Sampling Coordination of Business Surveys at Statistics Netherlands

Marc Smeets (mset@cbs.nl) and Harm Jan Boonstra (hbta@cbs.nl)

September 4, 2017

1 Introduction

A couple of years ago a coordinated sampling system for business surveys was taken into use at Statistics Netherlands with the purpose to centrally perform the sampling for the Dutch business surveys. Another purpose of the system is to lighten the (perceptual) survey burden of the enterprises by spreading the total survey burden over the enterprises as evenly as possible. This paper focuses on the methodology of the sampling system which is based on the former EDS-system introduced by Huis et al. (1994). Here EDS stands for Enquête Druk Systeem (Survey Burden System). The idea is that in a given group of surveys with a common sampling frame the spread of survey burden is realised by coordination of the sampling and by keeping the total built up survey burden for every enterprise in the sampling frame. The survey burden is then spread both over time and over the surveys in the group.

At this moment the sampling system performs the sampling for 17 business surveys, among which the Structural Business Statistics and the Short Term Statistics. Coordination of sampling can be applied in a group of surveys, where both stratified cross-sectional and rotating panel designs may occur. The spread of survey burden in the group is optimal when all surveys in the group use the same *basic* stratification, but under some restrictions it is allowed to depart from the basic stratification by using substrata.

The sampling algorithm is described in section 2. It is discussed under which conditions substrata can be used and how the algorithm takes into account population dynamics like deaths, births and (basic) stratum movers. In section 3 the results of a simulation study are presented where several aspects of the spread of survey burden are examined and where it is shown that no selectivity is introduced in the drawn samples. In section 4 some concluding remarks are given and some future challenges are discussed.

2 Sampling algorithm

In this section the underlying sampling algorithm of the coordinated sampling system is described. Assume that a group $\mathcal{G}$ of surveys with a common sampling frame U is given, in which both stratified cross-sectional and rotating panel designs may occur. Furthermore, it is assumed that the surveys in $\mathcal{G}$ use the same basic stratification. The use of substrata is discussed in subsection 2.2.

The algorithm produces a simple random sample s_h of n_h units in every basic stratum $h = 1, \ldots, H$. The size of the population U_h is denoted N_h, for $h = 1, ..., H$. For every panel in $\mathcal{G}$ panel rotation is performed with a user-defined and stratum-dependent rotation fraction v_h. And finally, for every survey in $l \in \mathcal{G}$ a stratum-dependent weight $W_{lh} > 0$ is available representing the survey burden caused by this survey.

The idea of the algorithm is to keep a vector of parameters (R_k, B_k, I_{pk}) for every unit $k \in U$, with $R_k \in [0, 1]$ a unique permanent random number (PRN), B_k the total built up survey burden and $I_{pk} \in \{0, 1\}$ the panel membership indicator for every panel $p \in \mathcal{G}$. Then the sample is drawn by selecting the first n_h units according to some ordering of (R_k, B_k, I_{pk}).

Before the first draw of a sample in $\mathcal{G}$, the sampling parameters (R_k, B_k, I_{pk}) are initialised as follows. For every $k \in U$ a unique random number R_k is uniformly and independently drawn from $[0, 1]$ and $B_k = I_{pk} = 0$, for every $p \in \mathcal{G}$. A sample for a cross-sectional survey or the first sample for a panel survey is then drawn by algorithm 1 and a subsequent sample for a panel survey, including panel rotation, is produced by algorithm 2.

Algorithm 1. Draw of a cross-sectional survey $l \in \mathcal{G}$

For every basic stratum $h = 1, \ldots, H$

1. Sort $k \in U_h$ by (i) B_k (increasing) and (ii) R_k (increasing).
2. Select the first n_h units in this ordering. These units form the sample s_h.
3. For every $k \in s_h$, let $B_k = B_k + W_l$.

