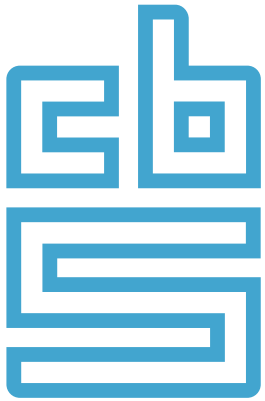


Coordinated Sampling: Theory, method and application at Statistics Netherlands (CBS)

Marc Smeets, Harm Jan Boonstra, Remco Paulissen and Anita Vaasen-Otten

Workshop on Coordinated Sampling for Business Surveys - 1 March, 2019

Introduction

- Aims of coordinated sampling system
  - Support sampling for Dutch business surveys.
  - Sample coordination: both positive (panels) and negative (even spread of total survey burden, both over time and surveys).
  - Facilitate micro monitoring the expected total survey burden that enterprises encounter by CBS.
- Current situation
  - Sampling and coordination for 25 Dutch business surveys.
  - Coordination over time for all surveys.
  - Coordination over surveys only for Structural Business Survey and Investment Survey.
- Methodology of coordinated sampling system
  - Based on former EDS system (Huis et al., 1994).
  - Sampling algorithm implemented in R-package SBS.

Conditions on applied sample coordination

- Support both stratified cross-sectional surveys and stratified rotating panel designs.
- Allow construction of (disjoint) groups of surveys over which sample coordination is applied independently.
- Both cross-sectional surveys and rotating panels can be combined in groups.
- No reduction of total survey burden.
- No guarantees are given to enterprises.
- Coordination is independent of response behaviour.

Basic principles of sampling method

- Based on a PRN method.
- Randomness guaranteed by
  - assigning a unique random number $R_k \in [0, 1]$ to enterprise k.
- Coordination realised by
  - keeping a survey burden value $B_k \geq 0$ for every enterprise k, representing the total built-up survey burden,
  - keeping the actual panel memberships $I_{pk} \in \{0, 1\}$ of the panels p in the group for every enterprise k.
- Sampling scheme:
  - select first units in specified ordering determined by values of (R_k, B_k, I_{pk}).

Initialisation of sampling algorithm

- Given group G of surveys with common sampling frame U.
- Both stratified cross-sectional and rotating panels can be combined in G.
- For every survey l in G a weight $W_{lh} > 0$ is available representing the survey burden caused by this survey in stratum h.
- Initialisation by assigning to every $k \in U$:
  - R_k: unique random number, uniformly and independently drawn from $[0, 1]$.
  - $B_k = 0$: total built-up survey burden in G.
  - $I_{pk} = 0$: panel memberships of panels $p \in G$.

Algorithm for cross-sectional surveys

Draw of stratified cross-sectional survey $l \in G$ with sample size n_h and weight W_{lh} in stratum h:

