Discussion paper

Sampling methods to better spread the response burden for small businesses

Marc Smeets and
Jonas Klingwort

September 2023

Content

Summary

Statistics Netherlands operates a coordinated sampling system to spread the response burden in business surveys. The monitoring of the response burden revealed that despite the applied sample coordination, several businesses are still heavily sampled each year (so called hotspots). Different sampling methods can be applied to improve further the spread of the peaking response burden in business surveys. Those methods considered include adjustments of sampling fractions, drawing balanced samples, and stratified sampling approaches. This paper evaluates the sampling method's effects on spreading peaking response burden in two simulation studies and compares them with the baseline method of ignoring the peaking response burden. The first simulation study considers a practical situation at Statistics Netherlands. Here, a preliminary drawn sample is considered. The different sampling methods are applied, and their effects in reducing the number of hotspots in the preliminary sample are evaluated. The second simulation study analyzes the method's longer-term effects regarding spreading the peaking response burden. Moreover, whether the sampling units are drawn with the prescribed inclusion probabilities in the different methods is studied. The results of these studies will show which sampling methods are appropriate to lighten the peaking response burden. The advantages and disadvantages of the methods will be explained and discussed in detail.

Keywords

Sample coordination, hotspots, response burden, balanced sampling, Poisson sampling, business surveys.

Reviewers

Harm Jan Boonstra, Remco Paulissen, Geert Gommans, Lars Burgerjon, Remco Kitzen, Lianne Ippel, Leanne Houben, Vera Toepoel.

1. Introduction

The requirement to reduce the (perceived) response burden is commonly faced in official statistics. Statistics Netherlands (SN) operates a coordinated sampling system, or EnquêteDrukSysteem[1] (EDS), for business surveys that allows spreading the response burden evenly among the businesses (Smeets and Boonstra, 2018). The spread of the response burden is realized by applying negative sample coordination based on permanent random numbers (PRNs) and considering the cumulated response burden of the businesses based on the number of received surveys. Despite the applied sample coordination, several businesses are still heavily sampled each year. This is mainly because of large sampling fractions but also because not all surveys make optimal use of the features of the EDS. This results in a peaking response burden for some individual businesses.

In response to a letter of advice from the business community (Commissie van Straalen, Strategische commissie betere regelgeving bedrijven, 2020), SN started a project in May 2021 to prevent, as far as possible, businesses from being sampled disproportionately often. This project focuses primarily on small and medium-sized businesses with 0-19 working persons (size classes 0-4). In a pilot study that ran from May 2021 to May 2022, it was monitored how often a business was sampled within the last twelve months (Gommans et al., 2022). In this pilot, businesses with 0-9 working persons (size classes 0-3) get the label hotspot if they have been sampled for more than three different surveys within the last twelve months. Businesses with 10-19 working persons (size class 4) get the label hotspot if they have been sampled for more than four surveys within the last twelve months. During the pilot, several sampling methods were evaluated to lighten the peaking response burden by reducing or spreading the number of hotspots as much as possible in collaboration with the statistical departments (Gommans et al., 2022, Ippel et al., 2023, and Subsection 2.2). The effects of removing hotspots on the precision of the survey estimates is analysed by Klingwort & Smeets, 2022.

Meanwhile, in Matei et al. (2023), a sample coordination method has been developed by means of balanced sampling that realizes a double control of the burden for businesses with a peaking response burden. This method also uses permanent random numbers and is based on correlated Poisson sampling. In Matei et al. (2023), this method is applied to Dutch business surveys, and in a simulation study it is shown that the variance of the number of hotspots can be reduced compared to Poisson sampling and Pareto sampling. This means that balanced sampling prevents the selection of large numbers of businesses with a peaking response burden. In this paper, we compare balanced sampling and a stratified approach with the methods used in the pilot, evaluate the methods, and investigate the longer-term effects of the methods on the number of sampled hotspots.

[1] Response burden spreading system

In a first simulation study, we investigate whether balanced sampling can be applied outside the EDS as an alternative for the methods used in the pilot. Here we examine the number of hotspots in a single sample and verify whether the prescribed sample sizes are realized. In a second simulation study we investigate the longer-term effects of the methods on the number of sampled hotspots. For both simulation studies data from the Research and Development (R&D) Survey are used.

2. Sampling methods

2.1 Balanced sampling

The sample coordination method developed by Matei et al. (2023) allows a double control of the response burden of the hotspots: once by drawing balanced samples to avoid clustering of hotspots in the sample and once by applying negative sample coordination to minimize the overlap between samples drawn in successive surveys. This sample coordination method is introduced as the targeted double control strategy.

In this paper, we consider the method of balanced sampling used in Matei et al. (2023), and we investigate whether it can be used outside the EDS as an alternative to the methods used in the pilot. A follow-up study (not included in this paper) will investigate whether the targeted double control strategy can be implemented in the EDS, taking into account the cumulated response burden of the businesses. In this section, we briefly explain the balanced sampling method developed by Matei et al. (2023). Balanced sampling is based on Poisson sampling. To each unit in the population a uniform random number is assigned, which is used for the successive sample selections. A Poisson sample is drawn as follows: generate random numbers $r_1, \dots, r_k, \dots, r_N \sim \mathrm{Unif}(0,1)$ and select unit k in sample S if $r_k < \pi_k$. Here $\pi_k = P(k \in S)$ is the inclusion probability of unit, or business, k. Note that the sample size n of a Poisson sample is random, with the expected sample size $E(n) = \sum_{k=1}^{N} \pi_k$.

Spatially Correlated Poisson (SCP) sampling (Grafström, 2012) is a modification of Poisson sampling, which provides spatially balanced samples with a fixed sample size. SCP sampling is applied to a population where each unit has associated geographical coordinates. The idea is that contiguous units may provide similar information and that a sample contains more information if the sample avoids clustering of contiguous units. By spatially balancing the sampled units are spread across the geographic area, realizing a geographic coverage over the entire population of the survey. Instead of applying SCP sampling to geographical coordinates, SCP sampling can also be applied to other characteristics of the units, resulting in well-spread samples with respect to these characteristics. Note that these characteristics should not depend on response indicators, which may lead to selectivity in the drawn samples. The targeted double control strategy applies SCP

sampling to a measure of the cumulated response burden to spread the response burden of the units in the sample. This is the so-called Adapted Spatially Correlated Poisson (ASCP) sampling method.

By using a binary variable representing the hotspot status of the units, ASCP sampling realizes a balanced sample with respect to the hotspot status. In this way, large numbers of hotspots in the sample can be avoided. The binary variable for unit k has value 1 if k is a hotspot and 0 otherwise. In this paper, we refer to ASCP sampling simply as balanced sampling, and we only consider the hotspot status as a measure for the cumulated response burden. Balanced samples can be drawn using the function scps_coord() in the R package BalancedSampling (Grafström, 2022). The function scps_coord() uses the population vectors of inclusion probabilities, random numbers, and the cumulated response burden of the units.

2.2 Sampling strategies applied in the pilot

The data collection department of SN draws the samples for the Dutch business surveys. For most of the surveys this is done by means of the EDS. More surveys are added to the EDS every year. The sampling process consists of two steps. First, a preliminary sample is drawn using the EDS. The sample is submitted to the statistical department, and a final sample is drawn after approval from the statistical department.

During the pilot, the data collection group tracked which units in the drawn samples became hotspots. If possible and in collaboration with the statistical departments, the hotspots are removed from the samples. In the pilot, the following methods were applied (Gommans et al., 2022):

1. *Remove hotspots from the drawn sample.* In this way, fewer questionnaires are sent out, and the realized sample size is smaller than the prescribed sample size.
2. *Replace hotspots with other units.* In this way, the prescribed sample size is maintained as much as possible. The following strategies were applied:
   a. Draw a larger sample first and then remove hotspots from the larger sample. Based on the number of hotspots in the preliminary sample, it can be determined how many additional units have to be drawn for the final sample. It may require a few iteration steps before the final sample can be drawn.
   b. Remove hotspots from the sampling frame and draw a sample with the prescribed sample size from the remaining units in the sampling frame.
   c. Replace hotspots in a rotating panel with other units during the panel rotation.

Note that by applying these methods, the inclusion probabilities of the units in the finally drawn sample are no longer equal to the prescribed inclusion probabilities of the units. In practice, this is generally not considered, but by weighting the observed response to the population totals this is corrected as much as possible.

The applied strategies in the pilot aim to lighten the peaking response burden either by reducing (method 1) or by spreading (methods 2a, 2b, and 2c) the number of sampled hotspots. In this paper, we want to compare balanced sampling with the strategies of the pilot in a simulation study applied to the R&D survey. Since the R&D survey is a cross-sectional survey and balanced sampling is a method of spreading the response burden, only methods 2a and 2b are relevant to the simulation study.

