ANALYSIS OF URBAN CLUSTER SIZE IN THE CANADIAN LABOUR FORCE SURVEY

Normand Laniel and Chris Mohl, Statistics Canada
Normand Laniel, Statistics Canada, R.H. Coats Bldg, 16-P, Ottawa, K1A 0T6, CANADA

KEY WORDS: Survey Costs, Sampling 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 such as employment and unemployment in the Canadian population.

The LFS follows a stratified multi-stage sample design. In large urban areas it uses an area frame and sometimes an apartment list frame when the number of apartment dwellings justifies it. The sample is selected in two stages in these areas. In small urban and rural areas an area frame is used and the sample is selected in three stages.

The LFS has a rotating panel sample where one sixth of the households are replaced each month. Any household remain in the sample for six consecutive months. This permits the production of efficient estimates of month-to-month changes and prevents undue respondent burden.

Data collection is done through personal interviews in the first month and by telephone interviews in subsequent months. Since March 1994, the collection is performed using laptop personal computers.

The LFS is undergoing a redesign since 1991. The redesign attempts to introduce new methodologies in order to balance the total cost of the survey with improved reliability of the resulting estimates. As well, it aims to enhance the LFS for use as a general vehicle for other household surveys (more details in Drew *et al.*, 1991 and Singh *et al.*, 1993).

The present paper focuses on the research aimed at the cost-variance optimization of the sample design for the area frame in large urban areas.

Section 2 of this document presents the sample designs compared in the study. The methodology of the study is discussed in section 3 and the results in section 4. The conclusions are presented in the final section.

2. SAMPLE DESIGNS

In the following paragraphs, the details of the current LFS sample design for the area frame in large urban centres are presented (for more details see Singh *et al.*, 1990). As well, the rationale and description of the design alternative are given.

2.1 Current Design

In the largest urban areas, the Area Frame has a two-level stratification. The primary strata are formed contiguous and compact using a modified version (Drew *et al.*, 1985) of the optimal stratification algorithm of Friedman and Rubin (1967). Within primary strata, secondary strata are formed non-contiguous and non-compact also using the optimization algorithm mentioned above.

In the other large urban areas, depending on its size, either a single level non-geographic stratification was carried out using the optimization routine or the formation of one to three geographic strata was done manually.

Two stages of sampling are used for the Area Frame. The first stage unit is a small geographic area called a cluster and the second stage unit is the dwelling. The cluster is often a city block, or a set of blockfaces, with a size of 20 to 100 dwellings. In some instances, it is a census enumeration area, in which case its size is around 250 dwellings. The level of geographic detail and census counts available from the 1981 census determined the type of first stage units. In urban areas where the blockface information was available in machine readable form, the clusters were formed automatically. Otherwise, the formation was done manually.

Under the current design, the clusters are sampled using the Rao, Hartley, Cochran (1962) random groups method. In this method, within each stratum the list of clusters is randomly split into n groups (normally six, but twelve in special cases) and one cluster is selected from each random group with probability proportional to size. The random group method has numerous advantages, the most important being the possibility of using the Keyfitz sample updating method to improve the efficiency of the sample design. The Keyfitz method is used to incorporate new selection probabilities of clusters while maximizing the retention of already selected units. The new selection probabilities reflect the change in size of clusters since census time as a consequence of the construction of new dwellings and/or the demolition of old ones within the cluster boundaries.

The second stage units (i.e., the dwellings) are selected systematically within the selected clusters.

The LFS sample design is self-weighting. This is achieved by setting the second stage sampling fraction within a selected cluster equal to the stratum overall sampling fraction divided by the selection probability of the selected cluster within its random group. In most urban areas there are, on average, 5 to 6 dwellings selected per cluster.

As mentioned in the introduction, each month one sixth of the dwellings in the sample are replaced. Consequently, when all dwellings in a cluster have been in the sample for six months the cluster has to rotate out. This cluster rotation is performed as follows. Each random group within a stratum is assigned a rotation number between 1 and 6. Initially selected clusters are retained for a random number of months to ensure that initial selection probabilities are preserved when cluster rotation occurs. Subsequently selected clusters are retained for a number of months equal to six times the inverse of the within cluster sampling fraction. Every six months, the sample of dwellings within a cluster is replaced by a new one until the retention period of the cluster is over. At this point the cluster is replaced by a new one from its random group.