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

Illustration for cross-sectional survey

Figure 2. Enterprises in random order

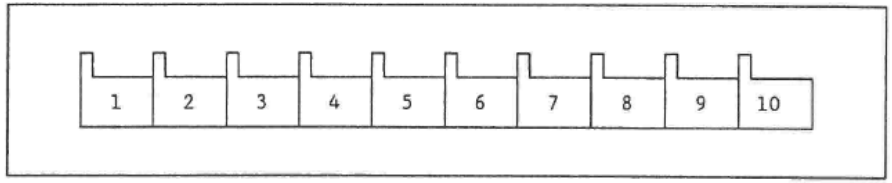


Figure 3. Before the second sample

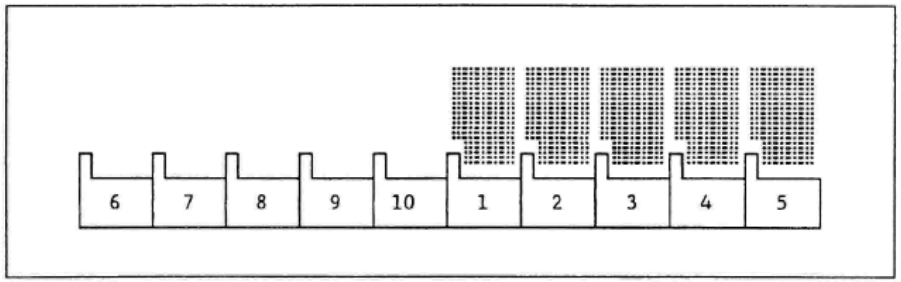


Figure 4. Before the third sample

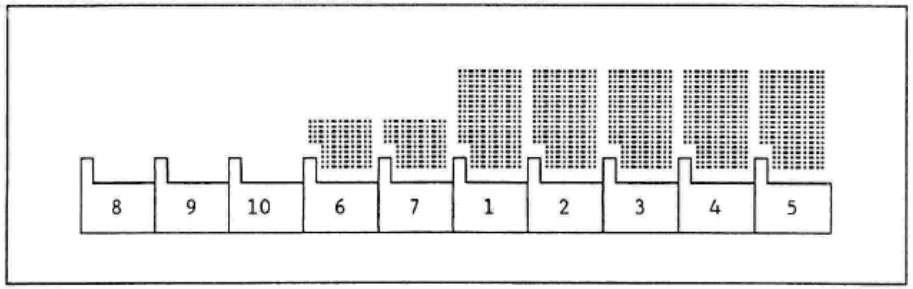
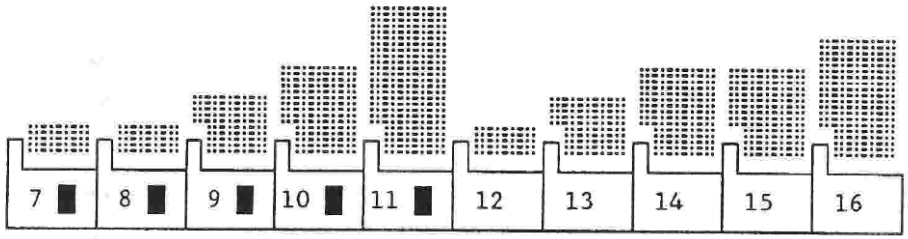


Illustration of panel rotation



Situations:

1. rotation fraction $v_h = 0.2$, sample size $n_h = 5$: 11 out, 12 in.
2. rotation fraction $v_h = 0.2$, sample size $n_h = 4$: 11 out.
3. rotation fraction $v_h = 0.2$, sample size $n_h = 3$: 10, 11 out.

Algorithm for rotating panels

Subsequent draw of stratified rotating panel $p \in G$ with sample size n_h, weight W_{ph} and rotation fraction v_h in stratum h:

1. Sort k by (i) I_{pk} (decreasing), (ii) B_k (increasing) and (iii) R_k (increasing).
2. Define $u_h = \text{round}(v_h m_h)$, with m_h number of units in panel. Remove last u_h units with $I_{pk} = 1$ from panel.
3. Adjust panel to get sample size n_h:
   - $m_h - u_h < n_h$? Add first $n_h - (m_h - u_h)$ units with $I_{pk} = 0$ to panel.
   - $m_h - u_h > n_h$? Remove extra $m_h - u_h - n_h$ units from panel (last units with $I_{pk} = 1$).
   - $m_h - u_h = n_h$? No adjustment.
4. Update I_{pk} and let $B_k = B_k + W_{ph}$ for every k with $I_{pk} = 1$

Population dynamics

- Assign appropriate (R_k, B_k, I_{pk}) to births and stratum movers
  - before every draw of a sample in G,
  - such that births, stratum movers and existing units have same joint distribution of (R_k, B_k, I_{pk}) in stratum h.
- Births
  - assign new $R_k \in [0, 1]$,
  - copy (B_k, I_{pk}) from existing unit j in h with R_j closest to R_k.
- Stratum movers
  - determine relative position of stratum mover in old stratum,
  - copy (B_k, I_{pk}) from existing unit j in new stratum closest to relative position. A new R_k close to R_j is assigned.
  - Possible orderings: (i) by R_k, (ii) by B_k,R_k or (iii) by I_{pk},B_k,R_k.
- For rotating panels updating the panel due to population dynamics is applied before panel rotation.

