# SOME METHODS FOR UPDATING SAMPLE SURVEY FRAMES[1] AND THEIR EFFECTS ON ESTIMATION

J.D. Drew, G.H. Choudhry, and G.B. Gray[2]

Frames designed for continuous surveys are sometimes used for ad hoc surveys which require selection of sampling units separate from those selected for the continuous survey. This paper presents an unbiased extension of Keyfitz's (1951) sample updating method to the case where a portion of the frame has been reserved for surveys other than the main continuous survey. A simple although biased alternative is presented.

The scope under Platek and Singh's (1975) design strategy for an area based continuous survey requiring updating is then expanded to encompass rotation of first stage units, establishment of a separate special survey sub-frame, and procedures to prevent re-selection of ultimate sampling units.

The methods are evaluated in a Monte Carlo study using Census data to simulate the design for the Canadian Labour Force Survey.

## 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 even 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.

[1] Adapted from a paper presented at the Annual Meeting of the American Statistical Association, August 14-17, 1978, San Diego, California, U.S.A.

[2] J.D. Drew, G.H. Choudhry, and G.B. Gray, Household Surveys Development Division, Statistics Canada.

In the case of sample designs based on area frames, a solution to the problem of out of data relative sizes lies in their periodic check by regularly scheduled field counts, followed by a revision of the selection probabilities, and finally a necessary change in the sample to reflect the new probabilities. Keyfitz [4] presented a method whereby revised selection probabilities could be incorporated 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, in particular, 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. This implies that unbiased variance estimates cannot be obtained.

Rao, Hartley, and Cochran [7] devised a sampling procedure referred to as the random group method in which unbiased estimates and their variances can be obtained while selecting one unit per random group. As suggested by Platek and Singh [6], the Keyfitz update procedure 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 but where a portion of the fsu's, excluding the selected one, is reserved exclusively for special survey use. The units are reserved by applying some known probability mechanism, and 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 extend the study of update strategy to 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 particular rotation scheme under consideration, it is shown that when units are reserved in the above manner, the probabilities of selection for the continuous survey remain unaffected prior to update. The unbiased updating procedure in Section (2) is adapted to accommodate the rotation scheme. As an alternative, a biased updating procedure, which approximates Working Probabilities by the revised probabilities of selection, is considered.

In Section (4), 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 (5), we report the results of a Monte Carlo study based on the random group design. This design is used by the Canadian Labour Force Survey in self representing areas.

## 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 stratum; 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

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

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

Let s denote a set of n units reserved for special surveys, and let S by any such set (note that S is a function whereas s is a realization), then Pr(s) is the probability of reserving the set s of units in any order. Let C denote the continuous survey. We have

$$\Pr(s) = \sum_{j \notin s} \Pr\ (j \text{ selected for } C).\ \Pr(s|j \text{ selected for } C)$$

$$= \sum_{j \notin s} P_j . \Pr(s|j \text{ selected for } C). \qquad (2.1)$$

The only restriction placed on methods of reserving units is that the computation of Pr(s) should be practical.

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 C should be:

$$p_i' = \frac{X_i'}{\sum\limits_{i=1}^{N} X_i'} \qquad i=1,2,\ldots,N. \qquad (2.2)$$

Note that the revised selection probabilities for the continuous survey are constrained by the non-selection of the reserved units. We therefore 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.,

$$\sum_{s}{}' \Pr(s) \left(\frac{p_i(2)}{1 - \sum\limits_{j \varepsilon s} p_j(2)}\right) = p_i' \qquad i=1,2,\ldots,N \quad , \qquad (2.3)$$

where $\Sigma'_s$ denotes the sum over all possible unordered n-tuples from (N-1) units, excluding unit i, and Pr(s) is defined by expression (2.1). Therefore, from (2.3) we have:

$$p_i(2) = \frac{p'_i}{\Sigma'_s \dfrac{Pr(s)}{1-\Sigma_{j\varepsilon s} p_j(2)}} \qquad i=1,2,\ldots,N. \tag{2.4}$$

The solution for $p_i(2)$'s can be obtained iteratively by using $p_i$ as initial values. Note that 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 the set s of reserved units, is:

$$\Pi'_{i|s} = \frac{p_i(2)}{1-\Sigma_{j\varepsilon s} p_j(2)} . \tag{2.5}$$

The posterior probability for the continuous survey to contain the ith unit as the selected one given that the set of s of units was reserved is

$$\Pi_{i|s} = \frac{Pr\ (i \text{ selected for C and the set s of unit reserved})}{Pr(s)}$$

$$= \frac{p_i\cdot(Pr(s|i \text{ selected for C}))}{Pr(s)} . \tag{2.6}$$

We now perform Keyfitz's type update based on (N-n) available units by comparing $\Pi'_{i|s}$ with $\Pi_{i|s}$ for $i \notin s$. In order to revise the conditional probabilities $\Pi_{i|s}$ to $\Pi'_{i|s}$, we undertake the Keyfitz updating procedure.

