COST-VARIANCE OPTIMIZATION FOR THE CANADIAN LABOUR FORCE SURVEY

G.H. Choudhry, H. Lee, and J.D. Drew, Statistics Canada

The Canadian Labour Force Survey (LFS) is a monthly household survey conducted by Statistics Canada to produce estimates for various labour force characteristics. It follows a stratified multi-stage rotating sample design with six rotation groups. Since its inception in 1945, the survey has undergone a sample redesign following each decennial census of population.

The 1981 post censal redesign effort included a research phase highlighted by Singh and Drew (1981), Singh, Drew, and Choudhry (1984) in which all aspects of the survey design were examined in an effort to improve the cost efficiency of the survey vehicle. This report deals with the research aimed at cost-variance optimization of the sample design.

Two important factors in the choice of a sample design are total cost and the reliability of the resulting estimates. The optimum solution can be obtained by minimizing either total cost or total variance when the other is fixed. Equivalently, the approach we have followed is one of minimizing the product of variance and cost for fixed sample size.

The cost-variance optimization was carried out in two steps. We first consider the optimization of the sample designs followed in each of the two major area types identified in the LFS design; i.e., the Self-Representing (SR) Areas or major cities, and Non-Self-Representing (NSR) Areas or smaller urban and rural areas. The scope of the optimization includes the allocation of sample to the two stages of the SR design (Section 2), and the consideration of alternatives to the old design in NSR areas (Section 3). For both types of areas variances are obtained empirically using data from the 1971 and 1976 Censuses, while cost models are developed using data from a time and cost study, and by means of a simulation study.

In Section 4, we consider the second stage of optimization, the allocation of sample to NSR and SR areas, taking into account the design improvements identified for each type of area. Finally, Section 5 summarizes the improvements identified, and their implications on the redesigned sample.

2. SR Design

The old SR design is a stratified two-stage area sample (Platek and Singh 1976). Each SR unit is stratified into a number of contiguous strata called subunits and each subunit is subdivided into clusters which are the primary sampling units (PSU's). The PSU's are selected using the random group method due to Rao, Hartley, and Cochran (1962) and at the second stage of sampling, a systematic sample of dwellings is taken in such a manner that the design becomes self-weighting. Let $1/W$ be the sampling rate in the stratum and n be the number of PSU's to be selected from the stratum. The N PSU's in the stratum are randomly partitioned into n groups so that the i-th random group contains N_i PSU's. Further, let M_j be the dwelling count for the j-th PSU in the stratum, M_o = the total dwellings in the stratum, and $m = M_o/W$ be the average sample size for the stratum. Then the average sample per selected PSU is $m/n = d$, where d will be called the average density for the stratum. Our objective is to obtain d which for a fixed m minimizes the product of variance and cost. For the optimization we obtain the total variance via the components of variance approach and consider a linear cost function as described below.

2.1 Variance Function

Let Y be the stratum total for a characteristic y of interest, and $\hat{Y}$ its estimate. If $N_i = N/n$, i.e., the number of PSU's in each of the random groups is the same, from the expression for the variance of $\hat{Y}$ given by Rao et al. (1962), the relative variance of $\hat{Y}$ defined by $\mathrm{Var}(\hat{Y})/Y^2$ can be written as

$$\text{Rel. Var}(\hat{Y}) = (W-1)\mu_2 + A(\mu_1 + \mu_2 - \mu_3) \qquad (2.1)$$

where

$$\mu 1 = \frac{1}{Y^2} \sum_{i=1}^{N} Y^2/\lambda_j - 1 ,$$

$$\mu 2 = \frac{1}{Y^2} \sum_{j=1}^{N} M_j S_j^2 ,$$

$$\mu 3 = \frac{1}{Y^2} \sum_{j} M_j S_j^2 / \lambda_j ,$$

$$A = (N - n)/(N - 1)n,$$

λ_j = normalized size of the j-th unit such that $\sum_j \lambda_j = 1$.