Basic and substratification

- Basic stratification: common stratification for surveys in G.
- Depart from basic stratification possible by use of substrata.
  - Assign/update parameters (R_k, B_k, I_{pk}) at basic stratum level.
  - Sampling is done per substratum.
  - Spread of survey burden is suboptimal.
- For cross-sectional surveys no restrictions.
- For panels:
  - Maximal 3 substrata in basic stratum h with fractions f_{h1}, 0, 1.
  - Panel indicator I_{pk} denotes imaginary panel.
  - Real panel can be derived from I_{pk}.

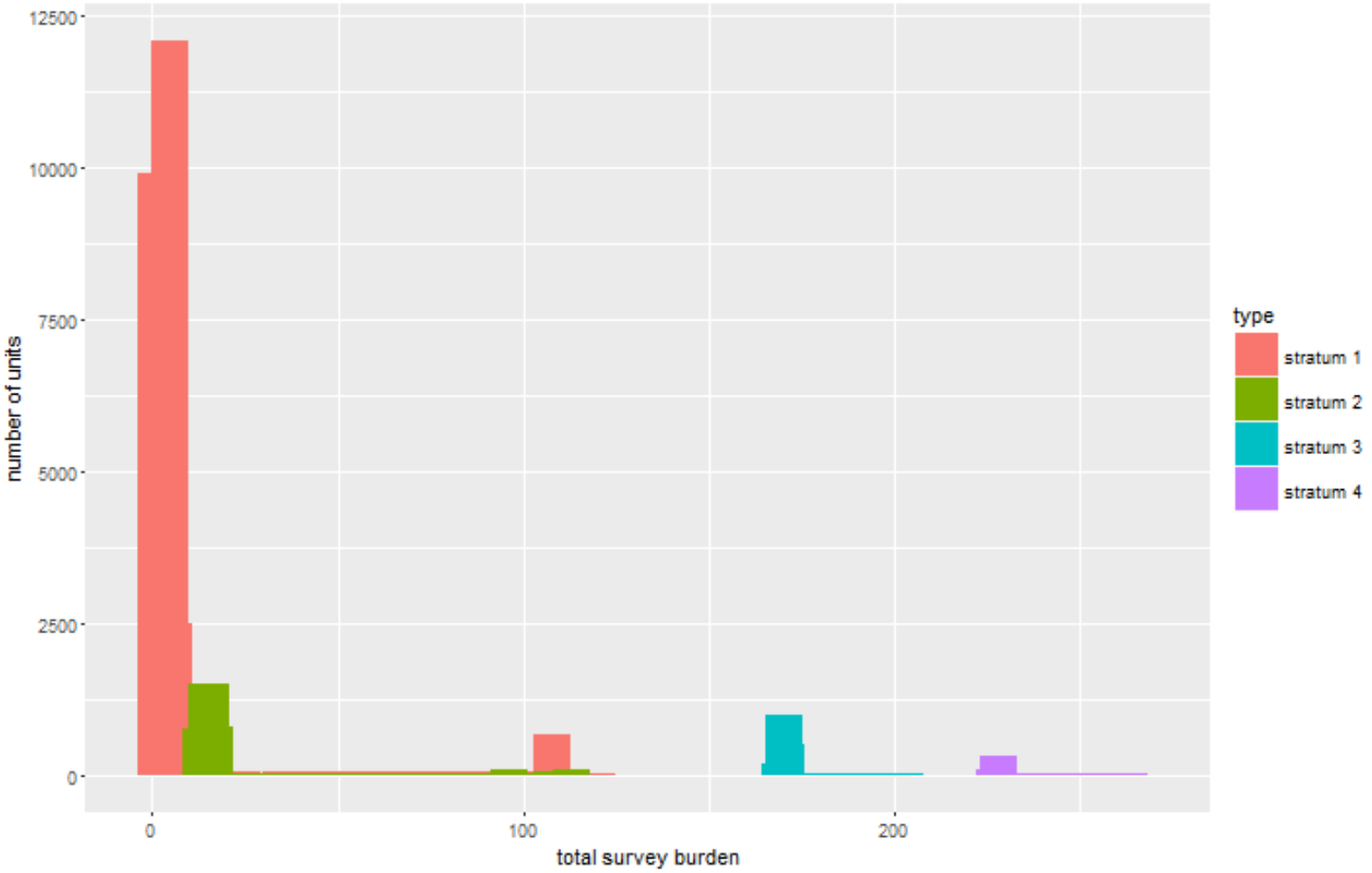
Some simulation results

- Coordination of sampling in group of 3 surveys
  - by simulating a series of 250 monthly draws,
  - from population with 100,000 units, 5 basic strata and
  - simulated population dynamics.
- Surveys with sampling fractions:

