COST MODELLING OF ALTERNATIVE SAMPLE DESIGNS FOR RURAL AREAS IN THE CANADIAN LABOUR FORCE SURVEY

Harold Mantel, Normand Laniel, Marie-Claude Duval and Jocelyne Marion, Statistics Canada
Harold Mantel, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6

KEY WORDS: Cost-variance study; Multi-stage sampling; Simulations.

1. Introduction

As part of a methodology redesign of the Canadian Labour Force Survey, two alternative sample designs were considered for rural areas: the current three-stage design and a proposed two-stage design. Here "rural" also includes small urban areas. Reduced clustering under the proposed design would lead to smaller variances and it would enhance the utility of the survey as a general household survey vehicle; however, data collection costs would likely be higher. It is also expected that the two stage design would make estimation for small areas easier, particularly estimation for small areas which were not planned in advance; however, we did not know how to measure the magnitude of this benefit. A general discussion of the role of the sample design in small area estimation is presented in Singh, Gambino and Mantel (1994).

In order to evaluate the relative efficiency of the two designs a cost-variance comparison was conducted, similar to the one reported by Choudhry, Lee and Drew (1985) for the last redesign of the Canadian Labour Force Survey. Census data were used to evaluate the relative variances under the two designs. For cost comparison a model of data collection costs was formulated in which these costs were divided into fixed costs, that would be the same under either design, and design dependent costs, which consisted of certain components of interviewers' travel. Detailed data on components of interviewers' travel under the current design were collected for the months of October and November 1992, and these were used to estimate components of the design dependent costs. Simulations of sampling and interviewer assignment formation were used to estimate ratios of components of the design dependent costs under the two designs.

Combining the estimated cost ratios with the separately calculated variance ratios, it was found that the overall efficiency of the proposed design is comparable to that of the current design. Based on these results and considerations of small area estimation it was decided to use a two stage design for most strata; however, we will continue to use a three stage design in strata which are sparsely populated.

2. Alternative Designs

The Canadian Labour Force Survey is a stratified multi-stage monthly household survey based on an area frame. Details of the design are given in Singh *et al.* (1990); here we describe briefly the current design for rural areas, which we call D0, and a proposed alternative, D1.

2.1 The Current Design. In the current design, there are three stages of sampling for rural areas. First, two or three primary sampling units (PSUs) are selected from each stratum using randomized probability proportional to size systematic (rppss) sampling. Strata and PSUs are generally contiguous collections of census enumeration areas (EAs). At the second stage, EAs are selected within each selected PSU using rppss sampling; the number of EAs selected varies across PSUs but is usually either 3 or 6. Occasionally the units selected at the second stage consist of more than one EA or part of a larger EA. After selection of EAs a list frame of dwellings in each sample EA is constructed in the field for use in the final stage of sampling in which dwellings are selected systematically within each EA, about ten dwellings per sample EA.

Each selected dwelling remains in the sample for six months. Personal visit interviewing is generally used for the first month, with telephone interviewing used whenever possible in months two through six. After the sixth month the dwelling rotates out of the sample and the next dwelling on the list rotates in. Each sample dwelling has an associated rotation group number, from one to six, which indicates the month in which the dwelling rotates out. Within PSUs, one sixth of the sample is in each rotation group. EAs and PSUs also rotate, though less frequently.

2.2. The Proposed Design. In the proposed sample design, D1, the first stage of sampling would be eliminated. EAs would be sampled directly from strata which could be somewhat smaller, on average, than strata under D0. Normally six EAs, but occasionally three or two, would be selected from each stratum using rppss sampling, and a systematic sample of about 10 dwellings would be taken from each EA. Each rotation group would be represented in one and only one sample EA in each stratum. For strata with six EAs selected there would be one sample EA in each rotation group; when two or three

EAs are selected, each sample EA would be split between, respectively, three or two rotation groups.

It is expected that D1 would have a lower design effect than D0, due to the elimination of a stage of sampling. Also, because the sample would be more evenly distributed geographically, it would likely be more appropriate for estimation for unplanned small areas, since the realized sample size in unplanned small areas would tend to be more stable.

