Survey Methodology, December 1985
Vol. 11, No. 2, pp. 95–110
Statistics Canada

Stratification in the Canadian Labour Force Survey

J.D. DREW, Y. BÉLANGER and P. FOY[1]

ABSTRACT

The use of a multivariate clustering algorithm to perform stratification for the Labour Force Survey is described. The algorithm developed by Friedman and Rubin (1967) is modified to allow the formation of geographically contiguous strata and to delineate heterogeneous but compact primary sampling units (PSUs) within these strata. Studies dealing with stratification variables, stratification robustness over time, and type of stratification are described.

KEY WORDS: Multivariate clustering algorithm; Geographic stratification; Continuous survey.

1. INTRODUCTION

The Canadian Labour Force Survey is redesigned after every decennial census of population and housing. The redesign which occured following the 1981 Census included an intensive program of research on various aspects of the sample design (Singh, Drew and Choudhry 1984). This report describes the portion of the research program dealing with stratification methods.

Because the LFS is used not only to provide information on labour force characteristics but also as a general design for various other household surveys, one of the principal objectives of the redesign was to increase the flexibility of the LFS for general applications. Stratification was considered a means of improving efficiency for general applications, as well as variables of particular interest to the LFS, through the application of more rigorous procedures than those used in the old design.

It was therefore decided to consider the use of multivariate clustering algorithms and to compare them with the methods used in the old design. A non-hierarchical algorithm developed by Friedman and Rubin (1967) was selected on the bases of the results of evaluations of various algorithms by Judkins and Singh (1981) as part of the redesign of the Current Population Survey of the U.S. Bureau of the Census. A description of the basic algorithm and of the extensions which we have developed appears in section 2.

Sections 3 and 4 describe the evaluation studies and the stratification eventually adopted in the two main types of area distinguished by the LFS sample design, namely non-self-representing units (NSRUs) and self-representing units (SRUs). Section 4 also describes how the algorithm was adapted to delineate the primary sampling units (PSUs) within the NSR strata.

Section 5 concludes with a number of observations on the possibility of adapting the new system to other applications.

2. STRATIFICATION ALGORITHM

The basic algorithm used for stratification is a non-hierarchical multivariate algorithm developed by Friedman and Rubin (1967). This choice is based on the results of studies

[1] J.D. Drew and Y. Bélanger, Census and Household Survey Methods Division, P. Foy, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

performed by Judkins and Singh (1981) and Kostanich, Judkins, Singh and Schantz (1981), who assessed a number of stratification algorithms for the Current Population Survey of the U.S. Bureau of the Census.

The latter modified the objective function of the algorithm for sampling with probability proportional to size (PPS), and we have added the capacity to formulate compact, contiguous strata. A more complete description of the following appears in Foy (1984).

2.1 The objective function of the algorithm

The algorithm is designed to partition the stratification units (census enumeration areas) into strata which are as homogeneous as possible with respect to a number of variables of interest that is, by minimizing the sums of the squares within each stratum.

The expressions for the sums of squares in the case of sampling with PPS are shown below after introduction of the following notation:

L = number of strata to form

N = total number of units (enumeration areas)

N_k = number of units in group (stratum) k; $(N_1 + N_2 + \ldots + N_L = N)$,

T_{jk} = size measure of unit j in group k,

$T_{.k}$ = size measure of group k,

$T..$ = total size,

${}_iX_{jk}$ = observed value of variable i for unit j in group k,

${}_iX_{.k}$ = total observed values of variable i in group k,

${}_iX..$ = total observed values of variable i,

W_i = weighting factor of variable i (see section 2.4 for further details),

p = number of variables of interest.

Thus, the expression of the total sum of squares with PPS, of variable i is given by

$$SCT_i = \sum_{k=1}^{L} \sum_{j=1}^{N_k} \frac{T_{jk}}{T..} \left(\frac{T..}{T_{jk}} {}_iX_{jk} - {}_iX.. \right)^2.$$

This is also the variance expression of the estimate of ${}_iX..$ when a unit is selected with PPS. The total sum of squares weighted for all variables is thus

$$SCT = \sum_{i=1}^{p} W_i SCT_i.$$

The within-group and between-group sums of squares are obtained respectively by the following expressions:

$$SCW_i = \sum_{k=1}^{L} \frac{T..}{T_{.k}} \sum_{j=1}^{N_k} \frac{T_{jk}}{T_{.k}} \left(\frac{T_{.k}}{T_{jk}} {}_iX_{jk} - {}_iX_{.k} \right)^2$$

and

$$SCB_i = \sum_{k=1}^{L} \frac{T_{.k}}{T_{..}} \left(\frac{T_{..}}{T_{.k}} {}_iX_{.k} - {}_iX_{..} \right)^2.$$

Their sums of squares weighted for all variables are given respectively by

$$SCW = \sum_{i=1}^{p} W_i SCW_i$$

and

$$SCB = \sum_{i=1}^{p} W_i SCB_i.$$

The within-group sum of squares of variable i, SCW_i, is also the variance expression of the estimate of ${}_iX_{..}$ when a stratum, and subsequently a unit of this stratum, is selected with PPS.

