SURVEY METHODOLOGY

TECHNIQUES D'ENQUÊTE

June - 1980 - Juin

VOLUME 6

NUMBER 1 - NUMÉRO 1

A Journal produced by
Methodology Staff
Statistics Canada

Préparé par les
méthodologistes de
Statistique Canada

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1980 Vol. 6 No. 1

A Journal produced by Methodology Staff, Statistics Canada

Preparé par les méthodologistes de Statisticque Canada

C O N T E N T S

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1980 Vol. 6 No. 1

A Journal produced by Methodology Staff, Statistics Canada

Preparé par les méthodologistes de Statistique Canada

Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed; however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department. Copies of papers in either Official Language will be made available upon request.

Politique de la rédaction:

La revue Techniques d'enquête veut donner aux personnes qu'intéressent les aspects pratiques de la conduite d'enquêtes, la possibilité de publier sur ce sujet dans un cadre candadien. Les textes pourront porter sur toutes les phases de l'élaboration de méthodes d'enquête: les problèmes de conception causé par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada. On pourra se procurer sur demande des exemplaires d'un article dans l'une or l'autre langue officielle.

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1980 Vol. 6 No. 1

A Journal produced by Methodology Staff, Statistics Canada.

Preparé par les méthodologistes de Statistique Canada.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 6th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested. Authors of articles for this journal are free to have their articles published in other statistical journals.

Présentation de documents pour publication:

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraitre un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes ménages et du recensement, Statistique Canada, 6^{e} étage, Edifice Jean Talon, Parc Tunney, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi. Les auteurs des articles publiés dans cette revue sont libre de les faire paraitre dans d'autres revues statistiques.

COMPARISON OF SOME RATIO TYPE ESTIMATORS FOR LARGE SCALE HOUSEHOLD SURVEYS[1]

M. Lawes and M.P. Singh[2]

In this paper three types of ratio estimators, namely combined, post-stratified and a generalized ratio estimator developed earlier by Singh (1969) and Naga Reddy (1974), are considered. Based on an empirical evaluation, their efficiencies are compared for two large scale household surveys, namely the Canadian Labour Force Survey and the Survey of Consumer Finances.

1. INTRODUCTION

Stratified multistage sample design is usually adopted for conducting large scale household surveys due to its operational convenience. Data available on auxiliary variables are utilized at various stages of sampling operations such as formation of strata and sampling units, determination of overall sample size and its allocation to different strata and stages of sampling, sample selection using unequal probabilities, etc. Auxiliary data needed for performing these sampling operations are usually at lower levels for example for each sampling unit or even for each element (i.e. blocks or enumeration areas) that comprise sampling units. Quite often, additional data are available from independent sources but at higher levels of aggregations (i.e. by census divisions or by provinces) and by different classifications

---

1 This paper was presented at the meeting of the Statistical Society of Canada, Montreal, May 27-30, 1980.

2 M. Lawes and M.P. Singh, Census and Household Survey Methods Division, Statistics Canada.

(i.e. population by age, sex, occupation, household, family type, marital status, etc.). Such data may be used efficiently at the estimation stage for constructing ratio or regression estimators.

In this paper, three types of ratio estimators are considered and their efficiencies are compared for two large scale household surveys, namely, the Canadian Labour Force and the Survey of Consumer Finances. The estimators examined are: a combined ratio estimator using data at province level, post-stratified ratio estimator using data by different sub-classes at province level and a generalized ratio estimator developed by Singh [1969] and Reddy [1974]. These estimators, along with simple survey estimator, are discussed in section 3 after giving a brief description of the two surveys in the following section. Using Keyfitz [1957] method for variance estimation, the expressions for the variances are given in section 4 under the usual approximation of ratio estimation. The efficiency comparison and analysis of results are then presented in the last section.

2. BRIEF DESCRIPTION OF THE SURVEYS

The estimation procedures have been applied to and evaluated for two surveys conducted by Statistics Canada, namely, the Canadian Labour Force Survey and the Survey of Consumer Finances. A detailed description of the Labour Force Survey is presented in Methodology of the Canadian Labour Force Survey (1977). Although the samples are selected from a common sampling frame, the type of data collected varies between the two surveys. For the Labour Force Survey described in sub-section 2.1, the majority of the data items are qualitative in nature dealing with the labour force activities of the repsondent during reference week whereas for the Survey of Consumer Finances discussed in sub-section 2.2, the data items are primarily quantitative in nature consisting of income amounts by source of income for the previous calendar year.

2.1 The Labour Force Survey

The Labour Force Survey is conducted on a monthly basis to collect information on the work activities and employment status of Canadians. The survey covers the civilian non-institutional population in the 10 provinces of Canada excluding residents of Indian Reservations and members of the Canadian Armed Forces. Each month a sample of about 62,000 dwellings is selected based on a multistage stratified probability sampling procedure. Information on the Labour Force activities of all eligible household members 15 years of age or over is collected by interviewers by either a personal or telephone interview. Within geographically contiguous economic regions the sample is selected independently within the following three types of areas:

a) Self-representing areas are comprised of the larger cities. Each city is divided into sub-units whose size varies from 1,000 to 12,000 households depending on the sampling ratio for the province. Sub-units so formed are subdivided into clusters which are usually city blocks. The clusters are randomly grouped into multiples of six groups, to facilitate the implementation of six-month household rotation scheme.

   One cluster per group is selected with probability proportional to size and a systematic sample of dwellings is selected from within selected clusters.

b) Non self-representing areas consist of rural and small urban areas. These areas are initially divided into strata containing on the average about 15 primary sampling units (PSUs). Two or more PSUs per stratum are selected with probability proportional to size. At the subsequent stages, subsampling is carried out independently within urban and rural parts of the selected PSUs.

c) Special areas are composed of areas or establishments which possess characteristics differing from the general population and which may require special interviewing techniques. Included in the special areas are hospitals, schools, hotels, military establishments and remote areas. Special areas make up about 1% of the sample frame.

2.2 The Survey of Consumer Finances

The Survey of Consumer Finances is conducted on an annual basis in the ten provinces of Canada with the exclusion of inmates of institutions and residents of Indian Reservations. There are two variants to the survey. In even numbered years the survey is conducted as a supplement to the April Labour Force Survey. A subsample of about 41,000 dwellings is selected from the Labour Force Survey sample for that month. The subsampling is generally carried out by selecting all sampled households in a subset of the selected clusters. The subject matter content for the supplement is restricted to basic questions on income and work experience during the previous calendar year. In odd numbered years an independent sample of about 17,000 dwellings is selected from the Labour Force Survey sampling frame following the basic sampling procedure described for the Labour Force Survey. In addition to information collected in the even numbered years, data on income related topics are collected in odd numbered years.

3. ESTIMATION PROCEDURE

Although the estimation procedures used for the two surveys differ in terms of minor details, the basic approaches are similar. Survey weights attached to respondent records are the product of six factors, the product of the first five being termed the subweight as detailed below:

a) the basic weight - In a probability sample, the sample design itself determines the weights which may be used to produce unbiased estimates. For both these surveys, using self-weighting design, the weights are the same within each type of area in a province and are equal to the inverse of the sampling ratio.

b) the rural-urban factor - This factor is relevant for non self representing areas, and adjusts sample distributions between rural-urban areas of selected PSU's within provinces to agree with distributions obtainable from census data.

c) balancing factor for non-response - The factors are calculated by the type of area (rural-urban) within primary sampling units and within subunits for self-representing areas.

d) the cluster weight - Within sampled clusters experiencing significant growth between the time of design and interviewing, subsampling of the cluster is carried out to avoid disruptions in the field operations. If subsampling has been carried out within a cluster, the cluster weight which is the inverse of the cluster subsampling rate, is applied to selected households in the cluster.

e) the sample size stabilization weight - To prevent uncontrolled growth in the sample size of the Labour Force Survey, a sample stabilization procedure is introduced into the survey operations under which a number of dwellings that are in excess of the predetermined level are dropped and a stabilization weight is used to compensate for the reduction in the total sample size. For the Survey of Consumer Finances, the sample stabilization capabilities have been used to effect a subsampling of households within selected clusters. The product of the above five factors is termed the subweight.

Mathematically, the subweight may be defined as:

$$Wphijc = Wph\ Fpj\ Bphij\ Cphic\ Spj$$

where

p	denotes province
h	denotes stratum
i	denotes primary sampling unit
j	denotes type of area (SRU, special, urban or rural)
c	denotes cluster
Wph	is the basic weight
Fpj	is the rural-urban factor
Bphij	is the non-response balancing factor
Cphic	is the cluster weight
Spj	is the sample size stabilization weight.

Simple survey estimates of aggregate totals can then be expressed as

$$\hat{X}_p^{(1)} = \sum_{hijc} \sum_{k\varepsilon(hijc)} Wphijc\ Xphijck \qquad (3.1)$$

where

$\sum_{k\varepsilon(hijc)}$ denotes the summation over all records corresponding to respondent sample units in the area identified by (hijc).

and Xphijck denotes the value for record k in (phijc).
For qualitative characteristics, Xphijk is an indicator variable with a value 1 if the record possesses the characteristic under question and zero otherwise. For quantitative characteristics, Xphijk is the value as reported on the record.

Simple survey estimates can be derived on the basis of these subweights. For qualitative characteristics, simple survey estimates of the total number of units possessing a characteristic are derived as the sum of the subweights on all records possessing the characteristic. For quantitative characteristics, the sum of the product of the subweights and reported values yields a simple survey estimate of the characteristic total.

As is well known, the efficiency of the final estimator compared to the simple survey estimator can be increased by utilizing additional information on a related characteristic in the form of a ratio estimation. We define below the three types of ratio estimators that are compared in this article. All of these use data on inter-censal population estimates available at the province level.

Let $\hat{P}_p$ and $\hat{X}_p$ denote simple survey estimates of the population and characteristic totals respectively at the provincial level. Note that $\hat{X}_p$ is the same as $\hat{X}_p^{(1)}$ defined in (3.1), the superscript has been dropped for the sake of convenience in defining the preceeding estimators, and $\hat{P}_p$ is determined by taking the simple sum of weights in (3.1), i.e. assuming x's to be 1. The auxiliary information, P_p the total population count for the province, is obtained from an external source independent of the survey data. The first type of ratio estimator considered is a <u>combined ratio estimator</u>.

$$\hat{X}_p^{(2)} = \frac{\hat{X}_p}{\hat{P}_p} P_p . \qquad (3.2)$$

The post-stratified ratio estimator utilizes auxiliary information available at sub-class or post-stratum levels within the province. Let the subscripts {a} denote a collection of post-strata. The auxiliary information, namely population totals for each of the post-strata, denoted by P_{pa}, are available from outside sources. With $\hat{P}_{pa}$

and $\hat{X}_{pa}$ denoting the simple survey estimates of population and characteristic totals respectively, the post-stratified ratio estimator can be expressed as

$$\hat{X}_p^{(3)} = \sum_a \frac{\hat{X}_{pa}}{\hat{P}_{pa}} P_{pa} . \qquad (3.3)$$

For both the Labour Force Survey and the Survey of Consumer Finances, the official survey estimates are based on a post-stratified ratio estimation procedure. This gives rise to the sixth factor in the weighting of the survey data which, when multiplied by the subweight, yields the final weight for survey records. The post-stratum adjustment factor is the ratio of the external population estimate for the stratum divided by the simple survey estimated population total for the post-stratum, i.e. referring to (3.3), for post-stratum "a" the post-stratum adjustment factor is $P_{pa}/\hat{P}_{pa}$. For the Labour Force Survey, the post-strata are defined on the basis of age by sex groupings while for the Survey of Consumer Finances, the post-strata are defined as labour force status by class of worker groupings within each province. The outside estimates for the Survey of Consumer Finances are derived from Census and Labour Force Survey sources.

The third type of estimator is a generalized ratio estimator. Like the combined ratio estimator, this estimator utilizes information only at the provincial level, with an additional factor which takes account of the reliability of the simple survey estimates and the correlation between estimates of the population and the characteristic totals. As before, let $\hat{P}_p$ and $\hat{X}_p$ denote simple estimates of the total population and characteristic total respectively at the provincial level, and the total population count P_p is available from outside sources. The generalized ratio estimator is then defined as

$$\hat{X}_p^{(4)} = \frac{\hat{X}_p \, P_p}{\alpha \hat{P}_p + (1-\alpha) P_p} \tag{3.4}$$

where α is a suitably chosen constant. The optimum value of α, obtained by minimizing the mean square error of $\hat{X}_p^{(4)}$, is given by

$$\alpha_o = \frac{CV(\hat{X}_p)}{CV(\hat{P}_p)} \rho(\hat{X}_p, \hat{P}_p) \tag{3.5}$$

where $CV(\hat{X}_p)$ is the coefficient of variation of $\hat{X}_p$
$CV(\hat{P}_p)$ is the coefficient of variation of $\hat{P}_p$
and
$\rho(\hat{X}_p, \hat{P}_p)$ is the correlation coefficient between $\hat{X}_p$ and $\hat{P}_p$.

It may be noted that unlike the combined and post-stratified ratio estimators, this generalized ratio estimator (using α_o) is unbiased to the first degree of approximation and has the same mean square error as that of the regression estimator (see Singh [1969] and Naga Reddy [1974]). As this estimator has been found to be highly efficient in small empirical investigations, it is included in this empirical investigation to study if the gain in efficiency is substantial in the context of large scale surveys.

4. VARIANCE ESTIMATION

For both the surveys considered here, the methodology for estimating the sampling variability is modelled after the Keyfitz method. To apply this method of variance estimation, a set of pseudo-replicates within each stratum is required. For the purposes of variance

estimation in the Labour Force Survey and the Survey of Consumer Finances in NSRU areas the n_h selected PSU's referred to as 'components' from stratum h are assumed to have been selected independently. In SRU areas, each subunit is split into two components for variance estimation purposes; this split is accomplished by dividing the selected clusters between the two components. In special areas, pseudo-strata are defined either as the design strata or as collapsed designed strata. Each pseudo-stratum is then sub-divided into from two to four components, the number depending on the sample sizes within each pseudo-stratum. We shall follow the notation presented in section 3 with the modification that the h denotes the strata or pseudo-strata and i denotes the components for variance estimation. The variance estimates for the estimators ($X^{(2)}$, $X^{(3)}$ and $X^{(4)}$) are based on the usual approximation to the variance of a ratio.

a) Variance estimate of the simple survey estimator:

$$\hat{Var}(\hat{X}_p) = \sum_{h\varepsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\hat{X}_{phi} - 1/n_h \sum_{i=1}^{n_h} \hat{X}_{phi})^2 \qquad (4.1)$$

where n_h is the number of components in stratum h.

b) Variance estimate of the combined ratio estimator:

Let
$$\hat{D}_{phi} = \hat{X}_{phi} - \frac{\hat{X}_p}{\hat{P}_p} \hat{P}_{phi}. \qquad (4.2)$$

Then

$$\hat{Var}(\hat{X}_p^{(2)}) = (\frac{P_p}{\hat{P}_p})^2 \sum_{h\varepsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} [\hat{D}_{phi} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{D}_{phi}]^2. \qquad (4.3)$$

c) Variance estimate of the post-stratified ratio estimator:

Let

$$\hat{D}_{phia} = \hat{X}_{phia} - \frac{\hat{X}_{pa}}{\hat{P}_{pa}} \hat{P}_{phia} \cdot \tag{4.4}$$

Then

$$\hat{Var}(\hat{X}_p^{(3)}) = \sum_{h\epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left[\sum_a \frac{P_{pa}}{\hat{P}_{pa}} \hat{D}_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_a \frac{P_{pa}}{\hat{P}_{pa}} \hat{D}_{phia}\right]^2 . \tag{4.5}$$

d) Variance estimate of the generalized ratio estimator:

Let

$$\hat{D}^*_{phi} = \hat{X}_{phi} - \frac{\hat{X}_p}{\alpha \hat{P}_p + (1-\alpha) P_p} \hat{P}_{phi} \tag{4.6}$$

Then

$$\hat{Var}(\hat{X}_p^{(4)}) = \left[\frac{\hat{P}_p}{\alpha \hat{P}_p + (1-\alpha) P_p}\right]^2 \sum_{h\epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \{D^*_{phi} - \frac{1}{n_h} \sum_{i=1}^{n_h} D^*_{phi}\}^2 \tag{4.7}$$

It should be noted that although α is estimated from the sample, its effect on the sampling variability is neglected. A brief development of the variance expressions for the ratio estimators is presented in Appendix A.

5. EFFICIENCY COMPARISON

5.1 Formulation

The information supplied by an estimator is usually measured by the inverse of its variance (or its mean square error). Thus, the efficiency of an estimator t_2 over another estimator t_1 is generally defined as the ratio of the inverse of their variances, that is $E_{12} = V(t_1)/V(t_2)$. In actual practice, however, the true variances V(t) will not be available. One may compute the efficiency E.. by taking the ratio of the estimated variances v(t) or alternatively by taking the ratio of the estimated relative variances rv(t)= v(t)/t. The latter approach may be more suitable in situations where the estimates t. themselves vary significantly from one estimation method to another. We have chosen the latter approach for comparison of efficiencies. Thus, E_{ij} is the estimated efficiency of estimator t_j over another estimator t_i and is defined as the ratio of the inverse of their relative variances. In particular,

$$E_{12} = \frac{rv(\hat{X}^{(1)})}{rv(\hat{X}^{(2)})}$$

is the efficiency of the combined ratio estimator over the simple survey estimator.

The efficiency E_{12} may be expressed in terms of the correlation coefficient between $\hat{X}_p$ and $\hat{P}_p$, and is presented along with estimates for the Labour Force Survey in Appendix B.

$$E_{23} = \frac{rv(\hat{X}^{(2)})}{rv(\hat{X}^{(3)})}$$

is the efficiency of the post-stratified ratio estimator over the combined ratio estimator.

E_{23} may be thought of as a post-stratification index as it provides a measure of the gain in efficiency as a result of post-stratification. The overall efficiency of the post-stratified ratio estimator over the simple estimator is:

$$E_{13} = \frac{rv(\hat{X}^{(1)})}{rv(\hat{X}^{(3)})} .$$

Finally,

$$E_{24} = \frac{rv(\hat{X}^{(2)})}{rv(\hat{X}^{(4)})}$$

is the efficiency of the generalized ratio estimator over the usual combined ratio estimator.

It should be noted that $E_{13} = E_{12}\ E_{23}$ and $E_{14} = E_{12}\ E_{14}$ so that the overall efficiencies of the various ratio-type estimators relative to the simple survey estimator may be viewed as contributions by successive changes in the application of ratio adjustment procedures.

6. OBSERVATIONS

The following observations can be made, based on the results of Tables 1A, 1B and 1C.

1) The magnitude of gain in efficiency varies from characteristic to characteristic, depending mainly upon the degree of correlation between the characteristic and the total population. Maximum gain is achieved at the national level for Employed (having correlation about 0.92) followed by Not In Labour Force (having correlation about 0.67) and Unemployed which has approximate correlation 0.40. At the provincial level, the magnitude of the efficiencies differs from province to province but again a similar trend is noticeable for various characteristics examined.

2) Each of the ratio type estimators considered here is generally more efficient than the simple survey estimator.

3) The combined ratio estimator which is the simplest to compute achieves most of the efficiency gains for all the characteristics. For example, in case of Employed, the combined estimator is on an average five times more efficient than the simple estimator, (ranging from 3.65 to 6.66 times) whereas the additional gain in using post-stratified estimator is only about 50%. For Unemployed, however, the gain in using the combined estimator over the simple estimator is only about 25% and the additional gain in using either of the other two estimators is negligible.

4) The generalized ratio estimator is always more efficient than the combined ratio estimator. Although more than 30% gains are achieved for certain provinces for the characteristics Employed and Not in Labour Force, the gain at national level is only about 12%.

5) Post-stratified estimator, which is presently used in the Labour Force Survey, is slightly more efficient than the other two ratio type estimators. It may, however, be noted that this estimator uses more detailed data, i.e. data at the post-strata (age-sex) level whereas the other two ratio estimators use supplementary data only at the aggregate level.

6) In respect of estimates of Unemployed in Saskatchewan (Table 1B) both the combined and post-stratified estimators lead to a loss in efficiency as compared to the simple survey estimator, whereas the generalized ratio estimator even in this exceptional case remains slightly more efficient. This indicates that in situations where the unemployment level is exceptionally low the use of post-stratified estimator could lead to substantial loss in efficiency.

7) In addition to basing the above conclusions on a larger set of data, the primary reason for presenting three years data in Tables 1 (A,B,C) is to study the performance of these estimators over a period of time. It should be noted that the redesigned sample using 1971 Census and other more recent data became the official source of labour force estimates starting from January 1976. The results in Table 1A thus relate to the time when the sample was most up-to-date.

Comparing the values of E_{12} in Table 1A, based on 1976 with the corresponding values in Table 1B based on 1978, it is noticed that the gain in the relative efficiency of ratio estimator as compared to the simple estimator increased in 1978 for all provinces with the exception of Prince Edward Island. For the provinces of Alberta and Saskatchewan, this gain is two to three times the gain in efficiency achieved in 1976 and as a result there is a significant gain at the national level as well. It may thus be concluded that as we move away from the redesign time and the sample becomes more and more out-of-date, the simple survey estimator becomes less and less efficient and consequently the relative efficiency of the ratio estimators shows a gradual increase.