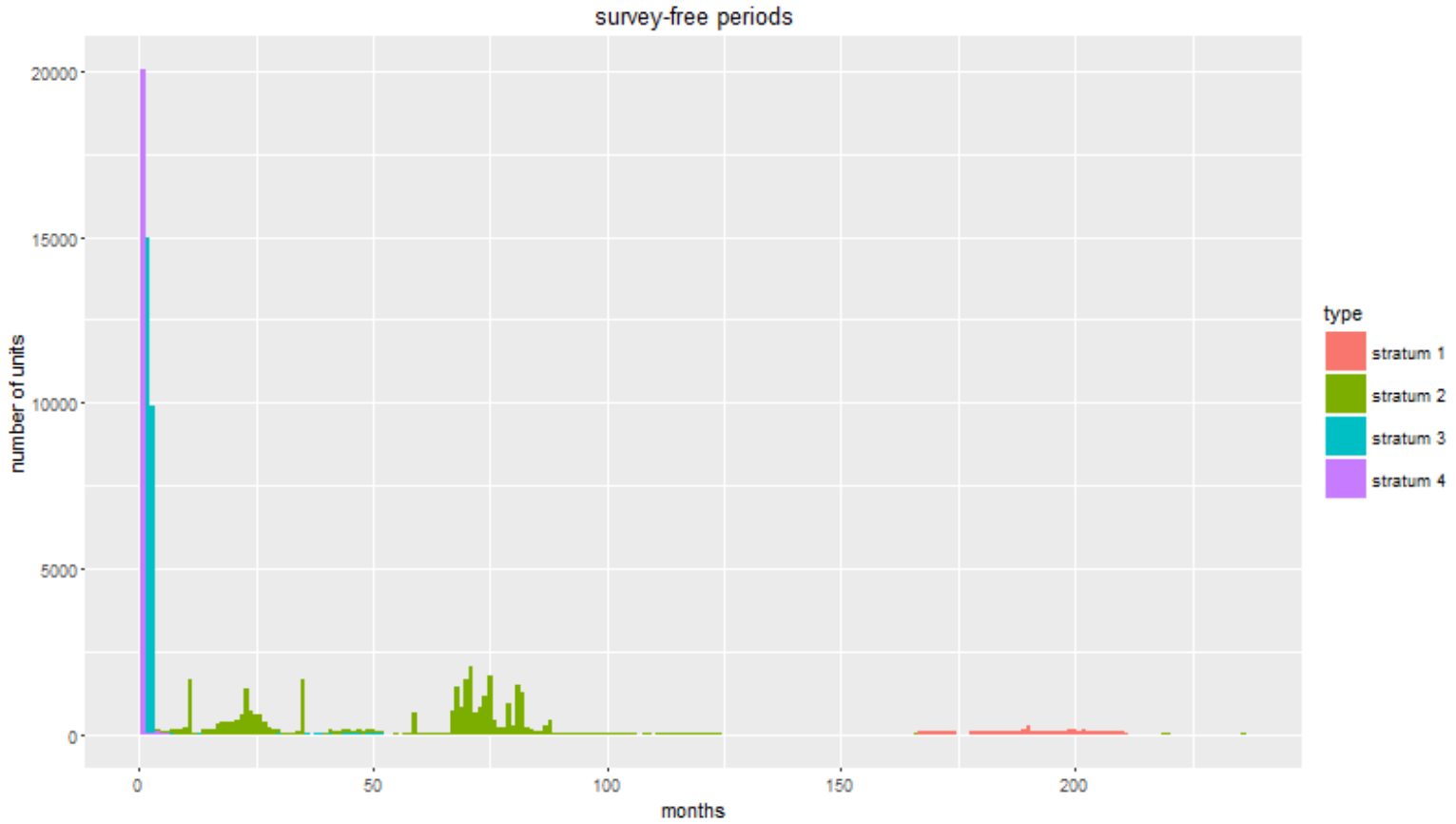
Survey	Frequency	Rotation	1	2	3	4	5
1 (no panel)	year	-	0.03	0.06	0.1	0.15	0.3
2 (panel)	month	0.1 (yearly)	0.02	0.06	0.1	0.15	0.3
3 (panel)	month	0.2 (monthly)	0.01	0.05	0.6	0.8	1

- Aspects of spread of survey burden:
  - survey-free periods,
  - length of stay in panel,
  - multiple draws in group at same time.

