SYNCHRONISED SAMPLING

Richard McKenzie and Bill Gross, Australian Bureau of Statistics
Bill Gross, Australian Bureau of Statistics PO Box 10, Belconnen, ACT, 2616, Australia
bill.gross@abs.gov.au

ABSTRACT

Synchronised Sampling is a Permanent Random Number method of selection to control sample rotation within repeated stratified surveys and to control overlap between different surveys. The sample in each stratum is specified as an interval and overlap is achieved by, for each stratum, constraining the selection interval to move within a survey range. While it does not automatically control rotation or overlap when the stratification of a survey changes, some control is achieved by careful choice of selection intervals and survey ranges. The paper describes the basic algorithms for synchronised sampling and the techniques used to set survey ranges and selection intervals when the stratification changes and assesses the effectiveness of the techniques.

Key Words: Coordination of surveys, Permanent Random Number sampling, Change of stratification

1. INTRODUCTION

Synchronised Sampling is a technique which has been used by the Australian Bureau of Statistics since 1983 to control sample rotation within surveys and overlap between surveys. It relies on assigning a permanent random number to each unit on the business register and selecting units whose random numbers lie in an interval. Rotation control is achieved by moving the interval to the right and overlap is achieved by, for each stratum, constraining the selection interval to move within a survey range.

The paper describes the basic algorithms for synchronised sampling and the techniques used to set survey ranges and selection intervals, especially after major changes in stratification. It also gives an assessment of how effective the techniques for setting survey ranges and selection intervals are.

2. DESCRIPTION OF SYNCHRONISED SAMPLING

Synchronised Sampling (Hinde and Young 1984) was developed in ABS in the early 1980's. The selection method was adapted from the JALES method (Atmer et al 1975). A more extensive description is given in Brewer et al (2000).

2.1. Basic Method

Each unit has a permanent random number from the uniform distribution on [0,1). The selections are specified as an interval in [0,1), whose start and end points lie on random numbers of units at the time of selection. The start point is in sample but the end point is not. To achieve a desired sample size, n, while allowing for births and deaths in the population, the start or end points are moved, but only to the right, to prevent units re-entering the sample.

For the first selection, a provisional start point, s_0 is set and the first n units at or to the right of s_0 are selected. They are described by the interval $[s_1,e_1)$. For the next selection, only the start and end points s_1 and e_1 need to be remembered and the number of units from the new population in the trial interval $[s_1,e_1)$ is compared with the desired sample size n. If it contains n units (i.e. there have been zero net births or deaths in the selection interval) then it remains the selection interval. If it contains more than n (i.e. there have been net births in the interval), the start point moves to the right, while if it contains less than n (i.e. there have been net deaths in the interval), the end point moves to the right. The same process applies if the sample size has changed from time 1 to time 2. For the case n=3, Fig 1 illustrates the first selection and Fig 2 shows selection after population births.

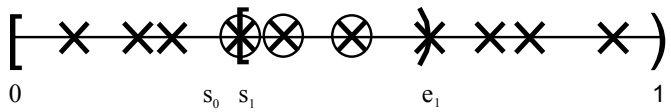
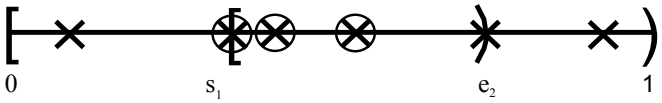


Fig 1: First selection. (⊗ : selected unit)



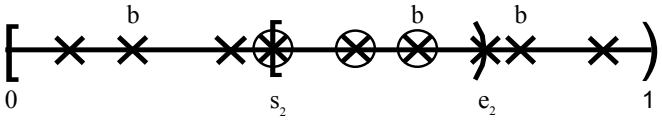
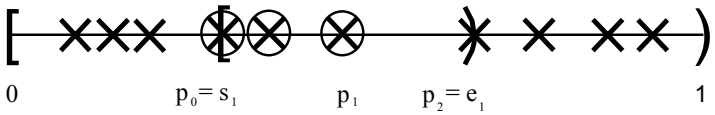


Fig 2: Selection after population births (b).

