Rethinking Sampling for UK Business Surveys

Markus Gintas Šova[1], Nathan Joseph Calvin Price[2] and Gareth G. James[3]

[1]Office for National Statistics, UK, e-mail: markus.sova@ons.gov.uk
[2]Office of the Chief Government Statistician, Zanzibar, e-mail: Nathan.Price.OCGS@gmail.com
[3]Office for National Statistics, UK, e-mail: gareth.james@ons.gov.uk

Abstract

The introduction of electronic data collection allows us to rethink how surveys are designed. We present two subsampling methods for business surveys to allow survey modularisation.
Keywords: business surveys, survey modularisation, PRN subsampling

1 Introduction

Traditionally, business surveys have consisted of paper questionnaires dispatched by post. The questionnaire has one or more (sometimes many) questions which the business completes by hand and then returns to the National Statistical Institute (NSI). The concept of *one* survey is thus closely associated with the design of *one* questionnaire.

More recently, the advent of electronic data collection over the internet has raised interesting possibilities. For example, there are some questions which occur in more than one survey, albeit sometimes with subtle variations. If these can be harmonised, then the surveys could be integrated into a single survey with some core questions asked of all selected businesses, and the remaining questions modularised and given to subsamples of the survey. Alternatively, if a survey has more than one question, electronic data collection would allow the survey to be *de-integrated*, with all selected businesses being asked only one question. This would be particularly useful for large strata of small businesses because it would result in a fairer short-term distribution of response burden.

This paper examines how subsampling can be applied to allow survey integration by modularisation. The following section briefly describes how the Office for National Statistics (ONS, UK's NSI) currently implements rotational sampling. Two subsampling methodologies are presented in section 3, with some concluding remarks in section 4.

2 PRN Sampling

Since 1994, ONS has used the Inter-Departmental Business Register (IDBR) as the sampling frame for most of its business surveys. The IDBR's key sampling methodology

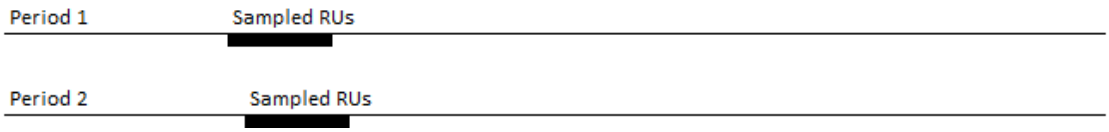
is stratified rotational sampling using permanent random numbers (PRNs). The IDBR's sampling units are called *reporting units* (RUs). Most RUs are enterprises, but some enterprises are split into two or more RUs for statistical purposes. When an RU is created on the IDBR, a PRN is generated whose value is permanently associated with the RU. From the point of view of a survey, each stratum consists of a set of RUs distributed along the PRN line (the set of all possible PRN values). For any stratum h a sample of size n_h is selected as the first n_h RUs on the PRN line whose PRN values are greater than or equal to a specified PRN start point, as shown in figure 1. Here the long thin line represents the PRN line, and the short thick line represents the selected sample which starts from the PRN start. If the PRN start is so large that there are insufficient RUs with a greater PRN value, the shortfall is made up for by selecting RUs from the beginning of the PRN line. Thus the PRN line is really a ring, but for clarity we shall continue to portray it as a line.

Figure 1: Representation of a PRN sample for one stratum

Because all the PRN values are independently generated from the same distribution, we have a simple random sample for the stratum.

After the sample has been selected, each stratum's PRN start is recalculated in preparation for the next survey period. This is done by moving the PRN start to the right so that a set number of RUs leave the sample. Over time the sample moves along the PRN line in a controlled way, as shown in figure 2. For further details see Ohlsson (1995).

Figure 2: Representation of a rotating PRN sample



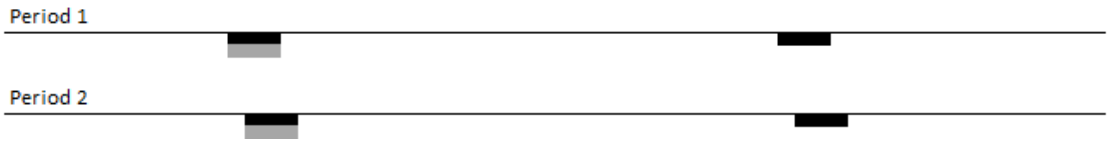
We note two important properties: Firstly, as the sample moves along the PRN line, every RU takes its turn at being sampled, ensuring a fair distribution of response burden over time. Secondly, the sample overlap between consecutive periods is controlled, giving a better variance of change than with independent sampling (see Lindblom, 2014).

3 Methodologies for Subsampling

ONS has started developing a new business register, called the *Statistical Business Register* (SBR), to replace the IDBR, giving the opportunity to specify one or more subsampling methodologies for survey modularisation. Ideally, we want the subsampling methodology to retain the key advantages of rotational sampling using PRNs, namely a fair distribution of burden over time and a controlled high overlap proportion between successive periods. James (2016) examines a number of subsampling methodologies. We present two of these as being particularly promising.