Further, it is noticed from Table 1C, based on 1979, that certain provinces show a decrease in the efficiency gain as compared to the gain in 1978 (Table 1B), and particularly so for the provinces of Alberta and Saskatchewan. This can be explained by the fact that a program of updating the sample was put in place during 1978/79 under which the samples from cities having larger growth are regularly updated and this program had maximum impact on the two provinces mentioned above.

It may, however, be noted that the additional gains of the other two ratio estimators remained relatively constant over time.

8) Until recently, only 18 age-sex groups (post-strata) were used in the LFS and therefore in Tables 1A, B, C, these groups were used for the empirical comparisons. Starting from 1979, there are now 40 age-sex groups used for the construction of post-stratified estimates. This increase was implemented primarily for the purpose of providing estimates for larger numbers or combinations of age groups. In table 2, a comparison is made between the two sets of post-strata. It is observed that the increase in the number of strata has no significant effect on the reliability of estimates. This, however, suggests that an optimum set of post-strata could be determined by repeating the study for various sets provided that corresponding population projections are available.

As noted in section 2, the Survey of Consumer Finances conducted in 1978 was a supplement to the LFS based on a sample of about 41,000 dwellings whereas the 1977 survey used a sample of about 17,000 dwellings selected independently from the LFS sampling frame. The estimated efficiencies for both survey years and especially for the 1977 survey, are subjected to larger sampling variability than the corresponding efficiencies as estimated for the Labour Force Survey data (particularly since the Labour Force Survey estimated efficiencies were averages over several monthly surveys). Despite the instability of the estimated relative efficiencies, the following patterns and relationships can be observed from Table 3 based on the 1977 and 1978 Surveys of Consumer Finances.

9) As in the case of labour force characteristics, the ratio estimators in this case as well, are more efficient than the simple survey estimator. The estimated gains in efficiency vary substantially from province to province. As well for some provinces there are large differences in the relative efficiencies between the two surveys.

10) Most of the gains in efficiency arise for the combined ratio estimator. The additional refinements due to the generalized ratio estimator and the post-stratified ratio estimator contribute relatively small gains in the efficiency.

11) The generalized ratio estimator is at least as efficient as the combined ratio estimator (See E 24) although the gain in efficiency over the combined ratio estimator varies considerably from province to province with observed gains at the national level of 6% and 11% for the 1977 and 1978 surveys respectively.

12) Although, at the national level, post-stratified ratio estimator results in a gain in efficiency as compared to the combined estimator (See E 23) (by 21% and 5% for 1977 and 1978 respectively), at the provincial level slight loss in relative efficiency is noticed for some province.

The post-stratified ratio estimator, however, ensures agreement of population totals between the survey and the control population totals derived from external sources (usually census data) within post-strata. This secondary benefit of the post-stratified ratio estimator is judged as being useful and important as it ensures a degree of consistency between the survey estimates and the external population totals down to the post-stratum level.

In summary, for both the Labour Force Survey and the Survey of Consumer Finances, considerable gains in efficiency, over the simple design based estimator, are realized by employing ratio-type estimation procedures. The combined ratio estimator provides most of the gains observed. The generalized ratio estimator generally provides minimal additional gains in efficiency over the combined ratio estimator, although for some provinces and for some characteristics, the gains

approach 30-40%. The post-stratification in the majority of situations, provides marginal improvements in the efficiency of the post-stratified ratio estimator over the combined ratio estimator. For the Labour Force Survey, the comparison of efficiencies of the estimators over time yielded interesting results showing an increase in the performance of the ratio-type estimators as the sample design deteriorated.

It should be noted that all these comparisons are based on the estimated relative variances and to that extent the degree of confidence in the conclusions would depend upon the stability of these variance estimates. In the context of the LFS, the variance estimates should be highly stable due to the large size of the sample as well as due to the averaging of estimates over survey months. However, for the SCF, the variance estimates may not be as stable due to smallness of the sample size and data being based on single surveys, and therefore, studies using data from additional surveys would be needed to provide greater confidence in the results.

ACKNOWLEDGEMENT

The authors wish to thank the referee for some helpful comments.

Table 1A: Relative Efficiency of the Ratio Estimators*
Period: 1976 (January-June Average)

Provinces	Employed			Unemployed			Not In Labour Force		
	E_{12}	E_{23}	E_{13}	E_{12}	E_{23}	E_{13}	E_{12}	E_{23}	E_{13}
Nfld.	2.62	1.19	3.10	1.12	1.02	1.17	2.10	1.42	2.92
P.E.I.	5.38	2.22	12.18	0.96	0.88	0.86	2.28	1.59	3.10
N.S.	3.31	1.28	4.24	1.23	1.02	1.25	2.02	1.37	2.76
N.B.	2.19	1.74	3.80	1.02	0.98	1.00	1.80	1.82	3.24
Que.	3.24	1.44	4.62	1.02	1.02	1.04	2.40	1.49	3.53
Ont.	3.35	1.49	4.97	1.06	1.02	1.08	1.80	1.56	2.82
Man.	3.72	1.39	5.24	1.10	1.04	1.14	2.50	1.49	3.76
Sask.	6.40	1.74	11.02	1.17	0.98	1.17	1.85	2.28	4.16
Alta.	5.20	1.49	7.78	1.06	1.04	1.10	1.74	1.49	2.59
B.C.	4.54	2.96	6.25	1.23	1.08	1.32	1.74	1.51	2.62
Can.	3.65	1.46	5.34	1.08	1.02	1.10	2.02	1.56	3.13

* The efficiency of the generalized ratio estimator over the combined ratio estimator (i.e. E_{24}) is not available for this period.

Table 1B: Relative Efficiency of the Ratio Estimators
Period 1978 (January, April, July, October average)

ovinces	Employed				Unemployed				Not In Labour Force			
	E_{12}	E_{23}	E_{13}	E_{24}	E_{12}	E_{23}	E_{13}	E_{24}	E_{12}	E_{23}	E_{13}	E_{24}
ld.	4.49	1.12	5.11	2.10	1.51	0.98	1.46	1.02	1.96	1.21	2.34	2.07
E.I.	3.61	1.61	5.76	1.39	1.21	1.19	1.42	1.04	0.96	1.85	1.77	1.49
S.	3.02	1.37	4.00	1.04	1.12	1.04	1.17	1.02	1.88	1.51	2.82	1.04
B.	3.84	1.32	5.15	1.17	1.28	1.08	1.37	1.04	1.49	1.54	2.25	1.28
e.	4.20	1.28	5.29	1.04	1.23	1.08	1.32	1.00	2.04	1.51	3.06	1.06
t.	4.97	1.61	11.16	1.12	1.25	1.06	1.32	1.08	1.49	1.93	2.82	1.17
n.	4.16	1.46	6.10	1.14	1.04	1.02	1.06	1.00	1.69	1.72	2.76	1.14
sk.	14.36	1.59	21.99	1.12	0.83	0.94	0.76	1.25	5.15	1.35	6.71	1.06
ta.	17.81	1.88	35.52	1.39	1.42	0.96	1.35	1.04	3.39	1.90	6.66	1.35
C.	4.37	2.82	14.90	1.08	1.19	1.10	1.30	1.00	1.59	2.72	4.93	1.08
n.	6.66	1.51	10.05	1.11	1.25	1.06	1.35	1.04	1.85	1.77	3.24	1.13

Table 1C: Relative Efficiency of the Ratio Estimators
Period: 1979 (January, April, July, October average)

Provinces	Employed				Unemployed				Not In Labour Force			
	E_{12}	E_{23}	E_{13}	E_{24}	E_{12}	E_{23}	E_{13}	E_{24}	E_{12}	E_{23}	E_{13}	E_{24}
Nfld.	3.04	1.33	3.97	1.04	1.23	1.08	1.32	1.02	2.60	1.34	3.98	1.06
P.E.I.	4.85	1.96	9.73	1.32	1.43	1.01	1.50	1.15	1.26	2.28	2.96	1.48
N.S.	4.38	1.51	6.37	1.17	1.19	1.06	1.26	1.05	1.69	1.65	2.80	1.19
N.B.	3.70	1.58	5.86	1.13	1.09	0.94	1.03	1.02	2.06	1.71	3.45	1.19
Que.	4.13	1.72	7.11	1.07	1.15	1.10	1.18	1.01	1.84	1.81	3.24	1.07
Ont.	7.28	1.56	10.64	1.07	1.30	1.06	1.37	1.03	1.71	1.65	2.81	1.11
Man.	5.84	1.91	11.17	1.21	1.15	1.04	1.21	1.07	1.22	1.88	2.71	1.20
Sask.	9.12	2.02	17.74	1.37	1.29	1.06	1.38	1.06	1.94	2.14	4.00	1.43
Alta.	9.84	1.68	15.40	1.10	1.16	1.01	1.17	1.04	1.71	1.87	3.19	1.10
B.C.	5.11	1.86	9.50	1.18	1.07	1.06	1.14	1.02	1.39	2.08	2.86	1.13
Can.	5.91	1.61	9.47	1.09	1.23	1.05	1.29	1.02	1.74	1.76	3.04	1.11

Table 2: Coefficients of Variation of LFS Estimates
(4 month average: Jan.79, April 79, July 79, Oct. 79)

Province	Employed		Unemployed	
	C.V.(40)	C.V.(18)	C.V.(40)	C.V.(18)
Nfld.	1.97	1.99	6.13	6.13
P.E.I.	2.12	2.15	8.67	8.79
N.S.	1.29	1.29	5.88	5.88
N.B.	1.37	1.37	5.37	5.37
Que.	0.75	0.75	3.68	3.67
Ont.	0.55	0.55	3.59	3.59
Man.	0.94	0.95	6.08	6.04
Sask.	0.94	0.96	7.80	7.78
Alta.	0.65	0.65	5.78	5.78
B.C.	0.87	0.87	4.74	4.80
Canada	0.32	0.32	1.87	1.87

Table 3: Relative Efficiency of Ratio Estimators*
Survey of Consumer Finances
Characteristic: Total Aggregate Income

Province	Period 1977, sample size = 17,000 approx.				Period 1978, sample size = 41,000 approx.			
	E_{12}	E_{23}	E_{13}	E_{24}	E_{12}	E_{23}	E_{13}	E_{24}
Nfld.	2.39	2.22	5.32	1.17	5.66	1.40	7.93	1.00
P.E.I.	2.11	0.88	1.86	1.00	2.67	2.13	5.70	1.42
N.S.	2.88	1.22	3.52	1.01	2.52	1.36	3.36	1.02
N.B.	5.02	1.79	8.98	1.04	3.55	1.60	5.67	1.15
Que.	5.54	1.44	7.98	1.06	3.57	1.09	3.88	1.07
Ont.	16.71	1.34	22.43	1.17	3.25	1.00	3.25	1.01
Man.	4.30	1.75	7.55	1.00	1.66	0.69	1.15	1.07
Sask.	4.50	1.45	6.55	1.00	11.22	0.94	10.53	1.42
Alta.	2.51	1.07	2.69	1.00	4.26	1.15	4.92	1.51
B.C.	2.63	1.53	4.04	1.00	6.42	0.97	6.44	1.01
Canada	9.76	1.21	11.81	1.06	3.80	1.05	3.99	1.11

* Labour force status x class of worker was considered as post-strata instead of age x sex groupings as in Table 1.

RESUME

Trois estimateurs par quotient sont considérés dans cet article. Il s'agit de l'estimateur combiné, de l'estimateur post-stratifié et d'un estimateur par quotient généralisé qui a déjà été proposé par Singh (1969) et par Naga Reddy (1974). Dans une évaluation empirique, on compare l'efficacité de ces estimateurs dans le contexte de deux enquêtes-ménages de grande envergure: l'Enquête sur la population active canadienne et l'Enquête sur les finances des consommateurs.

REFERENCES

[1] Keyfitz, V. (1957), "Estimates of Sampling Variance Where Two Units are Selected from Each Stratum", Journal of the American Statistical Association, 52,503-510.

[2] Cochran, W.G. (1963),"Sampling Techniques", 2nd Edition, John Wiley and Sons, New York.

[3] Reddy, V.N. (1974), "On a Transformed Ratio Method of Estimation", Sankhya, 36, 59-70.

[4] Singh, M.P. (1969), "Some Aspects of Estimation in Sampling from Finite Populations", Ph.D. Thesis (ch. 8), Indian Statistical Institute.

[5] Statistics Canada (1977), "Methodology of the Canadian Labour Force Survey" (1976), Catalogue No. 71-526.

[6] Turner, R. and Lawes, M. (1979), "Incomplete Data in the Survey of Consumer Finances", Statistics Canada (unpublished).

Appendix A

Considering the usual approximation for the variance of a ratio of statistics, the variance of the generalized ratio estimator can be expressed as:

$$\text{Var}\,(X_p^{(4)}) = \text{Var}\,\left(\frac{\hat{X}_p}{\alpha\hat{P}_p+(1-\alpha)P_p}P_p\right) \doteq P_p^2\left\{\frac{E(\hat{X}_p)}{E(\alpha\hat{P}_p+(1-\alpha)P_p)}\right\}^2\left\{\frac{\text{Var}(\hat{X}_p)}{[E(\hat{X}_p)]^2} - \right.$$

$$\left. - \frac{2\,\text{Cov}(\hat{X}_p,\ \alpha\hat{P}_p+(1-\alpha)P_p)}{E(\hat{X}_p)\ E(\alpha\hat{P}_p+(1-\alpha)P_p)} + \frac{\text{Var}(\hat{P}_p+1(1-\alpha)P_p)}{[E(\alpha\hat{P}_p+(1-\alpha)P_p)]^2}\right\}$$

$$= \left[\frac{P_p}{\alpha E(\hat{P}_p)+(1-\alpha)P_p}\right]^2\left\{\text{Var}\,(\hat{X}_p) - 2\alpha\,\frac{E(\hat{X}_p)}{\alpha E(\hat{P}_p)+(1-\alpha)P_p}\,\text{Cov}(\hat{X}_p,\hat{P}_p)\right.$$

$$\left. + \alpha^2\left[\frac{E(\hat{X}_p)}{\alpha E(\hat{P}_p)+(1-\alpha)P_p}\right]^2\text{Var}(\hat{P}_p)\right\}.$$

Now using the inevitable procedure of estimating $E(\hat{X})$ and $E(\hat{P})$ by $\hat{X}$ and $\hat{P}$ respectively, Var $(\hat{X}_p^{(4)})$ can be estimated by

$$\hat{\text{Var}}\,(\hat{X}_p^{(4)}) = \left[\frac{P_p}{\alpha\hat{P}_p+(1-\alpha)P_p}\right]^2\left\{\hat{\text{Var}}\,(\hat{X}_p) - 2\alpha\,\frac{\hat{X}_p}{\alpha\hat{P}_p+(1-\alpha)P_p}\,\hat{\text{Cov}}\,(\hat{X}_p,\hat{P}_p)\right.$$

$$\left. + \alpha^2\left[\frac{\hat{X}_p}{(\alpha\hat{P}_p+(1-\alpha)P_p}\right]^2\hat{\text{Var}}\,(\hat{P}_p)\right\}.$$

Now replacing $\hat{\text{Var}}\,(\hat{X}_p)$ and $\hat{\text{Var}}\,(\hat{P}_p)$ by expressions of the form (4.1) and with $\hat{\text{Cov}}\,(\hat{X}_p,\hat{P}_p)$ estimated by

$$\hat{Cov}(\hat{X}_p,\hat{P}_p) = \sum_{h\epsilon p} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} (\hat{X}_{phi} - 1/n_h \sum_{i=1}^{n_h} \hat{X}_{phi})\,(\hat{P}_{phi} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{P}_{phi})$$

the above expression reduces to

$$[\frac{\hat{P}_p}{\alpha\hat{P}_p + (1-\alpha)P_p}]^2 \sum_{h\epsilon p} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} \{\hat{X}_{phi} - \frac{\hat{X}_p}{\alpha\hat{P}_p + (1-\alpha)P_p} \hat{P}_{phi}$$

$$- \frac{1}{n_h} \sum_{i=1}^{n_h} (\hat{X}_{phi} - \frac{\hat{X}_p}{\alpha\hat{P}_p + (1-\alpha)P_p} \hat{P}_{phi})\}^2$$

$$\hat{Var}\,(\hat{X}_p^{(4)}) = [\frac{\hat{P}_p}{\alpha\hat{P}_p + (1-\alpha)P_p}]^2 \sum_{h\epsilon p} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} \{D_{phi} - \frac{1}{n_h} \sum_{i=1}^{n_h} D_{phi}\}^2 \qquad (1)$$

$$\text{with } D_{phi} = \hat{X}_{phi} - \frac{\hat{X}_p}{\alpha\hat{P}_p + (1-\alpha)P_p} \hat{P}_{phi} . \qquad (2)$$

With $\alpha=1$ the generalized ratio estimator reduces to the usual ratio estimator, i.e. $\hat{X}_p^{(2)}$

and

$$\text{Var}\,(\hat{X}_p^{(2)}) = (\frac{P_p}{\hat{P}_p})^2 \sum_{h\epsilon p} \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} (D_{phi} - \frac{1}{n_h} \sum_{i=1}^{n_h} D_{phi})^2 \qquad (3)$$

$$\text{where } D_{phi} = \hat{X}_{phi} - \frac{\hat{X}_p}{\hat{P}_p} \hat{D}_{phi} . \qquad (4)$$

Similarly for the post-stratified ratio estimator, the variance can be expressed as

$$\text{Var}\ (\hat{X}_p^{(3)}) = \text{Var}\ \Big(\sum_a \frac{\hat{X}_{pa}}{\hat{P}_{pa}} P_{pa} \Big)$$

$$= \sum_a \text{Var}\ \Big(\frac{\hat{X}_{pa}}{\hat{P}_{pa}} P_{pa} \Big) + \sum_{a \neq b} \sum \text{Cov}\ \Big(\frac{\hat{X}_{pa}}{\hat{P}_{pa}} P_{pa}, \frac{\hat{X}_{pb}}{\hat{P}_{pb}} P_{pb} \Big) . \qquad (5)$$

Using expressions (3) and (4) $\text{Var}\ \Big(\frac{\hat{X}_{pa}}{\hat{P}_{pa}} P_{pa} \Big)$ is estimated by:

$$\hat{\text{Var}}\ \Big(\frac{\hat{X}_{pa}}{\hat{P}_{pa}} P_{pa}\Big) = \Big(\frac{P_{pa}}{\hat{P}_{pa}}\Big)^2 \sum_{h \in p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \Big(D_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} D_{phia}\Big)^2 \qquad (6)$$

$$\text{with}\ \ D_{phia} = \hat{X}_{phia} - \frac{\hat{X}_{pa}}{\hat{P}_{pa}} \hat{P}_{phia} . \qquad (7)$$

Then for two different post-strata denoted by a and b,

$$\text{Cov}\ \Big(\frac{\hat{X}_{pa}}{\hat{P}_{pa}} P_{pa}, \frac{\hat{X}_{pb}}{\hat{P}_{pb}} P_{pb} \Big)$$

$$= P_{pa}\, P_{pb}\, \text{Cov}\ \Big(\frac{\hat{X}_{pa}}{\hat{P}_{pa}}, \frac{\hat{X}_{pb}}{\hat{P}_{pb}} \Big)$$

$$= P_{pa}\, P_{pb}\, \frac{E(\hat{X}_{pa})}{E(\hat{P}_{pa})} \frac{E(\hat{X}_{pb})}{E(\hat{P}_{pb})} \Big\{ \frac{\text{Cov}\ (\hat{X}_{pa}, \hat{X}_{pb})}{E(\hat{X}_{pa})\, E(\hat{X}_{pb})} - \frac{\text{Cov}\ (\hat{X}_{pa}, \hat{P}_{pb})}{E(\hat{X}_{pa})\, E(\hat{P}_{pb})}$$

$$- \frac{\text{Cov}\ (\hat{X}_{pb}, \hat{P}_{pa})}{E(\hat{X}_{pb})\, E(\hat{P}_{pa})} + \frac{\text{Cov}\ (\hat{P}_{pa}, \hat{P}_{pb})}{E(\hat{P}_{pa})\, E(\hat{P}_{pb})} \Big\} \quad \text{up to 2nd order terms.}$$

Again estimating $E(\hat{X}_{pa})$, $E(\hat{X}_{pb})$, $E(\hat{P}_{pa})$ and $E(\hat{P}_{pb})$ by $\hat{X}_{pa}$, $\hat{X}_{pb}$, $\hat{P}_{pa}$ and $\hat{P}_{pb}$ respectively,

$$\hat{Cov}\left(\frac{\hat{X}_{pa}}{\hat{P}_{pa}} P_{pa}, \frac{\hat{X}_{pb}}{\hat{P}_{pb}} P_{pb}\right) = \frac{P_{pa}}{\hat{P}_{pa}} \frac{P_{pb}}{\hat{P}_{pb}} \left[\mathrm{Cov}(\hat{X}_{pa}, \hat{X}_{pb}) - \frac{\hat{X}_{pb}}{\hat{P}_{pb}} \mathrm{Cov}(\hat{X}_{pa}, \hat{P}_{pb})\right.$$

$$\left. - \frac{\hat{X}_{pa}}{\hat{P}_{pa}} \mathrm{Cov}(\hat{X}_{pb}, \hat{P}_{pa}) + \frac{\hat{X}_{pa}}{\hat{P}_{pa}} \frac{\hat{X}_{pb}}{\hat{P}_{pb}} \mathrm{Cov}(\hat{P}_{pa}, \hat{P}_{pb})\right]$$

$$= \frac{P_{pa}}{\hat{P}_{pa}} \frac{P_{pb}}{\hat{P}_{pb}} \left[\sum_{h \epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\hat{X}_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{X}_{phia})(\hat{X}_{phib} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{X}_{phib})\right.$$

$$- \frac{\hat{X}_{pb}}{\hat{P}_{pb}} \sum_{h \epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\hat{X}_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{X}_{phia})(\hat{P}_{phib} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{P}_{phib})$$

$$- \frac{\hat{X}_{pa}}{\hat{P}_{pa}} \sum_{h \epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\hat{X}_{phib} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{X}_{phib})(\hat{P}_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{P}_{phia})$$

$$\left. + \frac{\hat{X}_{pa}}{\hat{P}_{pa}} \frac{\hat{X}_{pb}}{\hat{P}_{pb}} \sum_{h \epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\hat{P}_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{P}_{phia})(\hat{P}_{phib} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{P}_{phib})\right]$$

$$= \frac{P_{pa}}{\hat{P}_{pa}} \frac{P_{pb}}{\hat{P}_{pb}} \sum_{h \epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left[\{(\hat{X}_{phia} - \frac{X_{pa}}{\hat{P}_{pa}} \hat{P}_{phia}) - \frac{1}{n_h} \sum_{i=1}^{n_h} (\hat{X}_{phia} - \frac{\hat{X}_{pa}}{\hat{P}_{pa}} \hat{P}_{phia})\right.\}$$

$$\left. - \{(\hat{X}_{phib} - \frac{\hat{X}_{pb}}{\hat{P}_{pb}} \hat{P}_{phib}) - \frac{1}{n_h} \sum_{i=1}^{n_h} (\hat{X}_{phib} - \frac{\hat{X}_{pb}}{\hat{P}_{pb}} \hat{P}_{phib})\}\right]^2$$

$$= \frac{P_{pa}}{\hat{P}_{pa}} \frac{P_{pb}}{\hat{P}_{pb}} \sum_{h \epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} [\{D_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} D_{phia}\}\{D_{phib} - \frac{1}{n_h} \sum_{i=1}^{n_h} D_{phib}\}]^2 .$$

Now equation (5) is estimated by :

$$\hat{Var}\ (\hat{X}_p^{(3)}) = \sum_a \left(\frac{P_{pa}}{\hat{P}_{pa}}\right)^2 \sum_{h \epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(D_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} D_{phia}\right)^2$$

$$+ \sum_{a \neq b} \sum \frac{P_{pa}}{\hat{P}_{pa}} \frac{P_{pb}}{\hat{P}_{pb}} \sum_{h \epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} [\{D_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} D_{phia}\}\{D_{phib} - \frac{1}{n_h} \sum_{i=1}^{n_h} D_{phib}\}]$$

$$= \sum_{h \epsilon p} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} [\sum_a \frac{P_{pa}}{\hat{P}_{pa}} D_{phia} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_a \frac{P_{pa}}{\hat{P}_{pa}} D_{phia}]^2$$

which is equivalent to expression (4.5) in the body of the paper.