When a cluster is first introduced into the sample, a list of habitable dwellings within the cluster boundaries is set up in the field by interviewers. This is normally done a few months prior to its first month in the sample and during a non-survey week. Subsequently, in order to keep the sample representative of the population, every six months (i.e., at the time of dwelling rotation), the cluster list is checked and updated for any new dwellings constructed and/or old dwellings demolished. This list updating is usually done during the survey week at the time of performing personal visits in the cluster.

The estimation methodology of the LFS involves the use of a regression estimator where the auxiliary variables are age-sex groups at the province level and the population aged 15 years and older at the sub-provincial level.

2.2 Design Alternative

As mentioned above, with time the size of clusters changes as dwellings are added and deleted within their boundaries. These changes are not uniform across clusters; some decrease in size or do not change, while others are subject to small or large growth. The changes become more significant as we move further away from the time of the census on which the sample design is based. The result is a less efficient sample design. One solution that was used from 1978 to 1982, was to recalculate the cluster selection probabilities according to the up-to-date size measures and reselect the clusters by applying the Keyfitz update method within each random group. The method permitted the retention of 70% of the originally selected clusters. In other words, 30% new clusters had to be selected. Such a sample update involves costly activities in the field, namely, the counting of dwellings within all clusters, selected or not, and the listing of the newly selected clusters. During the last decade, due to tight budget constraints no funds were available to update the sample. Hence, one objective of this redesign was to design a sample more robust to cluster size changes in case the same funding scenario persists for the next decade.

One idea that was proposed to make the sample design more robust with respect to cluster size changes was to use clusters of larger size. The rationale for this proposition is that the relative size change should be smaller for larger clusters and thus such a design would be more efficient. In the current design the average cluster size is around 50 dwellings. A possible alternative for a larger cluster is to use the census Enumeration Area which has an average size around 250 dwellings in urban areas. As well, there is the possibility of using the Computer Assisted Districting Program (CADP) to form the clusters. That program was used to form EAs for the 1991 census in areas where blockface level data was available. With CADP the desired size of the cluster is specified by the user as an interval. It was then proposed to study three other sizes of cluster: 100, 150 and 200 dwellings. When new cluster sizes are larger and the number of clusters selected in the sample remain the same, the new listing costs are larger as well. To reduce these costs, given specific listing procedures, more dwellings have to be selected from each selected cluster. Consequently, the average number of dwellings selected per cluster (called the **density** hereafter) was varied from 5 to 16 for this study.

3. METHODOLOGY OF THE STUDY

The study had the objective to compare a design based on the current average cluster size of around 50 dwellings with designs based on four other average cluster sizes: 100, 150, 200 and 250. The criteria used to perform the comparison are the sampling variance of the sample design and the field costs, namely: enumeration, listing and list updating costs. The urban area of Ottawa as defined by the 1981 census was chosen to study the impact of different cluster sizes on the efficiency and costs. More details about the data, the sampling variance and the field costs are provided in the next sub-sections.

3.1 Variance

In this section, the calculation of the sampling variance is first discussed and then followed by a description of the data used in the calculations.

3.1.1 Variance Formula

To approximate the LFS regression estimator, a combined ratio estimator was used in this study where the auxiliary variable is the population aged 15 years and older. This estimator can be written as

$$\hat{Y}_c = \hat{R}_c X = \frac{\hat{Y}}{\hat{X}} X \ .$$

It is well known that the variance of this estimator can be approximated by

$$V(\hat{Y}_c) \approx V(\hat{Y} - R_c \hat{X}) = V(\hat{U}) = \sum_h V(\hat{U}_h) \ .$$

The following notation will be used in the remainder of the section:

N_h the number of clusters in stratum h,
N_{hg} the number of clusters in random group g of stratum h,
Z_h the sum of size measures over the clusters in stratum h,
z_{hi} the size measure of cluster i in stratum h (e.g., the number of dwellings in the cluster as per the 1981 census),
$U_h = Y_h - R_c X_h$ the stratum total for the transformed variable u,
$u_{hi} = y_{hi} - R_c x_{hi}$ the total for cluster i in stratum h for u,
M_{hi} the number of dwellings in cluster i of stratum h (e.g. sometime after the 1981 census),
$u_{hij} = y_{hij} - R_c x_{hij}$ the total for dwelling j in cluster i of stratum h for variable u,
$\bar{u}_{hi}$ the average of u per dwelling in cluster i of stratum h and
S_{hi}^2 the variance of u in cluster i of stratum h.