It is also expected that data collection costs would be somewhat higher under D1 than under D0, because of higher travel costs due to an interviewer's assignment being spread out over a larger area. The restriction that each rotation group appears in only one sample EA in each stratum would reduce costs since generally only the EA with dwellings rotating into the sample would need to be visited each month; however, more than one EA may need to be visited if, for example, there are vacants or non-respondents. As well, estimates of variance for D1 would be inflated since they would include a component due to month in sample bias.

3. Overall Strategy

In order to compare the efficiency of designs D0 and D1 a cost-variance comparison was carried out. The cost variance comparison was conducted separately in each of 14 out of the 71 economic regions (ERs) in the ten provinces. ERs are large subprovincial regions of similar economic structure which are the first level of stratification for the Canadian Labour Force Survey. The relative variances under the two designs could be calculated using census data. Obtaining data with which to compare the data collection costs under the two designs was more problematic.

As a first step models of data collection costs under D0 and D1 were constructed. Then data on current interviewer travel were collected to estimate components of the collection costs under D0. Finally, sampling and interviewer assignment formation was simulated under both D0 and D1 in order to estimate the ratio of various cost components under the two designs. An overall estimated relative data collection cost is then obtained as a weighted average of the estimated relative costs of the different components, with weights given by the estimated cost for each component under D0.

4. Cost Model

For comparison of data collection costs for rural areas under D0 and D1 we divide the overall data collection costs into two components, design dependent costs and fixed costs.

For fixed costs we assumed that all costs not associated with interviewer travel would be identical under the two designs. As well, it was assumed that costs for travel within EAs (second stage units) under D0 would be equivalent to costs for travel within EAs (PSUs) under D1.

The remaining components of travel cost are assumed to depend on the design. These components are:

i) travel between rural and major urban areas

ii) travel between penultimate stage sampling units (penSUs) within rural

iii) travel between the interviewer's home and work area within rural

The term "penultimate stage sampling" in *ii* refers to the stage of sampling just before the final stage of sampling dwellings. For both D0 and D1 penSUs correspond to EAs; however, these are secondary sampling units under D0 and primary sampling units under D1.

For cost comparison of the two designs we assume that both the average numbers and the average speeds of each type of trip *i* through *iii* are the same under both designs. The two designs may differ only in the average distance of each type of trip.

Differences between the two designs for average distance of trips of types *i* and *ii* would be caused by differences in the average spread of interviewers assignments.

Results of the special time and cost study (described in the next section) indicated that trips of type *iii* account for about 2/3 of overall design dependent costs. This component is also quite distinct from the others in terms of average distance of a trip and average speed, and it was expected that the impact of the competing designs on trips of type *iii* would be quite different from their impact on trips of types *i* and *ii*.

The final model for rural data collection cost under D0, which we denote by $C0$, is then

$$C0 = C_F + C_D + C_H \, , \tag{4.1}$$

and that for D1 is

$$C1 = C_F + f_D C_D + f_H C_H \, , \tag{4.2}$$

where

C_F is the total fixed costs,

C_D is the total cost for travel of types *i* and *ii*,

C_H is the total cost for travel of type *iii*,

f_D is the ratio of average distances between rural penSUs in interviewers assignments under D1 and D0, and

f_H is the ratio of average distances of rural penSUs in interviewers' assignments from their homes, under D1 and D0.

Data from the special time and cost study were used to estimate the components C_F, C_D and C_H. Simulations of sampling down to penSUs and interviewer assignment formation under D0 and D1 were used to estimate the cost ratios f_D and f_H.

5. Time and Cost Study

A special time and cost study was undertaken in October and November of 1992 to obtain data that could be used to estimate components of data collection costs. In particular, it was necessary to determine the current extent of travel within and between the various levels of sampling units.

A time and cost study with similar ambitions was undertaken in 1982 (Lemaître, 1983). In that study about 300 interviewers (25%) were asked to keep a travel diary in which they recorded details of each visit to a sample dwelling. The detail of the data obtained from this study was very useful for an investigation of the likely cost implications of alternative designs; however, there was some skepticism about the quality of the data since the procedure was quite disruptive for the interviewers and they may not have considered it to be an important part of their work.