μ_1, μ_2, and μ_3 are population parameters and are fixed for a particular characteristic. Reexpressing A $= (Nd/m - 1)/(N - 1)$, we have

$$\text{Rel. Var}(\hat{Y}) = \alpha_0 + \alpha_1 d \qquad (2.2)$$

where $\alpha_0 = (W-1)\mu_2 - (\mu_1 + \mu_2 - \mu_3)/(N-1)$

$\alpha_1 = \frac{N}{m}(\mu_1 + \mu_2 - \mu_3)/(N - 1)$

The values of α_0 and α_1 for unemployed for Halifax SR areas were obtained from 1981 census data and these are

$\alpha_0 = 0.019005, \qquad \alpha_1 = 0.0007972$

From (2.2), we observe that from the variance point of view, the value $d = 1$ (i.e., one dwelling per PSU) is optimum. However, since α_1 is very small as compared to α_0, the increase in the variance as d increases will be very small.

2.2 Cost Model

A simple cost model was considered to investigate the impact on cost as d is varied, using cost component data obtained during a special time and cost study described by Lemaître (1983).

In SR areas, first month interviews are conducted in person, while interviews for months 2-6 are done by telephone in 85% of cases. For the purposes of our cost model, we define the following set of parameters.

C_0 = fixed costs (interviewing time, plus home to area travels)

C_1 = average cost of dwelling-to-dwelling travel within the same PSU

C_2 = average cost of PSU-to-PSU travel

γ = number of PSU-to-PSU moves per selected PSU

g_1 = number of dwelling-to-dwelling moves per stratum

g_2 = $n\gamma$ = number of PSU-to-PSU moves per stratum

The total cost for m dwellings will be

$$T = C_0 + g_1 C_1 + g_2 C_2. \qquad (2.3)$$

Since the total number of moves depends on the sample size and the proportion of households interviewed in person, we can write

$$g_1 + g_2 = \theta m. \qquad (2.4)$$

By substituting for g_1 in (2.3) and replacing n by m/d, we have

$$T = C_0 + \theta m C_1 + (C_2 - C_1) m\gamma/d, \qquad (2.5)$$

and the cost per dwelling is given by

$$C = C_0/m + \theta C_1 + (C_2 - C_1)\gamma/d. \qquad (2.6)$$

From the time and cost study, estimates for components C_0/m, θC_1 and $(C_2 - C_1)$ γ/d for Halifax were 3.28, 0.28 and 0.22 for d = 5. While the cost model was refined further to take into account changes in C_1, C_2 and γ with different values of d, it is clear from the simple model that unit costs would decrease only marginally with increases in the density, due to the dominance of the fixed costs.

Combining cost and variance results, the finding was that the cost-variance efficiency increased monotonically with decreases in d, to the extent of about a 3.5% gain per unit reduction in d. However, in practice it was decided to retain the density of 5 for the redesigned sample, on the grounds that a lower density would have resulted in more selected PSU's with higher implementation and maintenance costs.

3. NSR Design

3.1 NSR Design Alternatives

Design Alternative D_0: Old Design (see Figure 1)

Key features of the old NSR design (Platek and Singh 1976) were:

(i) Stratification: Economic Regions (ER's) whose numbers varied from 1-10 per province served as major strata. Within ER's, from 1-5 geographically contiguous strata were formed, using industry data from the 1971 Census.

(ii) Primary Sampling Units (PSU's): These were delineated within strata, to be geographically compact areas similar to the stratum with respect to stratification variables, and with respect to the ratio of rural to urban population. The first stage of sampling was the randomized probability proportional to size systematic (RPPSS) method of Hartley and Rao (1962).

(iii) Within PSU Sampling: Urbans All urban centers assigned to selected PSU's were included in the sample. The second stage of sampling was a sample of blocks, following RPPSS sampling. The third and final stage of sampling was a systematic sample of dwellings.

(iv) Within PSU Sampling: Rurals The second stage of sampling was a RPPSS sample of EA's. EA's were then field counted to delineate clusters having from 3-20 dwellings. The third and fourth stages of sampling corresponded to an RPPSS sample of clusters and a systematic sample of dwellings.

Design Alternative D_1: Elimination of Cluster Stage of Sampling in Rurals

This design alternative is identical to D_0, except dwellings are selected directly within selected rural EA's.

Design Alternative D_2: Explicit Urban/Rural Stratification

In the old design, the maintenance of the stratum urban to rural population ratio at the PSU level required frequent discontiguity between rural and urban portions of PSU's, leading in turn to increased travelling costs.

Design alternative D_2 was formulated as follows:

(i) Stratification: Rural and urban portions of ER's as primary strata, with geographically contiguous secondary rural strata and secondary urban strata formed without geographic constraints.