Appendix B

The efficiency of the combined ratio estimator versus the simple survey estimator can be derived and expressed in terms of the correlation coefficient between the numerator and denominator of the ratio of characteristic to population totals (see for example Cochran, Section 6.8, p. 165). Specifically,

$$E_{12} = 1 - 2 \frac{\sqrt{rv(\hat{P}_p)}}{\sqrt{rv(\hat{X}_p)}} \left[\hat{\rho}(\hat{X}_p, \hat{P}_p) - \frac{1}{2} \frac{\sqrt{rv(\hat{P}_p)}}{\sqrt{rv(\hat{X}_p)}}\right]^{-1}$$

where $\hat{\rho}(\hat{X}_p, \hat{P}_p)$ denotes the estimated correlation coefficient between the simple survey estimates $\hat{X}_p$ and $\hat{P}_p$ and rv denotes the relative variance.

The following table presents the estimated relative efficiencies E_{12} and estimated correlation coefficients $\hat{\rho}$ for the Labour Force Survey for the three periods 1976, 1978 and 1979 considered in this report. (The values E_{12} and $\hat{\rho}$ are averages of the corresponding monthly estimates).

Table 4: Estimated Relative Efficiencies and Correlation Coefficients at the Canada Level.

Characteristic

Period	EMPLOYED		UNEMPLOYED		NOT IN LABOUR FORCE	
	E_{12}	$\hat{\rho}$	E_{12}	$\hat{\rho}$	E_{12}	$\hat{\rho}$
1976	3.65	0.853	1.08	0.245	2.02	0.708
1978	6.66	0.925	1.25	0.441	1.85	0.703
1979	5.91	0.915	1.23	0.318	1.74	0.676

NON-TEXTBOOK PROBLEMS IN THE REVISION OF A BUSINESS BASED EMPLOYMENT SURVEY[1]

Michael J. Colledge[2]

The paper illustrates several practical problems in the adaptation of statistical theory to survey design in the context of the revision of an employment survey programme.

1. INTRODUCTION

The practical problems in adapting statistical theory to survey design may often present more of a challenge than the associated theoretical development. The aim of the paper is to illustrate this point by describing a selection of practical problems encountered during the revision of an employment survey programme at Statistics Canada.

Section 2 of the paper provides a fairly comprehensive account of the environment within which the revision took place. Section 3 presents a selection of the problems arising. The progress which has been made towards resolving these problems, and some specific solutions, are described in Section 4.

---

[1] Adapted from a paper presented at the 1980 Annual Meeting of the Statistical Society of Canada, May 1980.

[2] Michael J. Colledge, Business Survey Methods Division, Statistics Canada.

It should be noted that the perspective given here of the revision project, the Employment Statistics Development Project, is essentially a personal one. A list of more definitive documents follows the main text. References to the list are indicated by a number in brackets, for example, [1].

2. ESTABLISHMENT-BASED COLLECTION OF EMPLOYMENT DATA

2.1 Concepts, Terminology

A more precise title to this paper would have referred to revision of "establishment-based surveys of civilian employment and payroll in the non-agricultural sector of the economy". The objective of such survey activity is, in brief, to measure levels and month to month trends in paid employment, hours and earnings in all sectors of the economy (except agriculture, hunting, fishing and trapping) and to publish data at detailed geographic and industrial breakdown.

The term "establishment-based" implies that the data are collected from business and institutional establishments, in contrast to the Labour Force Survey [1] which also collects labour statistics but via households. Thus the survey universe is the set of all employers, i.e. firms incorporated or unincorporated, institutions, government departments, agencies, etc.

In this context "establishment" has a very particular meaning: it refers to the "smallest unit that is a separate operating entity capable of reporting all elements of basic industrial statistics" [2]. For the purposes of collecting employment, hours and earnings data, establishments are sometimes further divided according to geographic location and reporting capability into "reporting units".

Each establishment is assigned a standard industrial classification (SIC) code according to the nature of its principal activity.

There are just under 300 SIC codes grouped into a number of "industry divisions" [2].

"Employment" refers to all civilians drawing pay for services rendered or for paid absence during the survey reference period for when an employer makes CPP or QPP and or UIC contributions. Thus it includes full time, part time and casual employees but excludes owners of unincorporated businesses, self employed persons and unpaid family workers. The "survey reference period" is, in principle, the last pay period. "Hours" refers to time actually worked plus hours of paid absence, e.g. holidays, vacation, sick leave. "Earnings" refers to gross pay during the pay period including commissions, bonuses, overtime pay, etc., but excluding employers' contributions to pension plans, travelling expenses, etc.

2.2 Data Collection Methods in Current Use

Employment and payroll data are presently collected by survey and from administrative sources by two essentially separate programmes under the control of Labour Division and Public Finance Division respectively of Statistics Canada [4]. The principal features of these programmes are described below.

(a) ES-1 Survey (Labour Division).

The ES-1 survey is a monthly census of larger firms in the commercial sector. In this context, "larger" refers to firms with 20 or more employees in any one month of the year, and the "commercial sector" excludes by definition education, health, welfare, public administration and defence. Approximately 50,000 questionnaires are mailed monthly from the Head Office (Ottawa). Also data for some government reporting units are obtained from the Public Finance Division programme. Non respondents are sent two reminders and finally contacted by telephone via the Regional Offices. The response rate at the time of publication is about 80%.

Returned questionnaires are clerically screened before entering the data capture and processing system. Edit, imputation and tabulation are automated. The monthly publication [3] is produced 3 months after the reference period. There is a preliminary release after about 60 days.

(b) ES-2 Survey (Labour Division).

The ES-2 survey is complementary to ES-1 for the commercial sector, i.e. it covers firms of less than 20 employees. The universe of 550,000 or so small firms is stratified into 17 industrial groups and 16 geographic areas. The monthly sample of some 37,000 is selected according to a panel rotation scheme. Within a stratum each firm is allocated to a panel. The number of panels depends upon the number of firms and the required stratum sampling fraction. In any given month 12 panels from each stratum are in sample. At the end of the month the oldest panel goes out of sample and a new one rotates in.

Questionnaires requesting employment data (not hours, earnings) are sent out by mail. They are followed where necessary by a single mailed reminder and, for non-respondents new to the sample, by Regional Office telephone contact. The response rate is about 75%. The employment figures are not published separately. They contribute about 25% of the total employment reported in a publication [5] which appears within 3 months of the reference period with a preliminary release after about 60 days.

(c) GAP Survey (Labour Division).

The GAP survey covers the non-commercial sector of the economy e.g. hospitals, education, institutions but excludes public administration and defence. It is a monthly census of about 6500 reporting units along the same lines as the ES-2 survey. The response rate is over 80% and the resulting data contributes some 16% of the total employment reported in the publication [5].

(d) Government Programme (Public Finance Division).

Government employment and earnings information is collected by 5 surveys on a monthly, quarterly or annual basis from about 4500 respondents, and from administrative sources. Response rates are 90% or better. Some data are exchanged at micro (i.e. reporting unit) level with the ES-1 and GAP surveys. At macro level the data provide the public administration and civilian defence component of publication [5], accounting for about 8% of the total employment. In addition there are 3 publications detailing employment and earnings for federal, provincial and local governments respectively [6, 7, 8].

2.3 Reasons for Revision

In brief, the major reasons which prompted Labour Division to undertake a revision of the programme were:

(i) to improve timeliness;
(ii) to reduce respondent burden, e.g. by decreasing the total sample size;
(iii) to extend coverage by providing employment, hours and earnings data for small firms, and for the non-commercial sector including public administration;
(iv) to investigate large scale use of data from administrative sources to supplement and/or replace survey data;
(v) to review the employment category breakdowns, and the data presentation, and to consider provision of additional information, e.g. labour turnover;
(vi) to examine the possibility of coordinating all data collection activities within a single programme;
(vii) to incorporate advances in data processing technology and to improve man/machine interfaces;
(viii) to correct inadequacies in the methodology of the small firm panel rotation scheme;

(ix) to consider change of survey reference period to be closer to that of the Labour Force Survey.

The relative importance of these factors has changed somewhat since the revision started. Increased consciousness of respondent burden has now made this perhaps the most significant single factor. On the other hand, timeliness considerations are a less compelling reason than previously, owing to improved performance of the ongoing surveys.

2.4 Revision Process: The Employment Statistics Development Project

(a) Phase I: Programme Objectives Team

The intention to revise employment data collection methods was announced at the 1976 meeting of the Federal-Provincial Committee on Labour Statistics. Subsequently, an interdepartmental team, the Programme Objectives Team, was established to review the current programme and to identify the information which should be produced by the revised scheme. Submissions were solicited from users of the data and were obtained primarily from the government sector. The team presented its final report [9] in May 1977.

(b) Phase II: Design Specification Team

A Design Specification Team, comprising senior personnel from several Divisions within Statistics Canada, was set up to develop the broad methodological framework within which the recommendations arising from Phase I could be met. The team completed its report [10] in August 1977.

(c) Phase III: Project Implementation Team

The Project Implementation Team began work in 1977. Its objectives were, and are, to design, build, test and hand over an operational

employment and payrolls survey programme developed in accordance with the recommendations of the Phase I and II reports. The strategy adopted has been to divide work along functional lines into tasks each of which is handled by a working group reporting periodically to the central team. Thus there are groups concerned with concepts, survey methodology, computer systems, manual processing, etc. The central team itself reports at regular intervals to the Design Specification Team which reviews progress and provides guidance. Starting from small numbers, the project now employs some 30 people on a full time basis, and will continue to do so until towards the end of the development scheduled for December 1981.

3. PROBLEMS ENCOUNTERED

During the course of Phase III implementation, numerous problems have been encountered, far too many to describe in a single paper. The following paragraphs present a selection of problems which are fairly specific to this type of survey and which seem particularly pertinent at this stage in the revision process. In an attempt to classify them in some sort of coherent order they have been put into four groups. Whilst such a categorization is not precise it does go part way towards identifying the underlying causes.

3,1 Universe Related Problems

(a) Number, Growth, Size of Units

There are over 600,000 employers in Canada and the number is growing. This results in the necessity to have a highly automated system, both for frame maintenance and for data collection. Even so it is a massive problem to keep the frame up to date and to process data in a timely fashion. Firms and institutions may range in size from zero

employees up to thousands, and the distribution is highly skewed. For example about 95% of firms have less than 20 employees (see table 1). On the whole, these small firms are less than enthusiastic about supplying data because they derive little benefit from the resulting publications, yet they cannot be ignored as they account collectively for nearly 25% of all employment. Large units may be more inclined to respond but in their case it is time consuming to define and maintain coherent reporting structures which cover all aspects of their operations.

(b) Instability

Small firms are subject to quite frequent changes in ownership, and/or industrial activity, which it is not always easy to track. Also, some have a seasonal form of operation hence only respond during certain months of the year. Even larger firms can be difficult to identify as stable units. Mergers, amalgamations, changes of legal name, etc., may all serve to render a list of companies and their reporting structure quickly obsolete. The most stable aspect of many firms is physical location, buildings and plant, but use of an area frame to take advantage of this, is very expensive. In any case some types of business would be difficult to locate, e.g. wholesalers who arrange shipments direct from producer to retailer without storage space of their own.

The major problems arising from the size, growth and instability of the universe are:

(i) undercoverage, for example omission of new units or of additions to existing ones;

(ii) duplication, for example inclusion of the same unit under different names, perhaps a legal name and an operating name;

(iii) inaccuracy of classification data, especially of industrial activity (SIC) and of size, due to initial misclassification or failure to detect change;

(iv) changes in classification data, which, even when identified, are not necessarily easy to handle.

3.2 Problems Related to User Requirements

(a) Tabulation Detail

The specification of tabulations at industrial classification (SIC) by province level defines some 3500 strata. In combination with a requirement for stratification by size into, say, 4 groups this implies 14,000 cells. Spreading a sample of even 70,000 over this many cells is likely to leave many empty and hence to give rise to an estimation problem. Various techniques exist for handling this type of problem, e.g. synthetic estimation but these are not yet textbook tools.

(b) Reliability

It was originally recommended that a minimum acceptable level of overall reliability should be set after consultation with the user community. The problem with this approach is twofold. Firstly, the user generally requests estimates should be "as reliable as possible", which means a census; secondly the non sampling errors are difficult to estimate.

(c) Continuity

One important criterion upon which users always agree is that they want "continuity" of published data. It translates in practice into the need to run revised and existing programmes in parallel for a sufficiently long period to generate a basis for comparison purposes.

3.3 Problems Related to Respondents

(a) Respondent Burden

Although there has been increasing awareness of the workload imposed by administrative and survey demands, especially upon small firms, this "respondent burden" is not yet well quantified. There are no definitive measures of the relative burdens associated with mail questionnaires as opposed to telephone interviews, with regular as opposed to ad hoc questionnaires, with being surveyed every month as opposed to rotating in and out. Even the number of questionnaires sent or interviews conducted is not necessarily an exact measure of respondent burden as some firms may actually request additional questionnaires or interviews to suit their reporting arrangements, or may submit computer printouts or machine readable data in lieu of questionnaires.

(b) Data Availability

Not all respondents have ready access to data broken down into the required categories such as full time/part time, hourly paid/ salaried, regular payments/special payments, etc. The availability and relevance of such breakdowns are intimately related to the particular industrial activity involved.

3.4 Problems Related to Implementation Environment

(a) Microdata Transfer

It is a policy of Statistics Canada to move towards data integration, i.e. collecting and maintaining a central body of data capable of meeting many different needs. One aspect of this policy is reflected in a recommendation that, in order to reduce survey taking activity, data from the revised programme should be usable at micro level by

other programmes. Making such provision introduces constraints on the data collection strategy.

(b) Data Collection Vehicles

As previously indicated, data relevant to the programme are presently collected by two Divisions within Statistics Canada: Labour Division, which also has responsibility for the revised programme, and Public Finance Division. Whilst Labour Division can arrange to gather data in precise accordance with the requirements of the revised programme, the data collected by Public Finance Division is primarily acquired to suit the publication requirements of that Division. Slightly different concepts and coverage are involved. For example the programme excludes Canadians working abroad and employees on strike, whereas Public Finance Division publications include them. Thus the arrangement of sharing responsibility for data collection presents certain problems quite apart from the obvious time lags associated with data transfer between Divisions. On the other hand it is in accordance with the general principle of maximizing internal usage of data. It means that a respondent receives only one questionnaire in place of two. Also, contact with the respondent is confined to a single point within Statistics Canada which enhances the prospects of building and maintaining good relationships.

4. PROGRESS TOWARDS SOLUTIONS

The problems outlined in section 3 are interrelated and it is difficult to treat their solutions separately. Thus, in this section, the various strategies adopted by the implementation team are described and are related back to the problems which they address.

A complete summary of the development work and progress up until February 1980 is contained in [11] and [12]. The list of test programme

milestones, which follows, gives some idea of the general status of the project. Not all the problems referred to have been fully solved; some "solutions" have introduced further problems.

4.1. Test Programme

The earlier part of the test programme was designed primarily to tackle the problems of respondent burden and data availability, for example by investigating to what data respondents might be expected to have ready access. The latter part of the programme is more concerned with checking the embryo revised system and providing the capacity to cope with problems related to user requirements, such as continuity of published data. The individual tests are as follows.

(i) Employment Compensation Test (August 1977): to investigate the terminology best understood by respondents; where respondents tend to keep their records; what forms of compensation are paid.

(ii) Field Test I (March 1978): to determine the availability of certain data items, e.g. overtime payments, data breakdown according to sex, etc., within various industry groupings.

(iii) Reference Period Test (May 1978): to check respondents' capacity to report data for an earlier week in the month; to compare data for early and late reference periods. This resulted in the retention of a late reference period.

(iv) Field Test II (September 1978): to check use of different questionnaires for different industries; to compare data collection by Head Office and by Regional Office, by mail and by telephone.

(v) Quality Control Test (Oct. 1978): to investigate problems arising from certain reporting practices, e.g. separation of executive and other payrolls.

(vi) Field Test III (January 1980): to test operational and systems procedures for the "front end" of the survey system; to collect data enabling comparison of figures for existing and revised programmes; results of this test are not yet available.

(vii) Parallel run (scheduled for 1981): to test, correct and fine tune the revised survey system; to provide 12 months of comparable data.

4.2 Universe Frame Maintenance Procedure

To handle the universe related problems, the frame maintenance system will be highly automated. It will be based, as for the current Labour Division surveys, partly upon feedback from survey process itself and partly upon the "Business Register Master File System". Although this system is outside the scope of the revision it is a cornerstone of existing and future activities and thus will be described.

The objective of the Business Register, developed at Statistics Canada over the past ten years, is to provide a frame suitable for all establishment based surveys, not just employment [13]. Unfortunately the system does not in fact maintain a "register", rather it keeps a list of units (firms or institutions) based primarily upon Revenue Canada payroll deduction (PD) information. New units or establishments may be indicated by new PD accounts, dead units by closed accounts, etc. As every employer is legally obliged to make regular deductions on behalf of employees and to remit them via a PD account the system provides, in concept, very good coverage of employment universe.

In practice, there is some undercoverage which can arise in various ways. Employers may be late in submitting their requests for new PD accounts, thereby causing a time lag effect. They may provide insufficient information to enable proper classification of units, or they may not even report at all. The extent of undercoverage is not easy to measure precisely; it is believed to be reasonably small.

The duplication problem is more difficult to tackle. Revenue Canada data is identified by PD account number. A firm or institution may have several PD accounts for one establishment, or may have a single

PD account for several establishments, thus the relationship between units and PD information is complex. Furthermore the only means of linking information associated with a new PD number to any unit already on the Business Register is by name and address. Slight variations in name and address prevent successful use of straightforward matching and a purpose built record linkage procedure [14] has been developed to accomplish the task. Even with this finely tuned procedure however, there are many situations where a proper linkage will not be established and a new unit will be inappropriately introduced to the Register. For example, a business may operate under a trading name but use its legal name or even an accountant's name when applying for a new PD account. The net result is quite severe duplication, evidence of which comes to light during the course of survey operations. An improved programme for Business Register Maintenance is currently being designed.

Assignment and maintenance of classification data items such as industrial activity (SIC) and size is also a source of unresolved problems. Initial assignment is based upon PD information which is somewhat inadequate for the purpose; maintenance through the PD system per se is negligible. With the objective of improving size classification, investigations are currently taking place into the use of other Revenue Canada data, in particular the annual "T4 supplementary" information which includes, amongst other items, the total earnings and total number of employees associated with every PD account for a year.