In figure 2, the sample is represented by a segment of the PRN line. Now consider several such segments of equal size, equally spaced along the PRN line. The set of these segments is the main sample. One (or more) of the segments is the subsample, whose RUs are to answer questions from a module. In principle, there can be as many distinct modules as there are segments. This method is more fully described and simulated by Price (2016). Figure 3 depicts how this Multiple Segment method works over time for a simple two segment example. Each segment is represented by a short thick black line. The subsample is represented by a thick grey line.

Figure 3: Representation of a rotating Multiple Segment sample and subsample



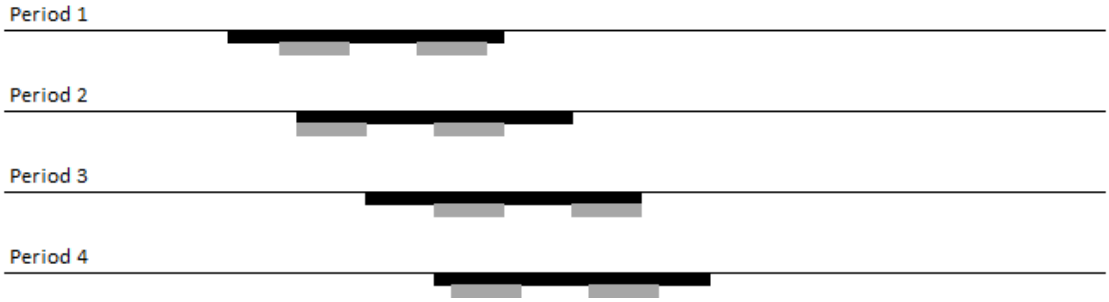
Each segment operates as a PRN sample. So the subsample has our desired properties of fair burden distribution and controlled high overlap. The equal sizes of the segments ensure that they move along the PRN line at the same rate (although with some sophistication this requirement can be eased). Otherwise, with segments moving at different speeds one segment will close in on another, resulting in RUs being rotated into the trailing segment only a few periods after being rotated out of the leading segment. Eventually the segments will collide, resulting in RUs being in the sample for twice as long as desired. The problems of encroaching and colliding segments can also be caused by births and deaths of RUs in the stratum. Therefore the Multiple Segment method requires the distances between segments to be monitored, and if they start getting too close then the rotation rates of individual segments may need to be temporarily amended. Price (2017a) offers and simulates solutions to this issue.

We now consider the main sample as a single segment on the PRN line, with a subsample represented by a subsegment. If this were moving at the same speed as the main sample then the overlap in the subsample between consecutive periods would be small, possibly even zero. Conversely, if the subsample were to move along the PRN sample at a slower rate than the main sample in order to maintain a high overlap, eventually the subsample would reach the start of the main sample and need to be repositioned at the end, resulting in a reduced subsample overlap at the time of the repositioning. This is shown in figure 4, where we see that there is a reduced subsample overlap between periods 2 and 3 and between periods 3 and 4. Furthermore, there are some RUs which do not get subsampled at all in this pass of the PRN line.

Suppose we now increase the number of subsamples whilst decreasing their size. Figure 5 shows two subsamples applied the scenario of figure 4. Only between periods 2 and 3 is the subsample overlap reduced. Further increasing the number of subsamples will lessen the reduction in the subsample overlap but will increase its frequency. We call this the *Stonehenge Method* due to it resembling a large stone being moved on wooden rollers.

Figure 4: The rotating subsample problem (the subsample moves more slowly)

Figure 5: Two rotating subsamples



4 Concluding Remarks

Of the two subsampling methods presented, that of Multiple Segments is elegant in its simplicity, with all subsamples having the desirable properties of a rotating PRN sample. However, the method requires monitoring with occasional intervention to ensure that no segment gets too close to another. The Stonehenge Method does not have this risk as it has only one segment. The subsample overlap is occasionally reduced, but this reduction can be lessened by having two or more subsamples. We have included both of these subsampling methods in the sampling specifications for the SBR.

References

James, G.G. (2016). *Options for sub-sampling.* ONS internal report.

Lindblom, A. (2014). On Precision in Estimates of Change over Time where Samples are Positively coordinated by Permanent Random Numbers. *Journal of Official Statistics* **30(4)**, 773 - 785.

Ohlsson, E. (1995). Coordination of Samples Using Permanent Random Numbers, in *Business Survey Methods*, eds. Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. & Kott, P.S. Wiley.

Price, N.J.C. (2016). *A PRN Sub-Sampling Specification.* ONS internal report.

Price, N.J.C. (2017a). *Multi-Segment Sampling: Making it Work.* ONS internal report.

Price, N.J.C. (2017b). *Overlap for GlassesCases-Subsampling.* ONS internal report.