Methods 2a and 2b can be seen as two different sampling schemes leading to a sample from which the hotspots have been replaced by other units as much as possible. In method 2a, this is accomplished by (iteratively) increasing the sampling fractions of non-hotspot units and drawing preliminary samples until the desired sample is obtained. In method 2b, this is accomplished by sampling only once from the thinned-out sampling frame. In both methods, the sampling fractions of the non-hotspot units are increased during the sampling process. Methods 2a and 2b are therefore considered as similar methods and we only apply method 2b in the simulation. We refer to this method as increasing the sampling fractions of non-hotspot units.

3. Outline of the simulations

This section describes the outline of the simulations. The purpose is to investigate whether balanced sampling can be used outside the EDS as an alternative to the sampling method in which the sampling fractions of the non-hotspots are increased (Section 2). Simulation study 1 focuses on the number of sampled hotspots when a single sample is drawn conditional to a given sample. Simulation study 2 focuses on the realized sampling fractions and the longer-term effects of the sampling methods on the sampled hotspots.

Both simulation studies are applied to the R&D survey. The R&D survey is an annual survey with the businesses as sampling units and a stratified sample design with equal inclusion probabilities within the strata (CBS, 2022). This means that the sampling methods are applied to each stratum separately. The whole sample is then obtained by combining the samples drawn in the strata. The strata of the R&D survey are defined by combinations of category, industrial classification according to SBI (Standaard Bedrijfsindeling), and size class, or GrootteKlasse (GK). The sampling frame consists of four categories, namely STP (STeekProef): sampling part; CON: CONGO (CONsistentie Grote Ondernemingen) units that are selected with probability 1; FR1 (FRaction 1): non-CONGO units that are selected with probability 1 and FR0 (FRaction 0): units that are excluded from sampling. Further, there are 99 SBI combinations and 6 size classes used in the stratification. The R&D survey collects data only from GK 4-9 and estimates GK 0-3 by using data from the WBSO register (Wet Bevordering Speur- en Ontwikkelingswerk data). Therefore, only hotspots in size class 4 are considered in this paper.

Since the sampling methods are applied to a stratified sample design with equal inclusion probabilities within the strata, the question arises whether balanced sampling leads to similar results as when the sample is evenly spread across separate substrata that are introduced for the hotpots and non-hotspot units. For this reason, we also consider the sampling method that defines a separate substratum for the hotspots and the non-hotspots and draws random samples in both substrata with the same sampling fractions.

For the simulation, we consider the sample design that was used for the R&D survey in 2020. This design was used in the beginning of 2021 to draw a sample for the R&D survey. At that time, the hotspot labels were available for all units in the population and no hotspots were removed from the sample. Furthermore, we only consider the sampling part (category STP) because spreading the response burden has no effect on the samples when the sampling fraction is 0 or 1. Even in category STP, there were a few strata with sampling fraction 1. These strata are not used in the simulation either.

Let U be the population of the R&D survey that is used in the simulation, that is, after removing the strata with sampling fraction 0 or 1, and let S be the given sample in U. The population size of U is $N = 33044$ and sample S consists of $n = 926$ units. There are $N_{\mathrm{HS}} = 211$ hotspots in the population. The strata are denoted by $1, \ldots, h, \ldots, H$, where stratum h has sample size n_h and population size N_h. The inclusion probabilities are equal to the sampling fractions in the strata: $\pi_k = f_h = n_h/N_h$ for all units k in stratum h. Table A.1 in the appendix shows the $H = 83$ strata of U, together with N_h, n_h, f_h, the number of hotspots in the population N_{HS}, the population fraction of hotspots $F_{\mathrm{HS},h} = N_{\mathrm{HS},h}/N_h$, the expected number of sampled hotspots $\mathrm{E}(n_{\mathrm{HS},h}) = n_h F_{\mathrm{HS},h}$ and the variance of the number of sampled hotspots $\mathrm{V}(n_{\mathrm{HS},h}) = n_h F_{\mathrm{HS},h}(1 - F_{\mathrm{HS},h})(N_h - n_h)/(N_h - 1)$. Note that $n_{\mathrm{HS},h}$ is hypergeometrically distributed with parameters N_h, $N_{\mathrm{HS},h}$, and n_h, from which the expectation and variance of $n_{\mathrm{HS},h}$ follow.

3.1 Simulation study 1: a practical situation

In the first simulation study, we simulate the following sampling process. First, it is assumed that the EDS draws a preliminary sample. Then, the different sampling methods are applied to reduce the number of hotspots in the preliminary sample. For the preliminary sample we use the given sample S of the R&D 2020 survey. We consider the following sampling methods:

1. *Ignoring hotspots*: sampling without taking into account the hotspots.
2. *Increasing fractions*: replace hotspots with other units by increasing the sampling fractions of non-hotspots.
3. *Balancing*: balanced sampling by taking into account the hotspots.
4. *Using substrata*: use separate substrata for the hotspots and non-hotspots.

The first method, that is ignoring hotspots, represents the current sampling method implemented in the EDS. In the first simulation, ignoring hotspots will

always lead to the preliminary sample S. This method is used as the reference in the simulation. The other sampling methods are described in Section 2. The sampling methods are implemented in the following way:

1. Generate uniform random numbers $0 < r_k < 1$ for all $k \in U$, such that $r_k < \pi_k$ if $k \in S$ and $r_k \geq \pi_k$ otherwise.
2. Apply the sampling methods:
   - *Ignoring hotspots*: draw a Poisson sample $S_1 (= S)$ with given r_k and π_k.
   - *Increasing fractions*: draw a Pareto sample S_2 with given r_k and sample size n_h, where the $N_{\mathrm{HS},h}$ hotspots in stratum h are removed from the population. If $n_h > (N_h - N_{HS,h})$, then Step 1 is applied and no hotspots are removed from the sample in stratum h.
   - *Balancing*: draw a balanced sample S_3 with given r_k and π_k and hotspot status.
   - *Using substrata*: Draw separate Pareto samples in the substrata of hotspots and non-hotspots with given r_k and let S_4 be the combination of these samples. The sample size $n_{h,1}$ in the substratum of hotspots in stratum h is obtained by rounding $f_h N_{\mathrm{HS},h}$ randomly and the sample size $n_{h,2}$ in the substratum of non-hotspots is $n_{h,2} = n_h - n_{h,1}$.

The values of n_h, N_h and $N_{\mathrm{HS},h}$ are taken from the R&D survey and are fixed in this simulation. Note that the implementation of ignoring hotspots in Step 2 with the random numbers generated in Step 1, indeed always lead to sample $S_1 = S$. The other methods are applied per stratum, leading to a sample per stratum. The samples S_2, S_3, and S_4 are then obtained by combining these samples from the strata. For methods 2 and 4 (increasing fractions and using substrata) Pareto sampling (Rosén, 1997) is used instead of Poisson sampling to guarantee samples with a fixed sample size. A Pareto sample of size n' from a population of size N' with given r_k and π'_k is obtained by selecting the n' units with the smallest values of

$$\rho_k = \frac{r_k(1-r_k)}{\pi'_k(1-\pi'_k)}.$$

In the simulation we draw samples $S_1^{(j)}$, $S_2^{(j)}$, $S_3^{(j)}$ and $S_4^{(j)}$ by repeating steps 1 and 2 in each run $j = 1, \ldots, R$. In total, we consider $R = 1000$ runs, where each run represents a sample draw. For this simulation, we only consider strata that contain at least one hotspot in the population because otherwise, the sampling methods would produce the same sample in each simulation run. So we only consider strata with $0 < f_h < 1$ and $N_{\mathrm{HS},h} > 0$. For convenience, the population for this simulation is also denoted by U and the corresponding preliminary sample by S. The population consists of $H = 48$ strata, has size $N = 17773$, and the sample S consists of $n = 598$ units.

For every sample $S_m^{(j)}$ drawn by method m in run j, we compute the sample size n and the number of hotspots n_{HS} in the sample, both for the whole sample and for the sample in stratum h.

We write

$$n_m^{(j)} = n\left(S_m^{(j)}\right), \qquad n_{\mathrm{HS},m}^{(j)} = n_{\mathrm{HS}}\left(S_m^{(j)}\right),$$
$$n_{h,m}^{(j)} = n_h\left(S_m^{(j)}\right), \quad n_{\mathrm{HS},h,m}^{(j)} = n_{\mathrm{HS},h}\left(S_m^{(j)}\right),$$

for $m = 1, \dots, 4$, $h = 1, \dots, H$ and $j = 1, \dots, R$.