Once again, we have the following result:

$$SCT_i = SCW_i + SCB_i, \ (i = 1, \ldots, p)$$

and

$$SCT = SCW + SCB.$$

The objective function of the stratification program is SCW, the within-group sum of squares weighted for all variables. We define the stratification index for variable i, I_i, as:

$$I_i = 100 \times \frac{SCB_i}{SCT_i} \qquad i = 1, \ldots, p.$$

A high index value indicates a good clustering.

2.2 Identification of the Best Clustering

One way of identifying the best clustering would be to generate all the possible partitions of N units into L groups and then simply select the one which minimizes the objective function. This approach is rarely feasible because the number of possible partitions may be unmanageably large.

Friedman and Rubin (1967) suggest the following algorithm. Begin with any partition of the N units into L groups. Consider moving a single unit to a group other than the one it is in. Move the unit to the group which offers the greatest reduction in the objective function. If no move will produce a reduction, leave the unit where it is. Using the partition thus created, we process the second unit in the same way, then the third, etc. The application of this procedure to each unit becomes an iteration which the authors describe as a *hill-climbing pass*. After several hill-climbing passes, the algorithm reaches a point at which no move of a single unit will produce a reduction in the objective function. This point is described as a local minimum of the objective function because it is dependent on the starting partition. Another starting partition might have achieved an even lower value of the objective func-

tion. To move beyond the local minimum, Friedman and Rubin describe two procedures, the *forcing pass* and the *reassignment pass*. By applying their algorithm to data described in their article, they obtain the highest known value of the objective function 10 times out of 14 runs from different starting partitions. They use another objective function, which is maximized. With some less well-structured data, the highest value was reached in 3 out of 11 runs, although it is impossible to be certain that this is the optimal solution. In their opinion, the forcing pass and reassignment pass methods are useful only on occasion. They have more confidence in the results obtained through the use of a number of starting partitions. This view is supported by Judkins and Singh (1981). We therefore decided to use the technique involving a number of starting partitions.

Because the algorithm moves only one unit at a time, calculation of the objective function is simplified. Following the initial calculation of the objective function, we merely recalculate the contribution to the objective function of the two groups involved in the move of the unit in question.

2.3 Contiguity

Previous LFS sample designs have used strata composed of contiguous geographic units; that is, each unit in a given stratum had to be touching at least one other unit in the same stratum. One of the main reasons was the assumption that such strata would retain the efficiency of the sample design for a longer period of time than if they were formed of discontiguous units.

In order to assess this assumption and to adopt the best possible stratification, we considered two means of taking geography into account in the stratification. The first method is described by Dahmström and Hagnell (1978), and consists of the use of centroids as variables of interest. This method uses two geographic variables (centroids), which are transformations of longitude and latitude. It yields compact strata, that is, strata in which the distance between units is made minimal by minimizing the usual within-group sum of squares of the centroids. However, the minimization is tempered by minimization of the other variables of interest. Moreover, there is no assurance that these strata will be composed of contiguous units.

The other method, which we describe as the contiguity vectors approach, is new. It guarantees contiguous, but not necessarily compact, strata. Studies described in section 3 dealt with the use of each of these methods in isolation or in combination.

2.3.1 Contiguity Vectors

To ensure the formation of contiguous strata, we proceeded as follows. Optimization is performed as described in the preceding section but beginning, in this case, with a starting partition which is contiguous, and permitting the movement of unit j from stratum A to stratum B only if, in addition to reducing the sums of squares, the following conditions are met:

i) unit j is contiguous to a unit in stratum B

ii) the movement of unit j to stratum B will not disrupt the contiguity of stratum A.

In order to verify these two conditions, it is essential that we know the links of contiguity between the units. Consequently, each unit must be assigned a contiguity vector containing a list of the units contiguous to it.

The first condition is easy to verify. In order to ensure that unit j is contiguous to a unit in stratum B, we must simply find one unit in its contiguity vector which is in stratum B.

The second condition is more difficult to verify. The principle is that a stratum is said to be contiguous if each pair of units in that stratum can be connected by a contiguous chain of units in that stratum. Suppose we want to move unit j from stratum A to stratum B. We therefore have to find, for each pair of units in the contiguity vector of unit j within stratum A, another link from among the units of stratum A. At this stage, the problem becomes like finding a path through a maze.