Algorithm 2. Subsequent draw of panel survey $p \in \mathcal{G}$

For every basic stratum $h = 1, \ldots, H$

1. Sort $k \in U_h$ by (i) I_{pk} (decreasing), (ii) B_k (increasing), (iii) R_k (increasing).
2. Let m_h be the number of units in the panel. If rotation is applied (mostly periodically), define $u_h = \text{round}(v_h m_h)$, otherwise $u_h = 0$. Remove the u_h last units with $I_{pk} = 1$ from the panel.
3. Adjust the panel to get the required sample size n_h:
   - If $m_h - u_h < n_h$, add the first $n_h - (m_h - u_h)$ units with $I_{pk} = 0$ to the panel.
   - If $m_h - u_h > n_h$, remove extra $m_h - u_h - n_h$ units from the panel (last units with $I_{pk} = 1$).
4. Let $I_{pk} = 0$ for every k that is removed from the panel and $I_{pk} = 1$ for every k that is added to the panel. For every k with $I_{pk} = 1$, let $B_k = B_k + W_p$.

2.1 Taking into account population dynamics

Changes in economic activity or reorganisations of the enterprises may lead to population dynamics in terms of deaths, births and (basic) stratum movers of the units $k \in U$. Deaths can easily be removed from the sampling frame. In order to prevent births and stratum movers from being systematically over- or underrepresented in the drawn samples, these units must be indistinguishable from the existing units with regard to (R_k, B_k, I_{pk}). In other words, in every stratum h, births, stratum movers and existing units must have the same joint distribution of (R_k, B_k, I_{pk}). Before every draw of a sample in $\mathcal{G}$ suitable values of (R_k, B_k, I_{pk}) are therefore assigned to the births and stratum movers.

For births a new PRN R_k is uniformly selected from $[0, 1]$ and values for B_k and for every panel $p \in \mathcal{G}$, also for I_{pk} are assigned that are appropriate to R_k. The appropriate values of B_k and I_{pk} are obtained by copying them from the existing unit in the stratum whose PRN is closest to R_k.

For stratum movers the relative position is taken over to the new stratum according to a user-defined ordering of (R_k, B_k, I_{pk}). This is realised by copying the values (B_k, I_{pk}) from the existing unit in the new stratum which is closest to the relative position of the stratum mover according to the same ordering. A new and unique PRN R_k is randomly chosen from the interval of PRNs determined by adjacent units of this existing unit. In this way it can be chosen whether the spread of survey burden has to be taken into account and whether it is more important to keep the built up survey burden or the panel membership. The possible orderings are: (1) only by R_k, (2) by B_k and R_k or (3) by I_{pk}, B_k and R_k.

2.2 The use of substrata

It frequently occurs that surveys in a group want to exclude specific enterprises in a basic stratum from being sampled or, on the contrary, select specific enterprises with probability 1. It is possible to depart from the basic stratification by using substrata. In order to prevent that selectivity is introduced in the drawn samples with respect to the substrata, the parameters (R_k, B_k, I_{pk}) have always to be assigned at the level of the basic strata, while the sampling is done per substratum.

For cross-sectional surveys, every basic stratum h can be divided into J_h substrata h_j with different sampling fractions. By applying the first two lines of algorithm 1 to every substratum h_j a sample of size n_{hj} is obtained. Then B_k is updated for the first n_h units in basic stratum h, where the units are sorted by B_k and R_k (both increasing) and $n_h = \sum_{j=1}^{J_h} n_{hj}$. Note that the values of B_k in h do not correspond anymore to the sampled units, implying that by the use of substrata the spread of survey burden is suboptimal.

For panels both B_k and I_{pk} must be defined at the level of basic stratum h, while at the same time I_{pk} must indicate the panel in every substratum of h. This is realised by allowing a restrictive form of substratification for panels, where basic stratum h is divided into maximal three substrata, one with an arbitrary sampling fraction f_{h1} (main substratum) and the others with fractions 0 and 1. Let the panel indicator I_{pk} represent an imaginary panel in basic stratum h. Sampling and rotation of this imaginary panel and updating the parameters (R_k, B_k, I_{pk}) is done by algorithm 2 with $n_h = f_{h1}N_h$. The real panel can be derived from the (imaginary) panel indicator by taking the units in the main substratum with $I_{pk} = 1$, supplemented with all units in the substratum with fraction 1.