Under the two-stage sampling scheme described in 2.1, as given in Choudhry *et al.* (1985) the variance from stratum h is

$$V(\hat{U}_h) = V_1(\hat{U}_h) + V_2(\hat{U}_h)$$

with

$$V_1(\hat{U}_h) = \frac{\left(\sum_{g=1}^{n_h} N_{hg}^2 - N_h \right)}{N_h(N_h - 1)} \left(\sum_{i=1}^{N_h} \frac{u_{hi}^2}{z_{hi}/Z_h} - U_h^2 \right)$$

and

$$V_2(\hat{U}_h) = \sum_{i=1}^{N_h} \left[W_h - 1 - \frac{\left(\sum_{g=1}^{n_h} N_{hg}^2 - N_h \right)}{N_h(N_h - 1)} \left(\frac{Z_h}{z_{hi}} - 1 \right) \right] M_{hi} S_{hi}^2$$

$$\text{where } S_{hi}^2 = \frac{\sum_{j=1}^{M_{hi}} (u_{hij} - \bar{u}_{hi})^2}{M_{hi} - 1} .$$

For this study we decided to estimate the theoretical variance above using the labour force data available from the census for a 1 in 5 sample. Using a tilde to indicate a census sample estimate, we have

$$\tilde{V}(\hat{U}_h) = \tilde{V}_1(\hat{U}_h) + \tilde{V}_2(\hat{U}_h).$$

Assuming that the sampling error attached to the census estimate of R_c is negligible, approximately unbiased estimates of the two components of variance are

$$\tilde{V}_1(\hat{U}_h) \approx \frac{\left(\sum_{g=1}^{n_h} N_{hg}^2 - N_h \right)}{N_h(N_h - 1)} \times \left(\sum_{i=1}^{N_h} \frac{\tilde{u}_{hi}^2 - \tilde{V}(\tilde{u}_{hi})}{z_{hi}/Z_h} - (\tilde{U}_h^2 - \tilde{V}(\tilde{U}_h)) \right)$$

and

$$\tilde{V}_2(\hat{U}_h) \approx \sum_{i=1}^{N_h} \left[W_h - 1 - \frac{\left(\sum_{g=1}^{n_h} N_{hg}^2 - N_h \right)}{N_h(N_h - 1)} \left(\frac{Z_h}{z_{hi}} - 1 \right) \right] M_{hi} \tilde{S}_{hi}^2$$

3.1.2 Data for Variance Calculation

To simulate the deterioration of selection probabilities due to changes in cluster size, we have used 1981 census data to determine the initial size of clusters and 1991 census data to reflect the change in size. Within Ottawa, only the areas where blockface level data was available were included in the study. As a result, a population of 132,800 dwellings is used in the study. The sample size determined at the time of designing the current sample for Ottawa was 390 dwellings. This is the size we have taken for this study.

In order to vary the density and the number of clusters selected, three stratifications were used as summarized in Table 1. Only three density values were used since previous investigations showed that the variance curves are close to linear. The strata were formed following the methodology of the current LFS sample design as described in sub-section 2.1.

The average cluster size in the current design for Ottawa is around 50 dwellings. To form clusters of average size 100, 150, 200 and 250, the current clusters were grouped with neighbours such that the desired

average size was achieved. The accepted tolerance, i.e., deviation from average, was 50%.