An algorithm has also been designed to create random starting partitions whose strata are contiguous.

2.4 Weighting of Variables

The weighting factors are of particular importance, since they determine the contribution of each variable to the cluster analysis.

It is usually preferable to standardize the variables by making the weighting factors inversely proportional to the total sum of squares of each variable. This standardization makes it possible to obtain a comparable contribution by each variable to the cluster analysis.

If, after standardization, we want to assign one or more variables greater importance in relation to the other variables in the optimization, we can do so by specifying a weight greater than 1 (normal). For example, a variable with a weight of 2 would have double importance. As described in section 3.2, we tested a number of combinations of weights for the geographic and non-geographic variables in an effort to obtain compact strata without unduly affecting the minimization of the other variables.

3. STRATIFICATION IN NON-SELF-REPRESENTING UNITS

3.1 Old Design (Platek and Singh 1976)

For the purposes of the LFS, each of Canada's ten provinces is divided into a number of economic regions (ERs), consisting of areas having similar economic structures. The boundaries of the ERs are determined in consultation with the provinces. These ERs are used as primary strata. The next stage in stratification is the partition of each ER into self-representing units (SRUs) and non-self-representing units (NSRUs). The self-representing units are cities in which the expected sample is large enough to represent at least one interviewer assignment; the NSR part make up the rest of the ER. Different sample designs are used in the SRUs and the NSRUs, because the population in the NSRUs is much more widely dispersed, necessitating a larger number of sampling stages. For the same reasons, we are retaining the concept of the SRUs and the NSRUs in the redesign.

In the old design, the NSR portion of each ER was stratified into a maximum of 5 contiguous strata with a population of between 36,000 and 75,000, based on the main characteristics of the 1971 census population, as described below and as discussed at greater length by Platek and Singh (1976).

The labour force was divided into 7 categories by industry. In each ER, the three largest industries were selected on the basis of specific criteria. The unit chosen for stratification was the combined municipality, which is the geographic region enclosed within a rural municipality and as such, often contains within its boundaries urban municipalities which are geographically smaller. By comparing, for each of these units, the proportions of the labour force working in each of the three categories with the corresponding proportions at the ER level, we identified the units showing a certain similarity which were grouped into strata. This comparison was done visually with graphics. Adjustments were occasionally necessary to satisfy the size and contiguity constraints.

Within each stratum, 12 to 15 PSUs were formed, all of them representative of the stratum in terms of the stratification variables, and of the ratio of rural to urban population. The rural parts of the PSUs were formed of contiguous EAs, and the urban parts were chosen to be as near to the rural part as possible. The sizes of the strata and the PSUs were determined so that, with two PSUs per stratum, the expected sample was equivalent to one interviewer's assignment size. On the basis of these criteria, and depending on the province, the population of the PSUs varied between 3,000 and 5,000 persons. Within the PSUs, sampling occurred in 2 or 3 stages.

3.2 Studies on Stratification during Redesign

Our studies were designed to produce conclusions which would assist in certain decisions relating to the following aspects of stratification: variables to be used, types of strata (wholly rural, wholly urban, or mixed), and the importance to be assigned to contiguity. Given the very limited time available for studies prior to the formation of the new strata and PSUs, and the general expectation that contiguous strata would be preferable over time to discontiguous strata, the first two aspects were given priority.

Some experimenting was required to find the best means of achieving contiguity, either by contiguity vectors, centroids or a combination of the two. However, following the redesign, a more detailed study was undertaken on the relative desirability of contiguous versus discontiguous strata.

3.2.1 Study on Variables and Type of Stratification

One constraint on the stratification method used in the old sample design was the limited number of stratification variables which could be taken into consideration (3 per ER).

With the new algorithm, this constraint is eliminated. In addition to the seven industry variables, we wished to determine the effect caused by the use of variables relating to the survey topic, such as employment, unemployment and income, and by such characteristics as education, housing and population. The latter characteristics have proven extremely efficient in similar studies performed by the U.S. Bureau of the Census for the Current Population Survey.

Table 1 describes the various options studied with respect to the choice of variables.

As regards the type of stratification, it was decided to study the effect of having separate strata for the rural and urban parts of the ERs, as an alternative to the mixed method of the old design.

The constraints on the sample design requiring PSUs to be approximately equivalent in population size, and the ratio between rural and urban population to remain generally the same for each PSU, frequently resulted in a lack of contiguity between the rural and urban parts of the PSUs. This led to an erosion in the presumed correspondence between the PSU and the interviewer assignment. Stratification into separate rural and urban parts, which could be substratified on an optimal basis, was, it was felt, a possible solution to this problem.

