SOME METHODS FOR UPDATING SAMPLE SURVEY FRAMES AND THEIR EFFECTS ON ESTIMATION

J.D. Drew, G.H. Choudhry, and G.B. Gray,
Statistics Canada

1. Introduction

Sample surveys frequently incorporate designs utilizing unequal probabilities of selection of units within strata. Since many characteristics are highly correlated with the relative sizes of the units, estimates based on such designs are in general more efficient than estimates based on designs where the sizes of the units are ignored. In continuous surveys, the sizes of the sampling units may change over time because of births and deaths of ultimate sampling units (e.g., construction or demolition of dwellings in the case of household surveys). An uneven rate of growth among the sampling units results in a decrease in the correlation between the characteristics being measured from the survey and the size measures, and consequently results in less efficient estimates than in the initial period.

Keyfitz [4] developed a method whereby revised selection probabilities could be incorporated into the sample while maximizing the probability of retaining the originally sampled unit in a stratum. More recently, Kish and Scott [5] adapted Keyfitz's procedure to other cases, for example, where units are shifted from one stratum to another. The chief drawback of the above methods is that they can be applied only to sample designs in which one unit is selected per stratum, so that unbiased variance estimates cannot be obtained.

Rao, Hartley, and Cochran [7] devised a sampling procedure we refer to as the random group method in which unbiased estimates and their variances can be obtained while selecting one unit per random group, and as suggested by Platek and Singh [6], the update procedure due to Keyfitz [4] may be applied to each random group.

In Section (2), we present an unbiased extension of Keyfitz's [4] sample updating procedure to the case where one first stage unit (fsu) is selected per stratum with unequal probability and without rotation of fsu's, but where a portion of the fsu's excluding the selected one is reserved exclusively for special survey use by some known probability mechanism. At the time of sample update, the continuous survey is restricted to the non-reserved portion of the frame. The method incorporates "Working Probabilities" following an approach similar to that used by Fellegi [1] in his PPSWOR selection procedure.

In Section (3), we consider a rotating sample in which the random group method is applied. After selecting one unit with pps in each random group for the continuous survey, a specified portion of the remaining units within each group is reserved with SRSWOR for special surveys. For the rotation scheme considered, it is shown that when units are reserved in the above manner, the probabilities of selection for the continuous survey remain unaffected. The unbiased updating procedure in Section (2) is adapted to accommodate the rotation scheme, and as well a biased updating procedure, which approximates Working Probabilities by the New Probabilities of selection, is considered as an alternative.

The reserved units from each random group within a stratum are merged together to form a special survey frame. Hartley and Rao's [3] randomized pps systematic method is employed to select samples from the special survey frame and an estimation procedure for special surveys is described in Section (4).

Since the design for self-representing areas in the Canadian Labour Force Survey [8] follows the random group method, for the purposes of evaluating the proposed updating schemes, and the procedure suggested for special surveys, a Monte Carlo study was carried out. The details and the results of the study are presented in Section (5).

2. Sample Update When A Portion Of The Frame Is Reserved: (Non-rotating Case)

Consider a stratum which has N first stage sampling units. A size measure X_i is associated with the ith unit in the population; i=1,2,...N. One unit from the stratum is selected for a continuous survey with pps where p_i, the probability of selecting unit i for the continuous survey, is given by

$$p_i = X_i / \sum_{i=1}^{N} X_i ; \quad i=1,2,\ldots,N.$$

For now, we assume that there is no rotation of fsu's for the continuous survey. Following selection of one unit for the continuous survey, some of the remaining fsu's are reserved for use by special surveys, by some known probability mechanism. At the time of sample updating, the continuous survey is to be restricted to the non-reserved portion of the frame.

Let S denote the set of n units reserved for special surveys, and $Pr(S=i)$ be the probability of reserving the units $\{i_1, i_2, \ldots, i_n\}$ in any order, then we have:

$$Pr(S=i) = \sum_{j \notin s} p_j . Pr(S=i \mid j \text{ selected for C}) \qquad (2.1)$$

where C denotes the continuous survey. The only restriction placed on methods of reserving units is that the computation of $Pr(S=i)$ should be practical.

Now at the time of update, revised size measures X_i' are obtained for each unit i=1, 2, ..., N. We require that the new probabilities of selection for the continuous survey should be:

$$p_i' = X_i' / \sum_{i=1}^{N} X_i' \qquad i=1,2,\ldots,N.$$

In order to revise the selection probabilities for the continuous survey and at the same time avoid selection of the reserved units, we define "Working Probabilities" $p_i(2)$, i=1,2, ..., N, such that the overall probability of

selecting unit i when averaged over all possible reserved sets of n out of (N-1) units excluding unit i should equal p_i', i.e.,

$$\sum_S' \Pr(S=i)\left(\frac{p_i(2)}{1-\sum_{j\in S} p_j(2)}\right) = p_i' \qquad (2.2)$$

$$i=1,2,\ldots,N.$$

where $\sum_S'$ is the sum over all possible unordered n-tuples from (N-1) units, excluding unit i. Therefore, from (2.2) we have:

$$p_i(2) = \frac{p_i'}{\sum_S' \frac{\Pr(S=i)}{1-\sum_{j\in S} p_j(2)}} \qquad (2.3)$$

$$i=1,2,\ldots,N$$

The solution for $p_i(2)$'s can be obtained iteratively using p_i's as initial values, although as N and n increase combinatorial difficulties quickly arise since $N\binom{N-1}{n}$ summations are involved for each iteration. The post-update conditional probability of selecting unit i, given that the set S=i is reserved, is:

$$\Pi_{i|S}' = \frac{p_i(2)}{1-\sum_{j\in S} p_j(2)} . \qquad (2.4)$$

Now the posterior probability for the continuous survey to contain unit i as the selected one given that the set S=i was reserved; viz., $\Pi_{i|S}$ is given by

$$\Pi_{i|S} = \frac{\Pr(\text{i selected for C and } S=i)}{\Pr(S=i)} \qquad (2.5)$$

We now perform Keyfitz's type update based on (N-n) available units by comparing $\Pi_{i|S}'$ with $\Pi_{i|S}$ for $i\notin S$. Therefore, the ith unit is selected with conditional probability $\Pi_{i|S}'$, and as a consequence of (2.2) and (2.4), the unconditional probability for unit i to be selected is given by p_i', i=1,2,...,N. Thus the updating scheme is unbiased. Moreover as only one unit is selected per stratum for the continuous survey, the variance is a function of the probabilities of selection of units and as such is unaffected by the reserving of units.