To summarize, frame problems have been addressed but not entirely solved. Efforts to improve the quality of the frame and to estimate for its inadequacies are continuing.

4.3 Choice of Sampling Unit and Take-All Stratum

Problems associated with respondents' diverse reporting procedures together with constraints imposed by the implementation policy

of data integration combine to influence the choice of sampling unit and take-all stratum in the following way. Firstly, although the concept of "establishment" is, in principle, common to all business surveys, the "reporting unit" is dependent upon the particular programme and may well differ from one to another. Thus the requirement that micro-data from the revised programme be usable by other surveys implies that it must be obtained from all reporting units associated with an establishment so that data transfers can take place at establishment level. This in turn suggests that sampling should be at establishment rather than at reporting unit level. A second argument against sampling reporting units stems from their possible non-homogeneity. As previously stated, some reporting units are created essentially for respondent convenience, e.g. to separate executive from general payrolls. They do not reflect a genuine stratification requirement and may be quite heterogeneous. On the other hand, previous experience suggests that establishments are not satisfactory sampling units. An establishment may have reporting units in quite different geographic strata and the assignment of a single sampling weight associated with the establishment to each of these reporting units sometimes gives completely inappropriate results.

The net result of these considerations is that all firms or institutions with more than one reporting unit (called "multis") have been defined as belonging to the "take-all" stratum, i.e. they fall automatically in sample. This is not a pretty solution, involving some 25,000 reporting units. However it is not quite as bad as may seem at first sight because it is the larger firms and institutions which tend to be multi-reporting units, and any reasonable sampling scheme would be sampling these with higher probability than the smaller "single" companies anyway.

Other types of units are assigned in small numbers to the take-all stratum too. For instance, units in certain strata are considered so heterogeneous as to require a census, government departments and

hospitals being an example. Some units are "carriers" and include in their reports data for other "carried" units. As it is not possible to obtain data from carried units separately, complication is avoided by allocating both carriers and carried to the take-all stratum. Respondents who have arranged to submit their data by computer tape or printout in lieu of a completed questionnaire are also assigned to the take-all stratum. It would be more burdensome to stop and restart their reporting arrangements as they rotated out and back into the sample than to leave them in the sample all the time.

4.4 Sample Size Determination and Allocation Procedure

The procedure for sample size determination and allocation has been assigned to cope with the user related problems of providing required tabulation detail at adequate reliability, within the context of a rather unstable and poorly classified universe. More specifically, the following considerations have been taken into account.

(i) Monthly estimates of employment, hours and earnings are required both at industry division by province level and at individual industry (SIC) by Canada level.

(ii) No specific indication of acceptable reliability has been given.

(iii) The overall sample size should be as small as possible.

(iv) Every unit in the universe is currently assigned an employment size coding, which is of somewhat dubious quality for the smaller units. No similar measure of size in terms of earnings currently exists.

(v) The sampling units have a very skewed distribution as regards employment and earnings.

(vi) Estimates of employment are available for all firms, but hours and earnings data is restricted to units of 20 or more employees.

(vii) The number of reporting units in the take-all stratum is significant and should be recognized in determining the required allocation of take-some units.

Starting with the skewness of size distribution, it is clear that some allowance for size has to be incorporated within the scheme. As the only measure readily available is an employment size code it has been decided to stratify using this code. Sampling with probability proportional to size was not seriously entertained because of the poor quality of the size coding.

The cumulative $\sqrt{f}$ technique was used to identify desirable stratum boundaries but the results had to be modified for practical reasons (essentially the absence of sufficiently detailed coding). The net result has been the decision to stratify into 4 size groups. In combination with 12 provinces and just under 300 SICs there is a total of some 14,000 "cells" at the finest level of stratification. Rather than attempt to allocate the sample directly to so many cells, it was decided to design for uniform reliability across all tabulation strata in the most important tabulation, namely, the one with an industry division by province breakdown. In essence, therefore, the scheme is to allocate according to size in each industry division, province stratum separately. As Neyman allocation is rather strongly dependent upon reasonable variance estimates, a more robust, "X-proportional" allocation procedure was preferred [15]. Under this scheme the sampling fraction is made proportional to the estimate of the average employment per unit with the size group. It is equivalent to Neyman allocation if the coefficient of variation of employment per unit is the same for each size group.

The value of reliability required uniformly across all tabulation strata is specified as an input parameter to the procedure in terms of coefficient of variation. This parameter setting determines, in conjunction with the estimate of total employment covered by take-all

units, the coefficient of variation required from the take-some sample in each tabulation stratum. Hence, given X-proportional allocation, the required sample size is determined at industry division by province by size level.

To define allocation at the finest (cell) level of stratification, proportional allocation across SIC, within industry division by province by size is used. In other words the sampling fraction which has been determined for a particular industry division by province by size stratum is applied to all SICs nested therein.

Given this allocation the resulting coefficients of variation for strata of the SIC, Canada level tabulation can be estimated. Of course the values are not uniform across SIC. Consideration is being given to an adjustment of sampling fractions to bring all coefficients of variation down to a specified acceptable level.

Direct application of the technique described above results in many cells at the finest level of stratification being allocated a sample of zero units, causing obvious estimation difficulties. To address this problem the procedure incorporates facilities to enforce a minimum selection of m units per cell (or as many units as are available for selection (if less than m). Setting m=2 to facilitate production of variance estimates results in a considerable increase in sample size, by, say, 3500 units in a total sample of 65,000. Of course even with m=2 there is no guarantee that units will respond, thus variance estimation may still be difficult. Collapsing by grouping sets of cells across similar SICs within the same industry division, province and size, coupled with the use of domain estimation is being considered as a means of curtailing these problems.

In summary, the sample size determination and allocation procedure will determine an overall sample size and an allocation for any specified level of uniform reliability of the industry division by province

employment tabulation. Rerunning the procedure with a variety of reliability settings generates a graph of sample size versus reliability; see graph 1 which is of considerable assistance in deciding how large the sample should be. Growth or compositional changes in the universe can be allowed for by recomputing the required sample size and allocation [16].

4.5 Sample Selection and Rotation Procedure

The scheme for sample selection and rotation has been designed with the problems of universe growth and change and of respondent burden foremost in mind. Firstly, as the number of units in the universe is increasing, use of a sample of fixed size would result in reduced reliability over a period of time. A fixed sample fraction, on the other hand would increase the sample more than necessary for specified reliability. Some form of sample size adjustment based on the output of the sample size determination procedure is preferable, and a rotation scheme can readily incorporate such provision. Secondly, the universe is constantly changing in composition. Thus, in the absence of rotation, a periodic redraw would be necessary with attendant risk of unwarranted jumps in the estimates. Again rotation can provide the means of adjustment on a monthly basis. Thirdly, there are a variety of rather subjective respondent burden considerations.

It is believed for example that for a respondent to be in the sample always is burdensome and should be avoided where possible by rotating the respondent out and another in, even though this does not reduce the total number of respondents. It is also felt that respondents are likely to experience more difficulty in completing the first questionnaire/interview than subsequent ones, i.e. that there is some sort of additional burden associated with the first month or two in sample. Thus it is unwise to rotate respondents rapidly out of and into the sample; they may actually prefer to remain in the sample all the time than to rotate out for short periods. The procedure in current use

for ES-2 survey is a panel rotation scheme. Within each stratum units are allocated to fixed panels, the number of which depends upon the required stratum sampling fraction. 12 panels are in sample at any given time, and at the end of each month one rotates out and another in. The advantages of such a scheme are its simplicity and the advance knowledge of exactly what units will be in sample. There are disadvantages. The design sampling fractions are fixed and cannot be adjusted to account for changing circumstances. New units are systematically allocated to panels, hence a new unit may be assigned to a panel which rotates out of sample one month later. Also, although new units can be systematically allocated, dead units occur at random. Thus in some strata the panels now vary considerably in size. This causes unevenness in the workload of the survey operations staff, and in the estimates.

The scheme proposed for the revised programme is not based on fixed panels. For each province by SIC by size stratum, the required sampling fraction is obtained from the sample size determination and allocation procedure. The numbers of units currently in sample and in the universe are derived from the frame file and the required stratum sample size is thereby determined. The units scheduled to rotate out of sample are identified and the units required to replace them and to produce the desired sample size are selected at random from those out of the sample. This simple strategy is augmented by controls to ensure that units stay in sample at least 12 months, that having rotated out they stay out for at least 12 months, that proper representation of new units occurs, and that changes in sampling fraction as generated by the sample size determination programme are phased in gradually [17]. Of course, in some strata there are insufficient units for any rotation at all.

In designing the rotation scheme no consideration has been given to the possibility of using composite estimation. It is not presently planned to exploit rotation in this manner as experience suggests the

potential gain in effeciency can be offset by increase in complexity and/or bias.

4.6 Data Collection Procedure

The strategy for data collection by Labour Division has been chosen taking into account respondent related problems. Although no quantitative measures have been developed, decisions based on empirical observation and past experience have been made in an attempt to reduce respondent burden.

There is evidence from the Office for Reduction of Paperburden that telephone interviews impose less of a workload than mail questionnaires. Furthermore, Field Test II results suggest that telephone interviews lead to higher response rates and more timely data collection. However the cost of telephone interviews is higher than for mail questionnaires and the number of calls required tends to increase with size of firm. Thus the collection strategy proposed for the revised programme is to obtain data by telephone interviews from Regional Offices for firms of less than, say, p employees. Present p=50. For larger firms mail questionnaires will be used. Smaller firms will be given the option of reporting by mail if they wish. Also, following practices developed in the existing programme, respondents will be able to submit computer produced printouts in place of questionnaires, and to separate executive and other payroll reports if they choose to do so.

Following investigations into the terminology best understood by respondents and the data items readily available to them, two new questionnaires have been designed corresponding to categories of industrial activity. Each one is tailored to the data items likely to be available to the respondent in that particular category. For example, manufacturing firms are asked for an hourly paid/salaried breakdown, whereas educational institutions are asked for a full time/part time division instead.

4.7 Linkage of Existing and Revised Systems

In order to tackle the problem of providing users with continuity of data it has been decided to run existing and revised systems in parallel for a time period of up to one year. At the same time sensitivity to respondent burden has prompted senior management to restrict the number of monthly questionnaires/interviews to one per reporting unit. This means that during the parallel run there will have to be a data exchange between the existing and the revised systems. Furthermore, as data exchange must be at micro level, the universe files for the two systems will have to be held precisely in step. The net result is, in effect, a single system considerably more complicated than either of its components, the price to be paid for meeting user and respondent demands. The complex details of system linkage and data flows have not yet been worked out and await a thorough evaluation of Field Test III results. It seems probable they will give birth to a whole new set of problems.

5. CONCLUDING REMARKS

The objective of this paper has been to indicate the general nature of problems arising during the course of a survey revision programme by citing specific examples. Some of the problems described have been more or less solved, others await satisfactory resolution. One feature most of them seem to have in common is that they are primarily practical, non-textbook, with no neat solutions.

It should not be imagined that all aspects of the development work have been covered. For example there has been no mention of data capture, of locating and correcting data errors, of detecting and allowing for outliers, of imputing for missing values, of estimation, etc. By and large, emphasis has been upon problems associated with the "front-end" of the survey system as this is where much of the work to date has taken place, but, even here, coverage is by no means complete.

RESUME

L'article décrit plusieurs problèmes pratiques portant sur l'adaptation de la théorie de la statistique à l'élaboration des enquêtes, dans le contexte de la révision d'un programme d'enquête sur l'emploi.

REFERENCES

[1] Statistics Canada, "The Labour Force", Catalogue 71-001, Monthly.

[2] Statistics Canada (1970), "Standard Industrial Classification Manual", Catalogue 12-501, 3rd Edition.

[3] Statistics Canada, "Employment, Earnings and Hours", Catalogue 72-002, Monthly.

[4] Statistics Canada, "Organizational Chart", Administrative Services.

[5] Statistics Canada, "Estimates of Employees by Province and Industry", Catalogue 72-008, Monthly.

[6] Statistics Canada, "Federal Government Employment", Catalogue 72-004, Quarterly.

[7] Statistics Canada, "Provincial Government Employment", Catalogue, 72-007, Quarterly.

[8] Statistics Canada, "Local Government Employment", Catalogue 72-009, Quarterly.

[9] Statistics Canada (1977), "Review of the Monthly Employment and Payroll Surveys Programme: Final Report of the Project Objectives Team", Labour Division.

[10] Statistics Canada (1977), "Employment Statistics Development Project: Report of the Design Specification Team", Labour Division.

[11] Cottrell-Boyd, T.M. (1979) "Employment Statistics Development Project: Report No. 6", Labour Division, Statistics Canada.

[12] Cottrell-Boyd, T.M. (1980), "Employment Statistics Development Project: Report No. 7", Labour Division, Statistics Canada.

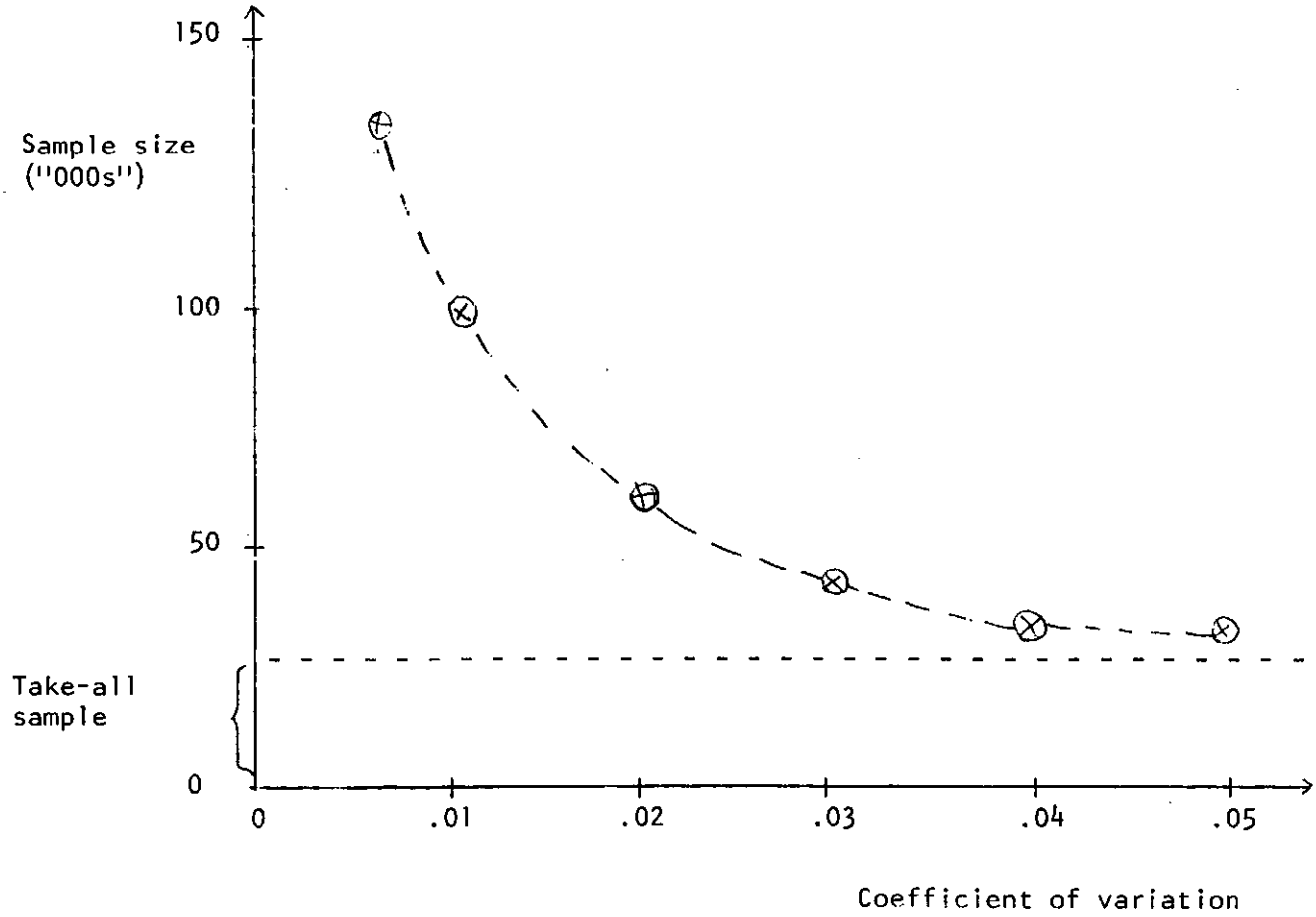
[13] Hubbard, M.R. (1979), "The Statistics Canada Business Register A Review and Discussion", Business Register Division, Statistics Canada.

[14] Jeays, T.M. (1977), "Link 2. A Record Linkage Utility for the Business Register Database, Version 1 Release 2", Systems Development Division, Statistics Canada.

[15] Srinath, K.P. (1979), "Methodological Research for the Monthly Survey", Business Survey Methods Division, Statistics Canada.

[16] Colledge, M.J., and Dinsdale, G. (1980), "Sample Size Determination for Parallel Run: Proposal", Business Survey Methods Division, Statistics Canada.

[17] Maranda, F. and Beckstead, D. (1980), "Specifications for Sample Selection and Rotation", Business Survey Methods Division, Statistics Canada.



GRAPH 1. Total Sample Size for Specified Reliability

(Results from Test version of sample size determination and allocation procedure).

TABLE 1: Distribution of Firms in Canada by Employment Size (approximate figures)

Employment size range	Number of firms
0 - 3	350,000
4 - 19	200,000
20 - 49	20,000
50 - 199	15,000
200 - 499	2,500
500 - 999	1,000
1000 - 1499	300
1500 - 2499	250
2500 - 4999	150
$\geq$ 5000	100
TOTAL	589,300

SAMPLE DESIGN OF THE MONTHLY RESTAURANTS, CATERERS AND TAVERNS SURVEY

M.A. Hidiroglou, R. Bennett, J. Eady and L. Maisonneuve[1]

Statistics on sales of establishments classified as restaurants, caterers and taverns have been collected since 1951. The sample has not been updated for births since 1968 and as a result, it is not representative of the current universe. This paper reports on several methodological aspects of the redesign. The sampling unit, sample design, sample size and allocation, data collection methods, edits and imputations, accumulations and calculations, frame and sample maintenance are described. The new survey will reduce manual procedures wherever possible. Collection, editing, imputation, tabulation and updating procedures will be completely computerized. Data collection will be decentralized and will take place via telephone.

1. INTRODUCTION

Establishments classified under Standard Industrial Classification Code 886 have been surveyed since 1951 through a monthly sample survey. This sample does not completely cover all establishments under SIC 886 since taverns, take-out restaurants, caterers and refreshment stands are excluded. The sample has not been updated for births since 1968 and the sample size presently stands at approximately 300 businesses.

---

[1] M.A. Hidiroglou, Business Survey Methods Division, R. Bennett, Merchandising and Services Division, J. Eady, Systems Development Division and L. Maisonneuve, Regional Operations Division, all of Statistics Canada.

As a result, it is not a representative sample of the current universe of around 30,000 which encompasses all establishments within SIC 886. This sample size fails to satisfy the needs of the Canadian System of National Accounts and the industry itself.

The frame for SIC 886 was updated and surveyed in 1976, 1977 and 1978. This operation was done under the heading of Restaurants, Caterers and Taverns Survey (RCTS). The RCTS ascertained the kind of business (KOB) and total sales of the businesses surveyed. The 1978 RCTS results were used to determine the sample size needed to support reliability criteria, provided by Merchandising and Services Division for a new monthly sample survey. This sample was drawn in December 1979 from an updated 1979 RCTS frame and data collection commenced in January 1980. This new monthly survey will run in parallel with the old monthly for a period of one year at which time the latter will be dropped. This new monthly survey departs in quite a number of ways from the old monthly survey. The old monthly survey was completely manually operated, the manual operations ranging from mailout to tabulation at head office. This new monthly survey will reduce manual procedures wherever possible. Collection, editing, imputation, tabulation and updating procedures will be completely computerized. Data collection will be decentralized and will take place via telephone by the Regional Offices.

Information from this survey will be used by members of the industry to compare their individual growth with that of the Province and in the creation of marketing techniques. It will be used by Federal and Provincial governments in the development of fiscal policies (cost sharing agreements) and to measure and forecast growth in the industry.

Other major users include the Restaurant Association of Canada, universities and tourism associations and agencies which are also involved in the development of the many facets of the food and beverage industry.

This paper reports on several methodological aspects of the redesign and will describe the sampling unit, sample design, the sample size and allocation, the data collection methods, edit and imputations, accumulations and calculations, tabulations, frame and sample maintenance.

2. COVERAGE

Eating and drinking establishments covered in this survey include all known businesses with establishments classified to the Standard Industrial Classification code (SIC) 886. These are broken down into the following KOB classification:

Type of Business	K.O.B. Code
Licenced Restaurants	88601
Unlicenced Restaurants Drive-in-Restaurants	88602
Take-Out Restaurants Refreshment Stands	88603
Caterers	88604
Taverns, Bars and Night Clubs	88605

The survey does not cover eating and drinking places owned by and operated as an integral part of hotels, motels and other accommodation businesses. As well, eating and drinking places classified to non-commercial establishments such as Armed Forces messes, private clubs,

legion branches, or service clubs were excluded from the survey. Finally, all eating and drinking places operated by establishments classified to an industrial sector other than Service Trades (e.g. manufacturing or retail trade) are excluded, since data for these are included in publications produced by surveys covering these areas.

