Survey Methodology, June 1985
Vol. 11, No. 1, pp. 33–50
Statistics Canada

Cost-Variance Optimization for the Canadian Labour Force Survey

G.H. CHOUDHRY, H. LEE, and J.D. DREW[1]

ABSTRACT

The cost-variance optimization of the design of the Canadian Labour Force Survey was carried out in two steps. First, the sample designs were optimized for each of the two major area types, the Self-Representing (SR) and the Non-Self-Representing (NSR) areas. Cost models were developed and parameters estimated from a detailed field study and by simulation, while variances were estimated using data from the Census of Population. The scope of the optimization included the allocation of sample to the two stages in the SR design, and the consideration of two alternatives to the old design in NSR areas. The second stage of optimization was the allocation of sample to SR and NSR areas.

KEY WORDS: Multi-stage designs; Sample allocation; Linear cost function; Components of variance.

1. INTRODUCTION

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. These redesigns serve to update the sample to reflect population changes. They also provide the opportunity to introduce improved sampling and estimation methodologies, and to respond to shifts in information needs to be satisfied by the survey.

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

The two important factors in the choice of a sample design are the 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 SR Areas or major cities, and NSR Areas which are the 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 NSR areas the old design is first evaluated empirically via a components of variance approach, and one stage of sampling in rural areas is identified for elimination. Subsequently the modified old design is compared to an alternative design featuring explicit rural/urban stratification from an overall cost-variance perspective. 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.

[1] G.H. Choudhry, H. Lee, and J.D. Drew, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

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 design (Platek and Singh 1976). Each Self-Representing Unit (SRU) 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 and $\sum_{i=1}^{n} N_i = N$. Let x_j and M_j, $j = 1, 2, \ldots, N$, respectively be the size measure and dwelling count for the j-th PSU in the stratum.

Define

$$\lambda_j = \frac{x_j}{\sum_{t=1}^{N} x_t}$$

and

$$\delta_{ij} = 1 \text{ if } j\text{-th PSU is in } i\text{-th group} \\ = 0 \text{ otherwise.}$$

Then $\pi_i = \sum_{j=1}^{n} \delta_{ij}\lambda_j$ is the relative size of the i-th group. Now define W_{ij}'s as

$$W_{ij} = \delta_{ij}\left[W\frac{\lambda_j}{\pi_i}\right] \text{ or } \delta_{ij}\left[W\frac{\lambda_j}{\pi_i} + 1\right] \tag{2.1}$$

such that $\sum_{j=1}^{N} W_{ij} = W$ for $i = 1, 2, \ldots, n$, where $[a]$ is the greatest integer less than or equal to a. Now select one PSU from each of the n random groups independently with probability proportional to W_{ij}'s and sub-sample the selected PSU j from the i-th group at the rate $1/W_{ij}$. Then the overall sampling rate within each of the random groups is $1/W$ so that the design becomes self-weighting with a design weight equal to W. The average sample size for the stratum is given by

$$m = \frac{1}{W}\sum_{j=1}^{N} M_j \tag{2.2}$$

$$= M_0/W$$

where M_0 is the total number of dwellings in the stratum. Let M_{ij} be the number of dwellings in the selected PSU j in the i-th group, then $m_i = M_{ij}/W_{ij}$ dwellings will be selected from the i-th group. The average number of dwellings selected from the i-th group for a given random grouping is $1/W \sum_j \delta_{ij} M_j$ and the average over all possible random groupings is m N_i/N since the expected value of δ_{ij} is N_i/N. If $N_i/N = 1/n$, i.e., the number of psu's in each of the random groups is the same, then the average sample per selected PSU is $m/n = d$(say), where d will be called the average density for the stratum. Since m is fixed, the sample of m dwellings can be elected by varying n and d such that the product *(nd)* remains equal to

m, the total sample size for the stratum. Our objective here is to obtain d which for a fixed sample size 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 in the following section.