First, it is verified whether the sampling methods draw samples with a fixed sample size in every stratum and whether the realized sample sizes are equal to the prescribed sample sizes. So, we need to check that

$$n_1^{(j)} = \cdots = n_4^{(j)} = n, \quad \text{for all } j = 1, \dots, R \text{ and}$$
$$n_{h,1}^{(j)} = \cdots = n_{h,4}^{(j)} = n_h, \quad \text{for all } j = 1, \dots, R \text{ and } h = 1, \dots, H.$$

Then, the reduction of the number of sampled hotspots by the three sampling methods is examined. For this purpose, we look at the distribution of the number of sampled hotspots in the simulation. For the whole sample frequency tables are made presenting the number of samples in the simulation with a certain number of sampled hotspots. For the strata, the distributions of the number of sampled hotspots $n_{\mathrm{HS},h,m}^{(j)}$ for $j = 1, \dots, R$ are examined per stratum. We will also compute the mean and the standard deviation of $n_{\mathrm{HS},h,m}^{(j)}$ of R samples for the methods $m = 1, \dots, 4$ and the strata $h = 1, \dots, H$.

3.2 Simulation study 2: longer-term effects

In the second simulation study, we investigate the longer-term effects of the sampling methods on the number of sampled hotspots, and we verify whether all units are drawn with the prescribed inclusion probabilities. For this simulation study, it is also assumed that the sampling methods are applied outside the EDS, so we do not take into account the sample coordination that is applied in the EDS. Again, the following sampling methods are considered:

1. *Ignoring hotspots*: sampling without taking into account the hotspots.
2. *Increasing fractions*: replace hotspots with other units by increasing the sampling fractions of non-hotspots.
3. *Balancing*: balanced sampling by taking into account the hotspots.
4. *Using substrata*: use separate substrata for the hotspots and non-hotspots.

We simulate a series of draws over time, where the periods represent the months $t = 1, \dots, T$. In each month t, it is determined which units in the population are hotspots, and the sampling methods are applied to prevent sampling of hotspots as much as possible. The sampling methods are implemented in the following way, where only the first step is different from simulation study 1:

1. Generate random numbers $r_1, \ldots, r_k, \ldots, r_N \sim \mathrm{Unif}(0,1)$
2. Apply the sampling methods:
   - *Ignoring hotspots*: draw a Poisson sample S_1 with given r_k and π_k.
   - *Increasing fractions*: draw a Pareto sample S_2 with given r_k and sample size n_h, where the $N_{\mathrm{HS},h}$ hotspots in stratum h are removed from the population. If $n_h > (N_h - N_{HS,h})$, then Step 1 is applied, and no hotspots are removed from the sample in stratum h.
   - *Balancing*: draw a balanced sample S_3 with given r_k and π_k and hotspot status.
   - *Using substrata*: Draw separate Pareto samples in the substrata of hotspots and non-hotspots with given r_k and let S_4 be the combination of these samples. The sample size $n_{h,1}$ in the substratum of hotspots in stratum h is obtained by rounding $f_h N_{\mathrm{HS},h}$ randomly, and the sample size $n_{h,2}$ in the substratum of non-hotspots is $n_{h,2} = n_h - n_{h,1}$.

In this simulation, all methods in step 2 are applied per stratum. The overall samples S_1, S_2, S_3, and S_4 are obtained by combining these samples from the strata. Now Pareto sampling is also used for the sampling method that ignores the hotspots to guarantee samples with a fixed sample size per stratum. In this simulation, the hotspot status is no longer fixed but is updated every month, as it was done in the pilot. We also use the same hotspot definition as in the pilot, that is, a unit in GK 4 is a hotspot when it is drawn five times or more in the past 12 months. When a sample has to be drawn for month t, a unit is considered to be a hotspot when it has been drawn four times or more in the months $t-1, \ldots, t-12$. In practice, there is a slightly different way of determining whether a unit is a hotspot. First a preliminary sample for month t is drawn, and based on the selections in the preliminary sample, it is determined whether a unit is labelled as a hotspot or not. The different approach would most likely not affect the results of the simulation.

For each method, we simulate a series with a length of 25 years, so $T = 300$ months. We draw samples $S_1^{(t)}$, $S_2^{(t)}$, $S_3^{(t)}$ and $S_4^{(t)}$ by repeating steps 1 and 2 for each month $t = 1, \ldots, T$. For every sample $S_m^{(t)}$, drawn by method m in month t, we compute the sample size n and the number of hotspots n_{HS} in the sample, both for the whole sample and for the sample in stratum h. We write

$$n_m^{(t)} = n\left(S_m^{(t)}\right), \qquad n_{\mathrm{HS},m}^{(t)} = n_{\mathrm{HS}}\left(S_m^{(t)}\right),$$
$$n_{h,m}^{(t)} = n_h\left(S_m^{(t)}\right), \quad n_{\mathrm{HS},h,m}^{(t)} = n_{\mathrm{HS},h}\left(S_m^{(t)}\right),$$

for $m = 1, \ldots, 4$, $h = 1, \ldots, H$ and $t = 1, \ldots, T$.

To investigate the longer-term effects of the sampling methods on the number of sampled hotspots, we look at the development of the number of sampled hotspots in the time series for the whole population and for the strata separately, so $n_{\mathrm{HS},m}^{(t)}$ and $n_{\mathrm{HS},h,m}^{(t)}$ for $m = 1, \ldots, 4$, $h = 1, \ldots, H$ and $t = 1, \ldots, T$.

The sampling fractions in the strata that are realized by the sampling methods in the simulation are computed as follows. A distinction is made between the full stratum and the subpopulations of hotspots and non-hotspots. The sampling fractions are computed for the whole population and for the subpopulations of hotspots and non-hotspots by averaging over time. The averaged sampling fractions are computed as follows:

$$\begin{aligned} \hat{f}_{h,m} &= \sum_{t=1}^{T} n_{h,m}^{(t)} / \sum_{t=1}^{T} N_h \\ \hat{f}_{\mathrm{HS},h,m} &= \sum_{t=1}^{T} n_{\mathrm{HS},h,m}^{(t)} / \sum_{t=1}^{T} N_{\mathrm{HS},h}^{(t)} \\ \hat{f}_{\mathrm{nHS},h,m} &= \sum_{t=1}^{T} n_{\mathrm{nHS},h,m}^{(t)} / \sum_{t=1}^{T} N_{\mathrm{nHS},h}^{(t)} \end{aligned}$$

Here $N_{\mathrm{HS},h}^{(t)}$ and $N_{\mathrm{nHS},h}^{(t)}$ are respectively the number of hotspots and non-hotspots in the population in month t and stratum h and $n_{\mathrm{HS},h}^{(t)}$ is the number of sampled non-hotspots in month t and stratum h by method m. Corresponding to the realized sampling fractions we also compute estimates for the 95% margins:

$$2\sqrt{\frac{f_h(1-f_h)}{\sum_{t=1}^{T} N_h}},\ 2\sqrt{\frac{f_h(1-f_h)}{\sum_{t=1}^{T} N_{\mathrm{HS},h}^{(t)}}} \text{ and } 2\sqrt{\frac{f_h(1-f_h)}{\sum_{t=1}^{T} N_{\mathrm{nHS},h}^{(t)}}},$$

for the whole stratum and the subpopulations of hotspots and non-hotspots, respectively. Here $f_h = n_h/N_h$ is the prescribed sampling fraction in stratum h.

4. Results

4.1 Simulation study 1: a practical situation

This subsection presents the results of the first simulation study. Table 4.1 displays the realized sample sizes by the considered sampling methods. We see that all methods produce samples of size 598 in each of the 1000 simulation runs. Figure 4.1 shows the mean (on the *x*-axis) and the variance (by color) of the realized sample sizes by the sampling methods over all simulation runs in each of the 48 strata (on the *y*-axis). From this, we conclude that all sampling methods draw indeed samples with a fixed sample size and that these sample sizes are equal to the prescribed sample sizes, both for the strata and the overall sample. This result is due to the way the sampling methods are implemented (Subsection 3.1) and the number of iteration runs will therefore not change the result.

Method	Realized sample size 598
Ignoring hotspots	1000
Increasing fractions	1000
Balancing	1000
Using substrata	1000

Table 4.1: Realized sample sizes over 1000 runs.