Total survey burden



Survey-free periods



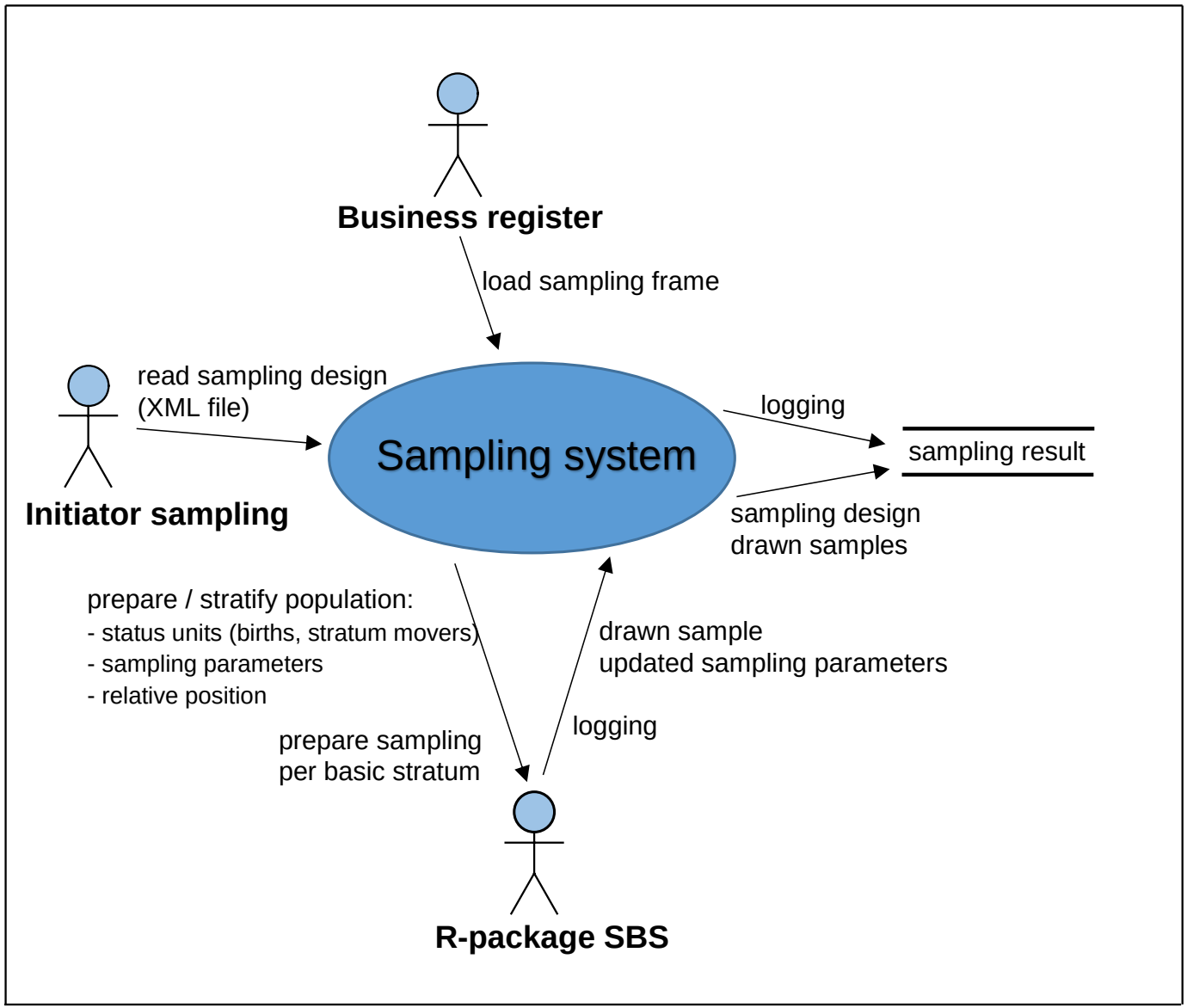
R-package SBS

- Functions for
  - drawing samples by survey burden sampling (SBS) or simple random sampling (SRS),
  - initialising and updating parameters (R_k, B_k, I_{pk}),
  - drawing panel samples, including panel rotation and updating for population dynamics.
- Main function apply_SBS() draws samples by SBS for given basic stratum.
- Sampling system
  - calls apply_SBS() per basic stratum,
  - keeps parameters (R_k, B_k, I_{pk}) and relative position of units per basic stratum,
  - determines status of units in basic stratum:
    0: existing units, 1 births, 2: stratum movers.
- Package not published on CRAN, but is available for those interested.

Example: use of function apply_SBS()

```
> library("SBS")
> UnitId
  UnitId StratumId       PRN SBV InSurvey Status  RP
1   1000         1 0.7879798 1.0    FALSE      0 0.4
2   1001         1 0.2322323 2.5     TRUE      2 0.6
3   1002         1 0.0000000 0.0    FALSE      1 0.0
4   1003         1 0.4784785 1.0    FALSE      0 0.2
5   1004         2 0.6562776 3.0     TRUE      0 0.8
6   1005         2 0.0000000 0.0    FALSE      1 0.0
> SubStratum
  StratumId Fraction NumUnits MinNumUnits
1         1     0.25       -1           3
2         2     0.50       -1          -1
> apply_SBS(UnitId, SubStratum, SB=1.0, IsPanel=FALSE,
InitializationModule = "SBV", ReturnAll=TRUE)
  UnitId StratumId       PRN SBV InSurvey Status        RP InclusionWeight
3   1002         1 0.3961745   2     TRUE      0 0.2857143         1.333
4   1003         1 0.4784785   2     TRUE      0 0.5714286         1.333