2.2. Control of Rotation

Planned rotation removes a unit after no more than R consecutive survey cycles in sample. R can be stratum specific. When a sample of size n is first selected, it is partitioned into rotation groups of sizes $n_1,..,n_R$ by projected starts, p_r, (r=0,..,R) where $p_0=s_1$, $p_R=e_1$, so that, at the time of initial selection, there are n_r units in each interval $[p_{r-1}, p_r)$. At the next selection, $[p_1, e_1)$ is used as the trial interval and the start and end points are moved as for the basic method to select n units. Hence s_2 is at, or to the right of p_1. New projected start points are set by dropping p_0, shifting the numbering of the rest and taking p_R as the new end point e_2. At the (R+1)'th selection, the start point s_{R+1} is at, or to the right of $p_R=e_1$, and so all units in the first sample have been rotated out. If there is a significant number of population births, then the start point may shift over more than one projected start, and the trial interval is modified to ensure that s_{r+1} is at, or to the right of p_r. For example, if $s_2 > p_2$, then the trail interval for the third selection is $[s_2, e_2)$ rather than $[p_2, e_2)$. Fig 3 illustrates a simple example.



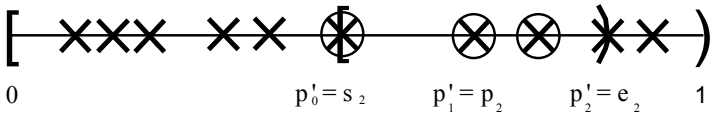


Fig 3: Rotation with pattern 2,1.

Due to planned rotation and to population changes, the selection interval moves to the right. At some time, there will not be enough units in the interval [s,1) and the additional units are selected to the right of 0, so that the selection interval becomes split into [s,1) and [0,e).

2.3 Minimising Overlap between Surveys

The aim is to prevent a unit being selected in more that one survey. Synchronised Sampling, operated correctly, can achieve this even if the surveys concerned have different stratification and different rotation rates, provided that the sum of the sampling fractions in a sub population is not too high. Protection from selection in other surveys is achieved at the expense of a decrease in the time before a unit may be reselected in the original survey. However, particularly in the 'small units' strata, where the sampling fractions are low, the reduction in the time before reselection will not appreciably increase the individual's perceived response burden.

Where there is rotation, the position of the selection interval changes from one cycle of a survey to another. To control overlap between surveys, each stratum of each survey is given a fixed interval, the survey range, and the selection interval is forced to remain within the survey range. The survey range should be large enough to hold the selection interval, plus enough unselected units to permit worthwhile rotation to occur.

In the simplest case where two surveys have the same stratification, the samples can be made disjoint by setting the survey ranges in each stratum to be disjoint. If the surveys have different stratifications then the constraints on the survey ranges are more complex. To ensure that the sample from a stratum in one survey is disjoint from the sample in another survey, the survey range for the stratum of the first survey must be disjoint from any of the survey ranges of intersecting strata of the second survey. It follows that groups of strata must be considered simultaneously - if two strata from the surveys intersect, then their survey ranges need to be considered simultaneously, and hence two strata linked by a chain of intersecting strata need to have their survey ranges considered simultaneously. To limit the length of such chains, some standard boundaries for industry and size are used. The process of allocating the position and length of the survey ranges for all the strata for all surveys is complex and is done manually with computer produced diagnostics. Hinde and Young (1984) describe how this was done when Synchronised Sampling was first used for a group of surveys. Section 3 describes the tools used to set survey ranges when a new survey is added to that group or when the stratification of an existing survey is changed.

2.4. Relationship with Stratification

While the same permanent random number is used for all surveys, at the system level, the method operates separately within each stratum in each survey. In particular, start points, survey ranges and rotation rates are stratum specific. As a consequence there is no necessary relationship between the stratifications for different surveys, although some standard boundaries are used for industry and size to make it easier to control overlap between surveys. There is also no system limitation on the number of surveys which can use Synchronised Sampling, although the total sampling fraction constrains the effectiveness of overlap control and the setting of survey ranges is more complex with a larger number of surveys.

When a unit changes stratum, Synchronised Sampling treats it as a death in the old stratum and a birth in the new stratum, so it cannot automatically take account of its selection status in the previous stratum. This is a major drawback when a survey is redesigned and the stratification is changed, since there are a very large number of stratum changes. However, it is highly desirable to maximise the common sample under the two stratifications, both to minimise the sample error on estimates of change and to minimise cost as new units in sample are more expensive to process. Synchronised Sampling can achieve this to a limited extent by appropriate choice of start points in the new stratification.