2.1 Variance Function

Suppose that we are interested in the total of a characteristic y for the subunit. Let y_{jh} be the y-value for the h-th household in PSU j where $h = 1, 2, \ldots, N$, then the total $Y = \sum_{j=1}^{N} \sum_{h=1}^{M_j} y_{jh}$ is estimated by

$$\hat{Y} = W \sum_{i=1}^{n} y_i \tag{2.3}$$

where y_i is the sum of the y-values for the m_i selected households from the PSU selected from the i-th group, $i = 1, 2, \ldots, n$. Ignoring the effect due to rounding involved in defining W_{ij}, the variance of $\hat{Y}$ is given by (Rao et al. 1962)

$$\mathrm{Var}(\hat{Y}) = A\left[\sum_{j=1}^{N} \frac{Y_j^2}{\lambda_j} - Y^2\right] + \sum_{j=1}^{N} M_j S_j^2 \left[W - 1 - \mathrm{A}\left(\frac{1}{\lambda_j} - 1\right)\right]. \tag{2.4}$$

where
$$Y_j = \sum_{h=1}^{M_j} y_{jh},$$

$$S_j^2 = \frac{1}{M_j - 1} \sum_{h=1}^{M_j} \left(y_{jh} - \frac{Y_j}{M_j}\right)^2,$$

$$A = \frac{\sum_{1}^{n} N_i^2 - N}{N(N-1)} \cdot$$

If $N_i = N/n$, i.e., all random groups have equal number of PSU's, then

$$A = \frac{N-n}{n(N-1)} \cdot$$

Relative variance of $\hat{Y}$ defined by $\mathrm{Var}(\hat{Y})/Y^2$ will be

$$\begin{aligned}
\text{Rel. Var}(\hat{Y}) &= A\left[\frac{1}{Y^2}\sum_{j=1}^{N} \frac{Y_j^2}{\lambda_j} - 1\right] + \frac{1}{Y^2}\sum_{j=1}^{N} M_j S_j^2 \left[W - 1 - A\left(\frac{1}{\lambda_j} - 1\right)\right]. \\
&= A\mu_1 + (W - 1)\mu_2 + A\mu_2 - A\mu_3 \\
&= (W - 1)\mu_2 + A(\mu_1 + \mu_2 - \mu_3) \qquad (2.5)
\end{aligned}$$

where
$$\mu_1 = \frac{1}{Y^2}\sum_{j=1}^{N} \frac{Y_j^2}{\lambda_j} - 1$$

$$\mu_2 = \frac{1}{Y^2} \sum_{j=1}^{N} M_j S_j^2 ,$$

$$\mu_3 = \frac{1}{Y^2} \sum_j M_j \frac{S_j^2}{\lambda_j} .$$

μ_1, μ_2, and μ_3 are the population prameters and are fixed for a particular characteristic. Since $m = nd$ and if we assume that $N_i = N/n$ then we can write A as

$$A = \frac{1}{N-1} (N \frac{d}{m} - 1)$$

and

$$\text{Rel. Var}(\hat{Y}) = (W - 1)\, \mu_2 + (N \frac{d}{m} - 1)\, \frac{(\mu_1 + \mu_2 - \mu_3)}{(N - 1)}$$

$$= \alpha_0 + \alpha_1\, d \qquad (2.6)$$

where

$$\alpha_0 = (W - 1)\, \mu_2 - \frac{(\mu_1 + \mu_2 - \mu_3)}{(N - 1)}$$

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

From (2.6), we observe that from reliability point of view, the value $d = 1$ (i.e., one dwelling per PSU) is optimum. But this will have impact on the cost as discussed in the next section. The values of α_0 and α_1 for unemployed for Halifax SRU were obtained from 1981 census data and these are

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

Since α_1 is very small as compared to α_0, the increase in the variance with the corresponding increase in d will be very small. Next we examine the effect on the cost due to varying the value of the average density d.

2.2 Cost Model

A simple cost model has been considered to investigate the impact on the cost as the density is varied. Due to telephone interviewing in the SR areas, personal visits are only required to a PSU during the rotation month and in cases where some households were without a telephone or did not agree to telephone interviewing.