2   1001         1 0.6684472   2     TRUE      0 0.7142857         1.333
1   1000         1 0.7879798   1    FALSE      0 0.1428571        -1.000
6   1005         2 0.4159730   2     TRUE      0 0.4285714         2.000
5   1004         2 0.6562776   3    FALSE      0 0.8571429        -1.000
```

Use of SBS by sampling system



Monitoring expected survey burden

- Determine expected annual survey burden for enterprise k by computing
  - the expected annual total survey burden: $\sum_l \pi_k^{(l)}$,
  - the corresponding variance: $\sum_l \pi_k^{(l)}(1-\pi_k^{(l)})$,
  - sum is taken over surveys l in scope of the sampling system,
  - computed for one year, so quarterly surveys count for four, biennial surveys for half.
- Compare with characteristics of enterprises, like complexity and importance (CSI-factor), size and industrial sector.
- Detect enterprises with extreme values and check whether sampling methods could be adjusted.
- Possible extensions: use weighted estimates, compute realised annual survey burden.

Extension to PPS Sampling

- Purpose: support sampling of more Dutch business surveys.
- Probability proportional to size (PPS) sampling:
  - inclusion probabilities π_k proportional to given size variable x_k,
  - $\pi_k = n x_k / \sum_{k \in U} x_k$ for sample size n.
- The following Dutch business surveys are rotating PPS panels
  - Service Producer Price Indices (SPPI, size: turnover),
  - Business Survey Netherlands (COEN, size: number of working persons).
- Adjust sampling algorithm for PPS such that built-up survey burden B_k is taken into account and sampling can be done with given values of (R_k, B_k, I_{pk}).