The study dealt with 11 economic regions from across Canada. The strata were defined on the basis of 1971 Census data, and assessed on the basis of 1981 census data. In performing the stratification, we used the 1971 Census enumeration areas as our stratification unit, except in Quebec and Ontario. For these two provinces, we selected census subdivisions, since the large number of EAs in certain ERs (up to 400) would have made execution of the computer programs extremely costly.

We used a conversion file between the geographic units of the two censuses to perform the evaluation based on the 1981 Census. The indices based on the 1981 data were considered more appropriate for evaluation purposes, since in fact the stratification data will be an average of 7 or 8 years old for the life of the sample design. Table 2 shows the indices based on both 1971 and 1981 census data.

Table 1
Stratification Options by Variables

Variables	Stratification option				
	1	2	3	4	5
Industries (7)[a]	x	x	x	x	x
Income		x	x	x	x
Employed		x	x		x
Unemployed		x	x[b]		x
Demography (2)[c]				x	x
Housing (4)[d]				x	x
Education (1)[e]				x	x

[a] number of persons employed in agriculture, forestry and fisheries, mines manufacturing, construction, transportation, services.
[b] double weighting on unemployment.
[c] population 15-24, population 55 and over.
[d] 1-person households, 2-person households, owned dwellings, total gross rent.
[e] secondary education.

For this study, we chose to form contiguous, compact strata, using contiguity vectors and centroids with an average weight of three (see subsection 3.2.2). The number of strata per ER was the same for all options.

The following conclusions were drawn from the results of the study, which are summarized in table 2.

Type of Stratification: Rural/urban stratification was far superior to total stratification in the case of the *agriculture* variable, which is not surprising. The same phenomenon was evident for the *manufacturing* variable, although it was less spectacular. For the *income* variable, rural/urban stratification was also initially more satisfactory, but it was not particularly robust (that is, the index deteriorated over time). Rural/urban stratification was preferable for the *unemployed* variable, while there was little difference for *employed*.

Stratification Variables: Option 4, in combination with rural/urban stratification, was clearly superior for the *unemployed* variable. As regards the other variables, option 5 was slightly more satisfactory than the rest for *employed* and *income*.

3.2.2 Study on Contiguity

As previously mentioned, it was decided to retain the concept of contiguous strata for the LFS. Such strata should be better for the production of small area estimates, because of their better geographic representation. In addition, it was felt that contiguous strata would maintain the efficiency of the sample design for a long period of time.

Table 2
Stratification indices for Option

Stratification variables	Total		Rural/Urban	
	1971	1981	1971	1981
	Unemployed			
7 industries	5.4	0.1	9.9	3.8
7 industries + income + employed + unemployed	5.2	2.3	10.2	3.4
7 industries + income + employed + unemployed × 2	7.4	2.3	10.2	5.3
17 variables	6.3	6.4	11.3	4.7
15 variables (excluding employed + unemployed)	3.6	0.1	9.8	9.0
	Employed			
7 industries	2.9	0.5	8.9	4.8
7 industries + income + employed + unemployed	8.8	2.7	8.6	3.2
7 industries + income + employed + unemployed × 2	9.1	2.8	13.1	2.2
17 variables	14.1	7.8	12.2	6.4
15 variables (excluding employed + unemployed)	6.3	1.6	11.4	3.7
	Income			
7 industries	7.4	5.7	18.9	9.5
7 industries + income + employed + unemployed	11.2	6.8	22.1	5.9
7 industries + income+ employed + unemployed × 2	10.3	6.8	28.3	9.5
17 variables	10.5	9.4	24.4	11.9
15 variables (excluding employed + unemployed)	21.0	5.3	28.9	4.5
	Agriculture			
7 industries	7.4	9.7	37.0	26.0
7 industries + income + employed + unemployed	7.6	7.8	40.0	28.7
7 industries + income + employed + unemployed × 2	8.6	7.9	43.2	31.0
17 variables	6.1	1.1	40.3	31.8
15 variables (excluding employed + unemployed)	7.0	0.4	42.7	29.0
	Manufacturing			
7 industries	14.7	8.5	16.9	13.2
7 industries + income + employed + unemployed	10.9	6.6	16.5	12.1
7 industries + income + employed + unemployed × 2	5.5	4.3	14.8	16.1
17 variables	12.5	13.5	13.3	10.7
15 variables (excluding employed + unemployed)	7.2	1.4	14.1	16.4

The next question was how to use the centroids or contiguity vectors, or a combination of the two, to obtain compact, contiguous strata without allowing the geographic constraints to affect minimization of the other variables unduly.

The study was performed with the same 11 economic regions. As anticipated, the use of contiguity vectors alone resulted in strata which were contiguous, but often irregular in shape. At the same time, the use of centroids alone, even with high weights, failed to provide any guarantee of absolute contiguity.