A breakdown of the interviewing cost by telephone and personal visit is available for individual interviewers from field operations, but further breakdown of the personal visit component of the cost was required to construct the cost model. For this purpose a special time and cost study was carried out in the field for a period of six months (February-July 1982) on a random sample of interviewers. The results from the analysis of time and cost data are documented in a report by Lemaitre (1983). For the purpose of our cost model, we define the following set of parameters

c_0 = Fixed costs
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 selectd PSU.

The fixed cost c_0 includes the time spent actually conducting interviews whether by telephone or in person and the travel cost from home to area and back. The fixed cost c_0 depends only on the total sample size m and not on n, the number of selected PSU's. Suppose that there are g_1 dwelling-to-dwelling moves and g_2 PSU-to-PSU moves made, then the total cost for m dwellings will be

$$T = c_0 + g_1c_1 + g_2c_2. \tag{2.7}$$

If n is increased then g_2 will also increase and g_1 will decrease and vice-versa but $(g_1 + g_2)$ should remain constant because the number of moves depends on the sample size m and the proportion of households interviewed by personal visit. Then we may write

$$g_1 + g_2 = \theta m. \tag{2.8}$$

From (2.8) we substitute g_1 in equation (2.7) and obtain

$$\begin{aligned} T &= c_0 + \theta m c_1 + g_2(c_2 - c_1) \\ &= c_0 + \theta m c_1 + n\gamma(c_2 - c_1). \end{aligned}$$

Now replacing n by m/d we have

$$T = c_0 + \theta m c_1 + \frac{m\gamma}{d}(c_2 - c_1)$$

and cost per dwelling C as a function of average ensity d is given by

$$C = \frac{c_0}{m} + \theta c_1 + \frac{\gamma}{d}(c_2 - c_1). \tag{2.9}$$

From Time and Cost Study the parameters c_1 and c_2 for Halifax were 0.78 and 2.51 respectively. These parameters were observed with average density equal to 5 but c_2 increases with d and c_1 decreases with d. Assuming that the average distance between the units is inversely proportional to the square root of the number of units in an area, we can replace c_1 by $c_1(5/d)^{1/2}$ and c_2 by $c_2(d/5)^{1/2}$ in our model so that the modified model becomes

$$C = \frac{c_0}{m} + \theta c_1\left(\frac{5}{d}\right)^{1/2} + \frac{\gamma}{d}\left\{c_2\left(\frac{d}{5}\right)^{1/2} - c_1\left(\frac{5}{d}\right)^{1/2}\right\}. \tag{2.10}$$

c_0/m is fixed per dwelling cost and does not depend on density and its value was 3.28 from Time and Cost Study. The parameter θ does not depend on the density either and was equal to 0.356 from Time and Cost Study. The parameter γ increases with density because the average number of visits to a PSU will increase due to higher density. We have approximated γ by

$$\frac{1}{6} + \frac{5}{6}(1 - p^d)$$

where p is the probability of telephone interview for a household in a non rotate-in PSU and the value of p was 0.85 as obtained from interviewers' data. From the cost model (2.10), the values of per dwelling cost for $d = 2, 3, \ldots, 10$ are given in Table 1 along with the relative variances and the products of these two which are the values of the objective function to be minimized.

Table 1

Value of Relative Variance, Cost per Dwellings and Objective Function for Various Densities (Unemployed)

Density	Relative Variance	Cost per Dwelling	Objective Function
2	0.0206	3.79	0.078
3	0.0214	3.79	0.081
4	0.0222	3.79	0.084
5	0.0230	3.78	0.087
6	0.0238	3.77	0.090
7	0.0246	3.76	0.092
8	0.0254	3.75	0.095
9	0.0262	3.74	0.098
10	0.0270	3.73	0.101

As expected, we observe that under the model considered here, the cost per dwelling decreases very slowly as the density increases since the fixed per dwelling cost (c_0/m) dominates in (2.10) due to telephone intervewing. From the previous section we had found that the increase in the relative variance is very small as the density increases. As a result our objective function is monotonically increasing but the loss in the cost-variance efficiency with increase in d is small. However it was decided to retain the old density of 5 for the redesigned sample on the grounds that 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 NSR 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 geographicaly 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. PSU populations ranged from 3,000 to 5,000. In the first stage PSU's were selected following the randomized probability proportional to size systematic (RPPSS) method of Hartley and Rao (1962). Within PSU's urban and rural parts were sampled separately.