Define conditionally increasing and decreasing sets of units I and D, such that

$$i \varepsilon I \qquad \text{if } \Pi'_{i|s} \geq \Pi_{i|s}$$

and $\quad i \varepsilon D \qquad$ otherwise.

If $i \varepsilon I$ retain the unit. If $i \varepsilon D$ retain the unit with probability $\Pi'_{i|s}/\Pi_{i|s}$ and if rejected, as it would be with probability $(1-\Pi'_{i|s}/\Pi_{i|s})$, select one unit from the set I with probability

$$\frac{\Pi'_{i|s} - \Pi_{i|s}}{\sum\limits_{i \varepsilon I} (\Pi'_{i|s} - \Pi_{i|s})} \qquad \text{for } i \varepsilon I.$$

Then $P_{i|s}$, the conditional probability of selecting unit i under Keyfitz's procedure given the set s of reserved units, will be:

$$P_{i|s} = \Pi_{i|s} \left(\frac{\Pi'_{i|s}}{\Pi_{i|s}}\right) = \Pi'_{i|s} \qquad \text{for } i \varepsilon D$$

$$P_{i|s} = \Pi_{i|s} + \sum_{j \varepsilon D} \Pi_{j|s} \left(1 - \frac{\Pi'_{j|s}}{\Pi_{j|s}}\right) \left(\frac{\Pi'_{i|s} - \Pi_{i|s}}{\sum\limits_{i \varepsilon I} (\Pi'_{i|s} - \Pi_{i|s})}\right)$$

$$= \Pi_{i|s} + \Pi'_{i|s} - \Pi_{i|s} = \Pi'_{i|s}. \qquad \text{for } i \varepsilon I$$

Therefore, at update the ith unit is selected with conditional probability $\Pi'_{i|s}$. Averaging over all possible reserves of n out of (N-1) units, excluding unit i, we obtain the overall average probability, $P_i$ for unit i to be selected following update, as:

$$P_i = \sum_s{}' \Pr(s)(\Pi'_{i|s})$$
$$= p'_i \text{ by (2.3 and 2.5)} \qquad i=1, \ldots, N.$$

Therefore the updating scheme is unbiased. Since only one unit is selected per stratum for the continuous survey, the variance is a function of the probabilities of selection of units and is unaffected by the reserving of units.

## 3. SAMPLE UPDATING WHEN A PORTION OF THE FRAME IS RESERVED: (ROTATING CASE)

The results of the preceding section are applied to the Platek and Singh strategy [6] for a continuous, area-based sample requiring updating. The scope under this strategy is expanded to the case where the continuous survey incorporates rotation of fsu's. Here, only self weighting designs are considered for the continuous survey, so that when a portion of the frame has been reserved, it is required that the reserving mechanism does not affect probabilities of selection of units for the continuous survey as the sample rotates.

For simplicity we have considered as a model 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: this design 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:

$$\left.\begin{array}{lll} & R_i \geq 1 & i=1,\ 2,\ \ldots,\ N \\ \text{such that} & \sum_{i=1}^{N} |R_i - R.\ p_i| & \text{is minimized} \\ \text{and} & \sum_i R_i = R & \end{array}\right\} \quad (3.1)$$

It should be noted that inverse sampling ratios in the form of integers are more convenient than non-integers for implementation in the field and for sample rotation.

Define R unique ordered samples within each random group as

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

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 $j|R_i$, where $\sum_{k=0}^{i-1} R_k < r \leq \sum_{k=1}^{i} R_k$ for some $i \varepsilon \{1,\ 2,\ \ldots,\ N\}$, and $j = \sum_{k=1}^{i} R_k - r + 1$; $R_o$ is defined to be zero. Then another random number $r_i$, $1 \leq r_i \leq R_i$ is generated and the systematic samples determined by the random starts $r_i$, $(r_i+1)$ mod $R_i$, ..., $(r_i+j-1)$ mod $R_i$ are respectively associated with the samples $j|R_i$, $(j-1)|R_i$, ..., $1|R_i^1$. After each pre-specified constant interval of time, rotation takes place into the next sample on the list. At the time of rotation into the next cluster, i.e. cluster

[1] $R_i$ mod $R_i$ is taken equal to $R_i$ instead of 0. This convention will be adopted throughout in this paper.