To facilitate the presentation, it is convenient first to define certain terms which will be used throughout the text. Statistics on businesses covered by the Merchandising and Services Division of Statistics Canada are normally presented under one of two standard concepts. One is the location concept, under which every physically separate place of business is classified to its own specific kind of business classification. The other is the establishment concept, whereby the classification is based on the smallest separate accounting entry capable of reporting all elements of basic industrial statistics. Under the establishment concept, the sales activities of two or more business locations, not all of which are necessarily in the same industrial sector or kind of business, may be measured. For the purposes of the Monthly Restaurants, Caterers and Taverns Survey (MRCTS), the establishment concept is used. As a result, the reported data may include business activities at more than one location and/or in more than one kind of business. For establishments which have more than one location, their sales will be reported at the province by kind of business. Hence, the sales for locations belonging to the same province and kind of business will be agglomerated. For SIC 886, Merchandising and Services Division has partitioned the universe into chain and non-chain organizations. An organization operating four or more trading locations under the same legal ownership at the establishment level is defined to be a chain organization. All other organizations not meeting the chain definition are classified as independent organizations.

Total Net Sales and Receipts includes receipts from the sales of meals and lunches, alcoholic beverages and other merchandise. Excluded from these figures are sales taxes collected by businesses for remittance to any government agency and non-operating income such as service and carrying charges on credit accounts, bank interest and interest on investments, rent (Real Estate only) etc.

3. FRAME

The basic strategy for creating the 1979 Restaurants, Caterers and Taverns survey frame was to update the 1978 frame with information provided by the Business Register and the Retail Trade Survey. The Business Register provided businesses coded to SIC 886 having employees. The Business Register which receives its information through Revenue Canada, keeps a record of every business having employees and for whom payroll deductions are made. The Business Register is updated monthly for new businesses, firms going out of business and any name, address or classification changes. The Retail Trade Survey provided businesses having no employees for whom tax deductions are made. These businesses are obtained by using a supplementary area sample list. The area sample list was obtained by selecting a number of areas from a list covering all the regions in Canada with the exception of the Yukon and Northwest Territories, as defined by the Labour Force Survey. Following the 1966 Census of Population, the selected areas were then completely enumerated by field representatives and an area list of all business locations was created. This list was then matched against the main components of the survey frame. Any establishment duplicated on the main file by either a single or multi-location establishment was removed from the area sample list. Every month since the original enumeration, one twelfth of the selected areas has been completely re-enumerated in order to locate any new firms which may have come into existence

during the preceding year. The establishments remaining on the area list represented not only themselves but also similar establishments in the areas not selected for enumeration. A sampling weight, equal to the inverse of the probability that an area containing a particular establishment was selected, is associated with that establishment. The 1979 RCTS frame was therefore made up of units on the 1978 frame which were still active in December 1979, some 200 units from the area component provided by the Retail Trade Survey and some 6000 new establishments provided by the Business Register. These establishments were introduced into the Business Register during the course of the 1979 year. These new establishments were assigned an imputed sales value determined at the province and kind of business level and based on sales values reported for the 1978 RCTS. The sample for the monthly Restaurants, Caterers and Taverns Survey was then drawn from the 1979 RCTS frame. The sample size determination, allocation and selection are described in the section that follows.

4. SAMPLE SIZE DETERMINATION, ALLOCATION AND SELECTION

As was mentioned earlier, the 1979 RCTS frame is made up of chain and independent organizations stratified by province and kind of business. The independents are derived from the Business Register and the Retail Trade Survey. Those independents provided by the Retail Trade Survey are themselves a sample of establishments with no employees. The independents which are provided by the Retail Trade Survey will be termed as area units. These area units and the chain organizations are automatically included in the MRCTS sample. The remaining independents are further substratified within their province and kind of business classification into a take-all and take-some substratum. The overall sample size and allocation to these substrata is done at the province and kind of business level by taking into account precision constraints. These precision constraints are

defined in terms of coefficients of variation at the provincial and kind of business level. The sample size from the MRCTS was determined by using sampling rates derived from the 1978 RCTS frame. The 1978 RCTS was basically a mail survey. A sample of units which had not responded to the mail questionnaire was surveyed by telephone at the Regional Offices. Units which indicated that they did not wish to respond to the survey as a result of either receiving the mail questionnaire or being followed up by the Regional Offices had their sales imputed. This imputation was done at the provincial by kind of business by type of unit (independents originating from the area file or independents originating from the Business Register) level. All the chains responded to the survey and, therefore, none of their sales were imputed. Units which responded to the mail questionnaire were assigned a weight of one as were units which at that time indicated that they did not wish to provide data. Units which responded to the Regional Office follow-up and those that declined to respond to the follow-up were assigned a weight inversely proportional to their selection probabilities. In order to facilitate the development for sample size determination, some notation is defined. Let R_1 be the set consisting of all responding units to the survey, including those that were imputed. Units belonging to this set are used to compute totals. Let R_2 be the set consisting of all responding units to the survey which were not imputed. Units belonging to this set are used to compute sampling variances.

Let

y_{gpkj} = the sales associated with the j^{th} reporting unit in the g^{th} group, p^{th} province and k^{th} KOB,

w_{gpkj} = the weight associated with the above unit,

Q_{gpk} = the number of units in the g^{th} group, p^{th} province and k^{th} KOB which belong to R_1,

M_{gpk} = the number of units in the g^{th} group, p^{th} province and k^{th} KOB which belong to R_2.

Note that the value for y_{gpkj} is zero for units which did not respond either to the mail or to the field follow-up. The weight w_{gpkj} is greater or equal to one for units which were followed up by Regional Office or which belonged to the area universe. For all other units, this weight is equal to one. p ranges from 1 to 12, denoting the ten provinces and the two territories. k ranges from 1 to 5, denoting the five previously defined kind of businesses. g ranges from 1 to 3, denoting the groups. These groups are chains [1], independents originating from the Business Register [2] and area units originating from the Retail Trade Survey.

The next step is to describe how the sample size and allocation are determined so as to satisfy simultaneously the provincial and kind of business requirements. Let c(p,.) and c(.,k) denote the required coefficients of variation for the p^{th} province and k^{th} KOB respectively. The marginal provincial and kind of business totals are:

$$\hat{Y}_{.p..} = \sum_{g=1}^{3} \sum_{k=1}^{5} \sum_{j=1}^{Q_{gpk}} y_{gpkj} \, w_{gpkj} \; ; \quad p=1, 2, \ldots, 12;$$

and

$$\hat{Y}_{..k.} = \sum_{g=1}^{3} \sum_{p=1}^{12} \sum_{j=1}^{Q_{gpk}} y_{gpkj} \, w_{gpkj} \; ; \quad k=1, 2, \ldots, 5.$$

These marginal totals are used to compute KOB within province and province within KOB coefficients of variation. These within coefficients of variation are computed in order that they be equal within the classification of interest. Hence,

$$cw(.,k) = [c(.,k) \quad \hat{Y}_{..k.}] \; / \; (\sum_{p=1}^{12} \hat{Y}^2_{.pk.})^{\frac{1}{2}} \quad ; \; k=1, 2, \ldots, 5;$$

and

$$cw(p,.) = [c(p,.) \quad \hat{Y}_{.p..}] / \; (\sum_{k=1}^{5} \hat{Y}^2_{.pk.})^{\frac{1}{2}} \quad ; \; p=1, 2, \ldots, 12;$$

where

$$\hat{Y}_{.pk.} = \sum_{g=1}^{3} \sum_{j=1}^{0_{gpk}} y_{gpkj} \; w_{gpkj} \; .$$

Using these within coefficients of variation, a compromise first-round coefficient of variation for the p^{th} province and k^{th} KOB is computed as

$$c_o(p,k) = \frac{1}{2} \; [cw(p,.) + cw(.,k)].$$

The first-round coefficients of variation are next revised to satisfy the simultaneous KOB and provincial requirements by using the following iterative formula :

$$c_r(p,k) = \frac{c_{r-1}(p,k) \quad c(.,k) \; c(p\,.) \; \hat{Y}_{..k.} \; \hat{Y}_{.p..}}{\{\sum_{p=1}^{12} (c_{r-1}(p,k) \; \hat{Y}_{.pk.})^2\}^{\frac{1}{2}} \; \{\sum_{k=1}^{5} (c_{r-1}(p,k) \; \hat{Y}_{.pk.})^2\}^{\frac{1}{2}}}$$

where r=1, 2, ..., 10. The iterative process revises the coefficients of variation at the province by KOB level so that they approximate in the best possible way the marginal provincial and KOB coefficients of variation. In our experience, the c_r (p,k) values stabilized within four to five iterations. Let d (p,k) be the chosen c_r (p,k) coefficient of variation by province - KOB. It is then revised to take the area file variability into account. The revised coefficient is

$$dm\ (p,k) = \frac{\{[\hat{Y}_{.pk.}\ d(p,k)\ ]^2 - v_{3pk}\}^{\frac{1}{2}}}{(\hat{Y}_{.pk.} - \hat{Y}_{3pk.})}$$

where

$\hat{Y}_{3pk.}$ = estimated area file total for the $(p,k)^{th}$ stratum,

and

v_{3pk} = estimated area file variance for the $(p,k)^{th}$ stratum.

For each KOB by province classification, the cutoff points for take-all take-some substrata along with the corresponding sample sizes are computed. Note that the dm (p,k) coefficients of variation are used as input. The algorithm which was given in Hidiroglou (1979) is now described in steps. The province and KOB subscripts will be dropped in order to facilitate the presentation.

(a) Within province (p) and KOB (k) and for a given level of precision dm, determine the sample size needed if simple random sampling without replacement were used with no take-all for the non-chain units, as

$$n(0) = \frac{\hat{N}_2^2 \; s_{2,M}^2}{(dm\ \hat{Y})^2 + \hat{N}_2 \; s_{2,M}^2}$$

where

$\hat{N}_2$ = estimated total number of units in the province - KOB excluding chain and area file units,

$$= \sum_{j=1}^{Q_2} w_{2j} \ ,$$

$\hat{Y}$ = total estimated population sales in the province - KOB excluding area file units but including chains,

$$= \sum_{g=1}^{2} \sum_{j=1}^{Q_g} w_{gj} \; y_{gj} \ ,$$

$s_{2,M}^2$ = estimated unweighted variance for the province-KOB, computed over the M units which belong to R_2 and which are not area or chain units.

$$= \frac{1}{M-1} \left\{ \sum_{g=1}^{M} y_{2,j}^2 - \frac{1}{M} \left(\sum_{j=1}^{M} y_{2,j} \right)^2 \right\}.$$

It has been assumed in the above variance formula that the y values are arranged in descending order. The subscript denotes that those units originally belong to the take-some substrata.

(b) Assuming that ℓ units had been included in the take-all substratum, the total number of units in sample for the $(p,k)^{th}$ stratum would be:

$$n(\ell) = \ell + \frac{(\hat{N}_2 - \ell)^2 \; s^2_{2,\, M-\ell}}{(dm\ \hat{Y})^2 + (\hat{N}_2 - \ell)\; s^2_{2,\, M-\ell}} \quad .$$

Note that $s^2_{2,\, M-\ell}$ and $\bar{y}_{2,M-\ell}$, the respective variance and mean of the $(M - \ell)$ smallest independent units can be obtained recursively as:

$$s^2_{2,M-\ell} = \frac{1}{(M-\ell-1)} \{(M-\ell)\; s^2_{2,M-\ell+1} - \frac{(M-\ell+1)}{M-\ell} (y_{2,\ell} - \bar{y}_{2,M-\ell+1})^2\}$$

$$\bar{y}_{2,M-\ell} = \frac{(M-\ell+1)\; \bar{y}_{2,M-\ell+1} - y_{2,\ell}}{(M - \ell)} \quad .$$

The boundary point between take-all and take-some is $y_{2,\ell}$.

The next step is to compute $n(\ell + 1)$. If $n(\ell + 1)$ is less than $n(\ell)$, the process is repeated till the inequality is reversed. That is

$$n\ (a - 1) > n\ (a),$$

and

$$n\ (a) \leq n\ (a + 1).$$

This inequality states that the iterative process is to be stopped. The number of units to be included in the take-all portion is "a" and the number of units to be included in the take-some is $n(a) - a$. The boundary point is then $y_{2,a}$ for the particular province and kind of business under study. The associated sampling rate is $f(a)$ where

$$f(a) = [n\ (a) - a] / (\hat{N}_2 - a).$$

Once all the boundary points and the associated take-some sampling rates had been determined for all the province by kind of business classifications, they were used as input for drawing the sample from the 1979 RCTS. Each province by kind of business classification (stratum) was substratified into chain, area, take-all, and take-some substrata. These substrata were created as follows. Establishments which were designated as chains were assigned to the chain substratum. Establishments which belonged to the area portion were assigned to the area substratum. All other establishments, the independents, were assigned to the take-all substratum if their sales exceeded the prescribed cutoff value $y_{2,a}$; otherwise, they were assigned to the take-some substratum. The process for selecting the sample within these province by kind of business strata was as follows. All establishments belonging to the chain, area and take-all substrata were selected with certainty. The take-some substratum was sampled at the rate f(a) using simple random sampling.

5. EDITS AND IMPUTATIONS

This survey will have the sales data collected at the eight regional offices by telephone. Preprinted data sheets will be used to record data provided by firms participating in the MRCTS. Each data sheet will have spaces to record up to three sales figures for firms having a multiple kind of business or having operations in more than one province. Supplementary data sheets are available if more than three spaces are required. The information that is on each preprinted sheet is the Business Register Identification (BRID) number, the name and address, telephone number, province and kind of business and the associated number of locations. If any of this information is not correct, revisions will be made on the data sheet. The collected information will then be captured at the regional office and subject to some simple edits which will have been programmed for the mini-

computer. Records that do not pass these simple edits will be rejected. The operator will then have to pass back the questionnaire to the interviewer for correction. Those edits will insure that numeric fields designated as such are numeric, that they are of the correct length and that they either satisfy a range or are equal to some prespecified values. Records which have passed these simple edits are then transmitted to head office and transformed into transaction records. The Regional Office then acts as an emittor of data and head office as a receptor of data. The information on these transaction records is passed on to the sample file by matching on BRID number. Regional Office interviewers will also send a status code for each sales value obtained. These status codes which are given in Appendix A are used to categorize the response status of in-sample units and to identify units which must be edited or imputed.

It should be noted that most units are expected to report their sales data on a monthly basis. There are, however, some units which can only report on an annual, quarterly or even a thirteen periods basis. For units which can only report on a yearly basis, the monthly imputed value will be the previous year's sales values divided by the number of months in operation. The same type of imputation will hold for units which can only report on a quarterly basis. The imputation required for units which can only report on a thirteen period basis is more complex. Each of the thirteen periods corresponds to approximately twenty-eight days. The opening and closing dates for the thirteen periods vary from business to business. Hence, the opening and closing dates for one of these periods could be found either in the same month or in two consecutive months. Initially the sales associated with the thirteen period units are prorated by the ratio of the number of days in the month of interest to the number of days in the period which has a closing date in that month. These imputed sales, which are preliminary, are computed provided that the closing date is at least ten days

into the month. Otherwise, the sales are imputed using a ratio trend applied to the previous month's sales. Revised figures for the month are computed whenever sales are available for consecutive periods which have a closing date and opening date in that month. These revised sales are a function of the number of days in the month, the closing date and sales for the period which ends in that month and the opening date and sales for the next period which starts in that month. In most instances, the union of two consecutive periods will completely overlap one month and hence the associated revised sale will be a function of the data linked with those two periods. There are, however, some months of the year for which two consecutive periods have their closing date in that month and for which the next period has its opening date in that same month. For these cases, the revised sales are a function of the data associated with those three periods. An algorithm has been developed for the aforementioned cases and cases which can be more complicated (such as changing response status from one month to the next).

5.1 Editing the Monthly Data

The edits will be applied to records which have specific status codes chosen by the subject matter user. These status codes will be supplied by province, KOB and type of unit. This flexibility will allow for changes or expansion for the status code list as the survey progresses. A monthly trend edit will be computed by province, KOB and type of classification. This edit will be used to identify units whose current to previous period trend differs significantly from the general trend with the edited cell. The edited cell is that level of province, KOB and group aggregation for which edits are computed. For instance, before any collapsing occurs, the basic edit cell is the province by KOB and group classification, where the groups are chains, take-alls,

take-somes and area units. This collapsing occurs if there are not enough units within an edited cell to compute a stable mean and variance of the monthly trends. For a given group and kind of business, the collapsing is done across specified provinces. There are four levels of collapsing. The higher the level, the more provinces are included to participate. The choice of province combinations for collapsing purposes is based on similarity of annual sales means for those provinces, groups and kind of business classification. These annual sales means were computed using the 1978 RCTS results. Note that the collapsing will be done automatically if need be and that the level of collapsing will be determined by a preset minimum number of units needed to compute the means and variances. Note that some units will not be edited. These are births, deaths, non-respondents or those units becoming active or inactive on account of the seasonal nature of the operations. The notation defined below does not differ much from the one given in section 4. The symbol t is added to indicate that there now exists a time element (month and year).

$N_{gpk}(t)$ = number of units in the population at time t, in the g^{th} group, p^{th} province and k^{th} KOB,

$m_{gpk}(t)$ = number of units at time t which are used to compute the trend, allowing for collapsing, for the g^{th} group, p^{th} province and k^{th} KOB,

$n_{gpk}(t)$ = number of units in sample at time t, in the g^{th} group, p^{th} province and k^{th} KOB,

$n'_{gpk}(t)$ = number of unimputed units in $n_{gpk}(t)$,

$w_{gpkj}(t)$ = weight at time t associated with the j^{th} unit in the g^{th} group, p^{th} province and k^{th} KOB, (note that stratum jumper weights are automatically counted under this notation, see page 76).

$y_{gpkj}(t)$ = sales value at time t for the j^{th} unit in the g^{th} group, p^{th} province and k^{th} KOB,

$L_{gpkj}(t)$ = number of locations at time t for the j^{th} unit in the g^{th} group, p^{th} province and k^{th} KOB.

The ranges of g, p and k have been defined in the previous section. Note that j ranges from 1 to $n_{gpk}(t)$. The current month to previous month trends are computed as

$$r_{gpkj}(t) = \frac{y_{gpkj}(t)\ L_{gpkj}(t-1)}{y_{gpkj}(t-1)\ L_{gpkj}(t)} .$$

The mean and standard deviation of these ratios are used to construct intervals of tolerance. If the computed ratios lie within the intervals, they are said to have passed the ratio edit, otherwise they are flagged as a possible outlier. The span of these intervals is controlled by specifying a constant which determines how many standard deviations are tolerable on each side of the mean. Units declared as outliers will have their sales values imputed as described in section 5.2.

5.2 Imputation for Non-response and Outliers

Some of the imputations used in this survey have been mentioned earlier. These are imputing for number of locations, if need be, and prorating of sales figures for units which have a reporting period other than monthly. The imputation procedure operates by multiplying a non-responding unit's previous month's data by a measure of trend computed from responding units (excluding outliers) whose business characteristics are similar. Note that similar units may be obtained by collapsing if there are not enough data to compute trends. The collapsing will be across provinces by KOB and type of business.

During the initial month of the MRCTS, the averaged monthly sales (total yearly sales divided by twelve) from the 1979 Annual RCTS will be used as the base values to impute missing data. As the survey progresses, these imputed values should eventually be replaced by the monthly responses. The trend will initially be computed at the province, KOB and group level. The trend will be computed as

$$r_{gpk}(t) = \frac{\sum_{j=1}^{m_{gpk}(t-1)} y_{gpkj}(t)\ w_{gpkj}(t)}{\sum_{j=1}^{m_{gpk}(t-1)} y_{gpkj}(t-1)\ w_{gpkj}(t-1)} .$$

Note that the range is up to $m_{gpk}(t-1)$ for both numerator and denominator. The reason for this is that only pairs $y_{gpkj}(t)$, $y_{gpkj}(t-1)$ with valid information are entered into the numerator and denominator. Excluded from the trend computations are units which have not passed the edit test, or which have a zero sales value for month t-1. The number of units used in the computation of the ratio trend will be increased automatically by collapsing cells if there are not enough

data points. As with the editing collapsing, there are four levels to which cells may be collapsed. It is hoped that the nature of data collection for this survey will greatly lower the need for a large number of imputations. The non-responding units are then imputed by multiplying the above trend by its previous value. If no previous value exists, then the mean of the imputation cell will be used. The imputation cell is understood to mean a level at which the KOB and group cross-classification may be collapsed on the basis of province. Units declared as outliers will have their imputed value and their actual value printed out for review by subject-matter specialists. The appropriate Regional Office will then be asked to confirm the sales value for the unit declared as outlier. If the reported sales value is confirmed the reported value will be chosen. If the same unit is declared an outlier for several succeeding months, its weight will be changed to one and it will therefore become part of the take-all substratum.

6. ESTIMATION AND SAMPLING VARIANCES.

Once the data have been edited and necessary imputations carried out, the sample file is ready for aggregation. The basic building block for the MRCTS estimation process is the province, KOB and group type. Revisions of estimates are possible because Regional Operations Division are able to update data for the month previous to the reference month.