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

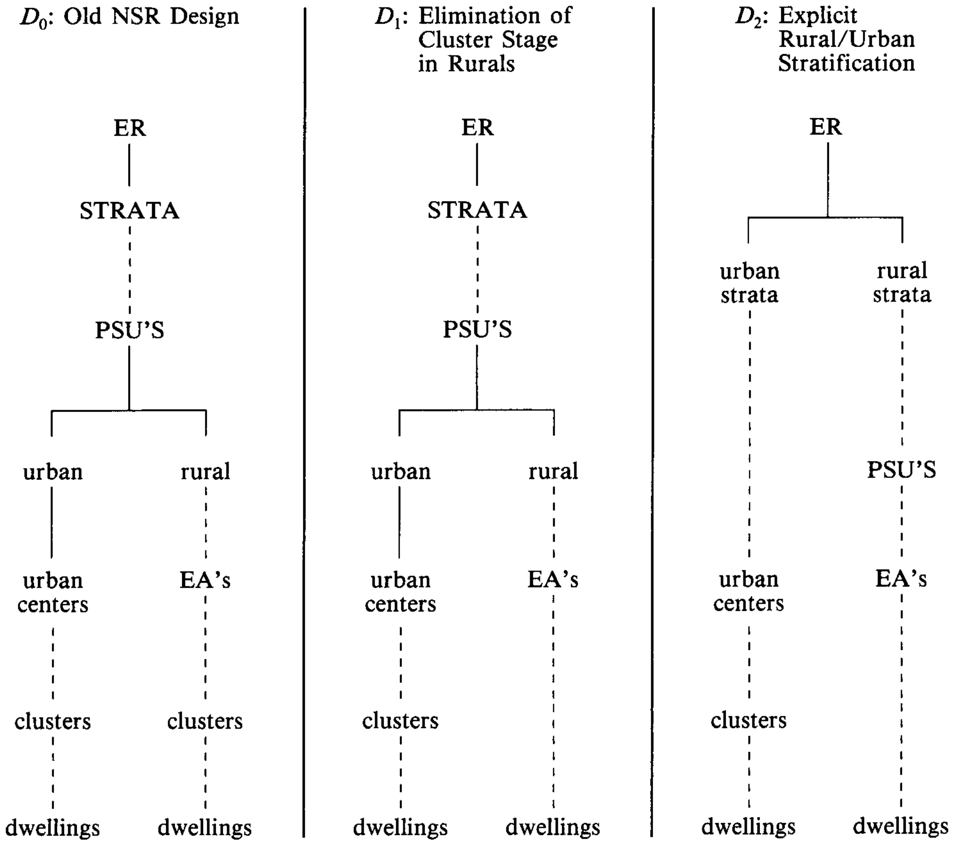

Figure 1. Representation of NSR Design Alternatives. (—— stratification, ----- stage of sampling)

iv) **Within PSU Sampling: Rurals** The second stage of sampling was a RPPSS sample of EA's. EA's were then field counted for the purposes of delineating 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

i) It would permit shortening of the lead time to select independent samples from the LFS frame to 7 months from 13 months, by eliminating the need for counting of EA's.

ii) Elimination of the clustering step would reduce sample maintenance costs.

iii) A priori, the reduction in the stages of sampling from 4 to 3 stages would translate into a reduced variance. it was expected that costs, on the other hand, would not be very much affected, particularly with the shift to telephone interviewing.

iv) At an early juncture in the redesign research program a field study was carried out on the operational implications of eliminating the cluster stage. Verification of EA listings a year later revealed no problems with the quality of listings, and analysis revealed no discernable impact on data collection costs.