By varying the weight of the centroids relative to the other variables, we found that a combination of a centroid weight of 3 and contiguity vectors offered a good compromise between compactness and non-geographic optimization.

3.3 Design Stratification

In view of these results and the superior results shown by a sample design using rural/urban stratification in a study on cost variance optimization (Choudhry, Lee, Drew 1985), we decided to use separate stratification for all economic regions except those in which either the rural or the urban population was too small to form at least one stratum. It was determined that each stratum should provide a sample of at least 90 dwellings, corresponding to the selection of two PSUs with a minimum take of 45 dwellings each. In cases where this requirement could not be met, we decided to proceed with overall stratification and thus to form mixed strata. This criterion led to the adoption of separate strata in over 2/3 of the ERs.

As regards the stratification variables, we compromised on a stratification based on the 15 variables of option 4 plus *employed*. *Employed* was added because its inclusion in option 4, as compared to option 5, improved the performance of the *employed* and *income* characteristics. For the same reason, *unemployed* was excluded as a stratification variable.

For the geographic constraints, it was decided to use the contiguity vectors in combination with a uniform centroid weighting of 3 in all economic regions.

A decision was also required as to the number of strata per ER. In practice, in most of the cases, there was no choice. According to the sample design, each PSU corresponds to one interviewer assignment, and we wanted to select at least two PSUs per stratum on order to produce unbiased variance estimates. Given these constraints, in almost 2/3 of the cases, only one stratum was formed with 2 or 3 selections, in the urban or rural parts or a combination of the two. In the other cases, stratification was performed in such a way as to permit the selection, again, of 2 or 3 PSUs per stratum. This decision was based on another study showing slight reductions in variance with this approach, as compared to the old sample design in which 4 to 6 PSUs were selected from each stratum (Choudhry, Lee, Drew 1985).

3.4 Study on Robustness of Contiguous and Discontiguous Strata

Robust strata are strata that maintain the efficiency of the sample design over time. Following redesign, a study was performed to determine whether contiguous strata would be more robust over time, as had been hypothesized.

The study dealt with three economic regions in Ontario, ERs 520, 540 and 580 (1981 numbering). For each of these regions, the results of the new stratification (selected for the redesign of the LFS), which consists of contiguous strata, were compared with a stratification without contiguity constraints. The strata were defined on the basis of the 1981 data, and evaluated on the basis of the 1971 data. For the contiguous strata, we used contiguity vectors with centroids, while for the discontiguous strata, we tested two options using centroid weights of 0 and 3 respectively. The stratification variables used were the same 16 variables described above (modified option 4).

The results are shown in table 3. We see that in general, the total index calculated on stratification is higher for the two options in which contiguity is not necessary, as might be expected (1981 column). However, these two options also give higher indices over time (1971 column).

Do we really need contiguous strata? Before answering this question, we would have to perform a more in depth study involving ERs from a number of provinces. Evaluation of stratification robustness would however pose certain problems. It is easy to evaluate robustness in Ontario, since stratification there is performed at the census subdivision level, which has changed very little since 1971. When stratification is performed at the level of the enumeration areas, which are very changeable, it is extremely difficult to obtain precise figures on robustness when the strata are neither compact nor contiguous.

However, should it prove that stratification without contiguity is more satisfactory, this could compensate for the possible problems involved in production of small area estimates. It could also open new horizons: once contiguity constraints are eliminated, why could we not begin by forming compact, but not necessarily contiguous, PSUs, and then grouping them into strata? This question also could only be answered by further and more detailed studies.

3.5 Formation of PSUs

The clustering algorithm was modified to permit PSU delineation in rural and mixed strata. In the rural strata in particular, the formation of the PSUs is conceptually very similar to stratification. The only difference relates to the fact that in stratification, we attempt to minimize the sums of squares of the geographic and non-geographic variables within each stratum, while in PSU formation, we want to minimize the sums of squares of the geographic variables (to obtain compact PSUs in order to reduce costs) and to maximize those of the non-geographic variables. The latter criterion enables us to obtain PSUs which are as heterogeneous as possible in terms of characteristics, so that they are all properly representative of the stratum during sampling.

There is, however, a conflict between the desired compactness of the PSUs and their heterogeneity, because of the tendency of adjacent units to possess similar characteristics. Because of low computer costs, we performed 3 delineations per stratum with centroid weights of 10, 15 and 20, relative to the other variables. The results of each delineation were then plotted on a graph whose axes are the centroids (see Figure 1). We then selected the best of the 3 delineations on the basis of the quality of variable optimization, as reflected by the stratification indices, and through reference to the graphs. A compactness index was also taken into consideration. In most cases, as it worked out, a centroid weight of 10 or 15 was selected.