3. Sample Updating When A Portion of the Frame is Reserved: Rotating Case

In this section, we apply the results of the preceding section to Platek and Singh's [6] strategy for a continuous, area-based sample requiring updating, expanding the scope under this strategy for special survey use of the frame and rotation of fsu's by the continuous survey.

For simplicity, we have considered a two-stage random group design with pps selection of fsu's (clusters), systematic selection of ultimate sampling units (dwellings) and sample rotation within and between fsu's -- the design which is used by the Canadian Labour Force Survey in large cities. The results can be generalized for designs with more than two stages of selection.

As before, we have N units within a stratum (random group) and a size measure X_i associated with each unit i= 1,2,...,N. We wish to sample within the stratum at the rate 1/R. Then we define cluster inverse sampling ratios as integers:

$$R_i \geq 1 \qquad i=1,2,\ldots,N$$

$$\text{such that } \sum_{i=1}^{N} \left| R_i - R\frac{X_i}{\sum X_i} \right| \text{ is minimized} \qquad (3.1)$$

$$\text{and } \sum_{i=1}^{N} R_i = R.$$

Define R unique ordered samples within each random group as

$$j|R_i \quad j=R_i, R_i-1, \ldots 2, 1;\ i=1,2,\ldots,N$$

consisting of a sampled cluster i to be systematically sub-sampled at the rate $1/R_i$ for j successive occasions before rotation of fsu's occurs. That is, we have the following set of R ordered samples

$$R_1|R_1,\ (R_1-1)|R_1,\ldots,\ R_N|R_N,\ldots,\ 1|R_N.$$

Initially one of the above samples is selected by generating a random number r, $1 \leq r \leq R$. Suppose the selected sample is $j|R_i$, then another random number r_i, $1 \leq r_i \leq R_i$ is generated and the systematic samples determined by the random starts r_i, $(r_i+1) \bmod R_i$, ..., $(r_i+j-1) \bmod R_i$ are respectively associated with the samples $j|R_i$, $(j-1)|R_i$, ..., $1|R_i$.[1] Rotation is achieved by advancing to the next sample on the list. At the time of rotation into the next cluster, i.e., cluster i* = (i+1) mod N, with sample $R_{i*}|R_{i*}$; a random number r_{i*}; $1 \leq r_{i*} \leq R_{i*}$ is generated and the systematic samples determined by the starts r_{i*}, $(r_{i*}+1) \bmod R_{i*}$, ... $(r_{i*}+R_{i*}-1) \bmod R_{i*}$ are associated with the samples $R_{i*}|R_{i*}$, $(R_{i*}-1)|R_{i*}$ ..., $1|R_{i*}$ respectively, and so on. In practice, random numbers r_i; i=1,2, ..., N are all generated at the time of initial introduction of the sample and the rotation schedule is created in terms of the actual systematic samples or starts.

Following this rotation scheme, the probability of selecting cluster i at any point in time is given by:

$$\Pr(i\in C) = p_i = R_i/R.$$

Given that cluster i is selected, the probability of each start being in the sample at any point in time is given by $1/R_i$, so that the overall

probability of selecting each start is 1/R, and consequently if y_{ik} is the characteristic total for start k in cluster i, then $R\, y_{ik}$ is an unbiased estimator of the group total Y.

We require that when a portion of the frame is reserved, the selection probabilities for the continuous survey remain equal to p_i, i=1,2,...,N, as the sample rotates. Assume that at time t-1 the entire frame is available. Between time t-1 and t one unit is reserved with equal probability from amongst the N-1 units not selected for the continuous survey at time t-1. Then the continuous survey will be in unit i at time t if:

i) the selected sample was one of the first R_i-1 samples of unit i at time t-1 ($\Pr = \frac{R_i-1}{R}$), since these would merely rotate into the next sample which would still be in i; or

ii) the selected sample was the last sample of unit i-2 at time t-1 ($\Pr = \frac{1}{R}$) and the unit i-1 was reserved ($\Pr = \frac{1}{N-1}$), then the sample would rotate into unit i; or

iii) the selected sample was the last sample of unit i-1 at time t-1 ($\Pr = \frac{1}{R}$) and the unit i was not reserved ($\Pr = 1 - \frac{1}{N-1}$) so that the sample would rotate into unit i at time t.

Therefore, by summing, we can obtain the probability for the continuous survey to be in unit i at time t, which is equal to R_i/R.

The posterior probability for unit i to be in continuous survey C given that unit j was reserved is given by:

$$\Pi_{i|j} = \frac{\Pr(i\varepsilon C,\ j \text{ reserved})}{\Pr(j \text{ reserved})}$$

$$= \frac{p_i \frac{1}{N-1}}{\sum_{i\neq j} p_i \frac{1}{N-1}} = \frac{p_i}{1-p_j} . \qquad (3.2)$$

It can be shown that in general when n out of N-1 are reserved with equal probability excluding the continuous survey selection, the probabilities of selection for the continuous survey are preserved, and the expression for the posterior probability $\Pi_{i|S}$ will simplify to:

$$\Pi_{i|S} = \frac{p_i}{1-\sum_{j\varepsilon S} p_j} . \qquad (3.3)$$

After reserving a portion of the frame, say one-third, following the above mechanism, which effectively would correspond to a stage of sampling for special surveys, a pps scheme can be adapted within the special survey frame. In Section (4), we provide a design and estimation procedure for special surveys utilizing the reserved portion of the frame, and in Section (5) we present empirical results which lend support to the suitability of this procedure.