$i* = (i+1) \bmod N$, with sample $R_{i*}|R_{i*}$; a random number $r_{i*}$; $1 \le r_{i*} \le R_{i*}$ is generated and the systematic samples determinded 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 $r_i$, $i=1, 2, \ldots, N$ 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:

$$Pr(i \varepsilon C) = R_i/R \doteq p_i \quad .$$

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 $(1/R_i)(R_i/R)$ or $1/R$. Consequently, since the design is self weighting, 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.

Now consider what happens to probabilities of selection when reserves are made from the frame, adopting the rule that if the unit that would have rotated is reserved, rotation will take place into the next unreserved unit. For simplicity we consider the case of one reserved unit. Since the probability of selecting a cluster at any point in time is given by $R_i/R$, we can assume with no loss of generality that at time t=0 cluster i is in the continuous survey, and that at time $t \varepsilon (0,1)$, one cluster, say $k \neq i$ is reserved with probability $p^*_{k|i}$. Then at t=1, the occasion of next rotation of the sample, the probability for cluster i to be in the sample for the continuous survey C, i.e. $Pr\ (i \varepsilon C | t=1)$ is given by:

$$\begin{aligned}\Pr(i\varepsilon C|t{=}1) = {} & \Pr(i\varepsilon C|t{=}0).\ \Pr(\text{cluster } i \text{ will not rotate out at } t{=}1)\\ & + \Pr(i-1\varepsilon C|t{=}0)\cdot\Pr(\text{cluster } i-1 \text{ will rotate out at } t{=}1)\cdot\Pr(\text{cluster } i \text{ not reserved})\\ & + \Pr(i-2|C\ t{=}0)\cdot\Pr(\text{cluster } i-2 \text{ will rotate out at } t{=}1)\cdot\Pr(\text{cluster } i-1 \text{ is reserved})\\ = {} & \frac{R_i}{R}\left(1-\frac{1}{R_i}\right)+\frac{R_{i-1}}{R}\;\frac{1}{R_{i-1}}\left(1-p^*_{i|i-1}\right)\\ & + \frac{R_{i-2}}{R}\;\frac{1}{R_{i-2}}\;p^*_{i-1|i-2}\\ = {} & \frac{R_i-1}{R}+\frac{1}{R}\left(1-p^*_{i|i-1}\right)+\frac{1}{R}\,p^*_{i-1|i-2}\ . \qquad (3.2)\end{aligned}$$

Now (3.2) equals $R_i/R$ if and only if $p^*_{i|i-1} = p^*_{i-1|i-2}$ for all i. This condition holds non-uniquely if one cluster is reserved with equal probability, excluding the unit selected for the continuous survey. The posterior probability for unit i to be in continuous survey C given that unit j was reserved is given by:

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

$$= \frac{p_i\,\frac{1}{N-1}}{\sum\limits_{i\neq j} p_i\,\frac{1}{N-1}} = \frac{p_i}{1-p_j}\ . \qquad (3.3)$$

Thus, the expression for $\Pi_{i|j}$ is simplified if one unit is reserved with equal probability.

In general, it can be shown that when n out of N-1 clusters 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 $\Pi_{i|s}$ simplifies to:

$$\Pi_{i|s} = \frac{p_i}{1 - \sum_{j \varepsilon s} p_j} \quad . \tag{3.4}$$

However, for the same reason that we have chosen a pps sampling scheme for the continuous survey, such a design in most instances would be advantageous for the special survey. Thus, instead of selecting one or more units specifically for a particular special survey with equal probability excluding the selection for the continuous survey, rather, our strategy will be to reserve a portion of the frame, say one-third, following the above mechanism for reserving fsu's and then to select units for the special survey from within the reserved portion following a pps·scheme.

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 overlaps of dwellings between the pre- and post-update samples.

Under this method of reserving fsu's, (2.1) and (2.3) reduce respectively to:

$$Pr(s) = (1 - \sum_{i \varepsilon s} p_i) \frac{1}{\binom{N-1}{n}} \tag{3.5}$$

and $$\sum_{s}{}' (1 - \sum_{i\varepsilon s} p_i) \frac{1}{\binom{N-1}{n}} \left(\frac{p_i(2)}{1 - \sum_{i\varepsilon s} p_i(2)}\right) = p_i' \qquad (3.6)$$

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

where $p_i'$ are defined in (2.2).