(ii) Sampling Within Rural Strata: Sampling in three stages as follows: RPPSS sample of PSU's (contiguous group of EA's similar to the stratum with respect to stratification variables); RPPSS sample of EA's; and systematic sample of dwellings.

(iii) Sampling Within Urban Strata: Sampling in three stages as follows: RPPSS sample of PSU's (individual or combined urban centers); RPPSS sample of clusters; and systematic sample of dwellings.

3.2 Variance Components Model

Design alternatives D_0, D_1 and D_2 were simulated using census data. Expressions for the variance components are given below:

Stage of Sampling	Variance Estimation Sampling	
1st	$V_{(1)} = V_{(1)}^{RPPSS}$	(3.1)
2nd	$V_{(2)} = W \sum_{i=1}^{N} \frac{V_{(2)i}^{RPPSS}}{W_i}$	(3.2)
3rd	$V_{(3)} = W \sum_i \sum_j \frac{V_{(3)ij}^{SRS}}{W_{ij}}$ if last stage	(3.3)
	$= W \sum_i \sum_j \frac{V_{(3)ij}^{RPPSS}}{W_{ij}}$ otherwise	
4th (where applicable)	$V_{(4)} = W \sum_i \sum_j \sum_k \frac{V_{(4)ijk}^{SRS}}{W_{ijk}}$.	(3.4)

The variance formula for RPPSS sampling are given by Hartley and Rao (1962). An algorithm due to Hidiroglou and Gray (1980) was used to calculate the joint probabilities required.

3.3 Cost Model

The cost model for design D_1 under personal interviewing was formulated as

$$C_{D_1} = F_0 + F_1 + F_2 + E_1 + E_2 \qquad (3.5)$$

where

F_0 = fixed fee for interviewing (fees are for time spent)

F_1 = fee for home to area, between PSU, and between secondary travel

F_2 = fee for within secondary (dwelling to dwelling) travel

E_1 = expenses associated with home to area, between PSU, and between secondary travel.

E_2 = expenses associated with dwelling to dwelling travel

All parameters are expressed in terms of per dwelling costs.

Under telephone interviewing, this was modified to

$$C^T_{D_1} = F_0 + \alpha(F_1 + F_2 + E_1 + E_2), \qquad (3.6)$$

where α is the factor by which travel time and mileage would be decreased under telephoning.

Now, under the assumption that D_2 would affect F_1 and E_1, say by a factor r, but would not affect other components we have,

$$C^T_{D_2} = F_0 + \alpha\, r\,(F_1 + E_1) + \alpha(F_2 + E_2). \qquad (3.7)$$

Parameters of $C^T_{D_1}$ and $C^T_{D_2}$ were estimated as follows:

F_0, F_1, F_2, E_1, E_2: These were estimated under D_0 from a special time and cost study (Lemaître 1983), carried out as part of the redesign research program. Since a field test of D_1 revealed no discernable differences in data collection costs between D_0 and D_1, these parameters were assumed unchanged under D_1.

α: Field testing of telephone interviewing carried out as part of the redesign research program did not have as an objective the estimation of cost savings. An estimated 10% reduction in total data collection costs was made by Regional Operations staff, which permitted calculation of α.

r: This parameter could not be estimated based on available data, rather a Monte Carlo simulation study was needed. The sample frames under D_1 and D_2 were simulated to the level of secondaries using Census data for each of the 11 study ER's. Fifty samples were drawn following each design, and the selected secondaries for each sample were grouped into geographically optimal assignments. If $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ are the average measures of within assignment geographic dispersion under designs D_1 and D_2, then r was estimated by $\bar{M}^{(2)}/\bar{M}^{(1)}$. The determination of optimum interviewer assignments, that is the minimization of the M-measure, reduces to a classification or clustering problem. Algorithms investigated included a transfer algorithm (Friedman and Rubin 1967), an exchange algorithm (Dahmström and Hagnell 1975), and a two cycle algorithm combining the above two which was adopted on the basis of its improved performance (Lee 1985).

3.4 Results of Cost-Variance Analyses

Variance Analysis: D_1 vs. D_0

Components of variance for 5 labour force characteristics were obtained for designs D_0 and D_1 using 1971 Census data for 5 ER's across Canada. Table 1 gives the % contribution from each stage of sampling to the total variance under D_0. It can be observed that 30-40% of the total variance under D_0 was due to the rural cluster (3rd) stage of sampling, and that under design D_1 20-30% variance reductions could be obtained.