If reserves are made in the above manner, there will be no bias of selection for the continuous survey prior to update. In the remainder of this section we show how the general method described in Section (2) can be adapted to the particular rotation scheme under consideration to achieve desired post-update probabilities while preventing dwelling overlaps between the pre- and post-update samples. Under this method of reserving fsu's, (2.1) reduces to:

$$\Pr(S=i) = (1-\sum_{i\varepsilon S} p_i)\frac{1}{\binom{N-1}{n}} \qquad (3.4)$$

By applying Keyfitz's sample updating procedure using conditional probabilities as described in Section (2), a cluster $i\notin S$ could be selected for the continuous survey with conditional probability $\Pi'_{i|S}$ as given in (2.3) so that when averaged over all possible reserves, the probability of selecting cluster i becomes p'_i. However, having retained a cluster in this fashion at update, it would be desirable to remain in the cluster only long enough so that sampling can be restricted to unused dwellings. This suggests a mapping (see Appendix A) from the possible pre-update samples into the possible post-update samples, such that following the rotation scheme, no overlap of dwellings would occur, and the required post-update probabilities achieved.

Revised cluster isr's R'_i, i=1,2,...,N can be obtained by replacing X_i by X'_i in (3.1). But, since we will be using a one to one mapping from the possible pre-update samples into the possible post-update samples to perform Keyfitz's type sample update as described in Appendix A, and there could be only $(R-\sum_{j\varepsilon S} R_j)$ possible pre-update samples, therefore we define post-update cluster isr's as integers $R_{i|S}(2) \geq 1$ for $i\notin S$ such that

$$\sum_{i\notin S} \left| R_{i|S}(2) - (R-\sum_{j\varepsilon S} R_j)\left(\frac{p_i(2)}{1-\sum_{j\varepsilon S} p_j(2)}\right)\right|$$

is minimized and that (3.5)

$$\sum_{i\notin S} R_{i|S}(2) = R - \sum_{j\varepsilon S} R_j .$$

Thus in this fashion, cluster $i\notin S$ will be selected with conditional probability $\frac{R_{i|S}(2)}{R-\sum_{j\varepsilon S} R_j}$ instead of $\frac{p_i(2)}{1-\sum_{j\varepsilon S} p_j(2)}$, which is only subject to error in rounding to integer sizes.

Since we will be sampling at the rate $R_{i|S}(2)$ instead of R'_i in the selected cluster i, $R(R_{i|S}(2)/R'_i)y_{ik}$ is an estimator for the stratum total, whose only bias is due to rounding to integers.

Due to the complexity involved in computing "Working Probabilities" and practical limitations of this method, a simple although theoretically

biased alternative is presented here. It was observed empirically for the case we considered, i.e., reserving one-third of the frame, that

$$p_i(2) \doteq p_i' \qquad i=1,2,\ldots, N$$

so that we now define the isr's $R_i'|_S \geq 1$ for $i \notin S$ by replacing $p_i(2)$ by p_i' in (3.5). Then $R(R_i'|_S/R_i')y_{ik}$ is the estimator for the stratum total.

4. Strategy for Use of Special Survey Frame

Within a stratum, the reserved units (clusters) from each random group are merged to form the special survey frame. If it were not necessary to provide a capacity for updating the frame and the sample, surveys other than the continuous survey could also use the frame, avoiding overlap with the continuous survey by merely spacing their selections at some interval from those for the continuous survey. However, at the time of update, whether via Keyfitz's method or an independent selection, the continuous survey selection could change resulting in conflict with samples selected for special surveys. On the other hand, if special survey is restricted to the same cluster in which the continuous survey selection happens to be, this has a disruptive effect on planning the rotation of the continuous survey resulting in increased rotation costs for the continuous survey. Also, in a stratum or collection of strata, the special survey may require a larger sample size than the continuous survey. If this is so, increasing the sample within fsu's is likely to be less efficient than selecting additional fsu's.

Since the sample size may vary for different special surveys, a randomized pps systematic design [3] is proposed within the special survey frame as this method is flexible with regard to the number of units selected [2]. Successive special surveys would, to the degree possible, utilize common fsu's to minimize listing costs; however, when the frame is updated, a completely independent selection would be carried out within the special survey frame, avoiding overlap at the dwelling level by means of the re-order mechanism described in Appendix (A).

Suppose that for each random group g, we select n_g clusters with SRS from the (N_g-1) available clusters excluding the continuous survey selection, where $g=1,2,\ldots,G$. Thus, within a sub-unit $n = \sum_{g=1}^{G} n_g$ clusters are reserved for the special survey frame.

Since the continuous survey is more likely to be in larger clusters, the overall probability of a cluster being reserved for the special survey frame decreases as the size of the cluster increases. An unbiased design which takes this into account is likely to be less efficient than a biased design which assumes that the probability of cluster i to be in the special survey frame is equal to n/N for all i. Under the latter assumption, for an overall sampling rate of $1/R_o$ from the sub-unit, $1/S_o=N/(nR_o)$ would be the equivalent sampling rate from the special survey frame. Define $S_o'=[S_o]$, then inverse sampling rates for clusters in the special survey frame are defined as integers $S_i \geq 1$ for $i \in S$ such that

$$\sum_{i\in S} S_i = S_o' \text{ and } \sum_{i\in S} \left|S_i - S_o' \left(\frac{X_i}{\sum_{i\in S} X_i}\right)\right|$$

is minimized, which partitions the special survey frame into S_o' systematic samples. If a special survey selects m of these samples and Y = total response from the m samples, then two estimators for the population total are considered:

$$\hat{Y}_{(1)} = \left(\frac{N}{n}\right) S_o' \, y/m, \qquad (4.1)$$

and

$$\hat{Y}_{(2)} = \left(\frac{X}{X_S}\right) S_o' \, y/m, \qquad (4.2)$$

where $X = \sum_{i=1}^{N} X_i$, $X_S = \sum_{i\in S} X_i$.