By applying Keyfitz's sample updating procedure using conditional probabilities as described in Section (2), a cluster $i\varepsilon s$ could be selected for the continuous survey with conditional probability $\Pi_{i|s}'$ given by:

$$\Pi_{i|s}' = \frac{p_i(2)}{1 - \sum_{j\varepsilon s} p_j(2)}$$

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 would be achieved.

The cluster isr's based on new sizes will be defined as before, with $R_i'$ replacing $R_i$ and $p_i'$ replacing $p_i$, N in expression (3.1).

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, we define post-update cluster isr's as integers $R_{i|s}(2) \geq 1$ for $i \notin s$

$$\left.\begin{array}{l} \text{such that } \sum_{i\notin s} (R_{i|s}(2) - (R - \sum_{j\varepsilon s} R_j).\, \Pi'_{i|s}) \\ \\ \text{is minimized and that} \\ \\ \qquad \sum_{i\notin s} R_{i|s}(2) = R - \sum_{j\varepsilon s} R_j. \end{array}\right\} \qquad (3.7)$$

Thus in this fashion cluster $i\notin s$ will be selected with conditional probability

$\dfrac{R_{i|s}(2)}{R - \sum_{j\varepsilon s} R_j}$ instead of $\Pi'_{i|s}$. Note that this computational procedure is only subject to error in rounding to integer sizes. In expression (3.6), to calculate working probabilities $p_i(2)$, $p'_i$ was taken as

$\dfrac{X'_i}{\sum_i X'_i}$ instead of $R'_i/R$ so that the effect due to rounding to integers is not introduced twice.

Since we will be sampling at the rate $R_{i|s}(2)$ instead of $R'_i$ in the selected cluster i, we will apply a compensating weight equal to the ratio $\dfrac{R_{i|s}(2)}{R'_i}$ at the estimation stage. As before, if $y_{ij}$ is the characteristic total for the selected sample k in cluster i, then

$R(\dfrac{R_{i|s}(2)}{R'_i})\ y_{ik}$ 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 biased alternative is presented here. It was observed empirically that, when $n/(N-1) \leq 1/3$

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

so that we now define the conditional probability of selecting unit i for the continuous survey C, given that the set s of units was reserved, as

$$\Pi^*_{i|s} = \frac{p_i'}{1-\sum_{j\varepsilon s} p_j'}$$

and we define the isr's $R'_{i|s} \geq 1$ for $i \notin s$ by replacing $R_{i|s}(2)$ by $R'_{i|s}$ and $\Pi'_{i|s}$ by $\Pi^*_{i|s}$ in (3.7).

Then $R(\frac{R'_{i|s}}{R'_i})\ y_{ik}$ is the estimator for the stratum total, and the mapping of pre-update samples into post-update samples is identical to the previous case.

It should be noted that if the number of post-update samples could be chosen as $R-\sum_{i\varepsilon s} R_i'$ instead of $R-\sum_{i\varepsilon s} R_i$, then the weights $\frac{R_{i|s}(2)}{R_i'}$ would in general be close to one, and the departure from a self-weighting design would be minimized. However, the mapping procedure for the case where the number of pre-update and post-update samples are not equal, becomes very complicated. Moreover, under this mapping, the probability of retaining the currently selected cluster will not be maximized as under Keyfitz's method.

## 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. Before presenting the methodology for the special survey frame, it should be pointed out that 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 the special survey is restricted to the same cluster in which the continuous survey selection happens to be, this may operationally link the continuous and special surveys to a degree that is detrimental to both. For instance, the special survey would be tied into the continuous survey's lead times for introduction of sampling units, while on the other hand, sporadic special survey use of the frame would have a disruptive effect on sample maintenance operations for the continuous survey.

Since the sample size may vary for different special surveys, a randomized pps systematic design [3] is proposed 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 wuld 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, ..., G. Thus within a sub-unit $n = \sum_{g=1}^{G} n_g$ out of $N = \sum_{g=1}^{G} N_g$ 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, let $1/W_o$ be the equivalent sampling rate from the special survey frame. Then

$$\frac{n}{N}\,(1/W_o) = 1/R_o$$

or

$$W_o = \frac{n}{N}\,R_o.$$

Define $W_o' = [\frac{n}{N}\,R_o]$.

A compensating weight, $\omega$, to offset the effect of rounding will be applied at the estimation, where

$$\omega = \frac{W_o'}{W_o} = \frac{W_o'}{\frac{n}{N}\,R_o} \quad .$$

Then inverse sampling rates for clusters in the special survey frame are defined as integers $W_i \geq 1$ for $i \varepsilon s$ such that

$$\sum_{i\varepsilon s} W_i = W_o' \quad \text{and} \quad \sum_{i\varepsilon s} (W_i - W_o' (\frac{X_i}{\sum_{i\varepsilon s} X_i}))$$

is minimized, which partitions the special survey frame into $W_o'$ systematic samples. Selection of M of these samples for a special survey corresponds to an $M/R_o$ sampling rate from the entire frame.