Actual gains might be less since for the study, the variables being estimated and the size measures referred to the same point in time whereas this would not be true in practice. No attempt was made to discount the gains, however, since the choice between D_1 and D_0 was clear both in terms of variances, and on operational grounds, where D_1 would reduce sample maintenance costs and shorten the lead time to select independent samples from the LFS frame. Further efforts were devoted hence to the choice between D_1 and D_2.

Variance Analysis: D_2 vs. D_1

In this study the number of ER's was expanded to 11, and study variables (employed and unemployed) were based on the 1976 Census, whereas size measures were based on the 1971 Census. Also variances were computed with ratio estimation based on total population.

The average variance efficiency of D_2 with respect to D_1 was 1.16 for employed and 0.97 for unemployed (Table 2).

Cost Analysis: D_2 vs. D_1

As expected the between PSU and between secondary component of interviewer fees and expenses were found to be higher under D_1 due to the frequent lack of contiguity between rural and urban portions of PSU's. The average reduction factor r in these components under D_2 was estimated to be 0.75 leading to an overall cost efficiency for D_2 vs. D_1 of 1.08 (Table 2).

Combined Cost-Variance Analysis: D_2 vs. D_1

Table 2 gives the relative cost-variance efficiencies of D_2 vs. D_1 under telephone interviewing. Since D_2 is 25% and 5% more efficient than D_1 for employed and unemployed respectively, it was decided to adopt D_2 in the 2/3 of ER's capable of supporting both urban and rural strata, and D_1 in the remaining cases.

3.5 Special 2-Stage Design for Prince Edward Island

For Canada's smallest province, Prince Edward Island, where sampling rates of 4% are required in order to produce reliable provincial data, design alternative D_3, an unclustered stratified sample of EA's and dwellings, was adopted on the strength of study findings showing that a slight loss in cost efficiency compared with D_2 was more than offset by sizable gains in variance efficiency.

3.6 Number of PSU's Selected Per Stratum

Under both designs D_1 and D_2, the sample yield per PSU was fixed at 55-60 dwellings to correspond to an interviewer's assignment. While there should be at least 2 PSU's per stratum to permit unbiased estimation of variance, some consideration was given to having 4-5 PSU's per stratum, to permit greater flexibility to reduce the size of the area sample. However, stratification to the point of 2-3 PSU's per stratum was adopted, based on variance reductions of 14.8% for employed and 5.4% for unemployed. The stratification procedures are described by Drew, Bélanger, and Foy (1985).

4. Cost-Variance Optimization between SR/NSR Areas

The next step in the cost-variance optimization was the optimization of the allocation of sample between SR and NSR areas. We used the simple cost and variance models considered by Fellegi, Gray, and Platek (1967), i.e.,

$$\text{cost:} \quad C = \sum_{j=1}^{2} C_j \frac{P_j}{W_j} , \tag{4.1}$$

$$\text{variance:} \quad V = \sum_{j=1}^{2} W_j P_j \sigma_j^2 , \tag{4.2}$$

where j = area type (= 1 for SR; = 2 for NSR)

C_j = unit (i.e., per person) cost

P_j = population

$1/W_j$ = sampling rate

σ_j^2 = unit variance

Fellegi et al. showed that if C is minimized with V fixed the ratio of the sampling rates is

$$\frac{W_1}{W_2} = \frac{\sigma_2}{\sigma_1} (C_1/C_2)^{\frac{1}{2}} \tag{4.3}$$

The other optimization criteria described in Section 1 also give the same ratio as above.

Unit costs and variances were estimated from historical survey data as modified to reflect structural changes in data collection methodology and sample design as described in Sections 2 and 3.

For Canada, the optimum allocation called for 67.1% of the sample in SR areas, as compared with 67.4% for a proportional allocation, and 53.2% for the old design. As a result of subprovincial data reliability constraints, the allocation adopted had 62.3% of the sample in SR areas. The cost-variance efficiency gain due to the sample re-allocation was sizable. To illustrate this, under uniform sampling rates for SR and NSR areas within provinces as in the old design, and assuming no structural changes in the sample design, the re-allocation would have resulted in an efficiency of 1.10%. In practice, the subprovincial requirements necessitated a departure from such uniform allocations.