The notation is the same as that given in Section 5.1. The basic estimate of total for any province, KOB and group type classification is

$$\hat{Y}_{gpk}(t) = \sum_{j=1}^{n_{gpk}(t)} y_{gpkj}(t) \quad w_{gpkj}(t) \ .$$

Note that some of the y values may have been imputed. The associated variance will be given by

$$v(\hat{Y}_{gpk}(t)) = (1 - \bar{f}_{gpk}(t))\ n'_{gpk}(t)\ s^2_{gpk}(t)$$

and $s^2_{gpk}(t)$ will not include imputed sales. This variance is computed as

$$s^2_{gpk}(t) = \frac{1}{n'_{gpk}(t) - 1} \{ \sum_{j=1}^{n'_{gpk}(t)} s^2_{gpkj}(t) - n'_{gpk}(t)\ \bar{z}^2_{gpk}(t) \}$$

where

$$z_{gpkj}(t) = y_{gpkj}(t)\ w_{gpkj}(t)$$

and

$$\bar{z}_{gpk}(t) = = \frac{1}{n'_{gpk}(t)} \sum_{j=1}^{n'_{gpk}(t)} z_{gpkj}(t) .$$

This variance allows for different weights within the same group type, province and KOB classification. Units within those classifications may have different weights on account of stratum jumpers. These stratum jumpers are units which change their kind of business during the course of the survey. They retain their original weight after changing KOB. The average correction factor is computed as

$$\bar{f}_{gpk}(t) = \left[\frac{1}{n_{gpk}(t)} \sum_{j=1}^{n_{gpk}(t)} w_{gpkj}(t)\right]^{-1} .$$

The variances associated with take-all or chain units are automatically zero. Once the estimation at the group type, province and KOB has been completed, estimation for higher levels of aggregation is additive over the proper sets. Furthermore, variance estimation for higher levels is also an additive operation which parallels the estimation process.

7. FRAME AND SAMPLE MAINTENANCE

The frame and the sample will be maintained with respect to births, deaths and amendments. This maintenance will be using as source the Business Register, the Retail Trade Survey area updating forms and information provided by the Regional Operations Division concerning the status of sampled establishments. The Business Register will provide births, deaths and amendments to the frame on a monthly basis. These updates will be for establishments which have employees. A sample of new establishments without employees will be provided on a monthly basis by the Retail Trade Survey. The Regional Operations Division will identify for head office establishments which are amended or are deathed. Amendments to sampled establishments can be changes in name, kind of business, Standard Industrial Classification or province. Deaths in the sample will be identified as establishments which have completely ceased their business operations. Amendments and deaths provided by the Regional Operations Division will be forwarded to the Business Register after the sample has been updated for these changes.

Updates originating from the Business Register are now discussed in greater detail since they will generate the bulk of changes to our frame and sample. These updates will be provided by the Current Update

File on monthly basis. This file is a record of all changes made to the Business Register in any operating cycle. It is essentially a file which contains a pair of records for every record that was added, updated or deleted in that cycle. The first record shows what was present before the change (blank, except for the Business Register Identification in the case of births) and the second record shows what was present after the change (blank, except for the Business Register Identification in the case of deletions). The Current Update File records will be put into three possible classifications. These are births, deaths and amendments. These changes will be applied to both the frame and the sample. Note that the Current Update File will be updated as well with information provided by the Regional Operations Division. These updates will be kind of business, Standard Industrial Classification, province or name and address changes. Amendments provided by the Current Update File refer to any change in information. The following changes will be considered as amendments: form of organization, survey status, activity status, registration status, kind of business, province, name and address, payroll deduction account number or telephone number. Amendments will be applied to establishments on the frame that match the Current Update File establishments. Births will be provided by the portion of the Current Update File which has a new Business Register Identification or a change of SIC to SIC 886. All establishments coded to SIC 886 will be included on the frame regardless of the SIC to which the head office may be classified. SIC 886 is a dynamic universe. Firms are constantly coming into existence, going out of business, merging with other businesses or splitting into two or more operations. This implies that these status changes should be very much kept up to date on both the sample and frame. It is for this reason that certain establishments should possibly be deactivated from the sample and master frame if they are found to be duplicates of incoming births. Finding these duplications is difficult for establishments on the frame which are not part of the sample. Deaths for these establishments will be

drawn to the attention of the Business Register at a later time. Note that it is only for establishments in the sample that the deaths can be quickly passed on to the Business Register and updated on the frame. Hence, for those establishments there will be little duplication on the frame.

The MRCTS is a survey which can be described as having a multiple frame, i.e. an area frame and a Business Register frame. To accommodate this multiple frame feature, weighting techniques such as the ones described in Hartley [1] could have been used. Since the area frame is small compared in terms of relative number of establishments with respect to the overall frame, the mutiple frame problem will be dealt with by unduplicating the two frames. The unduplication of the frames will be implemented by matching the area sample units to the Business Register derived units on a monthly basis. Any business establishment in the area list which is matched to active Business Register records updated via the Current Update File will be deactivated from the frame and sample. The corresponding Business Register records will then be activated.

The formal process including births from the Current Update File records onto the frame and sample is several fold. Births provided by the Current Update File will first be examined for completeness of information such as name and address, province, regional office code and kind of business. These records which are judged to be in scope and complete will then be added to the frame. Births whose sales values exceed the cutoff value associated with the birth's province and kind of business will be automatically introduced into the sample. Units which have sales values lower than the cutoff will be sampled systematically at a rate prespecified for the particular province and kind of business combination. The establishments with no employees which are provided by the Retail Trade Survey will also be examined for the completeness of information. These area records which are complete and in scope will then be added to both the sample and frame.

8. CONCLUSIONS

The results of the Monthly Restaurants, Caterers and Taverns Survey will be published on a monthly basis. It is anticipated that preliminary estimates will be out six weeks following the month of reference. The final estimates will be published four weeks later. The tabulations will be at the Canada level by kind of business, by provinces, and by chains versus independents for Canada and the provinces. The estimated change from current year to previous year will also be published for the mentioned tabulation levels after sufficient time has elapsed to permit these calculations. Estimates of the coefficients of variation will be tabulated but will most likely be used for internal purposes.

There are several possible problem areas which will have to be closely monitored in order to ensure integrity of this survey. These are universe deterioration, response rates and measurement error. The universe deterioration is first discussed. SIC 886 is one of the most volatile SIC's. Restaurants constantly go into and out of business within a short time span. It is for this reason that it is important to have an updating procedure which keeps up with the changes. For every birth that comes into the survey, it is therefore important to identify, if possible, the name or owner under which the unit was previously known. Establishing such links, if they exist, will reduce duplication in the frame and slow down its deterioration. These links may be established by Nature of Business reports and by using the Merchandising Divisional Master File's updating procedures. The updating procedures contain programs which can establish possible links between births and units already on the file. Another source of deterioration for the frame will be changes of KOB for units not in the sample Indeed, experience has shown that on the average there can be a 20% shift in KOB designation within one year. On a continuing basis all births will be contacted to obtain the kind of business before being placed on the frame.

Also, yearly sales attached with units not in sample will deteriorate over time on account of inflation. The effect will be that the cutoff boundaries established for the first year will have to be revised on the basis of sample results.

The response rate for this survey is anticipated to be high for a business survey, around 90%, as a result of the collection procedures utilized. Head Office will advise each Regional Office of the monthly response rate.

Another beneficial effect for field collection is that updates, be they name and address or changes of units into more or less KOB's, will be closely monitored for the sampled portion of the universe.

As Wolter et al [3] have reported, one potential source of bias for this type of survey will be non-sampling errors. Respondents who do not have book figures will provide an estimate of their sales. Awareness of response errors of that type can be brought to attention by keeping track of revisions to monthly sales.

ACKNOWLEDGEMENT

The authors are grateful to the referee for comments which have much improved this paper.

RESUME

Depuis 1951, on a cueilli des données sur les ventes des établissements qui sont classés comme restaurants, traiteurs et tavernes. La base d'échantillonnage n'a pas été mise à jour, en ce qui concerne les créations, depuis 1968; par conséquent elle ne réprésente pas bien la population actuelle. Cet article indique quelques concepts méthodologiques de la révision de cette base. L'unité d'échantillonnage, le plan de sondage, la taille et l'allocation de l'échantillon, les méthodes de la cueillette des données, la contrôle et l'imputation, les accumulations et les calculs, l'entretien de la base et de l'échantillon sont tous décrits. La nouvelle enquête réduira les operations manuelles dans la mesure du possible. Les procédures de cueillette, de contrôle, d'imputation, de totalisation et de mise à jour seront entièrement informatisées. La cueillette des données sera décentralisée et sera faite par téléphone.

REFERENCES

[1] Hartley, H.O. (1962), "Multiple Frame Surveys", Proceedings of the Social Statistics Section, American Statistical Association.

[2] Hidiroglou, M.A. (1979), "On the Inclusion of Large Units in Simple Random Sampling", Proceedings of the Survey Research Methods Section, American Statistical Association.

[3] Wolter, Kirk M.D., Isaki, Cary T., Sturdevant, Tyler R., Monsour, Nash J., and Mayes, Fred M. (1976), "Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys", Proceedings of the Business and Economic Statistics Section, American Statistical Association.

APPENDIX A

Provincial Codes

10	Newfoundland
11	P.E.I.
12	Nova Scotia
13	New Brunswick
24	Quebec
35	Ontario
46	Manitoba
47	Saskatchewan
48	Alberta
59	British Columbia
60	Yukon
61	N.W.T.

Status Codes

01	Acceptance
02	Estimate by Respondent
03	Respondent Unable to Estimate
04	Refusal
05	Closed -Out-of- Business
06	Increase due to Newly Acquired Liquor Licence
07	Increase/Decrease due to Special Events, Conventions, Seasonal Fluctuations, &c.
08	Increase/Decrease due to change in number of locations
09	Decrease due to temporary closure (Renovations, Fire, &c) or Increase following Re-opening
10	Closed - Seasonal
11	Out of Scope

REVERSE RECORD CHECK: TRACING PEOPLE IN CANADA[1]

J. -F. Gosselin[2]

The Reverse Record Check is the main vehicle used to assess the level of undercoverage in the Canadian Census of Population. A sample of persons is selected from sources independent of the current census and extensive tracing operations are undertaken to determine the usual address of each selected person as of Census day. Census records are then checked to determine whether or not each selected person was enumerated. The tracing is by far the most complex, costly and time-consuming operation associated with this study. It involves extensive use of administrative records as well as tracing in the field. This paper describes the various tracing methods used as well as the success obtained from each of them.

1. INTRODUCTION

The Reverse Record Check (RRC) method is generally recognized as one of the best procedures to evaluate the level of undercoverage in the Canadian Census. A frame containing all persons who should be enumerated in the current census is built up from the returns of the

---

[1] Adapted from a paper presented at the annual meeting of the American Statistical Association held in Houston, August 1980.

[2] J. -F. Gosselin, Census and Household Survey Methods Division, Statistics Canada.

previous census and intercensal birth and immigration registrations. A random sample is selected from each source and each selected person is traced to his/her current census address. Census documents are then checked to determine whether or not the selected person was enumerated.

The main advantage of this method lies in the fact that it does not involve any form of re-enumeration which generally leads to underestimates of coverage errors because of the strong tendency for persons missed in the census also to be missed in the re-enumeration process. However, one potential problem with this method is that there will always be a nucleus of selected persons who cannot be traced to their current census address and for which the enumeration status cannot be determined. Since the level of undercoverage among these persons is probably higher than average, it becomes extremely important to keep this group as small as possible in order to reduce the potential bias in the undercoverage estimates (this is a form of non-response).

The RRC method has been used successfully in Canada and is the main vehicle for assessing the level of undercoverage in the current Censuses of Population and Housing. It was first introduced in the 1961 Census using a small sample selected from the previous census and from which only national estimates could be produced reliably. In 1966 the method was used on a larger scale with a sample size of about 26,100 persons which included a sample of intercensal births and immigrants, and a sample of persons missed in the 1961 Census.

Both the 1971 and 1976 Reverse Record Checks were developed along lines very similar to the 1966 Check with improvements being made to the design and methodology. The main difference in 1976 was a further increase in and a redistribution of the sample to allow population undercoverage to be estimated at the province level.

Currently, a study of similar size is being planned for 1981. Reports on the 1966, 1971 and 1976 RRC may be found in [1], [2], [3] and [4]. In addition, [5] gives a comprehensive report on the 1976 Coverage Measurement Programme.

This method has been used successfully in Canada because of our ability to keep the proportion of untraced persons at a very low level through extensive use of administrative records as well as follow-up from the regions. The tracing operations are by far the most complex, costly and time-consuming operations associated with this study. This paper describes the various tracing methods used as well as the success obtained from each of them. A brief description of the methodology of the study will first be presented.

2. OVERVIEW OF THE METHODOLOGY

This study involves five major steps:

i. The construction of a frame of persons who should be enumerated in the census, based on sources independent of the current census.
ii. The selection of a random sample from each of these sources.
iii. The tracing of selected persons to determine the address of their usual place of residence on census day.
iv. The searching of census forms to determine whether or not the selected persons were enumerated at the address traced and a follow-up operation for cases not found.
v. The weighting of the sample data and the production of final results.

Each will now be discussed briefly.
The frame is constructed from four different sources (also called frames):

Census frame:	all persons enumerated at their usual place of residence in the previous census;
Birth frame:	all intercensal births;
Immigrant frame:	all intercensal immigrants;
Missed frame:	all persons missed in the previous census.

For the first three frames, records are available from which a sample can be selected. For the missed frame, no exhaustive list is available. However, those persons classified as missed in the census by the previous RRC are taken as a random sample from this frame.

These frames combined cover practically all persons who should be enumerated in the census. The major groups which are not covered are illegal immigrants, and persons missed in the previous census and not given a chance of selection in all of the previous RRCs. The latter group is probably becoming very small as more RRCs are conducted. The frame also includes some persons who legitimately could not be enumerated in the census, essentially emigrants and deaths since the previous census. However, these persons can be eliminated from the sample during the tracing operation and hence do not bias the results.

The sample design varies from frame to frame depending mostly on the nature of the lists or records available. The 1976 sample sizes are given in Table 8. Further details may be found in [4].

Since the address obtained at the time of selection is usually out of date, a tracing operation must be undertaken to determine the address of each selected person (SP) as of Census day. The tracing

methods used vary from frame to frame and include: registered letters sent to the last known address; searches of administrative files such as Family Allowance and Old Age Security records; telephone and field tracing conducted from the Regional Offices.

All cases traced to a possible census address undergo a Head Office searching operation whereby the current census documents are checked to determine whether or not the selected person has been enumerated at the address at which he/she was traced. Those who are found are automatically classified as 'Enumerated' and are considered finalized. For those cases not found, a follow-up operation is undertaken from Regional Offices whereby each selected person is contacted to verify his/her address as of Census day or to obtain other possible addresses, and to collect some basic data on characteristics of those missed in the census.

The sample data are then weighted to produce final estimates of population and household undercoverage.

3. TRACING OPERATIONS

In order to update addresses, a tracing operation is undertaken to determine the place of residence of each SP on Census day. The tracing system used consists of a series of individual operations or 'traces', carried out sequentially in a pre-determined order (i.e., cases not traced at a given stage undergo the next stage of tracing). The system used in 1976 was a slight variation of the one associated with the 1971 study. The tracing methods can be classified into three general types:

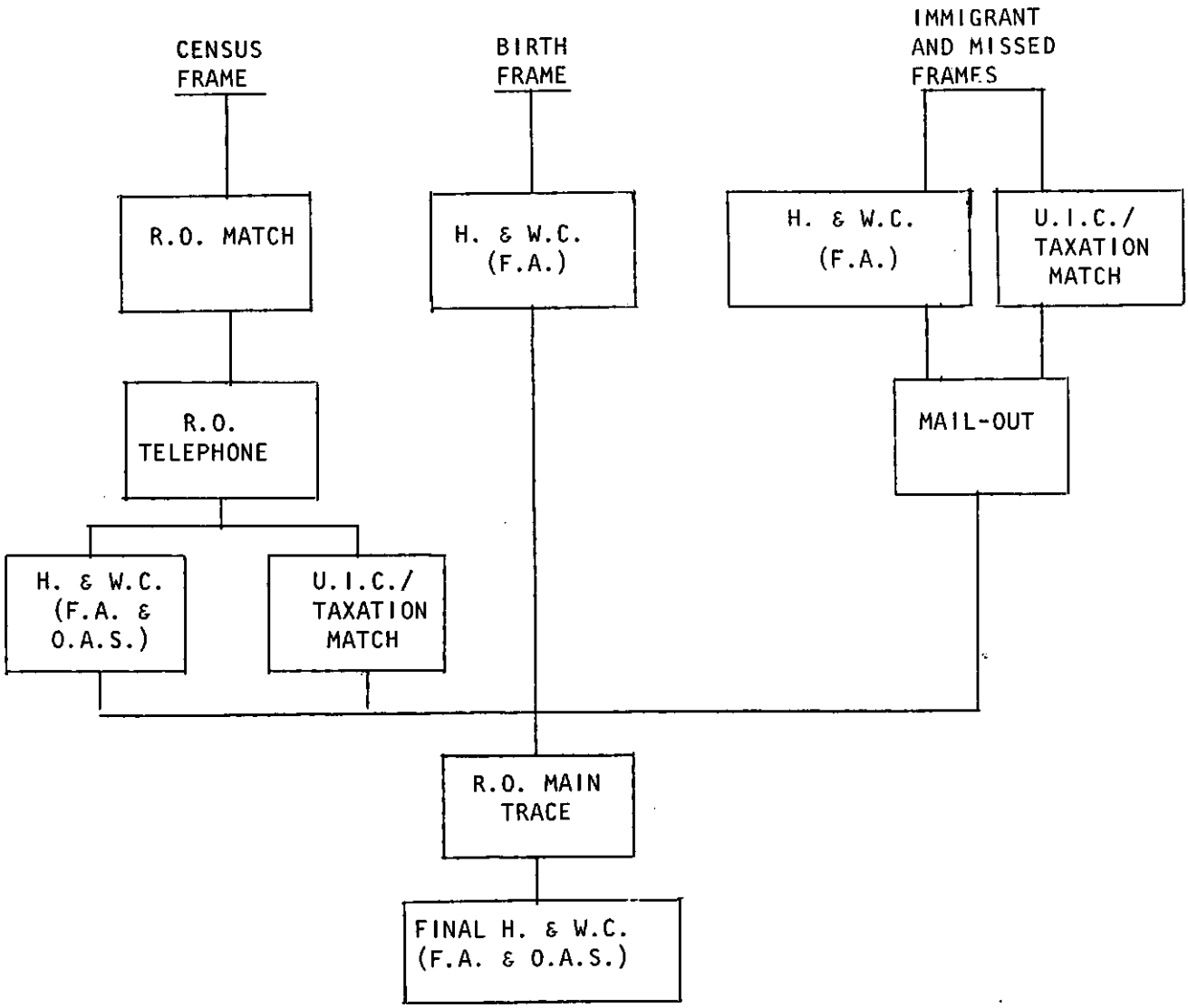
i. Operations that are conducted in or directed from Statistics Canada's eight Regional Offices (RO). These include a match with current census records built into the census Regional Office Processing operations (RO Match), as well as telephone and intensive field traces (Telephone Trace and Main RO Trace).

ii. Operations that make use of administrative records such as Health and Welfare Canada's Family Allowance and Old Age Security records, Unemployment Insurance Commission records, and Revenue Canada/Taxation records.

iii. Mail-out operations from Head Office.

As a general rule, attempts are made to maximize the use of existing administrative records because to do so often eliminates the need to contact the selected person and in general is much cheaper. However, complete and up-to-date records unfortunately tend to exist only for certain special groups of the population such as persons 65 or over. Hence there was a need to resort to the other two types of operations noted above.

DIAGRAM 1: THE 1976 REVERSE RECORD CHECK TRACING SYSTEM



H. & W.C. : Health and Welfare Canada
F.A. : Family Allowance
O.A.S. : Old Age Security
U.I.C. : Unemployment Insurance Commission

Diagram 1 illustrates the sequence of the various tracing operations for each frame. Individual traces will now be discussed.

3.1 R.O. Match

The first tracing operation undertaken for the census frame was the Regional Office Match. Its purpose was to look in the 1976 census records for nearly 28,000 persons selected from the 1971 census returns, to see whether they were still residing at, and had been enumerated at, their 1971 census address.

This operation was built into the census RO Processing operations in order to provide early access to census forms. In broad terms, this operation consisted of a search of the 1976 census records for those enumeration areas (EA) corresponding to the selected 1971 EAs to locate the selected persons.

As a result of this operation, cases were classified into five categories. Table 1, which presents these categories, indicates that 42% of the census frame sample was matched during this operation, which makes it very effective. This is particularly so, given that this operation made use of current census records, which eliminates the need for any further searching in Head Office as was required for other tracing operations. The small number of cases in category 2 had to be later re-checked since the 1971 information could not be matched exactly with the 1976 data (e.g., different initial or a small difference in the year of birth). Most of these were later re-classified as enumerated.

3.2 R.O. Telephone Trace

The purpose of the Telephone Trace was to contact selected persons who were not located during the RO Match or to obtain leads as to

the whereabouts of SPs through the use of the telephone from each Regional Office. Specifically the cases for which this operation was undertaken correspond to SPs classified in categories 3, 4 or 5 in the RO Match (see Table 1).