Let $y_m$ = response from mth selected sample.

Then $y = \sum_{m=1}^{M} y_m$ = total response from the sample.

Two estimators for the population total are considered:

$$\hat{y}_1 = \omega \; R_o \; y/M$$

$$= (\frac{N}{n}) \; W_o' \; y/M, \qquad (4.1)$$

and

$$\hat{y}_2 = \frac{(\frac{X}{N})}{(\frac{X_s}{n})} \; \omega \;\; R_o \; y/M$$

$$= (\frac{X}{X_s}) \;\; W_o' \; y/M, \qquad (4.2)$$

where $X = \sum_{i=1}^{N} X_i, \;\; X_s = \sum_{i \varepsilon s} X_i$.

The ratio adjustment $\frac{(\frac{X}{N})}{(\frac{X_s}{n})}$ 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 in the Monte Carlo studies 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 multi-stage stratified sample design [6]. In the self-representing areas consisting of large cities and metropolitan areas, accounting for over 2/3 of the country, 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 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 for seven Labour Force sub-units (strata) with varying growth rates between 1966 and 1971 Censuses.

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 to represent LFS clusters. However, in order that the distribution of cluster sizes within sub-units closely approximated 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 this study we have considered estimation of the following six characteristics:

i) Population,

ii) Number of Households,

iii) Number of Persons Employed,
iv) Number of Persons Unemployed,
v) Number of Persons Not in Labour Force,
vi) Total Income.

Five different methods were simulated 1,000 times independently within each sub-unit. A method is defined as a selection scheme associated with an estimation procedure. 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 select-on as in Method 1, a one-third portion from each random group was reserved with equal probability excluding the cluster selected for the continuous survey and the reserved clusters from each random group were merged together to form the special survey frame. 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 $y_{hr}^{(m)}$ = the estimate of $Y_h$ from the rth replication using method m; (r=1, 2, ..., 1,000; m=1, 2, ..., 5).

Then the average value of 1,000 estimates for method m, sub-unit h is given by:

$$\bar{y}_h^{(m)} = \frac{1}{1,000} \sum_{r=1}^{1,000} y_{hr}^{(m)} .$$

Combining all the 7 sub-units, the population total Y is given by:

$$Y = \sum_{h=1}^{7} Y_h,$$

and similarly combining the estimates for all sub-units, we have:

$$y_r^{(m)} = \sum_{h=1}^{7} y_{hr}^{(m)}$$

and

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

$$= \frac{1}{1,000} \sum_{r=1}^{1,000} y_r^{(m)}$$

Define the discrepancy of method m, $D^{(m)}$, to be the deviation of the average of 1,000 estimates, using method m, from the population total y, viz.

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

and % relative discrepancy by:

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

The estimate of standard deviation of $y_{hr}^{(m)}$ is:

$$\widehat{S.D.}(y_{hr}^{(m)}) = [\frac{1}{1,000} \sum_{r=1}^{1,000} (y_{hr}^{(m)} - \bar{y}_{h}^{(m)})^2]^{1/2}$$

Therefore, the estimate of the standard deviation of $y_r^{(m)}$ is

$$\widehat{S.D.}(y_r^{(m)}) = (\sum_{h=1}^{7} [\widehat{S.D.}(y_{hr}^{(m)})]^2)^{1/2} \quad ,$$

and the estimate of the standard deviation of $\bar{y}^{(m)}$ is

$$\widehat{S.D.}(\bar{y}^{(m)}) = \widehat{S.D.}(y_r^{(m)})/(1,000)^{1/2} \quad .$$

The estimated % coefficient of variation is then given as:

$$\widehat{C.V.}(\bar{y}^{(m)}) = 100\ \widehat{S.D.}(\bar{y}^{(m)})/\bar{y}^{(m)}$$

Within sub-unit h, define the efficiency of method m relative to method 1 as:

$$EFF_h\ (m \text{ vs } 1) = 100\ (MSE)_h^{(1)}/(MSE)_h^{(m)}$$

where

$$(MSE)_h^{(m)} = \frac{1}{1,000} \sum_{r=1}^{1,000} (y_{hr}^{(m)} - y_h)^2 \ .$$

Finally, define the overall efficiency for method m relative to method 1 as:

$$EFF(m \text{ vs } 1) = 100\ (MSE)^{(1)}/(MSE)^{(m)}$$

where

$$MSE^{(m)} = [\widehat{S.D.}(y_r^{(m)})]^2 + (\sum_{h=1}^{7} D_h^{(m)})^2 .$$

b) Analysis of Results

Although the primary purpose of the study was to evaluate the two updating schemes (i.e. methods 4 & 5) and the performance of the proposed special survey frame, it was also possible to study the gains resulting from updating the sample when the entire frame is available. Let us briefly then examine these gains.