Table 1. Number of Strata and Selected Units

Number of strata	Number of clusters	Density
14	84	4.7
8	48	8.2
4	24	16.4

In order to calculate the variance of the combined ratio estimator for the different designs, employment, unemployment and population (aged 15 years and older) data at the dwelling level was retrieved from the 1991 census sample databases. The census data was linked to clusters, which are defined using 1981 census data, by using blockfaces. When a blockface was new in 1991 it was linked to the nearest blockface existing in 1981 and then assigned to the cluster containing the old blockface. After this process, the distribution of the growth amongst clusters in the study from 1981 to 1991 was plotted and observed to be very similar to the distribution observed in the current LFS sample from urban areas.

3.2 Cost

Given that the alternative designs involve the use of different cluster sizes and that this means different cluster life lengths, the costs have to be evaluated for the whole life of the design, i.e. 10 years for the LFS. In this section, the cost models are first discussed, followed by a description of the data used in the calculations.

3.2.1 Cost Models

The components of cost considered for the study were enumeration, listing and list updating, hereafter denoted as C_E, C_L and C_U, respectively. The models used for the comparison are described below.

Choudhry *et al.* (1985) reports the results of an enumeration cost study where the number of dwellings selected per cluster was varied from 2 to 10. The number of clusters selected was decreased as the number of selected dwellings per cluster increased. The enumeration cost included the interviewing time and all travel components (i.e. home to area and back, cluster-to-cluster travel and dwelling-to-dwelling travel). They observed that enumeration cost was constant when the number of dwellings selected per cluster varied. In other words, the amount of clustering of the sample of dwellings had no impact on the total enumeration cost. Based on this finding, it was assumed for this study that the enumeration cost is directly proportional to the total number of dwellings selected in the sample. That is

$$C_E = c_E m F L_D$$

where c_E is the enumeration cost per selected dwelling; m is the total number of dwellings selected in the sample; F is the frequency of the survey (e.g. 12 months per year for the LFS); and L_D is the life of the design in years (e.g. 10 for the LFS).

Over the life of the design, the amount of listing of clusters, i.e., creation of a list of dwelling addresses within a cluster, depends on the clusters initially selected in the sample and those subsequently rotating into the sample. The former clusters tend to be of larger size than the average and the latter of smaller size. Consequently, the listing cost of initially selected clusters is larger than of those subsequently replacing them. However, having larger clusters in the initial sample also implies that it takes longer to replace them. That reduces the number of clusters to list over the life of the design. To account perfectly for the change in cost with time, one would have to incorporate the cluster selection probabilities in the cost models. For the purpose of this study we have simplified the cost models by assuming that all the clusters are of the same size. Another assumption that we have made is that the size of clusters does not change with time whereas in practice the average size increases with time. We believe that the following simplified models still provide a fair indication of the direction in which listing costs change with the cluster size.

Under the above assumptions, the listing cost can be decomposed into a cost independent of the cluster size and a cost dependent on the cluster size. The independent cost includes travel from home to cluster and back. The dependent cost includes dwelling-to-dwelling travel and writing down the dwelling addresses. The simple model is

$$C_L = (c_{0L} + c_{1L} M) n R_L$$

$$\text{with } R_L = 1 + (L_D - 1)/L_{clu,max} \quad \text{and} \quad L_{clu,max} = (M/\bar{m}) L_s$$

where c_{0L} is the average independent cost per cluster; c_{1L} is the average dependent cost per dwelling; n is the number of clusters selected in the sample; M is the (average) size of a cluster; R_L is the expected number of times clusters will have to be listed over the life of the design under the LFS rotation scheme; $L_{clu,max}$ is the average maximum life of a cluster in years; $\bar{m}$ is the average number of dwellings selected per cluster

or density; L_s is the life of a selection in years (e.g., ½ for the LFS).

To come up with a simple model for the list updating cost, we have made the same assumptions as for the listing cost. Furthermore, we have also assumed that the list updating of a cluster is performed during survey week (this is not the case in a few cases) and thus does not involve independent costs but only dependent ones (i.e. dwelling-to-dwelling travel and address writing). The simple list updating cost model is

$$C_U = c_U M n R_U \text{ with } R_U = L_D / L_U$$

where c_U is the average list updating cost per dwelling; R_U is the number of times clusters will have their list updated over the life of the design; and L_U is the updating period in years (e.g. ½ for the LFS). Note that clusters are listed a few months before they enter the survey; as a result their lists are updated during their first month in the survey.