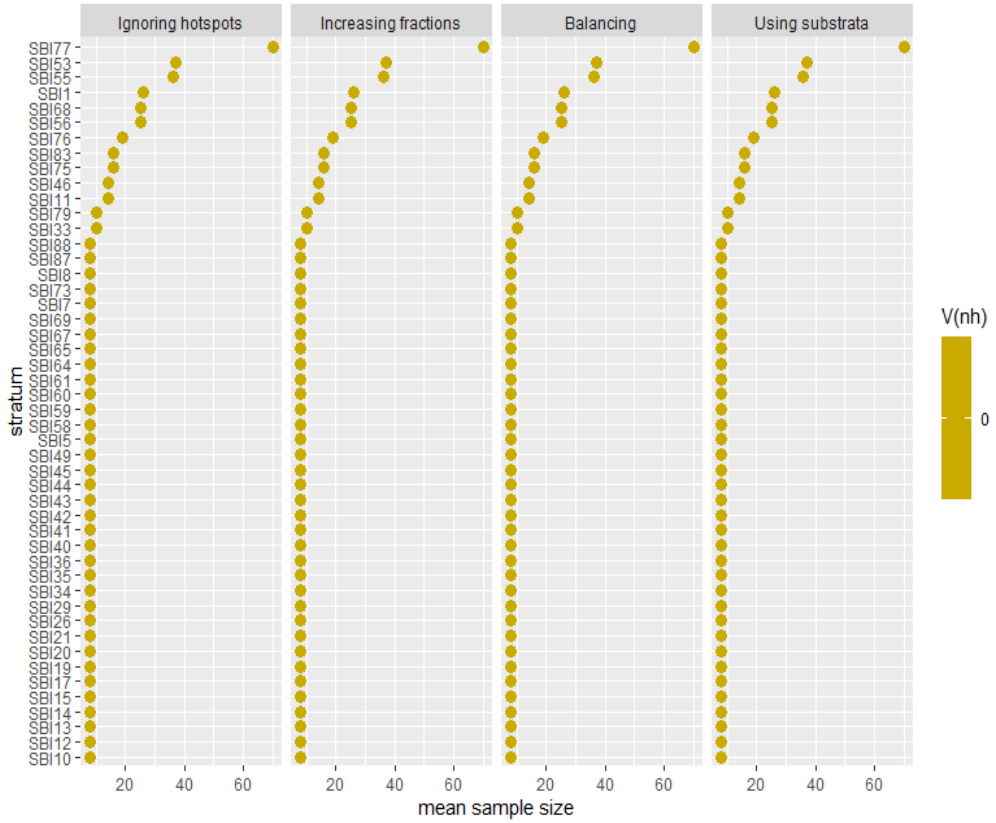


Figure 4.1: Mean of realized sample size. Color indicates the variance of the realized sample sizes in the simulation.

Tables 4.2 and 4.3 display the distribution of the number of sampled hotspots n_{HS} in the simulation for the considered sampling methods. When the hotspots in the preliminary sample are ignored there are 108 hotspots in the final sample. The method of increasing sampling fractions leads to the lowest number of 36 sampled hotspots in each simulation run. The number of sampled hotspots does not vary, because of the given sample and the given hotspot label which do not change during this simulation. When balanced sampling is applied, the number of sampled hotspots ranges from 68 to 76, with 72 being the most common. When substrata are used this number ranges from 47 to 62, where 55 is the most common value. So for the entire sample, the number of sampled hotspots can be reduced both by balanced sampling and by using substrata, where using substrata leads to fewer hotspots in the final sample than balancing. The largest reduction is achieved by increasing the fractions of the non-hotspots. Although this result seems at first sight to suggest to use the method of increasing the fractions, we will examine the

longer-term effects of this method in the second simulation study and address the disadvantages of this method in the discussion.

We now look at the number of sampled hotspots per stratum. Figure 4.2 shows the mean number of sampled hotspots (on the *x*-axis) per stratum (on the *y*-axis). The color indicates the population fraction of hotspots $F_{\mathrm{HS},h}$.

	Number of sampled hotspots									
Method	36	68	69	70	71	72	73	74	75	76	108
Ignoring hotspots											1000
Increasing fractions	1000										
Balancing		3	18	88	223	275	221	131	35	6	

Table 4.2: Number of sampled hotspots over 1000 runs (row sums are equal to 1000).

	Number of sampled hotspots														
Method	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62
Using substrata	1	3	7	26	60	57	134	145	172	123	102	67	31	20	7	5

Table 4.3: Number of sampled hotspots over 1000 runs by using substrata (row sums are equal to 1000).

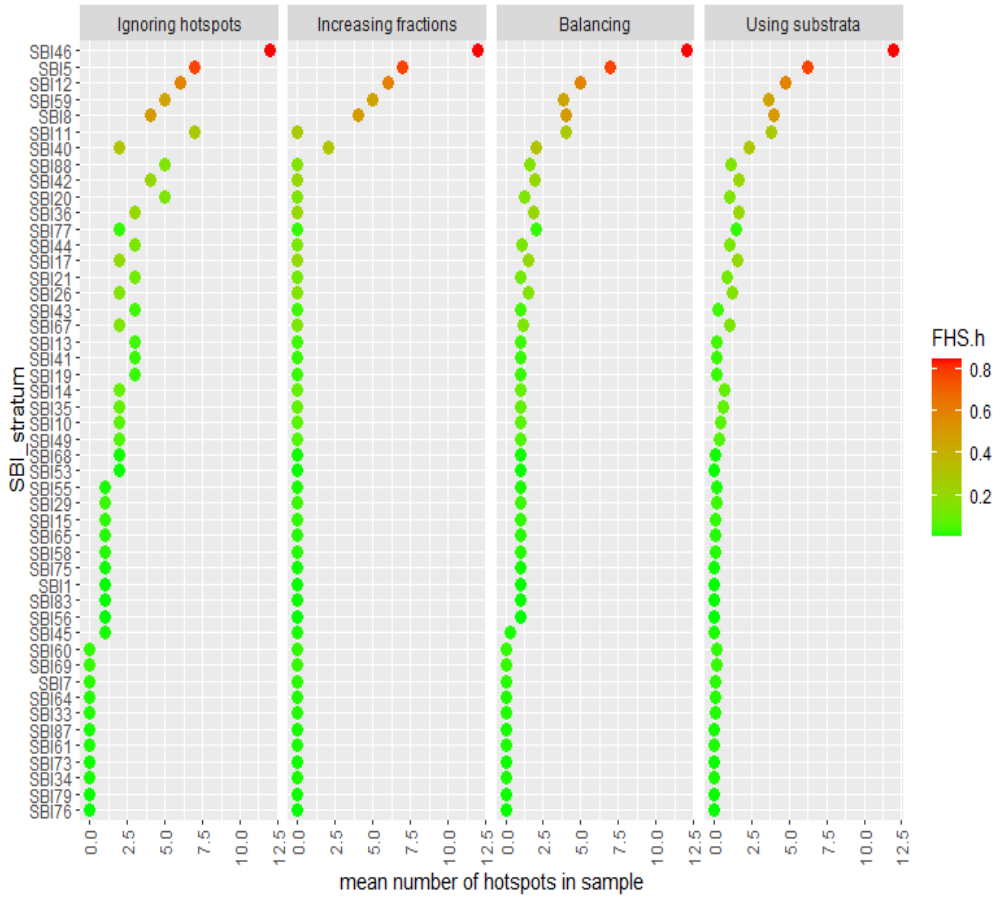


Figure 4.2: Mean number of sampled hotspots. Color indicates population fraction $F_{\mathrm{HS},h}$ of hotspots.

It can clearly be seen that the number of sampled hotspots in the simulation correlates with the fraction of hotspots in the population. In most strata, the number of sampled hotspots can be reduced to zero by increasing the fractions. When using substrata in 26 of the 48 strata, the mean number of sampled hotspots is close to zero, that is, in the lower and middle part of the displayed strata. For balancing this is only the case in 11 strata, that is, in the lower part of the displayed strata (from SBI 40 to SBI 56). In general, balanced sampling leads to fewer hotspots in the sample than ignoring the hotspots, but still more than using substrata and increasing the fractions.

Figures 4.3 and 4.4 display the mean reduction of the number of sampled hotspots compared to ignoring the hotspots. In Figure 4.3, the color indicates the expectation of the number of sampled hotspots $\mathrm{E}(n_{\mathrm{HS},h})$, and in Figure 4.4, the color indicates the variance $\mathrm{V}(n_{\mathrm{HS},h})$. The largest reductions are achieved in the strata with SBI 11, 20, and 88, which is the case for all three methods. In these strata, there is a (relatively) large number of hotspots in the preliminary sample (see Figure 4.2), and there are enough non-hotspot units in the population that can replace the hotspots in the sample, or in other words, that $N_h - N_{\mathrm{HS},h} \geq n_h$, which is equivalent to $f_h + F_{\mathrm{HS},h} \leq 1$.