In summary, the key properties of the method described above are:

- it controls overlap between a number of surveys
- it achieves pre assigned sample sizes in each stratum
- it copes with different stratification in different surveys
- it controls rotation by guaranteeing a maximum time in sample
- the rotation rate can be stratum specific
- it allows for births and deaths on the frame

Properties described in Brewer et al (2000) are:

- it gives very nearly simple random sampling within stratum
- it can control overlap for single establishment firms between an establishment survey and a firm survey
- it does not automatically control rotation or overlap when units change stratum

3. TOOLS FOR SETTING START POINTS AND SURVEY RANGES

Start points for selection intervals need to be reset when a survey is redesigned and the stratification changes markedly, and it is desirable to set these start points in a way which maximizes the common sample under the two stratifications. Under the old stratification the sample generally consisted of one interval in each stratum, but would be two intervals if the selection interval had split because it had recently reached the end of its allocated survey range, and recommenced taking new selection from the start of the survey range. Viewed in the new stratification, the old sample consists of a larger number of "partial" intervals – one or two for each old stratum which intersects with the new stratum. A "partial' interval contains some, but not necessarily all of the points between two numbers, and a selection interval in one old stratum is a partial interval in the new stratum because only units from that old stratum population are selected. Units in the interval from a different old stratum would not, in general, be selected, and these 'dilute' the selection interval and make it partial in the new stratum. Figure 4 illustrates a simple case where two strata are combined. As the sample, of size n, for the new stratum is specified by a start point and consists of the n units at, or to the right, of the start point, the aim is to find a start point which maximizes overlap between the old and new samples.

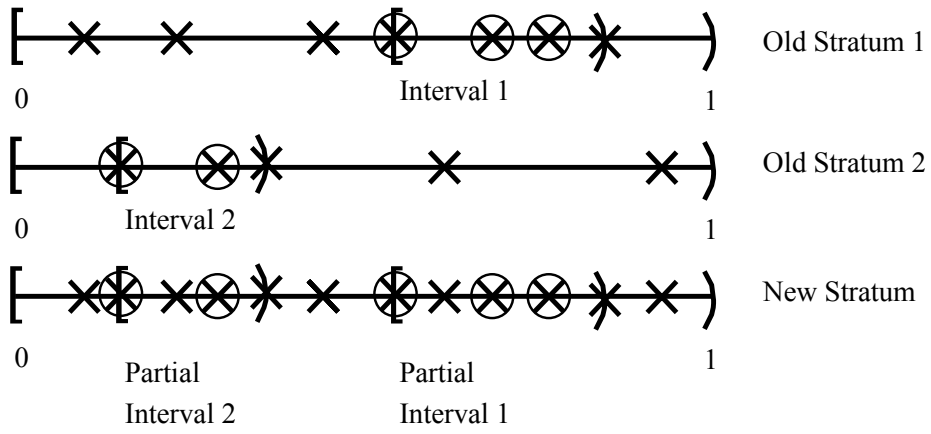


Fig 4 Selections when two strata are combined

The problem is similar when setting the survey ranges for a new or restratified survey, especially when the stratification is different from the existing surveys. Under the new stratification the survey ranges of the other surveys consist of a number of intervals or partial intervals. The aim, however, is to find an interval which minimizes overlap with the existing intervals and partial intervals. The survey ranges or selection intervals for the new strata could also be chosen to minimise overlap with the existing samples. For clarity, the rest of the description refers only to selection intervals.

In developing the tools to set start points, a key issue was how to represent and handle the partial intervals on the number line, especially when a number of different stratifications are involved. In the approach taken the realized number line is treated as a set of discrete numbers – the ones corresponding to units in the stratum - rather than as a continuum. Each unit in the new stratum is flagged, to indicate whether or not it was selected in any of the pre-existing surveys. Clearly, the unit's stratum in each pre-existing survey is used to find the appropriate selection interval and thence to set the selection flag for that survey. However, information about the old stratifications is not needed after that. In this way, any complicated relationships between the stratifications are not treated explicitly, but they are implicit in the distribution of the flags in the population.

