text{A.10})$$

By substituting A.4, A.5 and A.8 in A.2, A.6, A.7 and A.9 in A.3, A.2 and A.3 in A.10, and lastly A.2, A.3 and A.10 in A.1, we obtain an expression of the approximate variance of $\hat{Y}_{d,AB}^r$. The variances and covariances of the different estimators for area frame \mathbf{A} as well as \bar{Y}_d and S_d^2 are replaced by their values estimated from the data from SHS 2000 and SHS 2001, and the population and domain sizes are replaced by their values estimated using data from the 2001 Census in order to calculate an approximate variance. A value of $\alpha = 0.7$ was used for the evaluation; it was a satisfactory compromise for the ten variables for which it was necessary to estimate the mean for each of the frames evaluated. The results presented in Figure 4.1 correspond to the maximum of the CVs obtained using data from SHS 2000 and SHS 2001 for the substitution.

ACKNOWLEDGEMENTS

The authors wish to thank Guy Laflamme and Michel Latouche for their constructive comments which served to improve this article.

REFERENCES

- Arsenault, S. and Tremblay, J. (2001), "Méthodologie de l'Enquête sur les dépenses des ménages", Income Statistics Division, Statistics Canada catalogue No. 62F0026MIF-2001003.
- Duggan, J., Neusy, E. and Bélanger, Y. (2003), "Sample Design Issues in a Large-scale Multiple frame National Survey: The Canadian Component of the Adult Literacy and Life-skills Survey", *Proceedings of the Survey Methods Section*, SSC Annual Meeting.
- Fugère, D. and Lanctôt, P. (1985), "Méthodologie de détermination des seuils de revenu minimum au Québec", Ministère de la Main d'œuvre et de la Sécurité du revenu.
- Kalton, G. (2001), "Practical Methods for Sampling Rare and Mobile Populations", *Proceedings of the Annual Meeting of the American Statistical Association*.
- Kalton, G. and Anderson, D.W. (1986), "Sampling Rare Populations", *Journal of the Royal Statistical Society*, 149, pp. 65-82.
- Lavigne, M. and Michaud, S. (1998), "Aspects généraux de l'Enquête sur la dynamique du travail et du revenu", Income Statistics Division, Statistics Canada catalogue No. 75F0026MIF-1998005.
- Nadeau, C., Lapierre, B., Tremblay, J. and Gaudet, J. (2005), "Plan de sondage pour l'ajout d'un échantillon supplémentaire à l'Enquête sur les dépenses des ménages de 2003 pour le Ministère des Finances du Québec", Working Paper of the Household Survey Methods Division, Statistics Canada.
- Swain, L., Drew, J. D., Lafrance, B. and Lance, K. (1992), "The Creation of a Residential Address Register for Coverage Improvement in the 1991 Canadian Census", *Survey Methodology*, 18, pp.127-141.