The ratio adjustment in $\hat{Y}_{(2)}$ compensates for discrepancies in the size of the special survey frame relative to an n/N sub-sample from the frame, introduced as a result of sampling variability as well as the bias due to the assumption of simple random sampling for reserving units from the entire sub-unit. It was observed empirically that $\hat{Y}_{(2)}$ performed consistently better than $\hat{Y}_{(1)}$, therefore the estimator considered for the special survey frame in Section (5) is $\hat{Y}_{(2)}$.

5. Monte Carlo Study

a) Description

The Canadian Labour Force Survey follows a multistage stratified sample design [8]. In the self-representing areas consisting of larger cities a two-stage stratified sample design is employed. The strata consist of sub-units whose populations vary from 6,000 to 25,000 while fsu's (clusters) consist of city block faces, and ultimate sampling units consist of dwellings.

To evaluate the gains in the reliability of data as a result of updating procedures, and the suitability of the procedure suggested for special surveys, a Monte Carlo study was carried out where seven Labour Force sub-units with varying growth rates between 1966 and 1971 Censuses were chosen.

For the Census Enumeration Areas (EA's) comprising these sub-units, 1971 Census data was obtained at the individual level for the 1/3 sample of households which received a detailed census questionnaire. For the purpose of the study, institutions such as hospitals, and old age homes were excluded. For the most part, 1971 EA's were chosen as clusters, although to conform to the known distribution of cluster sizes by province and type of area for the LFS design, some of the larger EA's were sub-divided to form two or more clusters. The new size measures were obtained from the household counts pertaining to the 1/3 sample, while the corresponding old size measures were obtained by taking 1/3 of the dwelling counts for 1966 EA's and utilizing conversion tables from 1971 to 1966 EA's.

In the study we have considered estimation of the following six characteristics:

1) Population,
2) Number of Households,
3) Number of Persons Employed,
4) Number of Persons Unemployed,
5) Number of Persons Not in Labour Force,
6) Total Income. ($000's)

Five different methods, where a method is defined as a selection scheme along with an estimation procedure, were simulated 1,000 times independently within each sub-unit. The methods are described below:

Method 1 - Random group method using new size measures with complete frame available for the continuous survey.

Method 2 - Following selection for the continuous survey as in Method (1), a special survey frame was established following the reserving mechanism described in Section (3). Within the special survey frame, the design and estimation procedure described in Section 4 were followed.

Method 3 - Same as Method (1), but using old size measures.

Method 4 - Following selection by Method (3), one-third portion from each random group was reserved, and the sample was updated utilizing the "working probability" scheme described in Section 3.

Method 5 - Same as Method (4), except the sample was updated via the "revised probability" scheme described in Section 3.

Let Y_h = the characteristic total for sub-unit h based on the 1971 Census; (h=1,2,...,7),

and $Y = \sum_{h=1}^{7} Y_h$. Further, let

$y_{hr}^{(m)}$ = the estimate of Y_h from the rth replication using method m; (r=1,2,..., 1000; m=1,2,...,5) ,

$$\bar{y}_h^{(m)} = \frac{1}{1000} \sum_{r=1}^{1000} y_{hr}^{(m)} , \text{ and}$$

$$\bar{y}^{(m)} = \sum_{h=1}^{7} \bar{y}_h^{(m)} .$$

Define the discrepancy of method m as

$$D^{(m)} = \bar{y}^{(m)} - Y,$$

and % relative discrepancy by:

$$RD^{(m)} = 100\ (\bar{y}^{(m)} - Y)/Y.$$

The estimated standard deviation and % coefficient of variation of $\bar{y}^{(m)}$ are:

$$\hat{SD}(\bar{y}^{(m)}) = (1000)^{-1}\left[\sum_{h=1}^{7}\sum_{r=1}^{1000}(y_{hr}^{(m)} - \bar{y}_h^{(m)})^2\right]^{1/2}$$

$$\hat{CV}(\bar{y}^{(m)}) = 100\ SD(\bar{y}^{(m)})/Y$$

Define the overall efficiency for method m relative to method 1 as

$$EFF(m \text{ vs } 1) = (MSE)^{(1)}/(MSE)^{(m)}$$

where

$$(MSE)^{(m)} = 1000[\hat{SD}(\bar{y}^{(m)})]^2 + (\sum_{h=1}^{7} D_h^{(m)})^2 .$$

Efficiencies within a sub-unit are analogously defined.

b) Analysis of Results

It was possible from the study to evaluate the gains resulting from updating the sample when the entire frame is available. It can be observed from Tables (5.1) and (5.2) that with the exception of the characteristic unemployed, which is not very highly correlated with size measures, efficiencies tend to decrease (hence gains tend to increase) with decreasing correlation between the old and new size measures. Whereas, one might expect that in practice the greater the growth rate, the lower this correlation would be, sub-units 83112 and 95135 do not confirm these expectations. Even for areas of fairly moderate overall growth, substantial gains in simple survey estimates can result from updating as demonstrated by sub-unit 51201. However, due to the efficiency of techniques commonly utilized in estimation procedures for large scale surveys such as post-stratification by age-sex categories, the gains in precision for final survey estimates are likely to be smaller.

The performances of updating methods 4 and 5 and of the special survey frame relative to method 1 can be seen from an analysis of Tables 5.3 and 5.4.

From an efficiency point of view (Table 5.3) when one-third of the frame has been reserved, there is little difference between updating methods 4 and 5. Efficiencies under both methods are lowest for characteristics unemployed and not in labour force (91-93%). This small loss in efficiency for method 4 is most likely attributable to rounding to integer sizes, and to the departure from the self-weighting design, since otherwise, as noted in section (1), the variance under methods 1 and 4 should be identical. It seems plausible to attribute the loss in efficiency under method 5 to the same causes.