PRN methods for rotating PPS panels

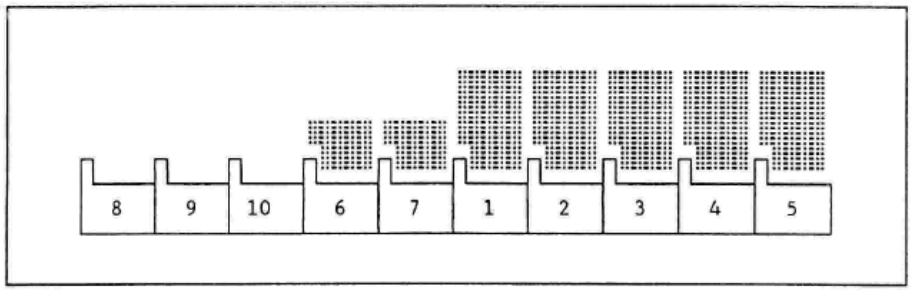
- Scholtus and van Delden (2016) investigated three PRN methods for the Dutch SPPI:
  - *Poisson sampling*:
    select k if $R_k \leq \pi_k$.
  - *Sequential Poisson sampling* (Ohlsson, 1995 & 1998):
    select n units with lowest $\rho_k = \frac{R_k}{\pi_k}$.
  - *Pareto sampling* (Rosén, 1997):
    select n units with lowest $\rho_k = \frac{R_k/(1-R_k)}{\pi_k/(1-\pi_k)}$.
- Panel rotation:
  use $r_k = (R_k - a) \bmod 1$ instead of R_k for moving $a \geq 0$.
- Scholtus and van Delden (2016): Pareto gives best results.

Strategy of PPS sampling algorithm

- Given (R_k, B_k, I_{pk}) for all k in stratum h.
- Use instead of R_k the relative position $r_k = i/(N_h + 1)$ of k with rank i in specified ordering determined by (R_k, B_k, I_{pk}).
- Determine $h = h_0 + h_1$, such that
  - $\pi_k = 1$ for $k \in h_0$ and $\pi_k < 1$ for $k \in h_1$,
  - select all units in h_0.
- Use Pareto sampling in h_1:
  - select n_{h1} units with smallest values of $\rho_k = \frac{r_k/(1-r_k)}{\pi_k/(1-\pi_k)}$.
- Update B_k and I_{pk} only by means of relative positions r_k.

Illustration of sampling algorithm

Figure 4. Before the third sample



For $x_k = (40, 25, 22, 20, 20, 12, 10, 10, 5, 5)$ and $n = 3$:

- rank of units: $(6, 7, 8, 9, 10, 4, 5, 1, 2, 3)$,
- $r_k = (0.54, 0.63, 0.72, 0.81, 0.90, 0.36, 0.45, 0.09, 0.18, 0.27)$,
- $\rho_k = (0.49, 2.19, 4.16, 8.18, 18.17, 2.11, 3.86, 0.46, 2.28, 3.85)$,
- Select units $1, 6$ and 8.
- Update B_k for units $8, 9$ and 10.

Cross-sectional PPS survey

Draw of cross-sectional PPS survey $l \in G$ with sample size n_h and weight W_{lh} in stratum h:

1. Sort k in h by (i) B_k (increasing) and (ii) R_k (increasing).
2. Determine relative positions r_k in h.
3. Determine $h = h_0 + h_1$ with take-all stratum h_0.
4. Determine ρ_k in h_1.
5. Select n_{h1} units with smallest values of ρ_k in h_1.
6. For n_h units in h with smallest r_k, let $B_k = B_k + W_{lh}$.

Some first simulation results

- Suppose in stratum we have
  $x_k = (40, 25, 22, 20, 20, 12, 10, 10, 5, 5)$ and $n = 3$.
- Simulate a series of t draws and repeat this R times.
- Realised fractions for k are estimated by
  - $\hat{\pi}_{kR}(t) = \frac{1}{R} \sum_{r=1}^{R} \iota\{k \in S_r(t)\}$,
  - $\iota\{k \in S_r(t)\}$ indicates whether k is selected in draw t and simulation run r.
  - under PPS: expectation π_k and variance $\pi_k(1 - \pi_k)$.
- Consider two sampling methods:
  1. Pareto sampling without sample coordination,
  2. Pareto sampling with sample coordination by means of survey burden values.