Design Alternative D_2: Explicit Urban/Rural Stratification

The old design with its separate sampling of urban and rural portions of PSU's featured an implicit urban/rural stratification. A drawback of the approach however was that 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.

In view of this problem with the old design, design alternative D_2 was formulated as follows:

i) **Stratification:** Rural and urban portions of ER's would constitute primary strata, which would be optimally sub-stratified to the point of having strata yields of 100-150 dwellings (i.e., 2-3 PSU's each corresponding to an interviewer's assignment). ER's not able to support at least one such urban and one such rural stratum (roughly ⅓ of ER's) were considered ineligible for D_2.

Secondary rural strata would be contiguous, while secondary urban strata would be formed without geographic constraints.

ii) **Sampling Within Rural Strata:** PSU's similar to the stratum with respect to stratification variables would be formed by grouping geographically contiguous EA's and will be selected by the RPPSS method. Second and third stages of sampling would be an RPPSS sample of EA's and systematic sample of dwellings.

iii) **Sampling Within Urban Strata:** Sampling would proceed 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 alternative D_0, D_1 and D_2 were simulated using census data. Expressions for the variance components are given below:

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

The variance formula and its computation method for the RPPSS sampling are described in Appendix A.

3.3 Cost Model

Whereas the cost model for the SR areas dealt with allocation of samples to 2 stages of sampling, here a cost model is needed to compare alternative NSR designs.

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

$$C_{D_1} = F_0 + F_1 + F_2 + E_1 + E_2$$

where F_0 = fixed fee for interviewing,
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.

Fees are compensation for the time spent and expenses for the distance covered. All Parameters are expressed in terms of per dwelling costs.

Under telephone interviewing, this was modified to

$$C_{D_1}^T = F_0 + \alpha(F_1 + F_2 + E_1 + E_2),$$

where α is the factor by which 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,

$$C_{D_2}^T = F_0 + \alpha r(F_1 + E_1) + \alpha(F_2 + E_2).$$

Parameters of $C_{D_1}^T$ and $C_{D_2}^T$ 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 (Lemaitre 1983), carried out as part of the redesign research program. Since the 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, which is described in Appendix B.

3.4 Results of Cost-Variance Analyses

Variance Analysis: D_1 vs. D_0

Components of variance for 6 labour force characteristics were obtained for designs D_0 and D_1 using 1971 Census data for 5 ER's across Canada. Table 2 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.

Table 2

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; $100\ (1 - \frac{V_{D_1}}{V_{D_0}})$

Characteristic	Percent Contribution to Total Variance from						Percent Variance Reduction;
		Urban		Rural			$100\ (1 - \frac{V_{D_1}}{V_{D_0}})$
	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

The 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 (as discussed in Subsection 3.1). 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 4).

Cost Analysis: D_2 vs. D_1

Values of all the parameters in the cost model are presented in Table 3 along with $C^T_{D_1}$ and $C^T_{D_2}$ and their ratio.

As expected the between PSU and between secondary component of interviewer fees and expenses are 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 as in Table 3 leading to an overall cost efficiency for D_2 vs. D_1 of 1.08 (Table 4).

Combined Cost Variance Analysis: D_2 vs. D_1

Table 4 gives the relative cost-variance efficiencies of D_2 vs. D_1 under telephone interviewing. In terms of overall efficiency, D_2 is 25% and 5% more efficient than D_1 for employed and unemployed respectively.

Based on these findings it was decided to adopt D_2 in the 2/3 of ER's capable of supporting both urban and rural strata, and design D_1 was adopted in the remaining cases.