The method of choosing the new start point is also simple. Each unit in a stratum population is examined as a possible start point and given a score which measures how well the sample with that start point meets the desired overlap criteria. In order to calculate the overall score for a possible start point, a score is first calculated for each population unit, by aggregating a weight for each survey in which it is selected. The weight is positive if maximum overlap with the survey is desired and negative if minimum overlap is sought and the magnitude of the weight reflects the importance of controlling overlap with a survey, e.g. because it places a high load on respondents. For the new stratum, the desired sample size is known, and so for each possible start point, the sample which would result can be determined. The overall score for a possible start point is the sum of the unit level scores over the sample it would generate. To maximise the common sample under new and old stratifications, the point after the last occurrence of the maximum score is chosen as the start point for the selection interval. The point after the last occurrence of the maximum is chosen because using the last occurrence leads to a bias, and using the next point reduces the bias. Flack et al (2000) consider a simple case where two strata are combined, sample is selected in only one of the old strata (stratum 1) and the sampling fraction in the new stratum is much greater than in old stratum 1. In this case the bias towards selecting from stratum 1 is of order 1, if the last maximum is used, but is of order 0, if the next start point is used.

In practice the criteria for overlap may be complex – maximizing common sample as well as avoiding samples or ranges for other surveys with different levels of importance – and the weights may not fully reflect this complexity. To help assess these more complex situations, the scores for each start point are plotted against random number. In the same plot, the samples for the other surveys involved are also shown. Figure 5 shows the graph for choosing a survey range which minimizes overlap with the samples from eight other surveys. The weight for each of the surveys is –1 and a range of length 0.39 was sought. The random numbers for the samples for the eight surveys (AWE, JVO, etc) are shown as short horizontal intervals to the right of the survey acronym. The maximum score is –1, which corresponds to the interval shown and contains only the second point in the EEH sample.

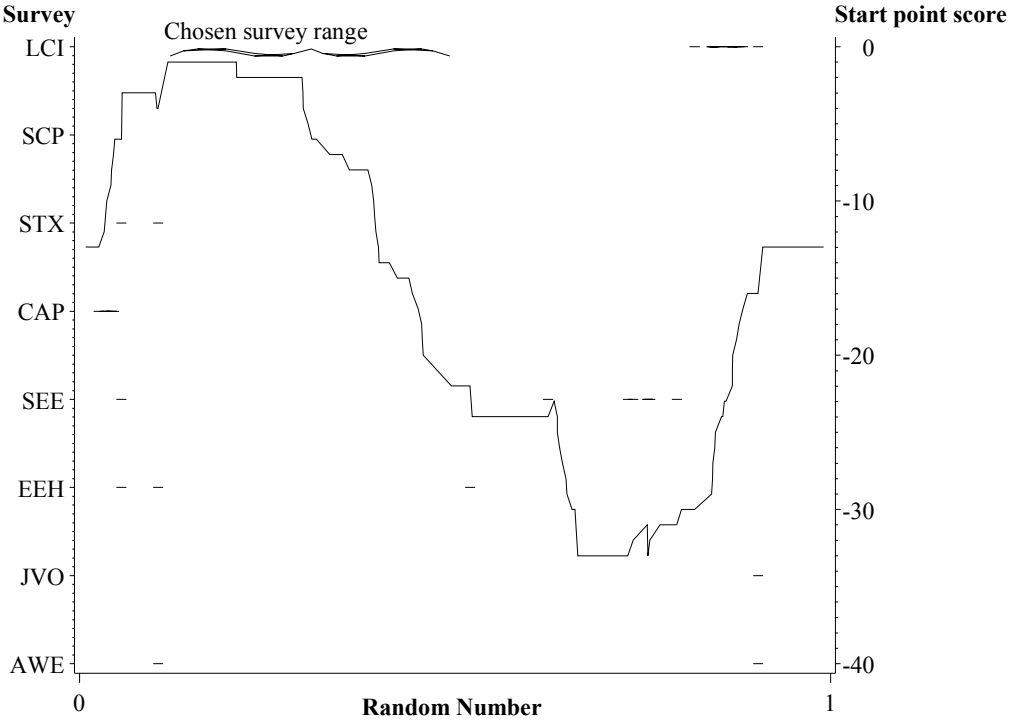


Fig 5: Graph of scores for potential start points, showing locations of existing survey samples