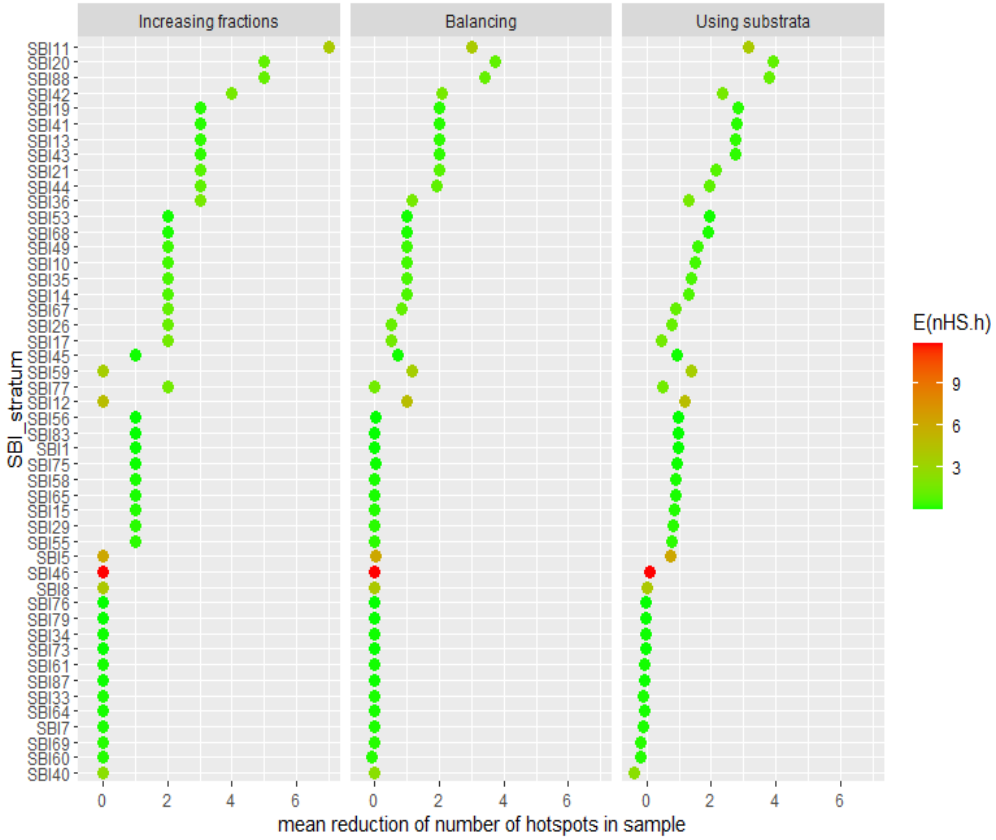


Figure 4.3: Mean reduction of sampled hotspots compared to ignoring hotspots. Color indicates expected number of sampled hotspots $\mathrm{E}(n_{\mathrm{HS},h})$.

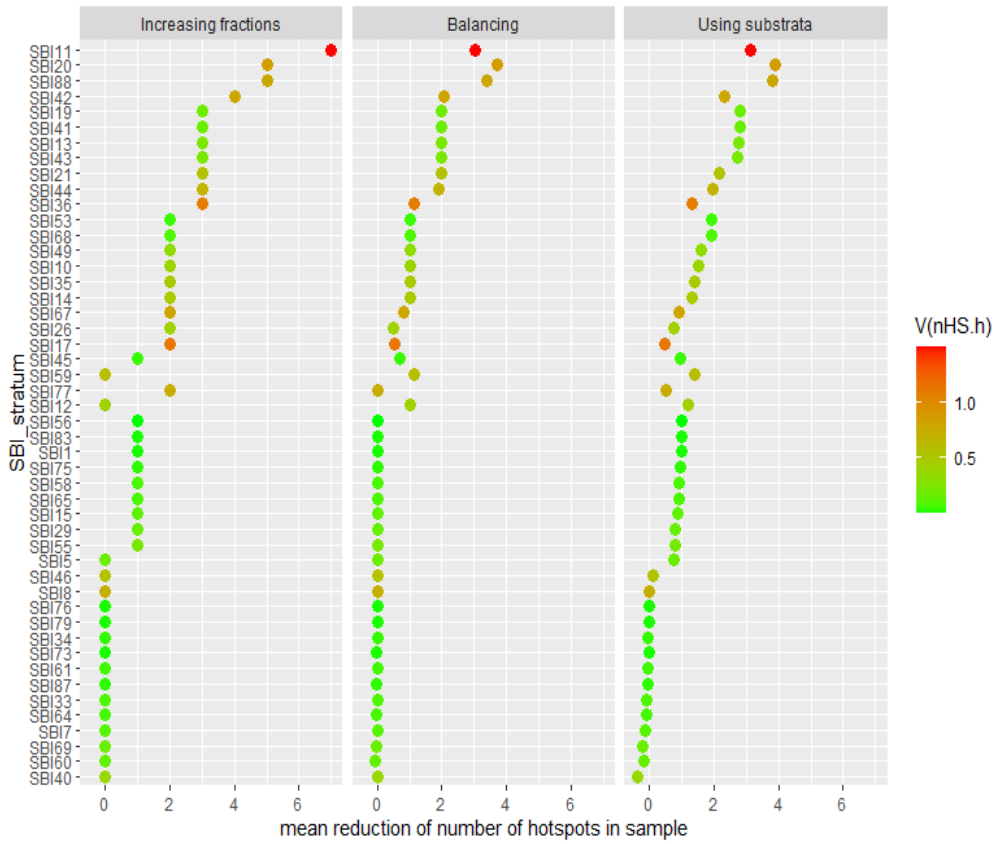


Figure 4.4: Mean reduction of sampled hotspots compared to ignoring hotspots. Color indicates variance of number of sampled hotspots $\mathbf{V}(\boldsymbol{n}_{\mathbf{HS},\boldsymbol{h}})$.

There seems to be a stronger correlation between the achieved reduction and the variance of the number of sampled hotspots (Figure 4.4) than between the achieved reduction and the expectation (Figure 4.3). In the stratum with SBI 46 , the number of hotspots in the preliminary sample is not reduced by the three sampling methods despite the high expected value of $n_{\mathrm{HS},h}$ (Figure 4.3). That is because in this stratum $f_h + F_{\mathrm{HS},h} > 1$ (Table A.1), so there are not enough non-hotspot units that can replace the hotspots in the preliminary sample.

Figures 4.5 and 4.6 show the mean reduction of the number of sampled hotspots compared to the method which increases the fractions when applying balanced sampling or using substrata. In Figure 4.5, the color indicates the expectation of the number of sampled hotspots $\mathrm{E}(n_{\mathrm{HS},h})$, and in Figure 4.6, the color indicates the variance $\mathrm{V}(n_{\mathrm{HS},h})$. We see that balanced sampling leads to fewer hotspots only in the strata with SBI 12 and 59. In both strata, the number of hotspots in the preliminary sample is relatively large (Figure 4.2), but there is no space in the population to manually replace the hotspots by non-hotspot units (Table A.1).

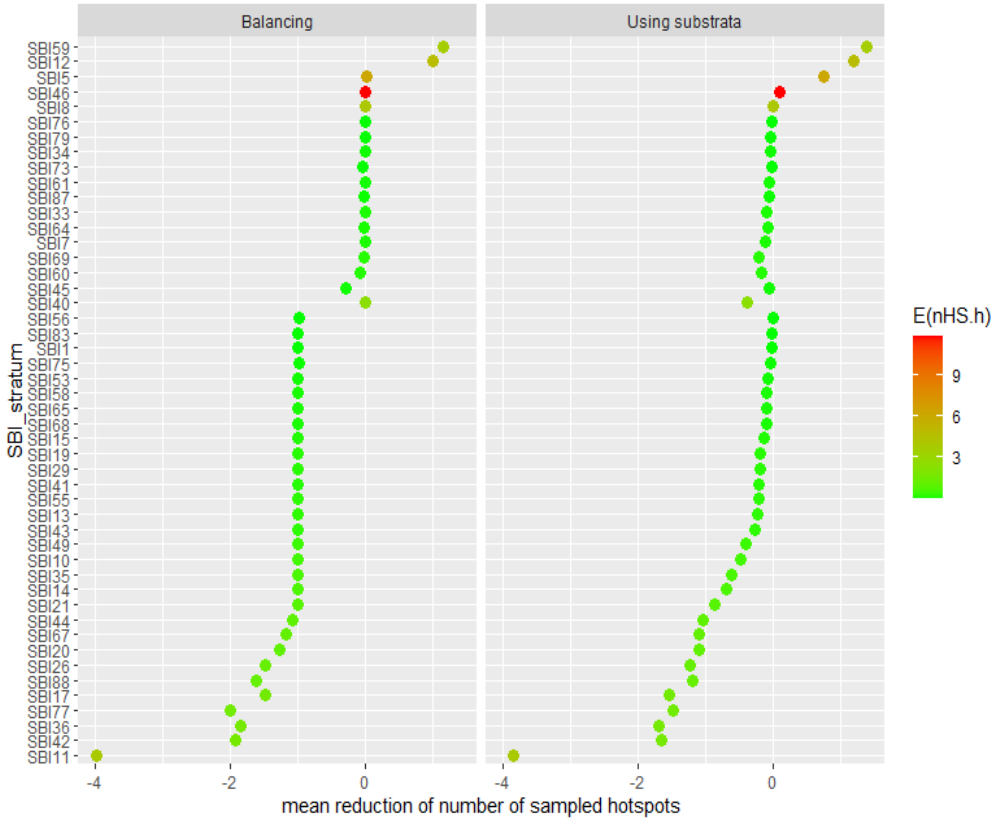


Figure 4.5: Mean reduction of sampled hotspots compared to increasing fractions. Color indicates expected number of sampled hotspots $\mathbf{E}(n_{\mathrm{HS},h})$.