For cases where the address was found but not the SP (category 3), a telephone call was made to the number at that address (which was noted during the RO Match from the census form) to get information on the whereabouts of the SP. If this was unsuccessful, an attempt was made to locate a telephone number by searching telephone directories either for the SP or for the head of his/her household in 1971. This latter procedure was also used for cases classified in category 4 or 5 in the RO Match. All leads provided by this initial contact were followed-up by telephone unless they were outside the Regional Office area jurisdiction.

Over 15,500 cases were sent to the Telephone Trace and the results are summarized in Table 2. Almost 59% of these cases were traced while a lead was obtained for an additional 32.5%. These were used in subsequent phases of tracing. In only about 9% did the Telephone Trace lead to a dead end. The Telephone Trace success rates for categories 3, 4, and 5 are shown in Table 3. It is worth noting that in the 1971 Reverse Record Check, the Telephone Trace was only undertaken for cases in category 3 in the R.O. Match. However, considering the tracing rate obtained, the extension of the operation to categories 4 and 5 was an overall success.

3.3 Tracing Using Health and Welfare Canada Records

Health and Welfare Canada (H. & W.C.) administers two programmes whereby payments are made to families with children aged 0 to 18 and to persons 65 years or over. These programmes are called Family Allowance (FA) and Old Age Security (OAS) respectively. Because

monthly payments are involved, one would expect these files to be relatively complete and up-to-date for the population that is covered by these programmes and hence these records are ideal for tracing purposes.

These files were used in the tracing of persons selected from all frames but at different stages of the operation. For the census frame, about 4,100 cases were sent to H. &W.C. These were cases which had not been located by the Telephone Trace and for which the selected persons either themselves were eligible for FA or OAS, or whose family contained an eligible member in the 1976 Census. A search was undertaken only in the last known province of residence and produced a success rate of slightly over 50% (see Table 4).

A very similar procedure was carried out for the immigrant (539 cases) and the missed (177 cases) frames with the exception that for these frames this search was actually carried out earlier, around census time using FA records only. The success rates obtained for these frames were about 64% and 75% respectively.

For the birth frame, the FA records were the primary source of tracing information. This was carried out in two phases. In the first phase, a search of records in the province of birth was undertaken. The few cases that were not found during the first phase were sent to neighbouring provinces. This was the second phase and was intended to cover cases that either had changed province before the census or for which the province of birth and residence differed. As one might expect the tracing rate for births was over 90%.

The overall success rate of this operation was 68%. One problem that was experienced was the presence of a relatively large number of cases (13%) which were found on the records that could not be considered as traced since the information provided was a postal

address such as rural route or postal box number. This type of address could not be used for searching purposes, although they were used as leads in further stages of tracing.

3.4 Matching with Unemployment Insurance Commission and Revenue Canada/ Taxation Records

The Unemployment Insurance Commission (UIC) maintains records of persons who have applied for a social insurance number (SIN) which is required for anyone entering the work force. Although this file is not kept up to date it was felt that the information could at least be used as a starting point particularly for immigrants for whom no precise address is available at time of selection. In addition this had the added feature of providing a SIN which could then be used to match with taxation records to obtain a more recent address.

The following cases were sent to UIC where a manual search of their records was undertaken:

(a) all persons selected from the immigrant and missed frames
(b) persons from the census frame not traced by the Telephone Trace and who did not qualify for the Health and Welfare Canada searches.

For those cases found, a computer match was carried out with the 1974 taxation records (i.e., the latest available file at that time) using the SIN obtained from UIC.

The results are summarized in Table 5. These indicate that a UIC record was found for nearly 75% of cases sent. Of those, more than 2/3 could be found on the taxation files giving an early 1975 address. Since neither source gave addresses which were current to the census, cases found on either of these records were not

considered traced. Rather, the information was used as leads in subsequent tracing stages (e.g. in a mail-out operation).

3.5 Mail-out

A Mail-out operation was undertaken for persons selected from the immigrant and missed frames who were not traced by Health and Welfare Canada. This was undertaken soon after Census day and involved about 1600 persons.

This operation was carried out in two phases. For the first phase, a registered letter was sent to the last known address. For a large proportion of cases, this address had been obtained from the UIC/ Taxation match, but in some cases it was the reported intended address at time of immigration. The latter was often very unreliable. The second phase of mailing was undertaken four weeks after the first mail-out. Reminder letters (also registered) were sent to persons who did not respond to the first letter. Also there were letters sent on the first mail-out that were undeliverable and therefore were returned by the Post Office. Of those which had a proper mailing address, a letter was sent to the householder at the address in an attempt to determine the whereabouts of the selected person.

The results of this operation are summarized in Table 6. The success rates for the initial and reminder letters were about 47% and 37% respectively. The householder letter produced rather poor results with less than 3% traced.

When taken as a proportion of the number of persons involved in this operation, the reminder letter added 8% of cases traced to the initial 47% produced by the first letter while the householder letter produced less than 1% cases traced. This resulted in an overall success rate of about 56% which is a definite improvement over the same operation

in 1971. This may be attributable to the introduction of the use of taxation records.

It should be noted that a Mail-out operation was also used for a very small number of cases from the census and birth frames where postal addresses were obtained from the Health and Welfare trace.

3.6 Main Regional Office Trace

The Main Regional Office Trace was an extensive telephone and field tracing operation carried out from the Regional Offices for about 5,500 cases from all frames not previously traced. This was the first tracing procedure common to all frames.

First, attempts were made to resolve most of the cases by telephone contact using city and local telephone directories to locate selected persons, or possible acquaintances, in order to obtain leads. Whenever possible, all leads were followed up by telephone. As a last resort, problem cases were sent to the field where all leads were to be explored.

In addition to relatives and neighbours, the following were used as sources of information: former landlords or employers, school or university files, social clubs, union offices, files of various government departments and agencies.

Nearly 80% of cases were traced by this operation. The variation by frames is shown in Table 7. This operation was very successful considering that, by that stage, we were down to a relatively small nucleus of persons for whom other means of tracing had already failed. It is also quite expensive as one might expect. Although no detailed cost figure could be obtained from the 1976 operation separate from other RRC field operations, the cost based on 1971 experience was probably of the order of $15-$20 per case sent.

3.7 Final Health and Welfare Canada Trace

This was the final tracing operation which involved cases not traced by the Main RO Trace and which were eligible for a FA or OAS search. The procedures were identical to the Main Health and Welfare trace (see section 3.3) with the exception that cases were sent not only to the province of their last known address but also to neighbouring provinces. A total of about 650 cases were involved in this trace of which about 42% were traced.

3.8 Overall Results of the Tracing

The overall results of the tracing operations are given in Table 8. The initial tracing rate was 96.4% and this rate ranged from a low of 90.2% for the immigrant frame to a high of 96.9% for the census frame.

This table also indicates the percentage of the original sample traced at various stages of the tracing operations. For the census frame, about 75% of the sample was traced in the R.O. Match or the Telephone Trace, while for the birth frame, almost 89% were located through the Health and Welfare search. For the other two frames, the mail-out operation was the most effective in tracing about 45% of the original sample.

As mentioned earlier, once a selected person was traced, a search of census documents was undertaken to determine whether or not he/she was enumerated at that address. Cases not found underwent a follow-up operation to verify their census address. As one might expect there were a number of cases that were traced but could not be contacted again during follow-up. These cases, albeit small in number, were re-classified as 'Untraced' which explain the fact that the final tracing rate was lower than the percentage initially traced

as shown in Table 8. The percentage distribution of the sample in the final status categories is presented in Table 9.

4. SUMMARY

The Reverse Record Check method has been used successfully in Canada to assess the level of undercoverage in the census. This is made possible by the implementation of a very effective but also very expensive and time-consuming tracing operation which results in a very high percentage of the original sample being traced. Currently a study of similar size is being planned for the 1981 census.

ACKNOWLEDGEMENT

The content of this paper is based largely on work done and reports prepared with the advice and co-operation of G.J. Brackstone and G. Théroux.

RESUME

La Contre - vérification des dossiers constitue le principal moyen utilisé pour évaluer le niveau de sous-dénombrement lors du recensement de la population du Canada. Un échantillon de personnes est choisi à partir de sources indépendantes du recensement en question, et des opérations intensives de dépistage sont mises en oeuvre afin de déterminer l'adresse habituelle de chaque personne choisie le jour du recensement. Les dossiers du recensement sont ensuite examinés afin de déterminer si chaque personne choisie a été recensée. Le dépistage représente de loin l'étape la plus complexe, côuteuse et fastidieuse de cette étude. Elle implique l'usage prononcé de dossiers administratifs, ainsi que de dépistage sur le terrain. Cet article décrit les diverses méthodes de dépistage utilisées, et indique le degré de succès de chacune de ces méthodes.

REFERENCES

[1] Muirhead, R.C., "1966 Census Evaluation Programme-Reverse Record Check", Technical Report, Sampling and Survey Research Staff, Dominion Bureau of Statistics, Canada, June 1966.

[2] Brackstone, G.J., Gosselin J.-F., "1971 Census Evaluation Project MP-1: 1971 Reverse Record Check Results Memorandum", CDN 71-E-23 (Part 1), Census Field, Statistics Canada, October 1974.

[3] Gosselin, J.-F., Théroux, G., "1976 Census Parametric Evaluation Programme, Reverse Record Check: Basic Results on Population and Household Undercoverage in the 1976 Census", Census and Household Survey Methods Division, Statistics Canada, January 1978.

[4] Gosselin, J.-F., Brackstone, G.J., "The Measurement of Population Undercoverage in the 1976 Canadian Census Using the Reverse Record Check Method", American Statistical Association Proceedings of the Social Statistics Section, 230-235, 1978.

[5] Statistics Canada, "Coverage Error in the 1976 Census of Population and Housing", 1976 Census Quality of Data Series, Catalogue 99-840, March 1980.

TABLE 1: RESULTS OF THE R.O. MATCH

Categories	%
1. Perfect match: SP found	42.0
2. Partial match: SP probably found	0.7
3. 1971 address/head of household found, SP not found	42.3
4. 1971 address found vacant in 1976	1.5
5. 1971 address/head of household not found in 1976	13.5

TABLE 2: RESULTS OF THE TELEPHONE TRACE

Results	%	
TOTAL TRACED		58.8
- Address found	54.0	
- Deceased	4.3	
- Emigrated	0.5	
TOTAL UNTRACED		41.2
- Lead obtained	32.5	
- Tracing failed	8.7	

TABLE 3: TELEPHONE TRACE SUCCESS RATE FOR CATEGORIES 3, 4 and 5 OF THE R.O. MATCH

Results of the RO Match	% Traced in the Telephone Trace
3. Address found, SP not found	58.1
4. Address found vacant in 1976	46.0
5. Address not found in 1976	62.5

TABLE 4: RESULTS OF HEALTH AND WELFARE TRACE

Results	Census Frame	Birth Frame	Immigrant Frame	Missed Frame	TOTAL
Total traced	50.5	90.8	64.2	75.1	68.0
Total untraced	49.5	9.2	35.8	24.9	32.0
- lead obtained or address incomplete*	22.8	2.4	3.2	5.6	13.0
- SP not found	26.5	6.8	32.6	19.3	19.0

*Rural route or post office box number only.

TABLE 5: RESULTS OF THE UIC/TAXATION SEARCH

Results	%
UIC record found	
- Taxation record matched	55.4
- Taxation record not matched	18.8
UIC record not found or match not unique	25.8

TABLE 6: RESULTS OF THE MAIL-OUT TO THE IMMIGRANT AND MISSED FRAMES

First Letter	- Total letters sent	1575
	% Traced	47.2%
Reminder Letter	- Total letters sent	336
	% Traced	37.5%
Householder Letter	- Total letters sent	416
	% Traced	2.6%
Summary Results	- Total selected persons	1575
	- Traced - First letter	47.2%
	- Traced - Reminder letter	8.0%
	- Traced - Householder letter	0.7%
	- Total traced by Mail-out	55.9%

TABLE 7: RESULTS OF THE MAIN REGIONAL OFFICE TRACE BY FRAME

Frame	% Traced
Census	80.1
Birth	71.0
Immigrant	72.8
Missed	84.8
TOTAL	79.6

TABLE 8: OVERALL RESULTS OF THE TRACING OPERATIONS - 1976

STAGES OF TRACING	FRAMES				
	Census	Birth	Immigrant	Missed	Total
SAMPLE SIZE	27,913	3,262	1,169	767	33,111
% TRACED BY STAGES[1]					
- R.O. Match	42.3	n.a.	n.a.	n.a.	35.7
- R.O. Telephone	32.9	n.a.	n.a.	n.a.	27.7
- Health and Welfare	7.5	88.9	29.6	17.1	16.4
- Mail-out	(-)	2.0	44.9	46.4	2.9
- Main R.O. Trace	13.5	4.4	24.5	30.5	13.3
- Health and Welfare (final)	0.9	1.6	(-)	(-)	0.8
% INITIALLY TRACED[2]	96.9	95.4	90.2	93.0	96.4
% TRACED-FINAL[3]	96.0	92.4	89.4	90.4	95.2

- Less than 0.1%

1 The UIC/Taxation operation is not listed since cases found are not treated as traced but are passed on to subsequent tracing operations.

2 The sum of the percentage traced by various stages is higher than the total initially traced since a number of cases were traced by two different operations (because of overlap) and therefore are included more than once in the individual percentages.

3. The difference between the final and the initial percentage traced are cases that were traced, not found in census documents, but that could not be contacted again during follow-up. These are treated as untraced.

TABLE 9: UNWEIGHTED PERCENTAGE OF CASES IN THE FINAL STATUS CATEGORIES

FINAL STATUS	%
Enumerated	88.2
Missed	2.5
Deceased	3.2
Emigrated	1.3
Tracing Failed (final)	4.8

1979 FARM EXPENDITURE SURVEY DESIGN AND ESTIMATION PROCEDURES

J.E. Phillips[1]

The Farm Expenditure Survey was developed to provide annual expenditure estimates for the Western Grain Stabilization Act which is an income stabilization program for grain farmers in the prairies and Peace River district of British Columbia. This paper describes the design of the 1979 survey which incorporated a stratified two-stage design in the area sample and a single take-all stratum in the list sample.

1. PURPOSE AND HISTORY

The Farm Expenditure Survey (FES) was developed to provide expenditure estimates for the Western Grain Stabilization Act which is an income stabilization program for grain farmers in the prairies and Peace River district of British Columbia. This area has been divided into 10 soil zones, which are made up of crop districts, by Agriculture Canada. The FES estimates are to be provided at both the soil zone and province level.

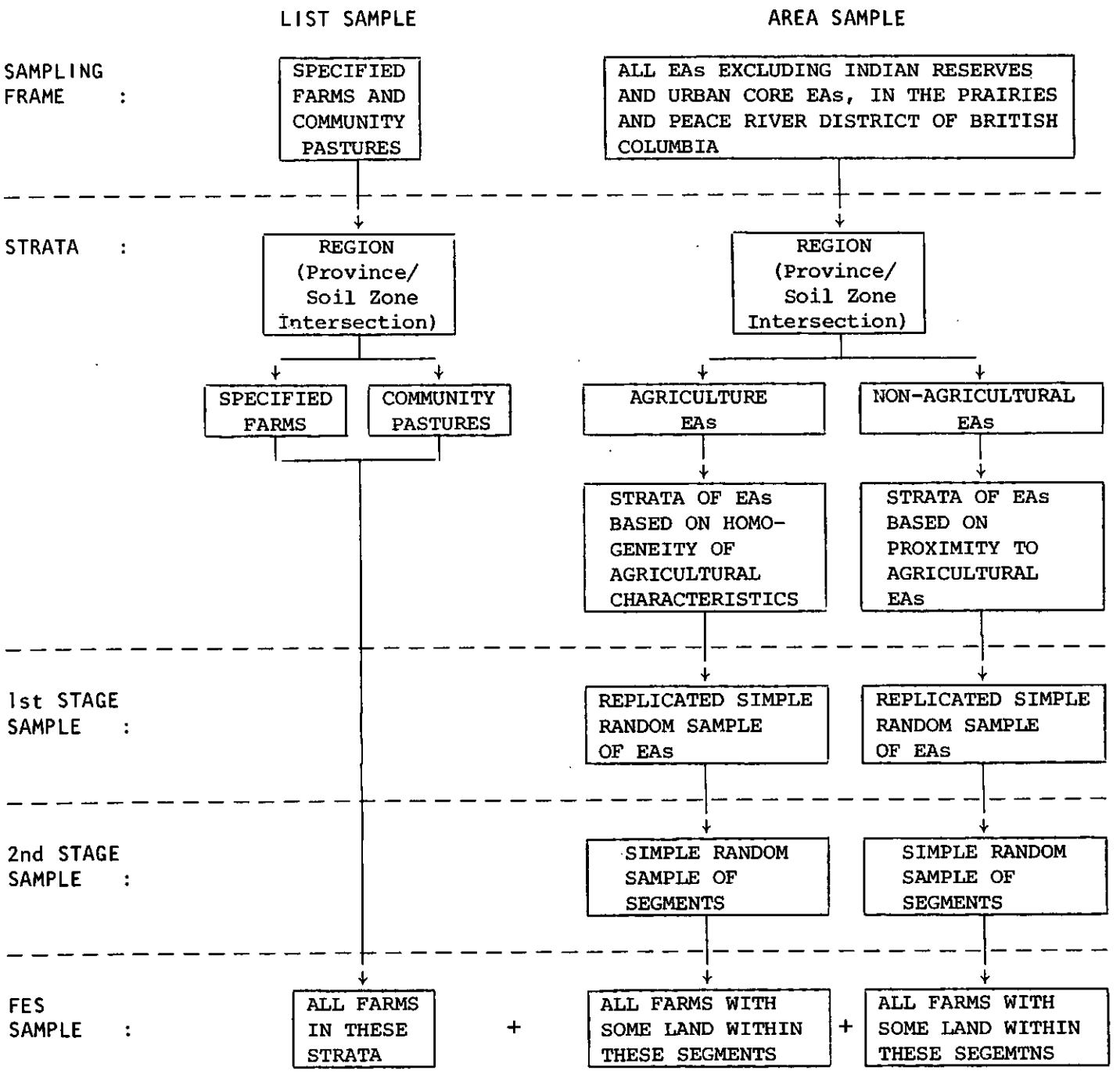
A pilot survey was carried out in March 1976, and full-scale surveys were carried out in March 1977 and 1978. This paper describes the design for the 1979 survey which incorporates the design used for the 1977 survey with some modifications which were introduced for the subsequent surveys.

2. OVERVIEW OF DESIGN

The diagram contained in Figure 1 summarizes the sample design of the FES. Four of the ten soil zones cross provincial boundaries. In order to produce provincial estimates these four zones were split into two distinct parts by provinces. Therefore, for the design, the

[1] J.E. Phillips, Institutional and Agriculture Survey Methods Division, Statistics Canada.

Figure 1: SAMPLE DESIGN FOR FES 1979



whole FES area was split into 14 regions which are intersections of soil zones and provinces. The sample design was independent within each region.

A list sample was used to pick up two types of farming operations. The first type was called specified farms. These were specialized farms which had extremely large values for items like crops or livestock compared to the other farms in the same region. These farms were included in the sample with a probability of 1.

The second type of farm operation picked for the list sample was federal and provincial community pastures, grazing associations and grazing reserves which had 2000 acres or more. Like the specified farms they were included in the sample with a probability of 1.

These community pastures were added in this way because they had caused problems in previous surveys in the enumeration and, later on, when the survey data was being imputed for missing or incorrect data. Any expenses incurred on pastures are paid for by the farmers themselves. Therefore, when enumerating these operations, it was not necessary to fill out the livestock and expense portion of the questionnaire since the farmers were already reporting this information. Since these operations are certainly not typical, because of large total land areas and no livestock or expenses, it would not be advisable to impute these values into other farms. Therefore it was decided to identify them before the enumeration, and treat them separately when imputing.

The design used for the area frame was a stratified two stage design within each region. The first stage was the selection of the primary sampling units (EAs) and the second stage was the selection of the secondary sampling units (segments). The EAs were separated into two major strata—agricultural and non-agricultural. An agricultural EA is one that has at least one farm with headquarters located inside the EA. The headquarters of a farm is defined as the main buildings or main gate.

The first stage involved further stratifying the two major strata into homogeneous groups of EAs. Then after determining the sample size to be allocated to each of these ultimate strata, a replicated simple random sample of EAs was selected from each.

The selected EAs were then divided into equal sized pieces of land called segments for the second stage. A simple random sample of segments was then selected from each EA.

Every farm that had some land located inside the selected segments and received sales of $250.00 or more from agricultural products in the previous year was enumerated. All the specified farms were enumerated. Also every community pasture was included in the sample; however these operations were not enumerated. Instead a questionnaire was mailed to the head offices in charge of the overall operation of these pastures. Whenever a specified farm or one of the community pastures in the list sample happened to fall in one of the segments selected for the area sample it was not enumerated again and therefore did not contribute to the segment estimate.

Table 1 of Appendix A shows the final sample allocation by region and the expected number of farms for the total sample.

3. SAMPLING FRAME

The target population was all farms in the 3 prairie provinces and in Peace River district of British Columbia which received $250.00 or more from the sale of agricultural products in the previous year.