4. EFFECTIVENESS OF ROTATION CONTROL AT A SURVEY REDESIGN

When the stratification and allocation of a survey change some rotation is unavoidable as the probabilities of selection change. In a sub-population where the probability of selection decreases then some of the sample must be rotated out to realize the lower probability of selection. For unit i, with probabilities of selection p_{1i} under the old design and p_{2i} under the new design, the maximum probability of it being in the common sample is $p_{max\ i} = \min(p_{1i},p_{2i})$. (Setting p_{1i} or p_{2i} to zero covers changes in scope.) The effectiveness of a method to maximise the common sample can be assessed by comparing the achieved common sample with the maximum attainable, $\Sigma_i\ p_{max\ i}$, the population total of $p_{max\ i}$.

The graphical method described in section 3 was used to set new start points for the 1999 redesign of the ABS Engineering Construction survey. The frame changed markedly from about 15,000 to about 26,000 with about 9,000 in common. The old design had a sample size of 2,500 from 129 strata, while the new design was for a sample of 3,000 from 265 strata. While the maximum attainable common sample was 1628 the start points chosen using the method gave a common sample of 501, or 30% of the maximum. As this is the best attainable using Synchronised Sampling, the example illustrates Synchronised Sampling's limited ability to maximise common sample at survey redesigns.

5. EFFECTIVENESS OF OVERLAP CONTROL

When assessing the effectiveness of a method for controlling overlap between a group of surveys it is useful to consider two aspects. The first is the expected load for a unit - defined as the sum of its probabilities of selection in the surveys in the group. The expected load is determined by the allocations for the surveys i.e. the sampling fraction chosen for each stratum.

The second is the distribution of the load (defined as the number of surveys in which a unit is selected) conditional on the expected load. It is this conditional distribution of load that the selection method affects. The aim of controlling overlap between surveys is to minimise the number of units which have a high load and the ability to achieve this is constrained by the expected load. For example, if the expected load is 1.5, then it is not possible for it to have a realised load which is only ever 0 or 1, because then the expected load on a unit would be less than or equal to 1. Hence there must be some chance of selecting the unit in more than one survey. More generally, it is easy to see that, if p_i and d_i are defined so that

p_i is an integer

$0 \leq d_i < 1$

and

expected load = p_i+d_i

then unit i must have a chance of being selected in p_i+1 surveys. A selection method which ensures that there is no chance of unit i being selected in more than p_i+1 surveys is one which results in

Pr(unit i selected in p_i+1 surveys) = d_i

Pr(unit i selected in p_i surveys) = $1-d_i$.

Such a selection method is optimal in terms of minimizing high load.

While such a method is optimal for minimizing high load, it also gives the minimum spread of load between units, which is desirable as it makes the distribution of load fair and equitable.

For a selection method, avoidable load is defined as the number of selections in multiple surveys in excess of the optimal p_i+1 for a unit with expected load p_i+d_i. More formally, if unit i with expected load p_i+d_i is selected in m_i surveys then its contribution to avoidable load is:

avoidable load (i) = 0 if $m_i \leq p_i+1$

= $m_i - (p_i+1)$ if $m_i > p_i+1$

The avoidable load for the method is the population sum, Σ_i avoidable $load_i$, of each unit's contribution. We take the avoidable load as the measure of how effective a selection method is in controlling overlap between a group of surveys.

In practice, there is a hierarchy of units used in business surveys (e.g. establishment as a sub-unit of a firm) and the notions of load and expected load can be modified to allow for using more than one level in the hierarchy. Load is attributed to the highest level unit in the hierarchy and is measured by the number of sub-units selected. Expected load is the sum of the probabilities of selection of sub-units.

The effectiveness of Synchronised Sampling in controlling overlap was assessed for medium sized businesses (defined as having between 20 and 200 employees) in the twelve ABS surveys which used Synchronised Sampling in 1999. In these medium sized businesses, the sampling fractions can be quite large and so they severely test the method's effectiveness in controlling overlap. Table 1 shows avoidable load for the businesses as well as the distribution of actual load against expected load. The shaded cells represent avoidable load, i.e. counts of business which in practice are selected more times than their expected load would suggest is necessary. It shows that Synchronised Sampling does not control overlap very well for these businesses. The proportion of total load which is avoidable is reasonable at 12% of total load, but where the expected load is less than 1, the avoidable load is high, at 17%.