Table 3

Values of Parameters in the NSR Cost Model and Relative Cost Efficiencies of D_1 vs. D_2 with Telephone Interviewing

ER	F_0	F_1	F_2	E_1	E_2	α	r	$C^T_{D_1}$	$C^T_{D_2}$	$C^T_{D_1}/C^T_{D_2}$
22	2.05	0.74	1.31	0.95	0.92	0.85	0.93	5.38	5.28	1.02
32	2.13	0.86	1.11	0.90	0.97	0.84	0.88	5.35	5.17	1.03
41	2.04	0.94	0.94	0.96	0.69	0.84	0.42	5.01	4.08	1.23
44	2.04	0.94	0.94	0.96	0.69	0.84	0.50	5.01	4.21	1.19
51	1.94	0.80	1.07	0.81	0.75	0.84	0.89	4.82	4.67	1.03
56	1.94	0.80	1.07	0.81	0.75	0.84	0.68	4.82	4.39	1.10
63	2.07	1.03	1.03	1.19	0.97	0.75	0.87	5.66	5.41	1.05
72	1.92	0.96	1.13	1.05	1.09	0.85	0.82	5.52	5.21	1.06
82	1.88	1.12	1.01	1.20	0.94	0.86	0.57	5.55	4.69	1.18
86	1.88	1.12	1.01	1.20	0.94	0.86	0.90	5.55	5.35	1.04
96	2.03	0.81	1.22	0.75	0.85	0.84	0.75	5.07	4.74	1.07

Table 4

Relative Cost-Variance Efficiencies of D_1 vs. D_2

	Variance Efficiency V_{D_1}/D_{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.

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, a stratified sample of EA's and dwellings, was considered as an alternative to D_2.

D_3 did not feature any clustering of the sample into geographically contiguous primaries designed to correspond to interviewers assignments, as it was hypothesized that given the high sampling rates, the increase in data collection costs might be more than offset by variance reductions due to elimination of a stage of sampling, and due to stratification gains resulting from having more strata (i.e., up to 4 times as many as under D_2).

Cost-variance study results showed the variance efficiency of D_3 vs. D_1 to be 2.39 for employed and 1.20 for unemployed, while costs under D_3 were only 8% greater. Hence, based on overall cost-variance efficiencies of 2.21 for employed and 1.11 for unemployed, D_3 was opted for.

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. In about half of the ER's, there was only enough sample for 2 or 3 PSU's to be selected. Further stratification in these cases was ruled out on the grounds that there should be at least 2 PSU's per stratum to permit unbiased estimation of variance.

For the remaining ER's, some consideration was given to having 4-5 PSU's per stratum, as this would permit greater flexibility to reduce the size of the area sample, for example, if a portion of the area sample at some time in the future were to be converted to a telephone sample under a dual frame set-up. 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 for these ER's. A detailed description of the stratification procedures followed can be found in Drew, Bélanger, and Foy (1985).

4. COST-VARIANCE OPTIMIZATION BETWEEN SR and NSR AREAS

The next step in the cost-variance optimization of the LFS design 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.,

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

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

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

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

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

Parameters were estimated as follows:

(i) **Unit costs:** Historical per dwelling costs by type of area were available. These were decreased by 10% for NSR areas, to take account of the estimated effect of a shift to telephone interviewing of all rotation groups except the rotate-in group for the redesigned sample.

(ii) **Unit variances:** Optimization was carried out with respect to the characteristic unemployed, for which variances were given by:

$$\sigma_j^2 = \beta_j \frac{u_j}{P_j}\left(1 - \frac{u_j}{P_j}\right); j = 1, 2 \qquad (4.4)$$

where β_j = design effect for unemployed, and u_j = unemployed.

Historical design effects by type of area were available, and were reduced to take into account of structural improvements in the respective NSR and SR designs as described in Sections 2 and 3. Unemployment levels were based on 1980-82 average LFS data, which seemed appropriate in light of medium term forecasts which were not calling for a return to pre-1982 recession levels of unemployment, and population counts were based on the 1981 Census.

Table 5 presents the percent of sample in SR areas under the following allocations: (i) old design, (ii) proportional allocation, (iii) optimum allocation under the assumed cost and variance model, and (iv) the allocation adopted for the redesigned sample. The optimum allocation could not be adopted because of subprovincial data reliability constraints. In most cases, the differences between the optimum allocation and the one adopted are small. The optimal allocation turned out to be quite close to proportional, and quite different from the allocation under the old design.