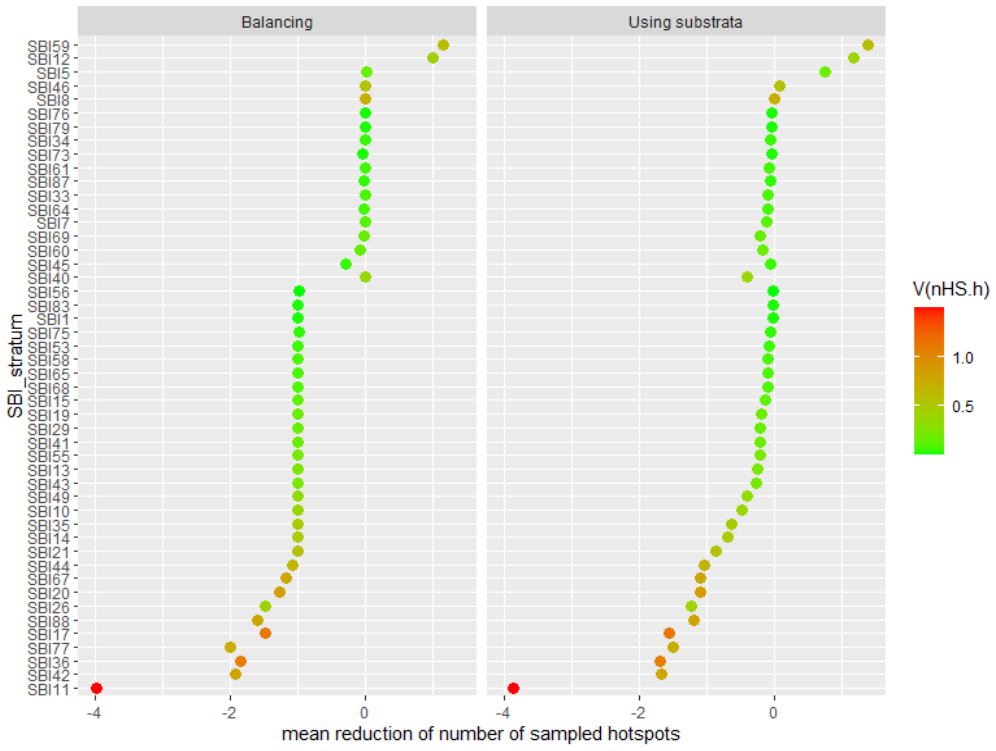


Figure 4.6: Mean reduction of sampled hotspots compared to increasing fractions. Color indicates variance of number of sampled hotspots $\mathbf{V}(n_{\mathrm{HS},h})$.

When substrata are used, we find similar results to balanced sampling. In the middle part of the displayed strata in Figure 4.5, roughly the strata with SBI 43, 13, 55, … , 56, using substrata leads to a mean reduction close to 0 hotspots, while balanced sampling selects one hotspot more than the method of increasing fractions. In these strata, the sampling fractions are small, and there are only a few hotspots in the population so the expected number of sampled hotspots $\mathrm{E}(n_{\mathrm{HS},h})$ is close to 0 (Table A.1). When using substrata, samples with a fixed sample size are drawn from the substratum of hotspots, where the sample size is the rounded number of $\mathrm{E}(n_{\mathrm{HS},h})$ (Section 3), which is mostly 0 for these strata in the simulation. This leads to a slightly smaller number of sampled hotspots in the simulation than using balancing. Balanced sampling draws a sample from the complete stratum with the prescribed sampling fraction. In practice, rounding is also applied for balancing, but the effect on the realized sampling fractions is smaller than in the substrata approach. This is because sampling is applied on the unit level for balancing and on the stratum level in the substrata approach.

In general, we can say that if there are enough non-hotspot units in the population that can replace the hotspots, increasing the fractions leads to fewer sampled hotspots than balancing and using substrata. In that case, a large reduction can be achieved when the number of hotspots in the preliminary sample is (relatively) large, which is for example the case for the stratum with SBI 11. Also, in Figures 4.5 and 4.6, we see a stronger correlation between the achieved reduction and the variance of the number of sampled hotspots (Figure 4.6) than between the achieved reduction and the expectation (Figure 4.5).

4.2 Simulation study 2: longer-term effects

In this subsection, we report the results of the second simulation study. Figure 4.7 shows the development of the number of sampled hotspots by the sampling methods for all strata in the population. The first 12 months are not shown, because there are no hotspots yet in the first year of the simulated series of draws given the hotspot definition. Table 4.4 gives a summary of the distributions of the number of sampled hotspots in the months 13-300 of the simulated series. The first thing to notice is that the series of balancing and using substrata are at the same level as the series of ignoring hotspots. The mean of the number of sampled hotspots is about 185 per month for these methods. The series of increasing fractions is at a lower level, with a mean of about 165 sampled hotspots per month. So there is a clear reduction in the number of sampled hotspots when the fractions are increased. If we look at the standard deviation (sd) in the series, we see that ignoring hotspots leads to the largest sd. This implies that the spread of the peaking response burden can be improved by increasing fractions, balancing, and using substrata. The smallest sd, and therefore, the best spread of the response burden is achieved by balanced sampling. Another interesting result is that the maximum number of sampled hotspots for increasing fractions (175 in month 233) is larger than the smallest value (168 in month 226) for ignoring hotspots. This means there is no difference between increasing fractions and ignoring hotspots in some months.

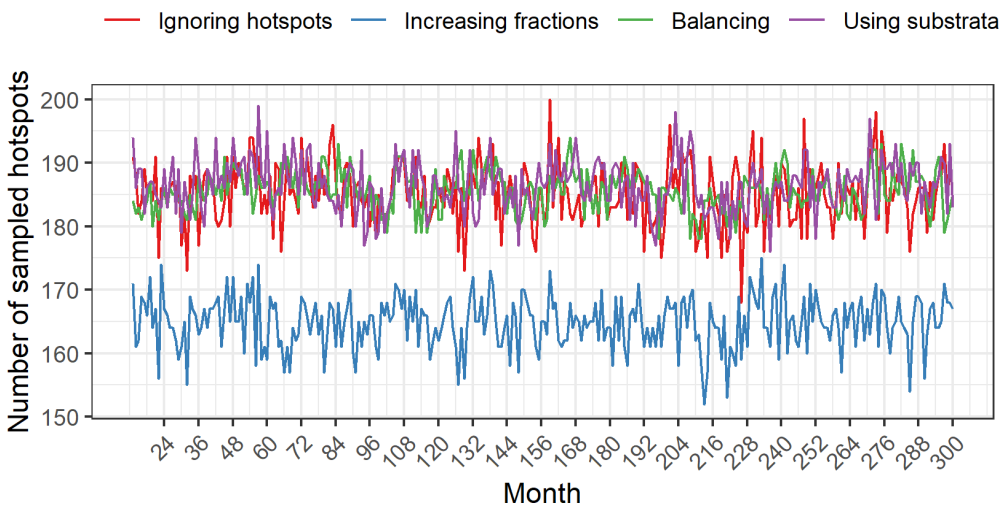


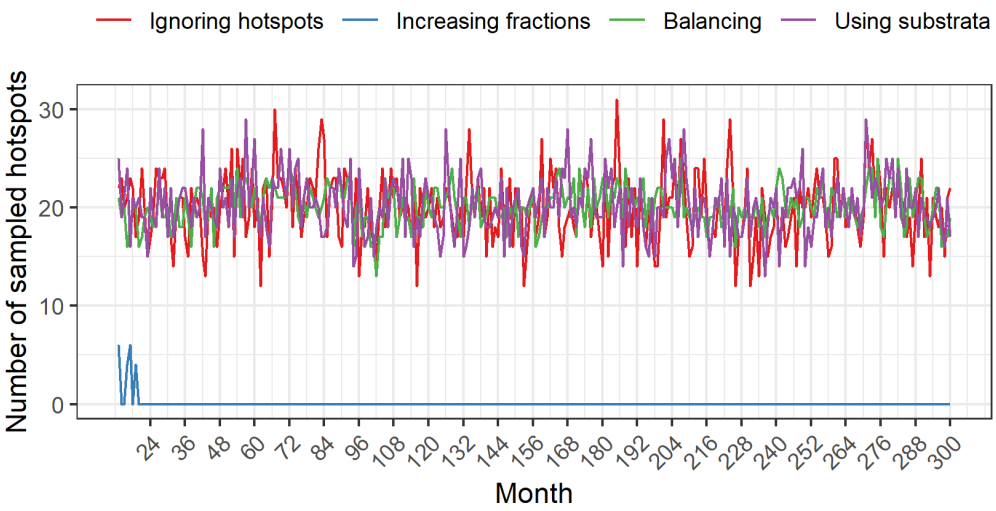
Figure 4.7: Time series of number of sampled hotspots for all strata in the population.