Table 1: Counts of Medium sized units by Expected Load and Actual Load

Expected Load	Actual Load (Number of times selected)							Avoidable Load	Total Load	% Avoidable Load
	0	1	2	3	4	5	6+			
[0,1]	25433	6941	1246	185	17	0	0	1667	10056	17
(1,2]	489	2180	1180	384	77	7	0	559	6035	9
(2,3]	17	141	457	298	106	24	5	169	2523	7
(3,4]	0	3	25	102	78	35	9	55	902	6
(4,5]	0	1	3	17	60	64	17	20	723	3
(5,6]	0	0	0	2	9	42	38	7	487	1
(6, ∞)	0	0	1	0	0	4	62	17	497	3
Total	25939	9266	2912	988	347	176	131	2494	21223	12

(Shaded cells contribute to avoidable load.)

While Table 1 shows the full range of expected load, it includes load in take all strata. However, no selection method can redistribute the load from take all strata, and so it is not useful to consider them in assessing the effectiveness of a selection method in controlling the distribution of load. Table 2 shows the avoidable load and distribution of load when take all strata are excluded, i.e. it applies to sampled strata. Units which had more than one sub-unit were also excluded from the analysis.

Table 2: Medium size simple units in sampled strata - Avoidable Load and counts by Expected Load and Actual Load

Expected Load	Actual Load							Avoidable Load	Total Load	% Avoidable Load
	0	1	2	3	4	5	6+			
[0,1]	24331	6729	1239	188	19	0	0	1672	9847	17
(1,2]	498	1151	717	251	61	5	0	388	3607	11
(2,3]	10	44	64	43	19	6	2	37	419	9
(3,4]	0	2	0	0	1	0	0	0	6	0
Total	24839	7926	2020	482	100	11	2	2097	13879	15

(Shaded cells contribute to avoidable load.)

Considering only simple units in sampled strata reduces the incidence of high expected load dramatically, and increases the percentage avoidable load slightly from 12% to 15%. As there is little change in the population of units with expected load less that 1, there is no change in the proportion of their total load which is avoidable.

ABS practice is to have survey ranges which are at least three times the sampling fraction. Hence Synchronised Sampling should control overlap very well if the expected load is less than 0.3, as it should be possible to have survey ranges which are disjoint if surveys use the same stratification. Table 3 shows detail of the avoidable load for units where the expected load is below 1.0. It indicates that, while the avoidable load is less for these low expected loads, it is still quite large. This is due to diversity in the stratification or sub-optimal setting of survey ranges.

Table 3: Medium size simple units in sampled strata - Avoidable Load for Expected Load below 1.0.

Expected Load	Avoidable Load	Total Load	% Avoidable Load
[0.0,0.1]	6	350	1.7
(0.1,0.2]	60	989	6.1
(0.2,0.3]	97	1178	8.2
(0.3,0.4]	124	1095	11.3
(0.4,0.5]	189	1328	14.2
(0.5,0.6]	228	1277	17.8
(0.6,0.7]	231	982	23.5
(0.7,0.8]	278	1062	26.2
(0.8,0.9]	265	936	28.3
(0.9,1.0]	194	650	29.8
Total	1672	9847	17.0

ACKNOWLEDGEMENTS

Tenniel Guiver and Anna Poskitt developed the programs for the method described in section 3. John Kieley, Lloyd Flack and Doug Lincoln applied them and computed the effectiveness of the methods for sections 4 and 5.

REFERENCES

Atmer, J.G., G. Thulin, and S. Bäcklund (1975).”Coordination of Samples with the JALES Technique,” *Statistik Tidskrift*, 13, pp. 443-450. (In Swedish with English summary).

Brewer, K.R.W., W.F. Gross, and G.F. Lee (2000). “PRN Sampling: The Australian Experience,” *ISI Proceedings: Invited Papers, IASS Topics, Helsinki August 10-18, 1999*, pp. 155-163.

Flack, L., W. Gross and A. Herning (2000). “Selection Bias when Maximizing Common Sample at Restratification using Synchronised Sampling,” unpublished report, Australian Bureau of Statistics

Hinde, R. and D. Young (1984) "Synchronised Sampling and Overlap Control Manual," unpublished report, Australian Bureau of Statistics.