Realised fractions

Table: realised fractions (in %) for $R = 20000$ and $t = 5, 10$

k	π_k	$\hat{\pi}^1_{kR}(5)$	$\hat{\pi}^1_{kR}(10)$	$\hat{\pi}^2_{kR}(5)$	$\hat{\pi}^2_{kR}(10)$	margins
1	71.01	71.30	71.38	73.20	72.81	0.64
2	44.38	45.00	44.91	42.08	42.18	0.70
3	39.05	39.05	38.81	39.06	39.37	0.69
4	35.50	35.41	35.35	35.85	35.99	0.68
5	35.50	35.47	35.70	36.35	36.10	0.68
6	21.30	20.80	21.19	21.29	21.32	0.58
7	17.75	17.42	17.36	17.84	18.09	0.54
8	17.75	18.07	17.76	18.28	18.04	0.54
9	8.88	8.80	8.88	8.20	7.88	0.40
10	8.88	8.67	8.67	7.87	8.22	0.40

Lengths of survey-free periods

Table: Lengths of survey-free periods for $R = 1$ and $t = 250$

	Method 1				**Method 2**			
k	min	mean	max	sd	min	mean	max	sd
1	0	0.43	6	0.82	0	0.25	1	0.43
2	0	1.18	10	1.62	0	1.50	2	0.87
3	0	1.68	12	2.04	2	2.33	3	0.47
4	0	2.04	15	2.64	2	2.34	3	0.47
5	0	1.78	13	2.25	0	1.50	2	0.87
6	0	3.71	17	4.25	2	2.34	3	0.47
7	0	4.01	21	4.95	9	9.00	9	0.00
8	0	4.22	21	4.27	2	3.98	6	2.01
9	0	10.30	33	9.63	9	9.00	9	0.00
10	0	8.94	60	10.05	9	9.00	9	0.00

Rotating PPS panels by SBS

Subsequent draw of rotating PPS panel $p \in G$ with sample size n_h, rotation fraction v_h and weight W_{ph} in stratum h:

1. Sort k in h by (i) I_{pk} (decreasing), (ii) B_k (increasing) and (iii) R_k (increasing),
2. Determine relative positions r_k in h.
3. Define $u_h = \text{round}(v_h m_h)$, with m_h number of units in panel.
4. Determine $h = h_0 + h_1$ with take-all stratum h_0.
5. Determine ρ_k in h_1.
6. Remove last u_h units in h_1 with $I_{pk} = 1$ from panel.
7. Adjust panel to get sample size n_{h1}:
   - $m_{h1} - u_h < n_{h1}$? Add $n_{h1} - (m_{h1} - u_h)$ units in h_1 with $I_{pk} = 0$ and smallest ρ_k to panel.
   - $m_{h1} - u_h > n_{h1}$? Remove extra $m_{h1} - u_h - n_{h1}$ units from panel (last units with $I_{pk} = 1$).
   - $m_{h1} - u_h = n_{h1}$? No adjustment.

Rotating PPS panel - continued

- Update of B_k and I_{pk}:
  8. Let $I_{pk} = 0$ for every k that is removed from the panel.
  9. Let $I_{pk} = 1$ for first $n_{h1} - (m_{h1} - u_h)$ units in h_1 with $I_{pk} = 0$.
  10. Let $B_k = B_k + W_{ph}$ for all units with $I_{pk} = 1$.

Future work and discussion

- Extensively testing the PPS algorithms
  - under population dynamics,
  - in combination with other surveys,
  - with substratification,
- Implementation of PPS sampling in R-package SBS and sampling system.
- Support PPS surveys COEN and SPPI by sampling system?
- Further extension to sampling designs like cluster sampling or multistage sampling.
- Discussion points:
  - How effective is sample coordination in the case of PPS?
  - Is it desirable to extend sample coordination to a larger group of surveys?
  - What are advantages and disadvantages for surveys to be involved in the sampling system?

References

- Ohlsson, E. (1995), Coordination of samples using Permanent Random Numbers. *Business Survey Methods,* John Wiley & Sons, New York, pp 153-169.
- Ohlsson, E. (1998), Coordination of PPS samples over time. *The 2nd International Conference on Establishment Surveys,* pp 255-264.
- Rosén, B. (1997), On sampling with probability proportional to size. *Journal of Statistical Planning and Inference* 62, 159-191.
- Scholtus, Sander and van Delden, Arnout (2016), *PPS Sampling with Panel Rotation for Estimating Price Indices on Services.* Proceedings of the Fifth International Conference of Establishment Surveys, June 20-23, 2016, Geneva.