Table 3

Stratification Indices by Geographic Constraints

Economic Region	No. of Strata	Geographic Constraints					
		Contiguity and Centroids (weight of 3)		Centroids (weight of 3)		None	
		1981	1971	1981	1971	1981	1971
520	2	32.2	28.5	30.2	30.1	34.5	27.0
540	3	21.8	14.1	24.9	17.8	35.2	26.8
580	4	22.8	18.9	41.4	33.7	43.7	38.5

Formation of the PSUs in the mixed strata led to an additional constraint. We wanted the proportion of urban population to be approximately the same in each PSU. Since we also wanted the PSUs to have approximately equal total populations, it was therefore necessary in some cases to split the large urban centres among a number of PSUs. The following solution was adopted:

1. The average number of parts of urban centres which a PSU will receive (N) is determined. This number depends on the proportion of the urban population in the stratum and on the number of urban units. In practice, it was set at 1 or 2. Certain strata without sufficient population or a sufficient number of urban units were reclassified as entirely rural strata.
2. The number of parts into which each urban centre will be divided is determined. The total number of parts must equal N times the number of PSUs and each urban centre is divided into a number of parts proportional to its population.

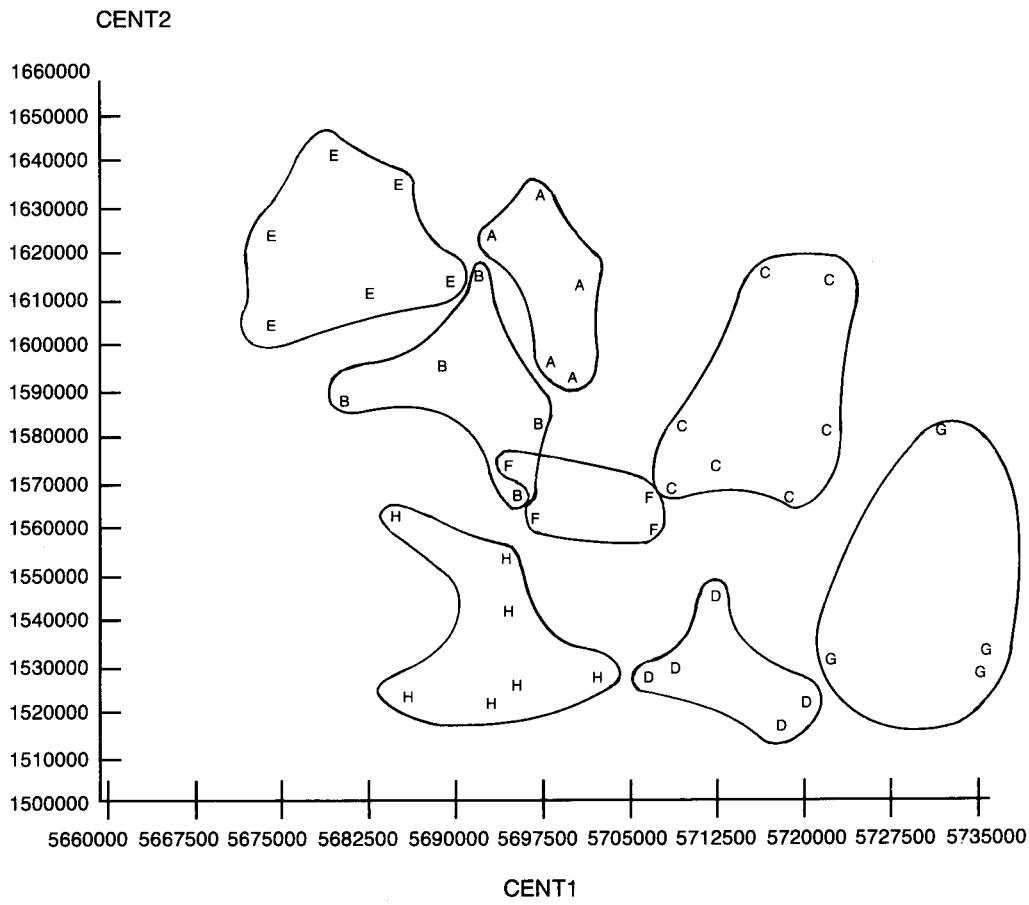

Figure 1. Example of PSU Delimitation. Each stratification unit is represented by a letter identifying the PSU to which it belongs. The PSUs are circled for clearer differentiation.

3. The optimal stratification program is applied, considering each part of an urban centre as a distinct stratification unit and adding the *urban population* variable to the other stratification variables. The weight assigned to this variable is adjusted to obtain the most evenly balanced rural/urban distribution possible within each PSU, without unduly disrupting compactness and overall optimization. This can be done only by trial and error. In practice, we found that a weight of 10 or 15 on urban population, relative to the other variables, produced satisfactory results.