5. Conclusions

The design changes taken as a result of the cost-variance studies include: elimination of a stage of sampling in NSR rural areas, adoption of a design featuring rural/urban stratification, adoption of a 2-stage NSR design in Prince Edward Island, increase the number of NSR strata to the extent that only 2 or 3 PSU's per stratum will be selected, and re-optimization of the allocation of sample bewteen NSR and SR areas. The near optimality of other design parameters established earlier by Fellegi, Gray and Platek (1967) was found to have remained unchanged, for example the number of dwellings to select per PSU in SR Areas.

The efficiency gains resulting from the changes permitted both a 7% reduction in the overall LFS sample size and substantial improvements in the reliability of subprovincial data (Singh et al. 1984) to be achieved with little or no impact on the reliability of provincial and national estimates. Table 3 gives the cost, variance and combined cost-variance ratios for the old sample (old design with 55,500 hhlds/month and no telephone interviewing in NSR's) vs. the redesigned sample (new design with 51,600 hhlds/month and telephone interviewing). The significant cost reductions are due to the shift to telephone interviewing in months 2-6 in NSR areas, and the sample size reduction. The overall cost-variance efficiency of the redesigned sample relative to the old sample was 1.16 (Table 3).

References

Dahmström, P., and Hagnell, M. (1975). Multivariate stratification of primary sampling units in multi-stage sampling with an application to SCB's general purpose sample. Research Report, University of Lund.

Drew, J.D., Bélanger, Y., and Foy, P. (1985). Multivariate clustering algorithm for stratification and its application to the Canadian Labour Force Survey. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

Fellegi, I.P., Gray, G.B., and Platek, R. (1967). The new design of the Canadian Labour Force Survey. Journal of the American Statistical Association, 62, pp. 421-453.

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

Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. Annals of Mathematical Statistics, 33, pp. 350-374.

Hidiroglou, M.A., and Gray, G.B. (1980). Construction of joint probability of selection for systematic PPS sampling. Journal of Royal Statistical Society, C29, pp. 107-112.

Lee, H. (1985). Cost simulation for the non-self-representing area design in Canadian Labour Force Survey. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

Lemaître, G. (1983). Some results from Time and Cost Study. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

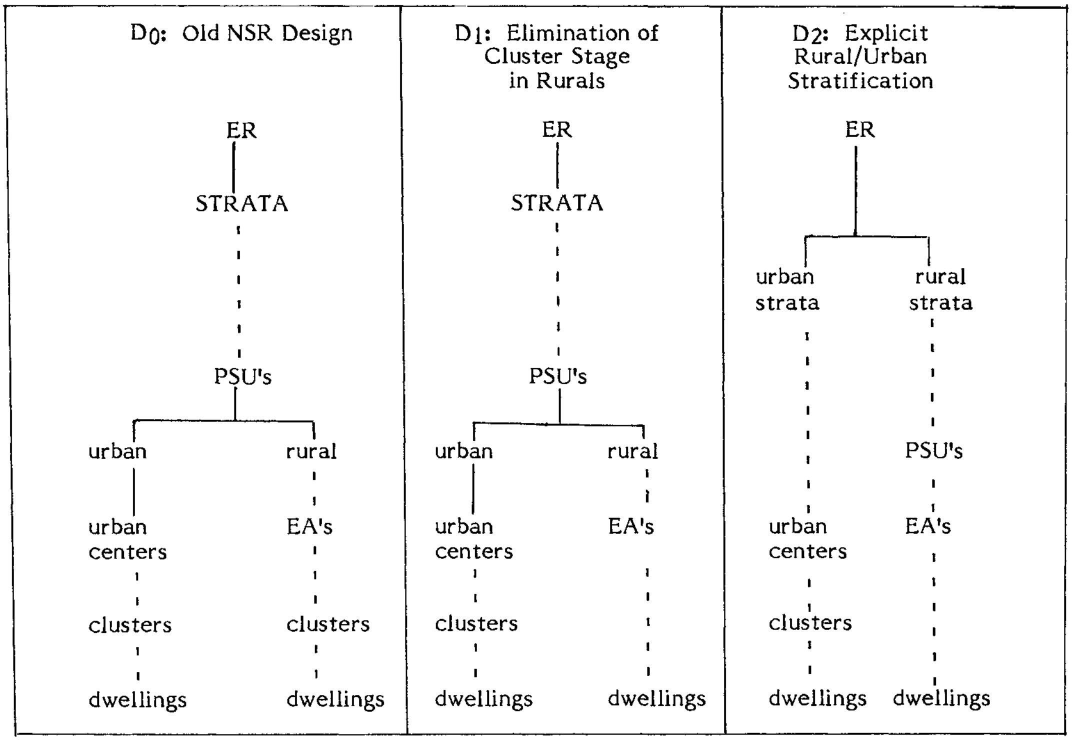
Platek, R., and Singh, M.P. (1976). Methodology of the Canadian Labour Force Survey. Catalogue No. 71-526, Statistics Canada.