3 Simulation results

By a simulation study it is shown that the sampling algorithm coordinates the sampling in a given group $\mathcal{G}$ of (imaginary) surveys without introducing selectivity in the drawn samples. Furthermore, it is shown that the survey burden is evenly spread over the units in the population. The tests are performed by simulating a series of draws (monthly) from an artificial population U with simulated population dynamics. Suppose that U is divided into 5 basic strata. At the beginning of the simulation U consists of 100,000 units and the sizes of the basic strata are given by Table 1. Births, deaths and stratum movers are simulated in such a way that U remains sufficiently stable over time.

Table 1: Population sizes of basic strata at beginning of simulation

Stratum h	1	2	3	4	5	total
N_h	83,288	12,688	3187	675	162	100,000

In $\mathcal{G}$ a cross-sectional survey is combined with two panels and there are 250 draws simulated. The surveys in $\mathcal{G}$ have U as the common sampling frame and the basic stratification as the common stratification. It is supposed that the surveys have the same weights $W_1 = W_2 = W_3 = 1$. The survey designs and the sampling fractions per basic stratum are given in Table 2. The use of substrata and the effect of using different weights is not examined in this simulation. The existence of possible selectivity in the drawn samples is

Table 2: Surveys in $\mathcal{G}$ with sampling fractions per basic stratum

Survey	Panel	Frequency	Rotation	Stratum	1	2	3	4	5
1	no	year	-		0.03	0.06	0.1	0.15	0.3
2	yes	month	0.1 (yearly)		0.02	0.06	0.1	0.15	0.3
3	yes	month	0.2 (monthly)		0.01	0.05	0.6	0.8	1

examined by computing the realised sampling fractions in every stratum, where the full population and the subpopulations of births and stratum movers are considered seperately. The fractions and the corresponding margins of error at 95% confidence are computed by

$$f_{i,h} = \sum_t n_{i,th} / \sum_t N_{i,th} \text{ and } m_{i,h} = 2\sqrt{\frac{p_h(1-p_h)}{\sum_t N_{i,th}}} \text{ for } i \in \{\text{full, births, movers}\}.$$

Here $n_{i,th}$ is the number of drawn units in stratum h at time t for subpopulation i with size $N_{i,th}$ and p_h is the sampling fraction according to the design. The results for survey 1 are given by Table 3. All fractions p_h turn out to be within the confidence intervals for the full stratum, the births and the stratum movers. From this we conclude that averaged over time no selectivity is introduced in the samples. Also over time there are no trends in the realised fractions (not shown here). The results for the other surveys are similar (also not shown).

Table 3: Mean of realised sampling fractions for survey 1

Stratum	fraction p_h	fraction $f_{\text{full},h}$ (95% margin)	fraction $f_{\text{births},h}$ (95% margin)	fraction $f_{\text{movers},h}$ (95% margin)
1	0.03	0.0301 (0.0003)	0.0288 (0.0047)	0.0317 (0.0124)
2	0.06	0.0600 (0.0009)	0.0488 (0.0168)	0.0546 (0.0127)
3	0.10	0.0999 (0.0023)	0.0667 (0.0516)	0.0909 (0.0355)
4	0.15	0.1506 (0.0059)	0.1250 (0.1785)	0.1127 (0.0848)
5	0.30	0.2989 (0.0152)	0.1429 (0.3464)	0.3846 (0.2542)

The spread of survey burden in $\mathcal{G}$ is investigated by computing the total survey burden for every unit, that is, the number of times a unit is drawn by one of the surveys in $\mathcal{G}$ during the whole simulation. Figure 1 shows the total survey burden per stratum for the units that exist the whole simulation and are no stratum mover. Stratum 5 is not shown, because of the small number of units in this stratum. Due to the larger fractions in stratum 3 and 4, the total survey burden in these strata is larger than in stratum 1 and 2. In stratum 3 and 4 the values are close to each other, implying that most of the units experience the same survey burden. In stratum 1 and 2 the total survey burden is concentrated around two values. In both strata the maximal survey burden is 120, the expected length of stay in the panel of survey 2. It implies that in stratum 1 and 2 units that are drawn in this panel hardly end up in the other surveys. From this, it can be concluded that the total survey burden is evenly spread across the units in the population.