Method	min.	mean	max.	sd
Ignoring hotspots	168	185.2	200	4.82
Increasing fractions	152	164.9	175	4.14
Balancing	178	185.6	195	3.25
Using substrata	176	186.4	199	4.06

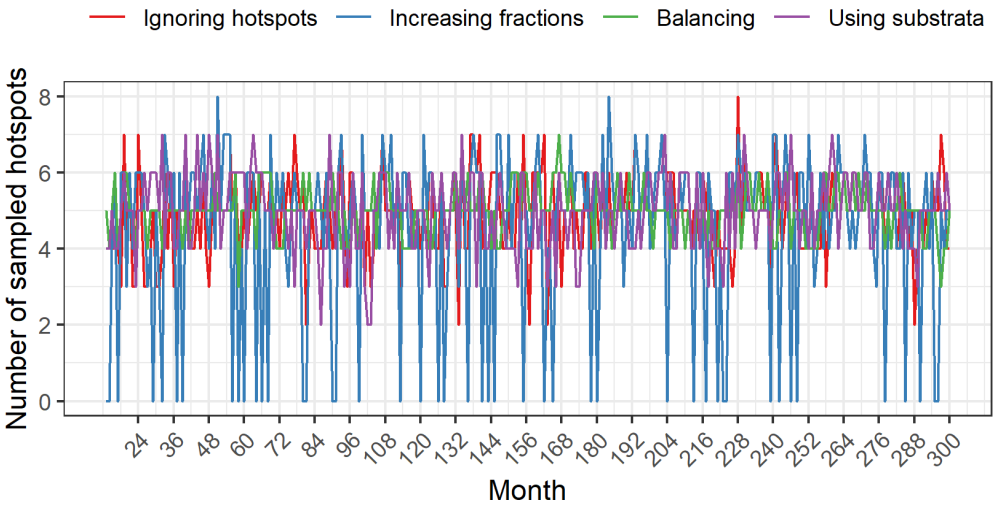
Table 4.4: Summary of number of sampled hotspots in the months 13-300 of the simulated series for all strata in the population.

Figure 4.8 and Table 4.5 show the results of the sampling methods in selections of strata with respectively small, medium-sized, and large sampling fractions f_h. For strata with very small sampling fractions $(0 < f_h < 0.04)$ no hotspots have been created in the simulated series of 300 months. In these strata, the number of sampled hotspots is always zero for all sampling methods (results are not shown). In strata with small sampling fractions $(0.04 \leq f_h \leq 0.41)$ there is room in the population every month to replace hotspots with non-hotspot units. The strongest reduction is then achieved by increasing the fractions, where the number of sampled hotspots in each month equals zero (Subfigure 4.8a and Subtable 4.5a). The numbers of sampled hotspots for balancing and using substrata are at the same level as for ignoring hotspots, implying that there is no reduction in the number of hotspots for these methods. Compared to ignoring hotspots and using substrata, balancing results in the smallest sd in these strata, and therefore, in the best spread of the peaking response burden.

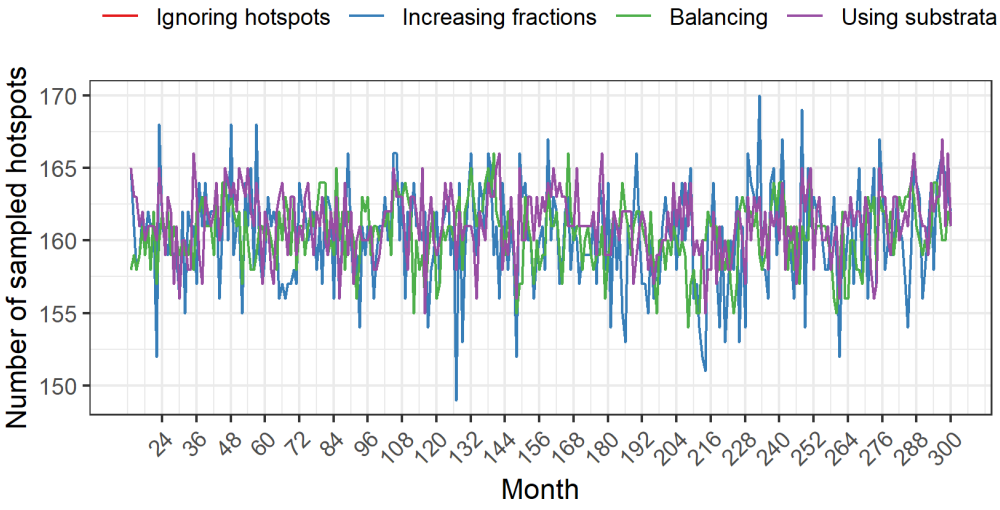
In strata with medium-sized sampling fractions $(0.41 < f_h \leq 0.48)$ the method of increasing fractions shows an interesting pattern (Subfigure 4.8b and Subtable 4.5b). In this case, the numbers of sampled hotspots are roughly at the same level for all sampling methods, while increasing fractions leads to the largest sd. By replacing sampled hotspots with non-hotspot units more and more units in the population are becoming hotspots. At a certain point, there is no more room to replace the sampled hotspots with non-hotspots in the population.



(a) Strata with small sampling fractions ($0.04 \leq f_h \leq 0.41$).



(b) Strata with medium-sized sampling fractions ($0.41 < f_h \leq 0.48$).



(c) Strata with large sampling fractions ($0.48 < f_h < 1$).

Figure 4.8: Times series of number of sampled hotspots for selections of strata by sampling fraction.

Method	min.	mean	max.	sd
Ignoring hotspots	12	19.96	31	3.53
Increasing fractions	0	0.07	6	0.60
Balancing	13	20.17	25	2.04
Using substrata	13	20.29	29	3.00

(a) strata with small sampling fractions ($0.04 \leq f_h \leq 0.41$)

Method	min.	mean	max.	sd
Ignoring hotspots	2	4.92	8	1.14
Increasing fractions	0	4.51	8	2.19
Balancing	3	5.00	7	0.64
Using substrata	2	4.98	7	0.99

(b) strata with medium-sized sampling fractions ($0.41 < f_h \leq 0.48$)

Method	min.	mean	max.	sd
Ignoring hotspots	149	160.3	170	3.44
Increasing fractions	149	160.3	170	3.44
Balancing	154	160.4	166	2.23
Using substrata	155	161.2	167	2.28

(c) strata with large sampling fractions ($0.48 < f_h < 1$)

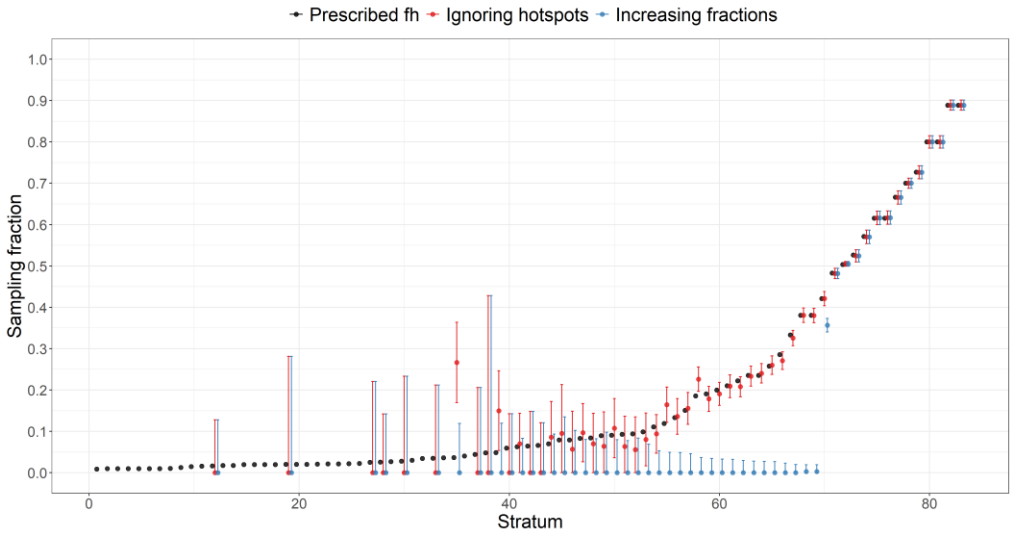
Table 4.5: Summary of number of sampled hotspots in the months 13-300 of the simulated series for selections of strata.

This leads to a worse spread of the peaking response burden and no reduction in the number of sampled hotspots is achieved. Also, for these strata, the optimal spread of response burden is obtained by balancing, with the smallest sd of 0.64.