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. It would be of interest to investigate this aspect further.

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 |
| population | 87.8 | 27.4 | 25.3 | 30.0 | 48.1 | 23.8 | 8.6 |
| households | 33.6 | 6.6 | 4.3 | 5.1 | 3.0 | 4.0 | 1.8 |
| employed | 78.3 | 37.3 | 58.6 | 39.0 | 29.9 | 24.6 | 13.5 |
| unemployed | 82.1 | 85.4 | 86.4 | 99.3 | 78.3 | 79.3 | 88.3 |
| not in LFS | 87.2 | 57.7 | 43.1 | 50.7 | 89.4 | 55.4 | 31.7 |
| income | 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.

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.

Table 5.3: Overall Efficiencies

| Characteristic | Method | | | |
|---|---|---|---|---|
| | 1 | 2 | 4 | 5 |
| population | 100 | 103.9 | 98.6 | 98.1 |
| households | 100 | 107.8 | 102.0 | 100.7 |
| employed | 100 | 101.1 | 101.5 | 100.4 |
| unemployed | 100 | 95.1 | 91.1 | 92.4 |
| not in LFS | 100 | 96.7 | 91.8 | 93.2 |
| income | 100 | 103.2 | 101.4 | 99.9 |

For 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%. The efficiencies do not appear to be appreciably affected by the procedure of reserving a portion of the frame, and then drawing the sample from the reserved portion as opposed to drawing the sample from the whole frame. This phenomenon seems to be attributable to both the design within the special survey frame and the proposed ratio estimator (4.2).

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

| Characteristic | Population value | Method | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 4 | 5 |
| population | 49,389 | .17<br>.1485 | - .12<br>.1458 | .00<br>.1497 | .11<br>.1500 |
| households | 14,264 | .07<br>.0512 | .01<br>.0493 | .01<br>.0507 | .02<br>.0510 |
| employed | 19,951 | .30<br>.1731 | - .45<br>.1719 | - .05<br>.1721 | .08<br>.1730 |
| unemployed | 1,615 | .35<br>.7391 | - .22<br>.7578 | .70<br>.7739 | .22<br>.7687 |
| not in LFS | 12,288 | - .10<br>.2414 | .30<br>.2454 | .52<br>.2515 | .53<br>.2495 |
| income ($1000's) | 250,547 | .08<br>.0972 | - .02<br>.0957 | - .06<br>.0965 | - .03<br>.0972 |

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 significant t-statistics at 95% level were obtained for the characteristic employed under method 2 and not in Labour Force for both methods 4 and 5, these biases appear nevertheless of no practical significance, being less than 1% of the population value. Also, it is worth noting that although we have not presented discrepancies for individual sub-units, these were calculated, and it was observed that no methods either under-estimated or over-estimated a characteristic for all sub-units.

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.

## RESUME

Les bases conçues pour des enquêtes permanentes servent parfois à effectuer des enquêtes spéciales qui nécessitent un échantillon distinct de celui de l'enquête permanente. Cet article présente une méthode sans biais de mise à jour d'une base de sondage, qui prolonge celle de Keyfitz (1951) en l'appliquant au cas où une partie de la base a été réservée à des enquêtes autres que l'enquête permanente. Une autre méthode, simple mais biaisée, est aussi exposée.

Les auteurs élargissent ensuite la portée de la technique de Platek et Singh (1975) sur la conception d'un échantillon permanent à partir d'une base aréolaire nécessitant des mises à jour, en incorporant à cette technique le renouvellement des unités d'échantillonnage de premier degré, l'établissement d'une base réservée aux enquêtes spéciales et des procédures visant à éviter de tirer deux fois la même unité finale.

Pour évaluer les méthodes proposées, les auteurs appliquent la méthode de Monte Carlo à des données du recensement, en simulant le plan de sondage de l'EPA.

## REFERENCES

[1] Drew, J.D., "Sample Update - A Mapping Procedure for Cases Where the number of Pre- and Post-Update Sample Are Not Equal", Internal Statistics Canada Technical Memorandum, Household Surveys Development Division (1978).

[2] 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.