In the urban strata, the PSUs were composed of urban centres. In some cases, small centres, relatively close together, were combined, without considering characteristic optimality.

Table 4 gives the average delineation indices for the PSUs in rural, mixed and urban strata. For the non-geographic variables, the lowest index represents the best delineation, while the opposite is true for the centroids. The results are clearly better, in terms of characteristic optimality, for the rural and mixed strata, in which the clustering algorithm was used. The high indices of the centroids show that the PSUs are relatively compact.

Table 4

Average PSU Delineation Indices

Variables	Type of stratum		
	Rural	Mixed	Urban
Agriculture	8.1	8.3	9.0
Forestry	21.8	24.5	35.9
Mines	20.6	36.0	57.0
Manufacturing	15.1	22.9	53.3
Construction	9.0	11.4	22.7
Transportation	9.9	12.8	22.7
Services	9.4	12.8	29.1
Employed	7.7	10.2	23.6
Unemployed[a]	13.6	14.2	18.6
Income	8.9	11.2	23.7
Population 15-24	9.4	13.4	29.8
Population 55+	7.4	13.9	34.5
1-person households	5.1	7.4	13.0
2-person households	7.9	11.9	28.1
Owned dwellings	6.8	12.5	29.4
Total gross rent	5.1	7.7	14.4
Secondary education	9.1	10.5	17.4
Total population[a]	3.2	4.0	10.5
Dwellings[a]	5.9	8.9	18.6
Centroid 1	91.6	92.7	99.2
Centroid 2	90.5	91.7	97.2

[a] Not used as a variable in optimization.

4. STRATIFICATION IN SELF-REPRESENTING UNITS

4.1 Old Design

The self-representing units of the old sample design corresponded to those cities large enough to yield an expected take equivalent to one interviewer assignment. The lower limit for SRUs varied from 10,000 persons in the Atlantic provinces to 29,000 in Quebec and Ontario.

The large SRUs were geographically stratified by grouping 3 to 5 contiguous census tracts (CTs), without any attempt to optimality. CTs are geostatistical units with populations between 3,000 and 5,000; because of their stability from one census to the next, they are practical operational units. It was felt that these strata would be efficient in estimating characteristics, and that their small size (between 10,000 and 15,000 persons) would permit sample updating in areas experiencing rapid growth, without disrupting the rest of the sample.

In addition to the area frame, an open-ended frame was set-up for apartment buildings in the large cities.

4.2 Study on stratification

Three large SRUs were considered in this study, namely Quebec City, Ottawa and Toronto. The stratification unit selected was the census tract. Because of operational constraints imposed by the stratification program, it was necessary to break Toronto up into six parts, corresponding generally to the city's major natural divisions. Stratification was carried out separately in each of these parts. The same 16 stratification variables finally selected in the NSR part were used.

Two main options were evaluated:

Option 1: Two-level stratification:
- contiguous, compact primary strata, with a centroid weighting of 3 and an expected take of approximately 150 dwellings.
- secondary strata - 4 or 5 per primary stratum, formulated without geographical constraints.

Option 2: compact stratification formulated with the use of centroids (weight of 3) and without contiguity vectors, comparable in size to the secondary strata of option 1.

Table 5 shows the results of the comparison between the old stratification and the two options studied. As in the NSR part, the strata were defined on the basis of 1971 Census data, and then evaluated on the basis of 1981 data.

We see that the two options studied consistently show better indices than the old stratification, with the possible exception of the first three variables, which, in any case, are of limited importance in cities. The old stratification nevertheless performed quite well, considering that it was carried out without any concern for optimality.

We also note that all three methods provide generally robust stratification over time, as reflected by the comparison between the indices for 1981 and 1971. Major exceptions to this rule, unfortunately, appear to be the employed and unemployed characteristics.

4.3 New Design

Given the similarity in results between the two options studied, it was decided to adopt two-level stratification (option 1) in large cities where the sample consists of 300 or more households, for the following reasons:

i) Contiguity in the primary strata gives us a suitable unit for sample updating.

ii) The primary strata can be used for the formation of interviewer assignments. The size of the strata was determined so that the sample within the geographic area, that is, the area frame sample plus the sample for the apartment frame, corresponds to two interviewer assignments (160 households in the city core and 120 elsewhere).

iii) Two-level stratification leads to better representation of the correlated response variance in variance estimates. In the old sample design, there was usually only one interviewer per stratum, resulting in an underestimate of this component of the variance. With non-geographic secondary strata, but geographic interviewer assignments, this problem will be less frequent.