3.2.2 Cost Data

For the enumeration cost, we have used the data from the Time and Cost Study (Mantel, 1994) which estimated c_E as being $6.20 per dwelling.

For the listing costs, the data provided to Head Office on a regular basis include kilometres, hours and other expenses but do not separate the dependent component from the independent one. However, from a discussion with those responsible for enumeration, reasonable assumptions were made which permitted us to approximate the components. As a result, c_{0L} was estimated as $15.13 per cluster and c_{1L} as $0.32 per dwelling.

The list updating cost is mixed with the enumeration cost and it is not possible to separate them. According to those responsible for enumeration the updating cost should be around one third of the listing dependent cost, thus we have used $c_U = 1/3 x c_{1L} = \$0.11$ per dwelling for this analysis.

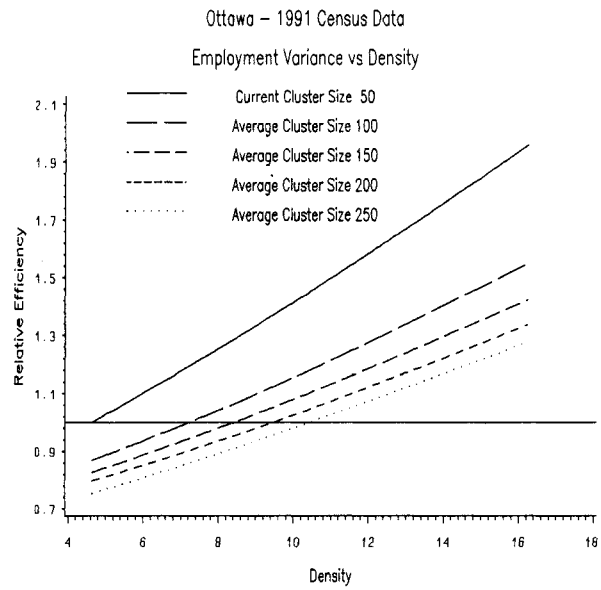
Since the cost data available is approximative, the results obtained from the cost models will provide an indication only of what would happen if the size of clusters is changed.

4. RESULTS

4.1 Variance Analysis

The result of the application of the variance formula in 3.1.1 to the data described in 3.1.2, is summarized in graphs 1 to 2. Graph 1 gives the relative efficiency, as compared to the current design parameters (i.e., cluster size of 50 and density equal to 4.7), for the characteristic employed. The efficiency is plotted as a function of the average size of the dwelling sample within cluster, i.e. the density. Note that since the size of the total sample of dwellings is fixed, the number of clusters selected decreases as the density increases which causes the variance to increase. As we were hoping, it is clearly observed that the variance becomes smaller as the cluster size gets larger. In particular, with a cluster size of 220 and a density value of 8 we get an efficiency gain of 9% at the mid-life of the design. The same analysis was done using 1986 census data, which corresponds to the start of the life of the design, and gave a 3% efficiency gain.

Graph 1



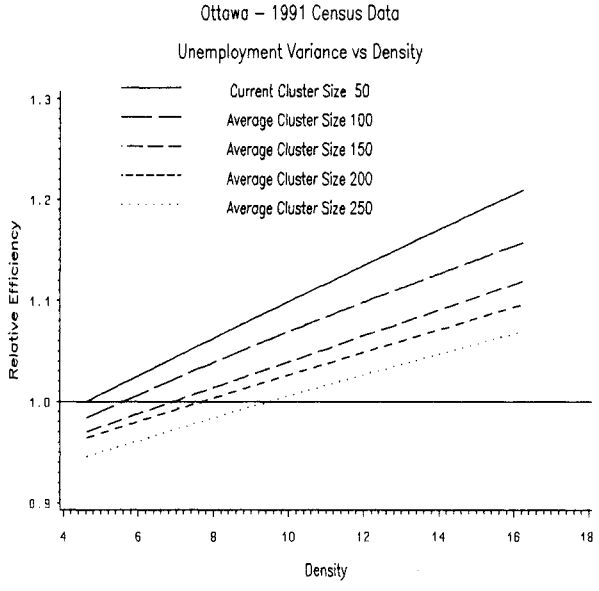
Graph 2 presents the results for the characteristic unemployed with the same layout as graph 1. The variance curves are much flatter for this variable. It suggests that we should expect an efficiency gain of about 1% at the mid-life of the design. Again, the same analysis was done using 1986 census data, which corresponds to the start of the life of the design, and gave no efficiency gain.