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

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

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

## APPENDIX (A)

## Operational Aspects of 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 with the rotation scheme as described in Section 3 (page 8).

At some point in time, revised household counts are obtained, and revised inverse sampling ratios $R_i'$; i=1, 2, ..., N; are defined as before so that $\sum_{i=1}^{N} R_i' = R$. Then the R unique ordered samples based on the revised sizes are:

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

Thus, 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

$$x_i/R_i \geq x_i'/R_i', \tag{A.1}$$

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 $R_i|R_i$, $(R_i-1)|R_i$, ..., $(R_i-R_i'+1)|R_i$ are mapped respectively into the samples $R_i'|R_i'$, $(R_i'-1)|R_i'$, ..., $1|R_i'$ and the samples $(R_i-R_i')|R_i$, $(R_i-R_i'-1)|R_i$, ..., $1|R_i$ are temporarily left unmapped.

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

c) Since $\sum\limits_{i \varepsilon D} (R_i-R_i') = \sum\limits_{i \varepsilon I} (R_i'-R_i) = f$, say, the unmapped pre-update samples in the decreasing clusters can be mapped in a one-to-one fashion into the available post-update samples in the increasing clusters. There are f! possible mappings. Ideally, we might choose that mapping which maximizes the time interval (i.e. number of rotation periods) before any post-update sample rotates 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 $i_\ell'$. 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 $a_\ell$ is the minimum time interval for mapping $\ell$, then the mapping $\ell^*$ for which $a_{\ell^*} = \max\{ a_1, a_2, \ldots, a_d\}$ is chosen. For a given $\ell$, 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 $i'_{\ell}$. There are $v_{\ell}$ unmapped samples in the decreasing cluster $i'_{\ell}$ - map all of these samples in the increasing cluster $i''_{k_1}$, $i''_{(k_1+1)\text{mod } e}$, $\ldots$ exhausting $w_{k_1}$ available samples in the increasing cluster $i''_{k_1}$ before proceeding to $i''_{(k_1+1)\text{mod } e}$ and similarly for $i''_{(k_1+1)\text{mod } e}$, $i''_{(k_1+2)\text{mod } e}$, $\ldots$usin as many of the increasing clusters as required. After mapping the $v_{\ell}$ samples from decreasing cluster $i'_{\ell}$ into increasing clusters $i''_{k_1}$, $i''_{(k_1+1)\text{mod } e}$, $\ldots$, the corresponding counts of available samples i.e. $w_{k_1}$, $w_{(k_1+1)\text{mod } e}$, $\ldots$ are adjusted. Next, take the decreasing cluster $i'_{(\ell+1)\text{mod } d}$ and find the first cluster $k_2 \varepsilon I$ with $i''_{k2} > i'_{(\ell+1)\text{mod } d}$ and as before map all the $v_{(\ell+1)\text{mod } d}$ unmapped samples in the decreasing cluster $i'_{(\ell+1)\text{mod } d}$ into the available samples in the increasing clusters $i''_{k_2}$, $i_{(k_2+1)\text{mod } e}$, $\ldots$ Repeat this process for clusters $i'_{(\ell+2)\text{mod } d}$, $i'_{(\ell+3)\text{mod } d}$, $\ldots$, $i'_{(\ell+d-1)\text{mod } d}$.