For the time and cost study described here interviewers were not asked to keep a separate diary of their visits to sample dwellings. Instead, the necessary information was collected on the household dockets (form F03) as the interviewers were going about their usual tasks. Information identifying the dwelling visited was already present on the form. The interviewers were asked to provide the bare minimum of additional information needed for reconstruction of the sequence of their personal visits and related travel distances and times. The additional burden to the interviewers was thus minimized at the cost of some additional complexity in the processing of the F03 forms and the data.

The data items collected for each personal visit to a sample dwelling were the day, the times of arrival and departure, and the vehicle odometer reading. The records could then be sorted by time within day and the time and distance of each leg of a trip could be determined by comparing successive visits. The time and odometer reading were also collected at the beginning and end of each trip, to allow reconstruction of the starting and finishing legs of a trip. Interviewers time was converted to a cost by assuming an average rate of $11.50 per hour, distances were converted using specified rates per kilometer that varied from province to province.

A second source of relevant information was the F85 forms that each interviewer fills out in order to be paid. These forms give hours, kilometers and other expenses for each contiguous period of work, typically a morning, afternoon or evening, broken down by project and task codes. These data allowed verification of total kilometers travelled as obtained from the F03 forms. It was also possible, by subtraction, to get a measurement of non-travel time, which includes time for telephone and personal interviewing, interviewer preparation, and possibly other minor tasks.

As mentioned above, one of the objectives in designing the data collection tool for this time and cost study was to minimize the extra work that interviewers had to do in order to record the data. For this reason the data was recorded in two separate areas of the F03 form. In area 9 of the F03 form, which was already being used to record time and date of attempted contacts, interviewers were asked to indicate which attempted contacts were personal visits (rather than telephone calls) and to record the time of arrival. The remaining information, odometer reading and time of departure, was recorded in area 50 which is normally reserved for supplementary questions. Data items from the two areas would be linked in the order that they appeared on the form to create complete records of each visit.

Unfortunately this minimization of additional data to be recorded also maximized the number of ways in which things could go wrong. In retrospect, it would have been better had all the data for each dwelling visit been collected in area 50. Looking to the future, with the growing use of computer assisted interviewing technology there is an enormous potential for using that technology to measure and monitor survey operations. The development of this potential could make collection of data such as these routine and would greatly improve their quality.

There were often inconsistencies in the number of personal visits indicated in area 9 and area 50. For October data from the rural parts of the 14 study ERs such inconsistencies were found for 955 out of 3,398 F03 forms for which some personal visitation was indicated. For November such problems occurred in 1,719 out of 3,600 cases. In most of these cases it was possible to reconstruct what the data should have been by comparing the sequence of attempted contacts in area 9 and area 50. A computer program was written to detect and correct the most common problems, a further 309 records were inspected and corrected manually, and one record was unusable.

The 6,998 F03 records for which some personal visitation was indicated yielded records of 11,442 incidents, where an incident means an interviewer either visiting a sample dwelling or leaving from or returning to their own home. 123 of these records had arrival times more than 5 minutes after departure; 78 had departure times more than an hour after arrival. These were examined manually and corrected where possible.

Missing values for either time of visit or odometer reading caused some difficulties in sequencing the visits of each interviewer. In other cases problems were apparent from very large travel times or distances or excessively high speeds. These cases were investigated and, whenever possible, corrected; sometimes missing values were substituted for bad data, a small amount of data was completely discarded.

Useable records of 9,838 legs of trips were obtained. About 11.5% of these had missing time and 2.6% had missing distance. Missing data were imputed separately within each ER and type of trip (12 classes) using a combination of ratio and hot-deck imputation.

The last step was reconciliation of the time and cost data from the F03 forms to the F85 claims data. The F85 data was taken to be correct and the time and cost data was ratio adjusted within interviewers to match the claimed kilometers of travel. For each interviewer other expenses and residual hours from the claims form were allocated to rural or major urban in proportion to the number of sample dwellings of each type for that interviewer. Kilometers and hours were converted to costs as described above and overall costs were summed to determine the components C_F, C_D and C_H in (4.1).