For the remaining characteristics, efficiencies are in the range 98-102%. The efficiency of the special survey frame drops to 95% for unemployed and 96.7% for not in LF, but for other characteristics, ranges from 101-108%. These 'high' efficiencies for the special survey frame seem to be attributable to both the design within the special survey frame and the proposed ratio estimator (4.2).

From Table (5.4), it can be observed that the % relative discrepancies are low in all cases. Comparing the % RD for the theoretically unbiased methods (1 and 4) with those of the other methods, suggests that the bias under methods 2 and 5 is not serious. It should be noted that while t statistics at 95% level were significant in a few cases, these biases appear nevertheless of no

Table 5.1: Correlations[1] and % Growth[2]

	Sub-unit						
	33102	83112	95135	51201	80114	53120	51110
Correlation	.87	.79	.78	.65	.63	.51	.48
% Growth	5.83	54.00	17.41	11.06	18.37	39.16	39.02

Table 5.2: Efficiency of Method 3 vs Method 1

Characteristic	Sub-unit						
	33102	83112	95135	51201	86114	53120	51110
1	87.8	27.4	25.3	30.0	48.1	23.8	8.6
2	33.6	6.6	4.3	5.1	3.0	4.0	1.8
3	78.3	37.3	58.6	39.0	29.9	24.6	13.5
4	82.1	85.4	86.4	99.3	78.3	79.3	88.3
5	87.2	57.7	43.1	50.7	89.4	55.4	31.7
6	93.3	42.1	46.2	35.4	26.5	26.5	10.8

[1] Correlation between old and new size measures

[2] % growth for the period between 1966 and 1971 Censuses

Table 5.3: Overall Efficiencies

Characteristic	Method			
	1	2	4	5
1	100	103.9	98.6	98.1
2	100	107.8	102.0	100.7
3	100	101.1	101.5	100.4
4	100	95.1	91.1	92.4
5	100	96.7	91.8	93.2
6	100	103.2	101.4	99.9

Table 5.4: % Relative Discrepancies/Estimated % Coefficient of Variation

Char.	Pop. Value	Method 1	2	4	5
1	49,389	.17	- .12	.00	.11
		.1485	.1458	.1497	.1500
2	14,264	.07	.01	.01	.02
		.0512	.0493	.0507	.0510
3	19,951	.30	- .45	- .05	.08
		.1731	.1719	.1721	.1730
4	1,615	.35	- .22	.70	.22
		.7391	.7578	.7739	.7687
5	12,288	- .10	.30	.52	.53
		.2414	.2454	.2515	.2495
6	250,547	.08	- .02	- .06	- .03
		.0972	.0957	.0965	.0972

practical consequence being less than 1% of the population value, and, as we also observed, lacking consistent direction from sub-unit to sub-unit.

In conclusion, we feel that Tables 5.3 and 5.4 demonstrate the overall suitability of the strategy we have presented, from the perspective of both the continuous survey and special surveys. We conjecture that under circumstances similar to those in the study, the two updating schemes will perform equally well, so method 5 should be preferred on the grounds of computational simplicity.

References

[1] Fellegi, I.P., "Sampling With Varying Probabilities Without Replacement: Rotating and Non-rotating Samples", Journal of the American Statistical Association, Vol. 58 (1963), pp. 183-201.

[2] Gray, G.B., "On Increasing the Sample Size (number of psu's)", Internal Statistics Canada Technical Memorandum, Household Surveys Development Staff, (1973).

[3] Hartley, H.O. and Rao, J.N.K., "Sampling With Unequal Probabilities and Without Replacement", Annals of Mathematical Statistics (1962), Vol. 33, pp. 350-374.

[4] Keyfitz, N., "Sampling With Probabilities Proportional to Size: Adjustment for Changes in the Probabilities", Journal of the American Statistical Association, Vol.46 (1951), pp. 105-109.

[5] Kish, L. and Scott, A., "Retaining Units After Changing Strata and Probabilities", Journal of the American Statistical Association, Vol. 66 (1971), pp. 461-470.

[6] Platek, R. and Singh, M.P., "A Strategy for Updating Continuous Surveys", Survey Methodology (Statistical Services, Statistics Canada), Vol. 1, No. 1 (June 1975), pp. 16-26.

[7] Rao, J.N.K., Hartley, H.O. and Cochran, W.G., "On a Simple Procedure of Unequal Probability Sampling Without Replacement", Journal of the Royal Statistical Society, Series B, Vol. 27 (1962), pp. 482-491.

[8] Statistics Canada (Household Surveys Development Division), "Methodology of the Canadian Labour Force Survey (1976)", Catalogue 71-526 occasional (published October 1977), pp. 33-38.

Footnote

1/ The convention R_i mod $R_i = R_i$ is adopted throughout the paper.

Appendix (A)

A Simple Method for Sample Update Using Keyfitz's Procedure

Consider a stratum having N units, with inverse sampling ratios R_i; i=1,2,...,N; defined according to (3.1), and[1] with the rotation scheme as described in Section 3 (page).

At some point in time, revised household counts are obtained, and revised inverse sampling ratios R_i'; i=1,2,...,N; are similarly defined. Then the R unique ordered samples based on the revised sizes are:

$$R_1'|R_1',\ (R_1'-1)|R_1',\ \ldots,\ R_N'|R_N',\ \ldots,\ 1|R_N'.$$

At the time of the next sample rotation, the probabilities of selection of clusters must be adjusted so that they are proportional to their revised isr's. Since we have the same number of post-update samples as the number of pre-update samples, a simple one-to-one mapping of pre-update samples into post-update samples can be defined such that:

i) Keyfitz's criteria of adjusting probabilities are satisfied.

ii) The post-update samples can be restricted to previously unselected dwellings, for which, if the same cluster is retained, a necessary but not sufficient condition is that

$$x_i/R_i \geq x_i'/R_i', \qquad (A.1)$$

where $x_i|R_i$ is the sample that would have resulted had there been no update and $x_i'|R_i'$ is the post-update sample. A further condition relates to the choice of the post-update start and is discussed later.