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

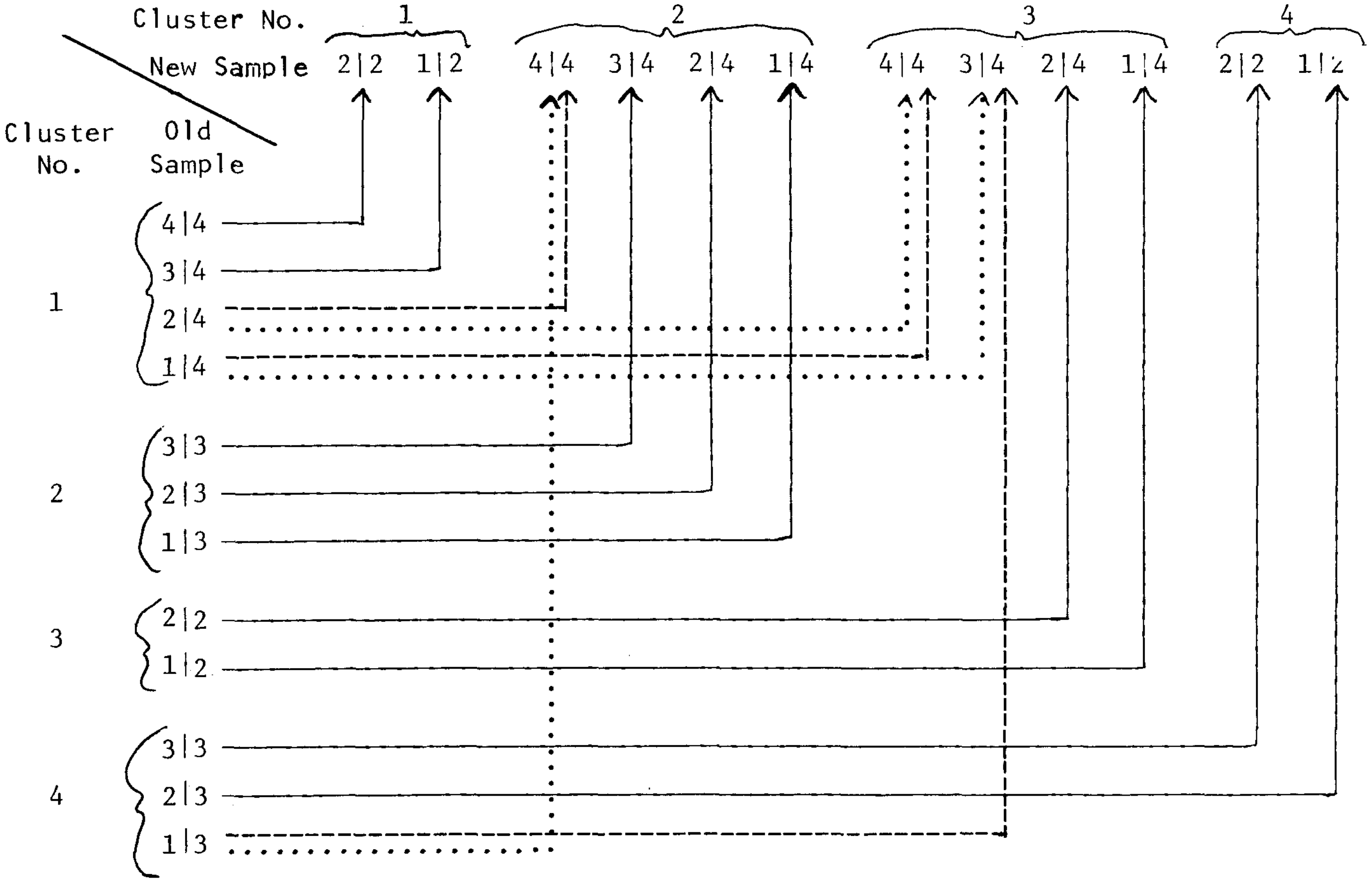

Fig. (1)

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 re-selection 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, re-selection 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 $r_i'$, $1 \leq r_i' \leq R_i'$ is chosen, and if the sample to be introduced is $j|R_i'$, then the systematic samples determined by the starts $r_i'$, $(r_i'+1)$ mod $R_i'$, ..., $(r_i'+j-1)$ mod $R_i'$ are associated with the samples $j|R_i'$, $(j-1)|R_i'$, ..., $1|R_i'$ 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 $R_i' \neq R_i$. 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 re-order 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

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

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

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

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

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

and assigning the maximum possible values to the elements of the matrix A in the order $a_{11}$, $a_{12}$, ..., $a_{1R_i'}$, $a_{21}$, ..., $a_{R_i1}$, $a_{R_i2}$, ..., $a_{R_iR_i'}$ subject to the constraints (A.2) and (A.3). Then the Pr(s→s') is simply given by $a_{ss'}/R_i'$ i.e. the matrix P will be defined as

$$P = \frac{1}{R_i'} A \tag{A.4}$$

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

| | | | Post-update Start | | | | |
|---|---|---|---|---|---|---|---|
| Pre-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 previous table, we find $Pr(1\rightarrow1) = \frac{6}{7}$, $Pr(1\rightarrow2) = \frac{1}{7}$ 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 re-order 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 $r_i'$, $1 \leq r_i' \leq R_i'$ is chosen and a rotation schedule comprising a sequence of systematic samples is determined in the same manner as prior to update.

## Re-ordering 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 on the following page, 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 $\ell_i$; $1 \le \ell_i \le L_i$, is determined. Then in the order $\ell_i$, $(\ell_i+1)$ mod $L_i$, ..., $(\ell_i+L_i-1)$ mod $L_i$, 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 at t=1, cluster i with $R_i = 6$, $r_i = 1$ was selected with the sample 6|6, and that $N_{i1} = 16$. 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 re-order 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 $N_{i1} = N_{i4} = 16$, 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 ≐ .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.