4.2 Cost Analysis

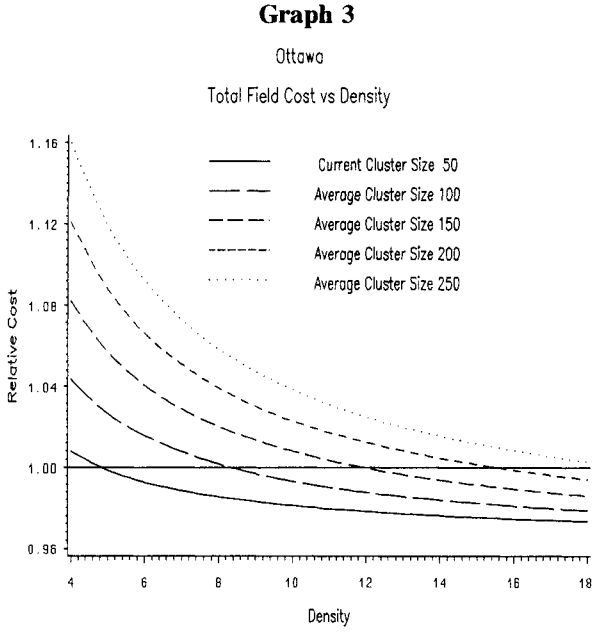
The curve of the total field cost as a function of the density has been produced for five cluster sizes using the cost models and data described in section 3.2. This is presented in graph 3 under the scenario of a 6-month updating period as in the current design.
The total field cost is relative to the cost of the current design. As expected, the curves show that the field cost decreases as the number of dwellings selected per cluster increases and as the cluster size decreases. If we were to choose a cluster size of 220 with a density equal to 8, used as an example in the variance analysis, the increase in cost would be around 5%. This increase in cost is essentially due to the list

updating cost component. This result is explained by the fact that the use of larger clusters means that for a given month more dwellings are involved in the list updating.

Graph 2



Graph 3



5. CONCLUSIONS

The variance analysis showed that the use of larger cluster sizes can improve the efficiency of the LFS estimates, significantly for employed and marginally for unemployed. Based on this finding a decision was made to use larger cluster sizes for the redesigned sample. Average sizes will be around 220 dwellings instead of being around 50.

The cost analysis revealed that larger cluster sizes may increase field costs due to the increase in list updating cost. As the estimates used for the cost model parameters were partly guesses based on experience, it was decided to simply monitor the field costs when introducing the new sample. If the monitoring shows that the costs are effectively higher then some measure to reduce them will be taken. For example, the length of the updating period could be decreased in areas where the growth of clusters is small or negligible. It is interesting to note that if the updating period was augmented to 12 months then the list updating cost using a size of 220 would be the same as the current cost.

As pointed out in section 3.2, the cost analysis was based under the assumption that all clusters are of the same size M. To improve the cost analysis one could use models that account for the variability of sizes and thus of selection probabilities. Another improvement to the cost analysis would be to obtain estimates of the parameters of the models from a special field study.

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.

Drew, J.D., Bélanger, Y. and Foy, P. (1985). Stratification in the Canadian Labour Force Survey. *Survey Methodology*, 11, 95-110.

Drew, D., Gambino, J., Akyeampong, E. and Williams, B. (1991). Plans for the 1991 Redesign of the Canadian Labour Force Survey. Proceedings of the Survey Research Methods Section, American Statistical Association.

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

Mantel, H. (1994). Time and Cost Study. Social Survey Methods Division, Statistics Canada, Ottawa.

Rao, J.N.K., Hartley, H.O., and Cochran, W.G.(1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, Series B, 24, 482-490.

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

Singh, M.P., Gambino, J. and Laniel, N. (1993). Research Studies for the Labour Force Survey Sample Redesign. Proceedings of the Survey Research Methods Section, American Statistical Association