6. Estimation of *f*

The factors f_D and f_H in (4.2) were estimated separately for each of the 14 study ERs. Following Choudhry, Lee and Drew (1985) this was done by a simulation study in which, under each of D0 and D1, sampling was simulated down to penSUs, rotation group numbers were assigned, interviewer assignments were formed, interviewers' homes were randomly located, and Monte Carlo averages of a measure of spread within assignments and distance from home to area were obtained.

The first step was specification of the designs to be simulated, *i.e.*, delineation of strata, PSUs and secondary stage units, and specification of sample sizes for each stage. For D0 the current design was used. Strata under D0 are compact and contiguous collections of EAs (Census Subdivisions in Quebec and Ontario), chosen to be as homogeneous as possible with respect to 16 stratification variables measured in the census. Furthermore, PSUs were also formed optimally as compact collections of EAs (not necessarily contiguous).

For design D1 also, strata were formed optimally within each study ER using the same stratification variables as had been used for design D0, subject to constraints on contiguity and size. However, in order to increase the utility of the proposed design D1 for small area estimation, the stratification units were Census Divisions, since it was thought that geographical domains which may become of interest in the future would likely be defined as Census Divisions or collections of Census Divisions. A stratum under design D1 then consisted of a small number of contiguous Census Divisions or sometimes a single Census Division. Therefore optimality considerations had a very limited role in stratification for design D1. Stratum sample sizes within ERs for D1 were proportional to number of dwellings and the ER sample sizes were close to what was expected under D0.

For the simulations samples were selected under D0 and D1 down to the EA level for the rural part of study ERs. Rotation group numbers were assigned at random to the sample EAs subject to the constraint that each rotation group was represented within each PSU for D0 and within each stratum for D1; if there were fewer than six EAs then some EAs would be assigned more than one rotation group number. As much as possible, the expected number of sample dwellings in each rotation group was balanced at the PSU (for D0), stratum and ER levels. The sample and rotation group assignments for major urban areas was taken as given rather than being simulated, since it was not expected that changing these samples would have much effect on assignment formation or travel in rural areas.

The next step was formation of interviewer assignments. The criteria for good interviewer assignments are that they should be geographically compact, balanced with respect to rotation groups, and of roughly equal size. We constructed a loss function

$$L = wL_1 + L_2$$

where L_1 measures average lack of compactness, L_2 measures average lack of balance with respect to rotation group, and w is chosen judgementally to balance the importance of compactness and rotation group balance. We used a stratification program to find assignment formations that minimized L subject to the constraint that assignment sizes were within 50% of the average within ERs.

To complete the simulation it was necessary to give interviewers a location for their home. For assignments that were partly in major urban centres we just assumed that the interviewers home was at the centroid of the urban part. For purely rural assignments the interviewers home was given a random location with a circular bivariate normal distribution centred at the assignment centroid and scaled so that the expected distance from the centroid matched the average home to work area distance obtained from the time and cost study. For Monte Carlo efficiency, this step was repeated independently ten times for each simulated sample.

The estimate of f_D in (4.2) was taken to be the ratio of the Monte Carlo average of the distances from sample rural EA centroids to respective assignment centroids under D1 to that under D0. Similarly, the estimate of f_H in (4.2) was taken to be the ratio of the Monte Carlo average of the distances between interviewers homes and assignment rural EAs under D1 to that under D0.

7. Results

Table 1 gives the estimated monthly cost, $C0$, of data collection in rural areas under design D0 for each of the 14 study ERs, as well as the estimated proportions of those costs which are of types F, H and D. Also given are the estimated cost ratio factors f_H and f_D, the estimated relative cost of D1 to D0,

$$C_{rel} = C1/C0 = (C_F + f_D C_D + f_H C_H)/(C_F + C_D + C_H),$$

the estimated relative variance for the characteristic "unemployed", V_{rel}, and the estimated relative efficiency of D1 to D0 for the characteristic "unemployed", $E_{rel} = 1/(C_{rel}*V_{rel})$.