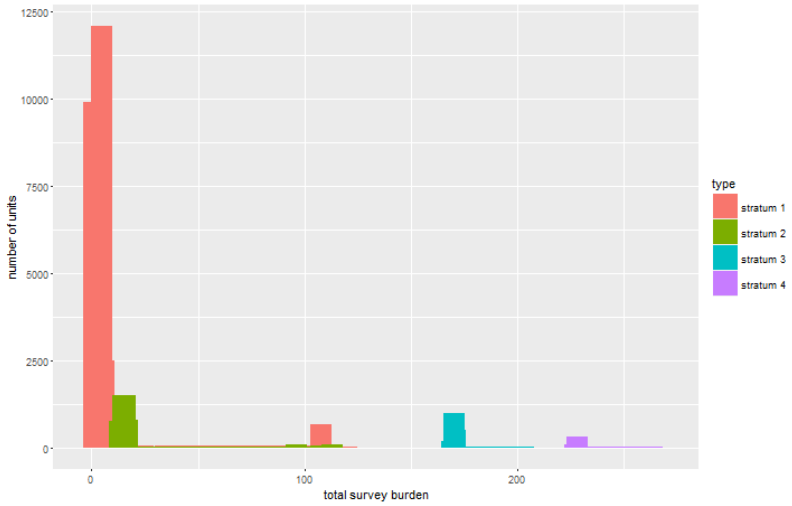

Figure 1: Total survey burden in $\mathcal{G}$ for units that exist the whole simulation and are no stratum mover

The spread of survey burden can also be examined on some other aspects. The lengths of stay in a panel could be examined to check whether the units do not stay too long in the panel given the rotation fraction. The spread of the burden over the surveys in $\mathcal{G}$ could be investigated by computing the length of the survey-free periods of the units in U, that is, the number of successive periods that a unit is not drawn for any of the surveys in $\mathcal{G}$. Another aspect is the occurrence that units are drawn for more surveys in $\mathcal{G}$ at the same time.

Here only the results are given for the survey-free periods in the simulation. The lengths of the survey-free periods are given by Table 4 for every stratum. Note that in stratum 5 there are no survey-free periods, because of the fraction 1 for survey 3. In stratum 1 all units have long survey-free periods. In the other strata the minimal survey-free period is only 1 month, which is caused by the larger sampling fractions in combination with the panels.

Table 4: Lengths of survey-free periods (months) for units that are drawn at least two times in the simulation and are no stratum mover

Stratum	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	167	180	190	189.5	200	211
2	1	24	68	53.1	75	236
3	1	2	2	2.6	3	52
4	1	1	1	1.0	1	6

4 Conclusions and future work

This paper discusses the methodology of the current coordinated sampling system for business surveys at Statistics Netherlands. The sampling system coordinates the sampling in a given group of surveys, where both stratified cross-sectional and rotating panel designs can be combined. The system also takes into account population dynamics. For stratum movers it can be chosen whether the built up survey burden has to be taken into account and whether it is more important to keep the built up survey burden or the panel membership of the stratum mover. By coordination of the sampling the total survey burden is spread as evenly as possible over the enterprises. The spread of survey burden is optimal when the surveys in the group use the same basic stratification. Cross-sectional surveys can depart from the basic stratification by defining substrata with different sampling fractions. For panels only three substrata can be defined in every basic strata, with an arbitrary fraction and the fractions 0 and 1. The extension to a general substratification for panels is one of the future challenges, but the restricted form of substratification will probably suffice in practice.

By a simulation study it is shown that sampling in the group can be coordinated without introducing selectivity in the drawn samples. By examining the total survey burden of the enterprises in the simulation, it is shown that the survey burden is evenly spread over the enterprises. Also some other aspects of the spread of survey burden are examined, such as the lengths of stay in the panel and the lengths of survey-free periods during the simulation.

The main purpose of the further development of the sampling system is to support more business surveys. Therefore, it has to be investigated whether sampling coordination can be applied to other sampling designs, like cluster sampling and Probability Proportional to Size (PPS) sampling, and under which conditions surveys with different sampling designs can be combined.

References

Huis, L. T. V., Koeijers, C. A. J., and de Ree, S. J. M. (1994). *EDS, Sampling system for the Central Business Register at Statistics Netherlands.* Technical report, Department of Statistical Methods, Statistics Netherlands.