3.1 Area Frame

The area frame was composed of all the 1971 Census Enumeration Areas (EAs) that covered the target area. All the urban core EAs (Census Metropolitan Area code = 1 or 2) and Indian Reserve EAs were eliminated from this frame. The census data was summarized at the EA level for use as the area frame. The design for the

area frame for the 1977 survey was based on the 1971 Census of Agriculture data. The 1976 census data became available in August 1977. This was too late to do a redesign for the 1978 survey based on this more up-to-date information. Also there was no information on expenditures collected in the 1976 Census and total sales was reported in broad ranges only. The 1971 Census provided data that was better suited to a design for an expenditure survey.

The 1976 Census data was used, however, to update the 1971 EA summary data on land acreages and livestock numbers. This information was used in the 1978 FES to determine whether an EA had changed from agricultural to non-agricultural and vice-versa and to update the stratification of the EAs. The same strata were used for the 1979 survey. Some of the EA boundaries had changed between the two censuses. This made assigning the 1976 EA data to the 1971 EAs more difficult. If there was a 1:1 correspondence between a 1976 and 1971 EA then all the 1976 data was assigned to that 1971 EA. If a 1976 EA corresponded to 2 or more 1971 EAs then the data was divided equally amongst the corresponding 1971 EAs.

After updating the EAs, the 1971 and 1976 values for total land, total cropland, the 6 major grains, total cattle, and total pigs were compared for each EA. If there were major changes the EA was assigned to a different stratum than in 1977—one which was more compatible with the updated information. Since some of the stratum population sizes were changed and the information was more up-to-date, the sample allocation between strata within the same region was recalculated for every region for the 1978 FES. This was achieved by determining the optimum allocation for several important items and then deciding on the final compromise allocation. The same allocation was used for the 1979 FES.

By using the 1976 census data to update the EAs some EAs were found to have changed from non-agricultural to agricultural and vice versa. Those EAs which had become agricultural since 1971 were assigned to the stratum with the lowest sales category. Those that had turned non-agricultural were put in the stratum which contained EAs which were in a municipality that did have some agricultural EAs in it.

The non-agricultural EAs were included in the area sample in order to cover that part of the population that had become agricultural since the 1976 Census was taken. Also for the Census, all of the farm's data was assigned to the EA in which the farm headquarters was located. Therefore it was possible for an EA to be classified as non-agricultural, i.e. to have no farm headquarters located inside its boundaries, and yet have some agricultural activity take place there.

3.2 List Frame

Since the FES was designed to produce estimates for expenditure items farms with large expenditure values were considered important. The 1976 Census did not collect data on expenditures but did ask for total sales by sales range. So all farms with sales of over $100,000 (the largest sales range) were examined. Then 4 more additional criteria were established for selecting specified farms. These criteria were based on total cattle, total pigs, total of crops plus summerfallow, and total chickens, and were established independently for each region. A rule for use with simple random sampling was employed to determine cut-off limits. In the optimal situation, any farm having a value for the item which exceeds the cut-off limit is then specified. The limits for each of the four items were calculated thus:

$$\text{limit} = \bar{Y} + \sigma \sqrt{\frac{N}{n}}$$

where $\bar{Y}$ is the population mean for the item in the region,
σ is the standard deviation for the item in the region,
N is the total number of farms in the region,
n is the number of farms in the sample for the region.

If there were too many farms exceeding these limits the limits were raised accordingly.

The community pastures and grazing associations selected for the list sample were all those in the prairies that were larger than 2,000 acres and for which we could get the exact land description. There were a total of 263, and most were located in Saskatchewan (see Table 1 of Appendix A). The list for Alberta was not completed because for many of the grazing associations the detailed land description was not available. This description was necessary because these operations covered such a large land area and would certainly have fallen in the area sample, causing overlap. The selected EAs were segmented so that none of the segments included any of these operations.

4. FIRST STAGE DESIGN

4.1 Stratification of Agricultural EAs

Within each region the EAs were divided into groups of homogeneous EAs called strata. Since the purpose of the survey is to produce expenditure estimates, it is best to stratify on expenditure items or those items highly correlated with them. The 1971 Census data was summarized at the EA level and this data was used for the stratification. The 1971 census provided data on only a few expenditure items but did have detailed sales data.

Several methods of stratification were applied and then compared by looking at the resulting variances of estimates for about 20 items. The four methods were:

1) a classification of EAs by setting limits on the value of certain physical items like crop acreages and livestock numbers,
2) a non-hierarchical cluster analysis on nine physical items (crops and livestock numbers),
3) a hierarchical cluster analysis on groups of EAs with similar values of sales,

4) a classification of EAs into 5 to 10 ranges of total sales.

The first three methods gave almost equivalent results with the third giving slightly better results for the financial items. The fourth method was poorer except for the total sales item.

In comparing the four methods, it was apparent that the first method, which was developed for the Agriculture Enumerative Survey (AES), was designed to measure physical items and so might not be best for measuring expenditure items. The cluster analysis employed in the second and third methods enabled more variables to be used in stratifying and took into consideration the size and specialization of the EAs. Of the two techniques the hierarchical cluster analysis (method 3) was better at determining the number of clusters or strata to have, and in defining these clusters.

The method chosen was the hierarchical cluster analysis on groups of EAs with similar values of total sales (method 3). The first step in this method was to stratify the EAs into groups by total sales. These stratum boundaries were set to optimize the estimation of total sales. They were determined after applying rules for optimum boundaries under equal allocation to each stratum and also under optimum allocation to each stratum. In most regions there were 3 or 4 of these major strata. The stratum of EAs with lowest sales was left as one stratum representing the marginally agricultural areas. The other 2 or 3 major strata were then sub-stratified using the hierarchical cluster analysis.

The variables used for this analysis were the percentage of sales for 7 items to the total sales for the EA. The 7 items were wheat, oats, other grains, other crops, cattle, other livestock, and other sales. The hierarchical clustering program looks at all the EAs in the group and initially pairs off similar EAs. Then it joins similar pairs together and continues this iterative process of joining groups of EAs together until all EAs are in one group again. A

diagram of this whole process is produced, so that at any stage you can determine how many groups there were and what EAs were in them. The groups are joined together according to merging criteria. Three different merging criteria were used in every major stratum. These criteria were,

1) minimum increase in the within group sum of squares,
2) mean within group squared deviation in the new cluster is minimal, and
3) within group sum of squares in the new cluster is minimal.

The results of all three were then compared and a decision on the number and content of strata was made.

There were 3 exceptions to the stratification process. Region 1 in British Columbia, and regions 4 and 5 in Alberta were too small to stratify, and so there is just one stratum in each of these regions. Tables 3.1 to 3.3 of Appendix A give a description of the strata within a few regions.

4.2 Stratification of Non-agricultural EAs

The non-agricultural EAs were divided into 2 strata in each region except for region 1 in British Columbia and 8 in Manitoba which were left as one stratum each because they were too small. For the other regions the strata were defined according to whether or not the EA was in a municipality which had some agricultural EAs in it. Stratum 12 included EAs which were in a municipality with some agricultural EAs and stratum 11 included those that weren't.

Both regions 4 and 5 in Alberta had so few of these EAs that it was decided not to sample from them at all, and so there were no non-agricultural strata in these regions. The bias introduced by doing this was thought to be negligible. Tables 3.1 to 3.3 of Appendix A identify the non-agricultural strata in the regions given.

4.3 Sample Allocation

A small portion of the sample was assigned to the non-agricultural strata. The replicates in these strata had one EA each and there were usually only 2 replicates per stratum. The expected number of farms from these EAs was quite low.

Once the agricultural strata were defined in each region the next step was to determine what sample size was best for each region as a whole and then for each stratum within the regions. First the sample size (no. of EAs) was determined for each region so that expected coefficients of variation for important items were in the 3-5% range. These sample sizes were coverted to the expected number of farms, by using the average number of farms enumerated per EA from the 1976 AES. A total sample of approximately 9,500 was desired and so the sample sizes were adjusted to achieve this.

Next the sample for each region was allocated amongst the strata in the region. This was determined by a compromise of the optimum allocations for several important items. Also the sample size for a stratum had to be divisible into equal-sized replicates. The allocation to the strata for some regions is given in Table 3.1 to 3.3 in Appendix A.

4.4 Replication

A replicate for the FES was an independent simple random sample of EAs selected without replacement from a stratum. Each stratum had at least two replicates and these were selected independently with replacement. This meant that an EA could be selected only once within a replicate but it could be selected in more than one replicate. One reason for replicating the sample was that it simplified the variance calculation. Also, having replicates in each stratum made it easier to spread rotation of the sample between and within the strata, since a whole replicate could be

rotated out without affecting the other replicates in the stratum. Retaining some common replicates between years allowed for better estimates of change over time and also for some consistency between the stratum estimates over several survey periods.

4.5 Sample Selection

For the 1977 FES a simple random sample was selected for each replicate within strata. However for the agricultural strata five different samples were selected for each region. The estimates for 14 items for each of the 5 samples were compared with the census total for the region. The sample chosen for the survey was the one that compared the best with the census total over the most items.

This same procedure was used for the 1978 and 1979 surveys when a rotation of the sample was done. (Details on rotation are given in Section 7.)

5. SECOND STAGE DESIGN

Every EA that was selected in the 1st stage was divided into equal-sized pieces of land called segments. In the prairies and Peace River district of British Columbia most of the land is laid out in 1 square-mile areas called sections. This made segmenting the EAs easier and wherever possible the segments were made up of 3 sections (3 square miles). This procedure also aided in identifying and locating the segments during field enumeration.

For EAs that were not sectioned off or that were towns, the segments were formed by following natural boundaries such as highways, rivers, railway lines, etc.

From every selected EA, one segment was randomly selected, without replacement, for every 30 segments in the EA.

6. ESTIMATION PROCEDURES

The list frame specified farms and community pastures were selected with a probability of 1. Therefore they were given a raising factor (or blow-up factor) of 1 and their values were added directly into the corresponding province and region estimates.

For the area frame, every farm that had land located inside a selected segment and that received $250 or more from the sale of agricultural products in the previous year was enumerated. In order to produce an overall estimate for the region and then for the province, the data for these farms were blown-up to represent first the EA in which they were located and then the stratum that each belonged to. This was done by multiplying the data by a raising factor for each replicate.

The raising factor has 2 components. The first, which blows up the data to the EA level, is the inverse of the probability of selecting the segment.

The second component, which blows up the data to the stratum level, is the inverse of the probability of selecting the EA.

The final raising factor is just component 1 x component 2.

There are three types of estimates possible with this design. The FES uses the weighted estimate. For this estimate all sample farms are used in the calculation of estimates, but their data is multiplied by a weight which is calculated at the farm level thus:

$$\text{weight} = \frac{\text{total land operated inside the segment}}{\text{total land operated}} .$$

The disadvantage of this method is that since the weight is calculated from the total land operated it might not be ideal for items that are

not highly correlated with land. The other two estimates possible have definite disadvantages which are explained in the footnote[1].

Within a region, the estimate for the i^{th} replicate in stratum h is

$$\hat{Y}_{hi} = \sum_{j=1}^{n_{hi}} y_{hij} \times \text{weight} \times \text{raising factor}$$

where y_{hij} = value of the item for the j^{th} farm in replicate i of stratum h

n_{hi} = number of sampled farms in the i^{th} replicate of stratum h.

The estimate for stratum h is

$$\hat{Y}_h = \sum_{i=1}^{r} Y_{hi}/r \quad \text{where} \quad r = \text{number of replicates in stratum h.}$$

The region and province estimates are just the sum of all the strata estimates in the same region or province.

The variance of the estimate is calculated at the stratum level. It is the mean of the sum of squared deviations of the replicate estimates from the stratum estimate divided by the number of replicates minus 1.

$$\hat{\text{Var}}\ (\hat{Y}_h) = \sum_{i=1}^{r} \frac{(\hat{Y}_{hi} - \hat{Y}_h)^2}{r(r-1)}$$

The province and region variances are just the sum of all the strata variances in the same province or region. The stratum variance for

[1] The open estimate uses only those sample farms which had their headquarters located inside the segment. No weight is applied. The disadvantage of this method is that only about half the farms are used which means that the estimates will be less precise.

The closed estimate uses all sample farms but the data used is just for that portion of the farm inside the segment. This estimate is not feasible for the expenditure items collected for the FES since it would be very difficult to allocate a portion of expenses to just that part of the land inside the segment.

the specified farms and community pastures strata is zero since a complete census is done for the list frame.

The percentage coefficients of variation for estimated items are calculated at both the region and province levels and allow comparison of the variances for different and unrelated items.

7. ROTATION OF THE SAMPLE

Since the FES is to be run every year, part of the sample is rotated out and replaced each year so that the same farmers are not continually asked to respond. It is not feasible to change the whole sample every year because good estimates of change over the years are only possible if part of the sample is retained. 30% of the 1977 sample was rotated for the 1978 survey and 20% for the 1979 survey. It is expected that 25% will be rotated out in the following two years. This means the entire sample will be replaced by the 1981 survey.

The rotation for the 1979 survey was spread as equally as possible between the strata. The replicates to rotate out were randomly selected from the replicates that had been in the survey since 1977.

It should be noted here that since it is entire replicates that are rotated, an EA could be rotated out in an old replicate and then rotated back in a new replicate. Should this happen, however, a different segment is selected at the second stage of selection where possible.

8. SUMMARY

The FES will continue to run annually as is required by the Western Grain Stabilization Act. Until 1982 or 1983, its design will remain more or less fixed, possibly with minor modifications made each year. Once the 1981 census results are available, however, there will likely be a complete redesign of the survey to make use of the more up-to-date data source.

RESUME

L'enquête sur les dépenses agricoles a été developpé pour fournir des estimations annuelles des dépenses afin de répondre aux exigences de la Loi de stabilisation concernant le grain de l'Ouest. Cette loi établit un programme de stabilisation des revenus des cultivateurs de céréales dans les prairies et dans le district de Peace River en Colombie-Britannique. Cet article décrit le plan de l'enquête de 1979, qui a utilisé un plan d'échantillonnage stratifié à deux degrés pour l'échantillon choisi de la base aréolaire et une seule strate à tirage complet pour l'échantillon choisi d'une liste.

REFERENCES

[1] Phillips, J., Serrurier, D., Smith-Doiron, K. (1976), "The 1976 FES Sample Design", Technical Report, Institutional and Agriculture Survey Methods Division, Statistics Canada.

[2] Serrurier, D. (1976), "The 1977 FES Sample Design", Technical Report, Institutional and Agriculture Survey Methods Division, Statistics Canada.

[3] Phillips, J. (1978), "The 1978 FES Sample Design", Technical Report, Institutional and Agriculture Survey Methods Division, Statistics Canada.

[4] Anderberg, M.R. (1977), "Cluster Analysis for Applications", New York, Academic Press Series in Probability and Mathematical Statistics, No. 9.

[5] Hartigan, J.A. (1977), "Clustering Algorithms", New York, Wiley Series in Probability and Mathematical Statistics.

APPENDIX A

TABLE 1: Sample Allocation by Region

Region		# of Specified Farms	# of Community Pastures	Agricultural Sample		Non-Agricultural Sample		Total Expected # of Farms
Prov	Soil Zone			EAs	Expected # of Farms	EAs	Expected # Of Farms	
BC	1	3	-	30	162	2	1	166
ALTA	1	4	1	80	512	4	3	520
	2	10	1	151	1012	8	6	1029
	3	16	13	120	840	8	6	875
	4	7	5	25	185	-	-	197
	5	3	5	15	135	-	-	143
	6	12	0	98	666	4	3	779
	Total	52	25	489	3350	24	18	3445
SASK	4	8	71	92	708	4	3	790
	5	5	60	126	895	8	6	966
	6	11	58	122	817	7	5	891
	7	10	25	129	955	5	4	994
	Total	34	214	469	3375	24	18	3641
MAN	8	3	7	84	538	2	1	549
	9	8	9	180	1332	8	6	1355
	10	3	8	74	326	4	3	340
	Total	14	24	338	2196	14	10	2244
Total	.....	103	263	1326	9083	64	47	9496

TABLE 2: 1979 FES Specified Farms
- Limits for Selection

(among those with sales > $100,000)

Region Soil Zone/ Province	Total Cattle	Total Pigs	Crops and Summerfallow	Total Chickens
01/BC	1,606	1,000	13,620	-
01/ALTA	1,454	1,760	11,595	11,193
02/ALTA	2,416	2,654	9,963	71,500
03/ALTA	5,100	2,500	10,231	80,000
04/ALTA	2,990	1,600	11,860	18,000
05/ALTA	2,480	1,000	11,210	15,800
06/ALTA	1,517	1,586	10,900	35,000
04/SASK	1,557	4,400	10,383	22,700
05/SASK	2,300	5,300	-	42,000
06/SASK	1,000	1,022	-	26,320
07/SASK	1,060	1,200	11,460	64,000
08/MAN	980	1,210	7,025	-
09/MAN	1,350	5,000	7,904	60,000
10/MAN	1,500	1,400	-	12,000

TABLE 3: Variables Used in the Stratification of Agricultural EA's

The number in brackets denotes the relevant field number on the 1971 Census of Agriculture Form 6.

Seven variables, expressed in terms of percentage to total sales (227), have been used in the hierarchical cluster analyses:

1. Wheat sales (207)
2. Oats sales (208)
3. Other grains sales (209)
4. Other crops sales (210 to 216)
5. Cattle sales (217)
6. Other livestock sales (218 to 222)
7. Other sales (227 minus sum of variables 1 to 6 = 223 to 226)

The following summaries appear also in strata specifications:

Grains other than wheat (variables 2 and 3)
Grains (variables 1 to 3)
Total crops (variables 1 to 4)
Livestock other than cattle (variable 6)
Total livestock (variables 5 and 6)
Others (variable 7)

TABLE 3.1: First Stage Design Specifications
Soil Zone 2 - Alberta

Stratum Specifications		Stratum Code	No. of Replicates	No. of EAs Per Replicate	Population Size	Sample Size
Total Sales $	% Sales					
AGRICULTURAL EA's						
≥1,000,000	Livestock Other than Cattle ≥ 30% or Others ≥ 50%	1	2	2	9	4
	Cattle ≥ 60%	2	5	2	17	10
	Remaining EAs	3	4	3	23	12
<1,000,000 & ≥ 500,000	Others ≥ 30%	4	4	2	23	8
	Cattle < 40%	5	3	3	20	9
	Remaining EAs	6	7	4	66	28
< 500,000 & ≥ 25,000	Cattle ≥ 60%	7	6	3	59	18
	Cattle ≥ 30%	8	9	4	116	36
	Remaining EAs	9	6	4	49	24
< 25,000	--	10	2	1	85	2
Total					467	151
NON-AGRICULTURAL EA's						
No Agricultural EA in Same Municipality		11	2	1	65	2
At Least one Agricultural EA in Same Municipality *OR* EA was Agricultural in 1971		12	6	1	134	6
Total					199	8

TABLE 3.2: First Stage Design Specifications
Soil Zone 6 - Saskatchewan

Stratum Specifications		Stratum Code	No. of Replicates	No. of EAs Per Replicate	Population Size	Sample Size
Total Sales $	% Sales					
AGRICULTURAL EA's						
≥ 750,000	Cattle ≥ 50%	1	4	2	15	8
	Remaining EAs	2	4	4	37	16
< 750,000 & ≥ 400,000	Cattle ≥ 40%	3	4	3	43	12
	Grains ≥ 60%	4	4	3	33	12
	Remaining EAs	5	6	4	61	24
< 400,000 & ≥ 20,000	Cattle ≥ 40%	6	4	3	49	12
	Grains Other Than Wheat ≥ 30%	7	5	4	72	20
	Remaining EAs	8	4	4	65	16
< 20,000	--	9	2	1	76	2
Total					451	122
NON-AGRICULTURAL EA's						
No Agricultural EA in Same Municipality		11	2	1	74	2
At Least One Agricultural EA in Same Municipality *OR* EA was Agricultural in 1971		12	5	1	98	5
Total					172	7

TABLE 3.3: First Stage Design Specifications
Soil Zone 9 - Manitoba

Stratum Specifications		Stratum Code	No. of Replicates	No. of EAs Per Replicate	Population Size	Sample Size
Total Sales $	% Sales					
AGRICULTURAL EA's						
≥ 900,000	Livestock Other Than Cattle ≥ 35%	1	5	2	16	10
	Cattle ≥ 35%	2	4	4	29	16
	Remaining EAs	3	4	3	22	12
< 900,000 & ≥ 400,000	Livestock Other Than Cattle > 25% OR Others ≥ 25%	4	7	5	83	35
	Grains ≥ 45%	5	5	5	50	25
	Remaining EAs	6	9	3	69	27
< 400,000 & ≥ 20,000	Crops Other Than Grains ≥ 40%	7	2	2	21	4
	Total Crops < 30%	8	3	4	75	12
	Remaining EAs	9	9	4	116	36
20,000	--	10	3	1	90	3
Total					571	180
NON-AGRICULTURAL EA's						
No Agricultural EAs in Same Municipality		11	2	1	42	2
At Least One Agricultural EA in Same Municipality *OR* EA was Agricultural in 1971		12	6	1	124	6
Total					166	8

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1979 Vol. 5 No. 1

A Journal produced by Methodology Staff, Statistics Canada.

Preparé par les méthodologistes de Statistique Canada.

C O N T E N T S

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/décembre 1979 Vol. 5 No. 2

A Journal produced by Methodology Staff, Statistics Canada.

Preparé par les méthodologistes de Statistique Canada.

C O N T E N T S