Such a mapping (non-unique) can be carried out as follows:

a) If $i\varepsilon D$, i.e. $R_i' < R_i$, then the samples $R_i|R_i,\ (R_i-1)|R_i,\ \ldots,\ (R_i-R_i'+1)|R_i$ are mapped respectively into the samples $R_i'|R_i',\ (R_i'-1)|R_i',\ \ldots,\ 1|R_i'$ and the samples $(R_i - R_i')|R_i,\ (R_i-R_i'-1)|R_i,\ \ldots,\ 1|R_i$ are temporarily left unmapped.

b) If $i\varepsilon I$, i.e. $R_i' \geq R_i$, then the samples $R_i|R_i,\ (R_i-1)|R_i,\ \ldots,\ 1|R_i$ are mapped respectively into the samples $R_i|R_i',\ (R_i-1)|R_i',\ \ldots,\ 1|R_i'$, leaving the samples $R_i'|R_i',\ (R_i'-1)|R_i',\ \ldots,\ (R_i+1)|R_i'$ as available samples.

c) Since $\sum_{i\varepsilon D} (R_i-R_i') = \sum_{i\varepsilon I} (R_i'-R_i) = f$, say, the unmapped pre-update samples in the decreasing clusters can be mapped in a one-to-one fashion into the available post up-date samples in the increasing clusters. There are f! possible mappings. Ideally, we might wish to choose that mapping which would maximize the time interval (i.e. number of rotation periods) before any post-update sample would rotate back into its corresponding pre-update cluster and begin re-using dwellings. However, evaluating all f! mappings will not always be practical, so we suggest the following procedure:

Let $D = \{i_1', i_2', \ldots, i_d'\}$ define the set of decreasing clusters ordered by increasing serial numbers, and $v = \{v_1, v_2, \ldots, v_d\}$ be the corresponding changes in their number of samples. Define $I = \{i_1'', i_2'', \ldots, i_e''\}$ and $w = \{w_1, w_2, \ldots, w_e\}$ analogously for the set of increasing clusters.

For each $\ell = 1, 2, \ldots, d$, the procedure described below determines a mapping beginning with the decreasing cluster i_ℓ'. The minimum time interval in which a post-update sample will rotate back into its corresponding pre-update cluster and begin re-using dwellings is also obtained for each mapping. If a_ℓ is the minimum time interval for mapping ℓ, then the mapping ℓ^* for which $a_{\ell^*} = \max\{a_1, a_2, \ldots, a_d\}$ is chosen.

For a given ℓ, the mapping is defined as follows:

Find the first cluster $k_1\varepsilon I$ with $i_{k_1}'' > i_\ell'$; that is, the first increasing cluster which will rotate into the sample after cluster i_ℓ'. There are v_ℓ unmapped samples in the decreasing cluster i_ℓ' - map all of these samples in the increasing cluster $i_{k_1}'', i_{(k_1+1) \bmod e}'', \ldots$ exhausting w_{k_1} available samples in the increasing cluster i_{k_1}'' before proceeding to $i_{(k_1+1) \bmod e}''$ and similarly for $i_{(k_1+1) \bmod e}'', i_{(k_1+2) \bmod e}'', \ldots$ using as many of the increasing clusters as required. After mapping the v_ℓ samples from decreasing cluster i_ℓ' into increasing clusters $i_{k_1}'', i_{(k_1+1) \bmod e}'', \ldots$, the corresponding counts of available samples i.e. $w_{k_1}, w_{(k_1+1) \bmod e}, \ldots$ are adjusted. Next take the decreasing cluster $i'_{(\ell+1) \bmod d}$ and find the first cluster $k_2\varepsilon I$ with $i''_{k2} > i'_{(\ell+1) \bmod d}$ and as before map all the $v_{(\ell+1) \bmod d}$ unmapped samples in the decreasing cluster $i'_{(\ell+1) \bmod d}$ into the available samples in the increasing clusters $i_{k_2}'', i_{(k_2+1) \bmod e}'', \ldots$ Repeat this process for clusters $i'_{(\ell+2) \bmod d}, i'_{(\ell+3) \bmod d}, \ldots, i'_{(\ell+d-1) \bmod d}$.

The following example for the case where we have 4 clusters with old and new isr's as given in Table (A.1) illustrates the procedure.

Table (A.1)

Cluster No.	Old isr	New isr
1	4	2
2	3	4
3	2	4
4	3	2
	12	12

The set of decreasing clusters D = {1,4} and the corresponding changes in isr's, i.e. V = {-2,-1}, and similarly for the set of increasing cluster I = {2,3}, W = {1,2}. Fig. (1) below shows the mapping of pre-update samples into the post-update samples.