The cost constraints associated with the computer time involved forced us to deal with certain SRUs on an individual basis. In fact, the Montreal region was divided into seven independent parts, during stratification. The same was done with Toronto (5 parts), Winnipeg (2 parts), Calgary (2 parts), Edmonton (2 parts) and Vancouver (3 parts). These divisions were made on the basis of natural criteria as suggested by the geography of these regions.

In large SRUs, apartment buildings existing at the time the sample design was developed were sorted by the primary strata in which they were physically located in order to achieve an implicit stratification of this sample.

Table 5

Comparison of Three Stratification Methods (SRUs)

Variables	Old Design		Two-level Stratification (Option 1)		Compact Stratification (Option 2)	
	1971	1981	1971	1981	1971	1981
Agriculture	5.5	2.9	3.2	1.8	3.4	1.8
Forestry	2.2	2.3	2.1	1.7	2.2	2.3
Mines	7.6	4.9	8.5	4.1	7.6	4.0
Manufacturing	34.7	35.0	36.6	34.1	39.1	35.0
Construction	32.5	29.6	39.7	30.1	42.4	33.4
Transportation	9.2	6.8	18.0	11.6	20.0	11.6
Services	29.5	27.5	45.8	33.1	46.7	32.1
Employed	15.1	8.0	31.4	14.1	32.8	12.6
Unemployed[a]	14.6	5.7	14.9	6.7	15.5	7.1
Income	39.4	38.6	51.8	29.8	53.6	48.0
Population 15-24	9.6	15.2	12.5	17.5	13.3	14.9
Population 55+	27.9	18.3	34.0	20.8	32.6	18.5
1-person households	20.3	19.2	36.3	33.8	37.8	35.0
2-person households	21.9	20.3	40.3	30.9	40.1	30.2
Owned dwellings	20.3	15.3	29.7	22.9	32.1	24.9
Secondary education	32.6	42.4	50.3	47.9	51.6	49.1
Population 15+[a]	27.0	8.2	38.0	13.4	37.6	12.0
Dwellings[a]	21.8	18.5	41.7	33.8	42.1	34.3

[a]Not used as a stratification variable.

In medium-sized SRUs, where the sample was not large enough to justify two-level stratification, optimal strata were simply constructed by means of the stratification program, without the application of geographic constraints.

The smallest SRUs, those not broken into block faces for census purposes, were manually stratified, without any attempt at optimality.

Finally, we might note that the phase-in period of the new sample produced a further constraint. For large SRUs, core areas were defined as consisting of complete old-design strata that were unaffected by boundary changes. By having strata in the new design respect these core areas, we ensured that during phase-in, the new sample in core areas represented the same geographic area as the old, which permitted gradual replacement of the old sample by the new without the need for a costly parallel build up of new sample (Mayda, Drew, Lindeyer 1985).

5. CONCLUSIONS

Use of multivariate clustering algorithm enabled us to develop a very general stratification, thus strenghtening the LFS in its role as a general household survey. In addition, automation of the various stages of stratification in the NSR and SR parts, and delineation of the PSUs in the NSRUs, led to a significant reduction in the cost and time required to redesign the sample.

The system is documented (Foy 1984) and can be used for the stratification of other surveys. It may also be used in situations requiring the definition of statistical or administrative regions, using a full range of variables.

For the LFS, one aspect requiring further research relates to the selection of contiguous or discontiguous strata, and the implications of discontiguous strata on sample design.

ACKNOWLEDGEMENTS

The authors would like to thank Sylvie Trudel and Marc Joncas for their assistance in carrying out the studies mentioned in this report, and the members of the LFS Sample Redesign Committee for their valuable suggestions. They are also grateful to the referee for his helpful comments.

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.

DAHMSTRÖM, P., and HAGNELL, M. (1978). The formation of strata using cluster analysis. Internal document, Department of Statistics, University of Lund, Sweden.

FOY, P. (1984). Stratification program for the Canadian Labour Force Survey: User's guide. Internal document, Census and Household Survey Methods Division, Statistics Canada.

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

JUDKINS, D.R., and SINGH, R.P. (1981). Using clustering algorithms to stratify primary sampling units. *American Statistical Association Proceeding of the Section on Survey Research Methods*, 274-284.

KOSTANICH, D., JUDKINS, D.R., SINGH, P.R., and SCHANTZ, M. (1981). Modification of Friedman-Rubin's clustering algorithm, for use in stratified PPS sampling. *American Statistical Association Proceedings of the Section on Survey Research Methods*, 285-290.

MAYDA, F., DREW, J.D., and LINDEYER, J. (1985). Phase-in of the redesigned Labour Force Survey. Internal document, 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.

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.