Table 5

Percent of Sample in SR Areas within Provinces for (1) Old Sample, (2) Proportional Allocation, (3) Optimum Allocation, and (4) Redesigned Sample

Province	Old Sample	Proportional Allocation	Optimum Allocation	Redesigned Sample
Newfoundland	41.8	51.3	42.6	44.6
Prince Edward Island	26.6	32.8	32.8	28.9
Nova Scotia	37.3	57.4	58.8	51.9
New Brunswick	49.5	52.5	47.4	53.6
Quebec	56.8	74.8	71.6	68.9
Ontario	62.5	79.1	78.8	75.0
Manitoba	54.1	71.0	76.4	56.4
Saskatchewan	44.7	51.8	62.1	56.8
Alberta	60.0	68.6	72.6	62.3
British Columbia	58.0	78.0	74.6	69.7
Canada	53.2	67.1	67.4	62.3

Table 6

Relative Efficiency of the Redesigned Sample Allocation with Respect to the Old by Province (Unemployed)

Province	Cost Ratio $(=\frac{C^{(O)}}{C^{(N)}})$	Variance Ratio $(=\frac{V^{(O)}}{V^{(N)}})$	Rel. Eff. $(=\frac{C^{(O)}V^{(O)}}{C^{(N)}V^{(N)}})$
Newfoundland	1.00	1.00	1.00
Prince Edward Island	1.01	1.02	1.03
Nova Scotia	1.04	1.14	1.18
New Brunswick	1.01	0.98	0.99
Quebec	1.03	1.06	1.09
Ontario	1.04	1.08	1.12
Manitoba	1.01	1.03	1.04
Saskatchewan	1.05	1.06	1.12
Alberta	1.01	1.01	1.02
British Columbia	1.02	1.09	1.11
Canada	1.03	1.07	1.10

The projected gains resulting solely from the re-allocation process under the assumption of fixed (old) provincial sample sizes and uniform sampling rates within the two area types are presented in Table 6. For this table, the unit costs and variances described above were used in determining the total costs and variances, $C^{(O)}$, $C^{(N)}$, $V^{(O)}$, $V^{(N)}$, under the old and new allocations respectively. The new allocation would have resulted in a 3% decrease in total cost and a 7% decrease in total variance of unemployed and for a combined relative efficiency (as defined in Table 6) of 1.10. Had it not been for the subprovincial data requirements, an efficiency gain of 1.12 could have been achieved under the optimal allocation.

The actual efficiency gains for the redesigned sample vs. the old sample are considered in the following section.

5. CONCLUSIONS

The changes in the LFS design taken as a result of the cost-variance studies are the following: 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 in 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 between 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 a 7% reduction in the overall LFS sample size and achieved the required reliability of subprovincial data (Singh et al. 1984) without impacting on the reliability of provincial and national estimates. The only exceptions were the provinces of Quebec and Manitoba, where greater subprovincial data demands

Table 7
Relative Efficiency of the Redesigned
vs. the Old Sample for Unemployed

Province	Cost Ratio* $(= \frac{C^{(O)}}{C^{(N)}})$	Variance Ratio $(= \frac{V^{(O)}}{V^{(N)}})$	Rel. Eff. $(= \frac{C^{(O)}V^{(O)}}{C^{(N)}V^{(N)}})$
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
Quebec	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 not included.

necessitated a slight loss in provincial data reliability. Table 7 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 7).

APPENDIX A

Variance Formula and Computation Method for RPPSS Sampling

Suppose that a sample of size n is selected by the randomized PPS systematic sampling from N units. Let p_i be the normalized size measure of the i-th unit such that $\sum_{i=1}^{N} p_i = 1$. The Horvitz-Thomson estimator of the total Y for a characteristics y is given by (Horvitz and Thomson 1952):

$$\hat{Y}_{HT} = \sum_{i \epsilon S} \frac{y_i}{\pi_i},$$

Where S = the selected sample of size n

y_i = y-values of i-th unit

π_i = np_i, the probability that the i-th unit is in S.

and its variance is

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \sum_{i<j} (\pi_i \pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 ,$$

where π_{ij} is the joint probability that both the i-th and j-th units are in S. Hartley and Rao (1962) gave an asymptotic formula for π_{ij}'s.