Mapping of Pre-update Samples into the Post-update Samples

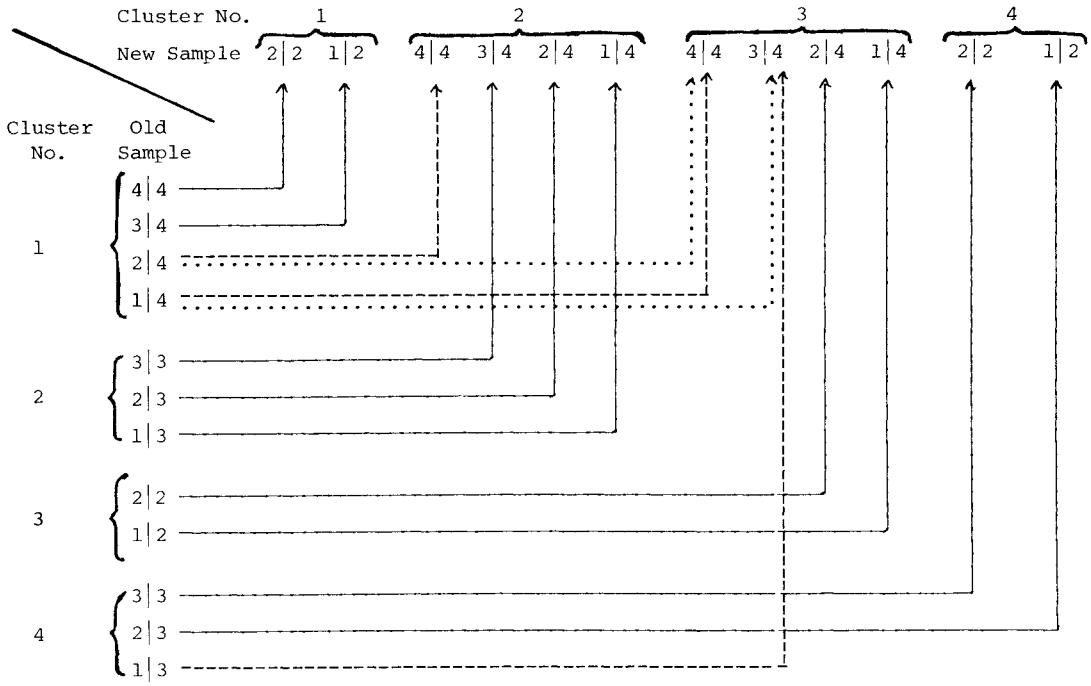

Fig. (1)

The solid lines correspond to the pre-update samples being mapped into the post-update samples in the same cluster, i.e. the cases where old selected cluster is retained. The unmapped pre-update samples in the decreasing clusters can be mapped into the post-update available samples in the increasing clusters starting from the decreasing cluster 1 (broken lines) or starting from the decreasing cluster 4 (dotted lines). The minimum time interval for the reselection of dwellings for the mapping indicated by broken lines is 3 and for the mapping indicated by dotted lines this time interval is 5. In the former mapping (broken lines) the minimum time interval corresponds to the pre-update sample $1|3$ in cluster 4 being mapped into the post-update sample $3|4$ in cluster 3, in which case following use of the samples $3|4$, $2|4$, $1|4$ in cluster 3, reselection of dwellings in the pre-update cluster 4, would occur with sample $2|2$. In the latter mapping (dotted lines) time interval corresponds to the pre-update sample $1|4$ in cluster 1 being mapped into the post-update sample $3|4$ in cluster 3. Thus the mapping indicated by dotted lines will be used.

Clearly, under the above mapping scheme:

i) The clusters are selected with probability proportional to their revised isr's as required.

ii) Each post-update sample is equally likely so that under the rotation scheme these probabilities will be preserved.

iii) Keyfitz's conditions on rejection and retention of clusters hold, and

iv) The condition necessary to avoid re-selection of dwellings also holds.

Having identified the post-update sample in the preceding mapping process, it remains to determine post-update random starts. The following 3 contigencies arise:

i) At the time of update the old cluster is rejected and a new cluster i is selected. Then a random start r_i', $1<r_i'<R_i'$ is chosen, and if the sample to be introduced is $j|R_i$, then the systematic samples determined by the starts r_i', $(r_i'+1) \bmod R_i', \ldots, (r_i'+j-1) \bmod R_i'$ are associated with the samples $j|R_i'$, $(j-1)|R_i', \ldots, 1\ R_i$ respectively.

ii) The previously selected cluster i is retained and $R_i' = R_i$. In this case the sequence of rotation within i remains unchanged.

iii) The previously selected cluster is retained

and $R_i' = R_i$. In this case, we require a mapping of the old starts into the new starts such that the overall probability for each new start equals $1/R_i'$, and such that the number of dwellings to be used under the post-update starts never exceeds the number of dwellings used prior to update. The first condition ensures unbiased selection at the start level, while the second condition allows us to reorder the dwellings, as described later, such that no dwelling re-selections occur.

Let $Pr(s \to s')$ denote the probability that the pre-update start $s(s=1,2,\ldots,R_i)$ will be mapped into the post-update start $s'(s'=1,2,\ldots,R_i')$. Thus we need to determine an $R_i \times R_i'$ matrix P so that $Pr(s \to s')$ is given by $P_{ss'}$, where

$$\sum_{s'=1}^{R_i'} P_{ss'} = 1 \text{ for all } s$$

$$\sum_{s=1}^{R_i} \frac{1}{R_i} P_{ss'} = \frac{1}{R_i'} \text{ for all } s',$$

and the condition necessary to prevent re-selection of dwellings also holds. This can be achieved by determining an $R_i \times R_i'$ matrix A such that

$$\sum_{s'=1}^{R_i'} a_{ss'} = R_i' \quad \text{for all } s \qquad (A.2)$$

$$\sum_{s=1}^{R_i} a_{ss'} = R_i \quad \text{for all } s', \qquad (A.3)$$

and assigning the maximum possible values to the elements of the matrix A in the order $a_{11}, a_{12}, \ldots, a_{1R_i'}, a_{21}, \ldots, a_{R_i 1}, a_{R_i 2}, \ldots, a_{R_i R_i'}$ subject to the constraints (A.2) and (A.3). Then the $Pr(s \to s')$ is simply given by $a_{ss'}/R_i'$ i.e. the matrix P will be defined as

$$P = \frac{1}{R_i'} A \qquad (A.4)$$

The probabilities $P_{ss'}$ $(s=1,2,\ldots R_i, s'=1,2, \ldots, R_i')$ defined by (A.4) will always map the old start with largest permissible probability into the smallest new start at each step beginning with the old start 1, then old start 2, and so on up to old start R_i.