As expected, the estimates of the cost ratio factors f_D and f_H, and of the relative cost C_{rel} are all greater than 1. There is wide variation among the estimated cost ratio factors, with f_H varying from 1.07 to 1.93, and f_D varying from 1.14 to 1.99. There is a positive correlation (.63) between the estimates of f_H and f_D, which is to be expected. The proportions of data collection costs attributable to home to area travel and to other design dependent costs also vary widely, with $C_H/C0$ varying from .12 to .30 and $C_D/C0$ varying from .03 to .17. There is also a moderate negative correlation (-.43) between f_H and $C_H/C0$, as well as a fairly weak negative correlation (-.17) between f_D and $C_D/C0$. These negative correlations tend to diminish the variation in C_{rel}, the relative data collection costs, which varies from 1.03 to 1.21.

Table 1: Estimated cost components, relative cost factors, relative costs and relative variances for comparing D1 to D0

ER	$C_F/C0$	$C_H/C0$	$C_D/C0$	$C0$	f_H	f_D	C_{rel}	V_{rel}	E_{rel}
020	0.76	0.17	0.07	1802	1.13	1.14	1.03	0.86	1.13
220	0.67	0.22	0.11	4916	1.17	1.64	1.11	0.88	1.02
310	0.73	0.19	0.08	3492	1.12	1.36	1.05	0.81	1.18
320	0.74	0.18	0.08	2904	1.07	1.39	1.05	0.66	1.44
330	0.69	0.19	0.12	3081	1.19	1.41	1.09	1.02	0.90
340	0.71	0.19	0.10	2550	1.16	1.20	1.05	0.84	1.13
350	0.67	0.16	0.17	2843	1.18	1.51	1.12	0.93	0.96
411	0.77	0.14	0.09	1637	1.93	1.89	1.21	0.83	1.00
510	0.76	0.21	0.04	4152	1.31	1.62	1.09	0.96	0.96
560	0.74	0.20	0.07	1977	1.27	1.32	1.08	1.06	0.87
630	0.62	0.30	0.08	2177	1.12	1.86	1.10	0.95	0.96
720	0.63	0.29	0.07	1906	1.26	1.52	1.11	0.97	0.93
820	0.80	0.17	0.03	3460	1.21	1.64	1.06	0.93	1.01
960	0.82	0.12	0.07	2158	1.59	1.99	1.13	0.95	0.93

The estimated relative variances of D1 to D0, V_{rel}, also vary widely, from .66 to 1.06. As mentioned earlier, it was expected that variances under D1 would be smaller than those under D0 because of smaller design effects due to the reduced clustering under D1. This is most often the case, but not for study ERs 330 and 560 in which V_{rel} is larger than 1.

One possible explanation for the variance under D1 being larger than that under D0 is the much more limited role of optimality considerations in the stratification for D1, as described in Section 6.

A second possible explanation for V_{rel} being larger than 1 in some ERs is that strata and PSUs under D0 may in some cases have been more robust with respect to optimality than strata under D1. 1981 census data were used in the specifications of both designs D0 and D1; however, the relative variances were calculated using 1986 census data.

The estimated relative efficiencies of design D1 to design D0 for the characteristic "unemployed" vary widely, from .87 in ER 560 to 1.44 in ER 320. In 6 out of the 14 study ERs the estimated relative efficiency is greater than 1, suggesting that for those ERs design D1 is more efficient. The raw average of the estimates is 1.03. The weighted average, using *C*0 as the weight, is 1.04. Thus, there seems to be very little difference in the overall efficiencies of designs D0 and D1; however, for individual ERs there may be an appreciable difference.

Based on the similar overall efficiencies of designs D0 and D1 for the 14 study ERs, and considering the enhanced utility of design D1 for small area estimation, it was decided to use design D1 wherever feasible. The only exceptions are those strata which are very sparsely populated, in which a three stage design will be used.

References

Choudhry, G.H., Lee, H., and Drew, J.D. (1985). Cost-Variance Optimization for the Canadian Labour Force Survey. *Survey Methodology*, **11**, 33-50.

Lemaître, G. (1983). Results from the Time and Cost Study. Internal report, Statistics Canada.

Singh, M.P., Drew, J.D., Gambino, J.G., and Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey*. Statistics Canada, Catalogue number 71-526.

Singh, M.P., Gambino, J., and Mantel, H. (1994). Issues and Strategies for Small Area Data. *Survey Methodology*, **20**, 3-22.