An exact formula by Connor (1966) is also available but quite involved. Recently Hidiroglou and Gray (1980) developed a computer algorithm using a modification of Connor's formula due to Gray (1971), which was used in our study and compared with the Hartley-Rao approximation. It was found that the Hartley-Rao approximations are very close to the exact values for $N \geq 16$. We decided to use the Hidiroglou-Gray algorithm for $N < 16$ and the Hartley-Rao approximation for $N \geq 16$ considering exponential increase in computation with the algorithm as N increases.

APPENDIX B

Cost Simulation of D_2 vs. D_1

In order to estimate r, the ratio of fees and expenses for travel from home to area, between PSU's, and between secondaries under NSR design alternatives D_2 and D_1, a Monte Carlo study was carried out. 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 M-measure for a given sample was defined in the following manner. Suppose that k interviewers cover an ER and $G_i = \{U_{ij}; j = 1, 2, \ldots, n_i\}$ is the i-th interviewer's assignment, with n_i second stage sampling units. Let (x_{ij}, y_{ij}) be the population centroid of U_{ij} defined in Euclidean coordinates. The M-measure for the ER is defined as

$$M = \sum_{i=1}^{k} M_i ,$$

$$M_i = \sum_{j=1}^{n_i} \{(x_{ij} - \bar{x}_i)^2 + (y_{ij} - \bar{y}_i)^2\}^{1/2} ,$$

where $(\bar{x}_i, \bar{y}_i)$ is the center of G_i, i.e., $\bar{x}_i = 1/n_i \sum_{i=1}^{n_i} x_{ij}$; $\bar{y}_i = 1/n_i \sum_{j=1}^{n_i} y_{ij}$.

The determination of optimum interviewer assignments, that is the minimization of the M-measure, reduces to a classification or clustering problem. The following clustering algorithms were investigated:

i) **Friedman-Rubin (1967) Transfer Algorithm**

This non-hierarchical algorithm which was adopted for stratification of the LFS sample (Drew et al. 1985), starts with a random partitioning of units and proceeds towards a local optimum by moving one unit at a time from one cluster to another if the move reduces M. It also checks that size constraints are not violated before moving a unit. An approximation to the global optimum is achieved by taking several initial random starts. A disadvantage of the Friedman-Rubin algorithm in this case was that the strict size constraints required in order to have approximately equi-sized assignments, restricted the movement of units between clusters.

ii) Dahmström-Hagnell (1975) Exchange Algorithm

This algorithm is similar to the Friedman-Rubin algorithm, except that it is based on exchanging pairs of units between clusters as opposed to transfering individual units. Hence it works better under strict size constraints.

iii) Combined Algorithms

Define a cycle of a combined algorithm as application of the exchange algorithm, followed by the transfer algorithm. Then we considered both single and two cycle combined algorithms.

The combined two cycle algorithm worked best, requiring the smallest number of random starts and the least computing cost to achieve the same level of optimality as the other algorithms. Performance of the 1 and 2 cycle combined algorithms based on 21 replicates is summarized below.

	One Cycle				Two Cycle		
	No. of Random Starts				No. of Ramdon Starts		
	1	2	4	10	1	2	4
M-measure*	336.18	329.19	325.65	325.51	327.55	325.69	325.51
Standard Deviation	15.84	15.45	15.67	15.69	16.10	15.67	15.69
Computing Cost ($)	5.94	11.24	21.67	53.90	8.17	15.12	29.38

* Average over 21 replicates.

REFERENCES

CONNOR, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

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., 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 (in preparation).

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, 421-453.

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

GRAY, G.B. (1971). Joint probability of selection of units in systematic samples. *Proceedings of American Statistical Association*, 271-276.

HARTLEY, H.O., and RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 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, 107-112.

HORVITZ, D.G., and THOMSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

LEMAITRE, 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, 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, 127-140.