The matrix A which defines the mapping for the case $R_i=6$ and $R_i'=7$ is given in Table (A.2)

Table (A.2)
Matrix A to Obtain the Probability for Post-Update Start Given the Pre-Update Start

Pre-Update Start	Post Update Start 1	2	3	4	5	6	7
1	6	1	0	0	0	0	0
2	0	5	2	0	0	0	0
3	0	0	4	3	0	0	0
4	0	0	0	3	4	0	0
5	0	0	0	0	2	5	0
6	0	0	0	0	0	1	6

From the above table, we find $Pr(1 \to 1) = \frac{6}{7}$, $Pr(1 \to 2) = \frac{1}{7}$ etc. It can be easily checked that if the mapping for the case $R_i=6$, $R_i'=7$ is given by the above matrix A, then the mapping for the case $R_i=7$ and $R_i'=6$ will be given by A^T where A^T is the transpose of matrix A, and this is true in general.

It can be readily verified that the mapping of pre-update starts to post-update starts combined with the earlier mapping of pre- to post-update samples, ensure that the number of dwellings to be used following update in retained clusters is less than or equal to the number unused prior to update. All that is required is to reorder the dwellings so that previously selected dwellings all appear under post-update starts that will not be used.

Before considering the re-ordering, it should be noted that in all cases for future clusters rotating into the sample following update, a random start $r_i', 1 \le r_i' \le R_i'$ is chosen and a rotation schedule comprising a sequence of systematic samples is determined in the same manner as prior to update.

Reordering of Dwellings

The cluster isr, R_i, and the number of dwellings N_{it} in cluster i at time t determine the number of dwellings that will be selected under each start in the cluster. If $b_{it} = [\frac{N_{it}}{R_i}]$ and $Q_{it} = N_{it} - R_i . b_{it}$, then the first Q_{it} starts have $b_{it}+1$ dwellings and the remaining ones have b_{it} dwellings. A schema or incomplete matrix is defined by N_{it} and R_i, as illustrated below for the case $N_{it}=16$, $R_i=6$.

Starts	1	2	3	4	5	6
Dwellings	X	X	X	X	X	X
	X	X	X	X	X	X
	X	X	X	X		

Fig.(2)

Ordinarily the dwellings are loaded row-wise into this schema, viz.

Starts	1	2	3	4	5	6
Dwellings	1	2	3	4	5	6
	7	8	9	10	11	12
	13	14	15	16		

Fig.(3)

so that the dwellings 1, 7, and 13 would be selected with start 1, etc. New dwellings are added in a row-wise fashion, expanding the size of the matrix.

If the isr is changed to R_i' at update with a post-update start of r_i', then the reorder would work as follows.

The dwellings under the unused starts are listed column-wise from left to right from the above schema, say there are L_i such dwellings. A random number ℓ_i; $1 \leq \ell_i \leq L_i$, is determined. Then in the order ℓ_i, $(\ell_i+1) \bmod L_i$, ..., $(\ell_i+L_i-1) \bmod L_i$, the unused dwellings are loaded column-wise into the schema under new isr beginning with the column r_i' and proceeding to the first column of the schema after the end of the last column is reached. Taking the remaining starts in the order in which they were used, dwellings are similarly loaded starting from the position following the last unused dwelling.

To illustrate, consider that t=1, cluster i with $R_i = 6$, $r_i = 1$ was selected with the sample $6|6$, and that $N_{il} = 16$. At t=4, the sample is updated, so that $r_i^* = 4$, where r_i^* is the start that would have resulted had there been no update. Say we have $R_i' = 7$, then the required mappings specify respectively that (i) the post-update sample should be $3|7$, and (ii) the post-update start should be $r_i' = 4$ with probability 4/7 and $r_i' = 5$ with probability 3/7. Say we have $r_i' = 4$. From Fig.(3), the dwellings under the old unused starts (i.e. starts 4, 5, and 6) are {4, 10, 16, 5, 11, 6, 12}. Say $\ell_i = 3$, then the following reorder would result.

New Starts	1	2	3	4	5	6	7
Dwellings	7	8	9	16	11	12	10
	13	14	15	5	6	4	1
	2	3					

Fig.(5)

After using starts 4, 5, and 6 rotation would take place into the next cluster.

It should be noted that if r_i' had been chosen as a random integer between 1 and R_i', then we could have had $r_i'=1$ in which case under the post-update starts 1, 2, 3 a total of 8 dwellings are to be selected whereas $L_i = 7$; that is a dwelling re-selection would have occurred.

It can be demonstrated with the above example that the re-order procedure is slightly biased for selection at the dwelling level. Given the pre-update sample $3|6$, the unused starts can be {1, 2, 3}, {2, 3, 4}, {3, 4, 5}, {4, 5, 6}, {5, 6, 1} , or {6, 1, 2}, with equal probability where r_i^* is the first start in each case. For $N_{il} = N_{i4} = 16$, the dwellings under each of these starts are all determined. The mapping of starts at update takes: $r^*_i = 1$ to $r_i = 1$ with probability 6/7 and to $r_i' = 2$ with probability 1/7, after which in each case 3 dwellings out of the 9 dwellings under pre-update starts {1, 2, 3} will be selected with equal probability; $r_i^* = 2$ to $r_i' = 2$ with probability 5/7 after which 3 out of 9 dwellings are selected with equal probability, and $r_i^* = 2$ to $r_i' = 3$ with probability 2/7 after which 2 out of the 9 dwellings are selected with equal probability, etc. The overall probabilities at time t=4 are .14484, .14749, .14749, .13955, .13690, .13690 for dwellings under pre-update starts {1, 2, ..., 6} respectively; whereas under the new isr of 7, the post-update probabilities of dwellings should each equal $1/7 \doteq .14286$. Given the choice between the inherent risks of respondent burden resulting from dwelling re-selections, and the slight selection bias at the dwelling level due to re-ordering, the latter has been deemed preferable.