Rao, J.N.K., Hartley, H.O., and Cochran, W.G. (1962). A simple procedure of unequal probability sampling without replacement. Journal of Royal Statistical Society, B24, pp. 482-491.

Singh, M.P., and Drew, J.D. (1981). Research Plans for the Redesign of the Canadian Labour Force Survey. Proceedings of the Section of Survey Research Methods, American Statistical Association Meetings.

Singh, M.P., Drew, J.D., and Choudhry, G.H. (1984). Post '81 Censal Redesign of the Canadian Labour Force Survey. Survey Methodology, 10, pp. 141-156.

FIGURE 1
Representation* of NSR Design Alternative



* legend ____ stratification
----- stage of sampling

Table 1
Percent Contributions to the Total Variance from Stages of Sampling for the Current Design and Percent Reduction in the Total Variance Due to Eliminating Cluster Stage of Sampling in Rural Areas

Characteristics	Percent Contribution to Total Variance from						Percent Variance Reduction $100(1 - \frac{V_{D_1}}{V_{D_0}})$
		Urban		Rural			
	1st stage	2nd stage	3rd stage	2nd stage	3rd stage	4th stage	
LF Population	14.5	12.9	10.8	5.8	40.5	15.5	30.5
Employed	21.2	11.2	10.4	6.3	35.0	15.8	27.1
Unemployed	12.6	15.8	16.6	4.8	33.0	17.2	24.8
Not in LF	24.7	11.9	10.7	4.8	32.9	15.1	22.9
Employed Agr.	42.4	1.0	0.8	12.3	30.8	12.6	20.4
Employed Non-Agr.	23.3	12.7	11.9	5.6	31.7	14.8	21.8

Table 2
Relative Cost-Variance Efficiencies of D_1 vs. D_2

	Variance Efficiency V_{D_1}/V_{D_2}		Cost Efficiency $C^T_{D_1}/C^T_{D_2}$	Relative Cost-Variance Efficiency $V_{D_1}C^T_{D_1}/V_{D_2}C^T_{D_2}$	
ER	Employed	Unemployed		Employed	Unemployed
22	1.09	0.93	1.02	1.11	0.95
32	0.91	0.72	1.03	0.94	0.74
41	1.14	0.86	1.23	1.40	1.06
44	1.39	1.14	1.19	1.65	1.37
51	0.96	1.01	1.03	0.99	1.04
56	1.12	1.51	1.10	1.23	1.66
63	1.35	1.06	1.05	1.41	1.11
72	1.00	0.91	1.06	1.06	0.96
82	1.09	1.01	1.18	1.27	1.19
86	1.20	1.05	1.04	1.25	1.09
96	1.38	1.05	1.07	1.48	1.12
All*	1.16	0.97	1.08	1.25	1.05

* Weighted average by population size

Table 3
Relative Efficiency of the Redesigned vs. the Old Sample for Unemployed

Province	Cost Ratio* $\left(= \frac{C^{(0)}}{C^{(N)}}\right)$	Variance Ratio $\left(= \frac{V^{(0)}}{V^{(N)}}\right)$	Rel. Eff. $\left(= \frac{C^{(0)}V^{(0)}}{C^{(N)}V^{(N)}}\right)$
Newfoundland	1.19	1.00	1.19
Prince Edward Island	1.10	1.13	1.24
Nova Scotia	1.22	1.04	1.27
New Brunswick	1.17	0.99	1.16
Québec	1.15	0.95	1.09
Ontario	1.13	1.03	1.16
Manitoba	1.17	0.96	1.12
Saskatchewan	1.23	1.02	1.25
Alberta**	1.15	1.00	1.15
British Columbia	1.15	1.01	1.16
Canada	1.17	0.99	1.16

* Based on the redesigned sample with telephone interviewing and the old sample with personal visit interviewing in NSR areas.

** Supplementary sample of 1300 households not included.