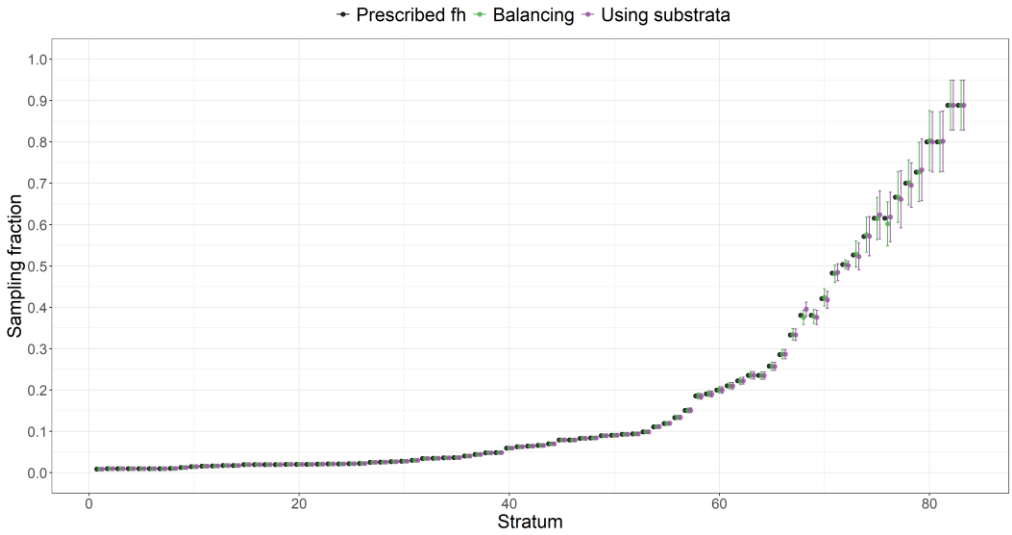
When the fractions become larger ($0.48 \leq f_h < 1$) there is no room in the population to replace the sampled hotspots with non-hotspot units. In that case, increasing fractions gives exactly the same results as ignoring hotspots (see Subfigure 4.8c, Subtable 4.5c, and also Section 3). For this reason the red line of ignoring hotspots cannot be seen in Subfigure 4.8c, because the blue line of increasing fractions has overlaid it. For these strata, the number of sampled hotspots is at the same level for all sampling methods, and the smallest sd is obtained by balanced sampling.

As a final result, we examine the realized sampling fractions in the simulation study. It already follows from the way the sampling methods are implemented (Section 3), that the realized sampling fractions $\hat{f}_{h,m}$ per stratum h are equal to the prescribed sampling fractions f_h for each sampling method m. This is also the case in the simulation study (results are not shown).

Figure 4.9 shows the realized sampling fractions $\hat{f}_{\mathrm{HS},h,m}$ in the subpopulation of hotspots per stratum h together with the 95% confidence intervals that are based on the 95% margins introduced in Subsection 3.2. The strata on the *x*-axis are ordered by the fractions f_h. In strata with very small fractions $(0 < f_h < 0.04)$ no hotspots have (yet) been created in the simulated series of draws. Accordingly, no realized sampling fractions or confidence intervals can be computed. These results are not shown.



(a) Ignoring hotspots and increasing fractions.



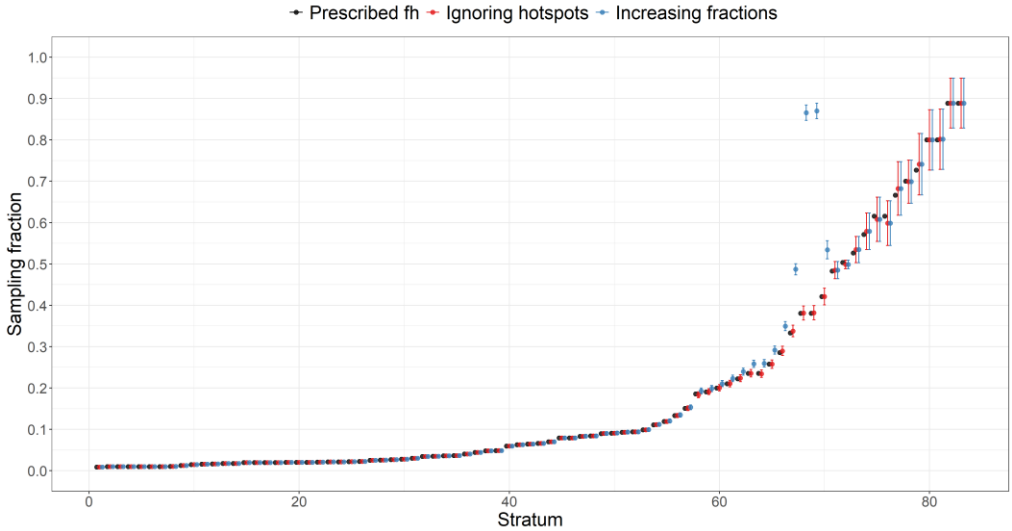
(b) Balanced sampling and using substrata.

Figure 4.9: Realized sampling fractions $\hat{f}_{\mathrm{HS},h,m}$ with 95% confidence intervals in the subpopulation of hotspots per stratum h and method m.

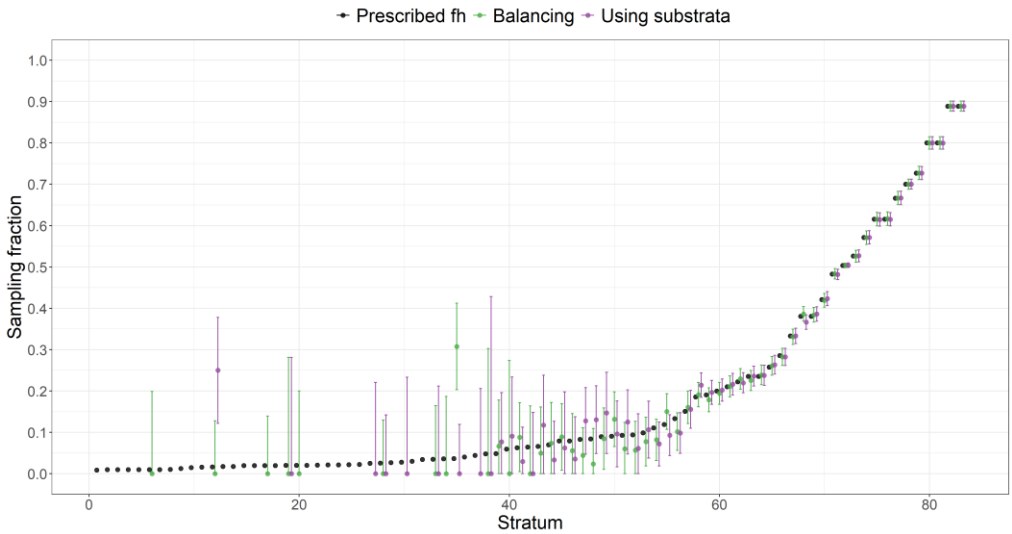
Subfigure 4.9a shows that when the hotspots are ignored, there are four strata where the prescribed fractions are outside the 95% confidence intervals of the realized fractions. These are the strata with index 35, 39, 55, and 58 in the plot which correspond to those strata with SBI 41, 52, 10, and 67. When increasing fractions is applied, the realized fractions in the subpopulation of hotspots are too low in 23 strata. These are the strata with index 47-48 and 50-70 in the plot. They correspond to the strata with SBI 43, 81, 58, 65, 84, 33, 17, 10, 49, 35, 67, 98, 29, 36, 14, 71, 88, 44, 21, 25, 100, 50 and 42, which are mainly strata with small

sampling fractions ($0.08 \leq f_h \leq 0.41$). For balancing and using substrata, there is only one stratum where the fraction f_h is outside the 95% interval (Subfigure 4.9b), which is likely due to chance. For balancing, this is the stratum with index 35 (SBI 41), and for using substrata the stratum with index 12 (label 12).

In the subpopulation of non-hotspots the realized fractions can be computed in every stratum and for all sampling methods. The results are shown in Figure 4.10. For ignoring hotspots, balancing, and using substrata the realized sampling fractions $\hat{f}_{\mathrm{HS},h,m}$ are close to the prescribed fractions f_h. All fractions are in the 95% confidence intervals. From Subfigure 4.10a it follows that for the method of increasing the fractions there are 12 strata where the realized fractions are too high, and f_h is outside the 95% interval. These are the strata with index 59-70, which correspond to those strata with SBI 98, 29, 36, 14, 71, 88, 44, 21, 25, 100, 50 and 42. The sampling fractions in these strata range from 0.19 to 0.42 (medium-sized sampling fractions).



(a) Ignoring hotspots and increasing fractions.



(b) Balanced sampling and using substrata.

Figure 4.10: Realized sampling fractions $\hat{f}_{\mathrm{nHS},h,m}$ with 95% confidence intervals in the subpopulation of non-hotspots per stratum h and method m.

From the results of this simulation study, we conclude that for a specific range of small sampling fractions a reduction in the number of sampled hotspots can be achieved by increasing the fractions, but with a risk of introducing selectivity in the drawn samples. The realized sampling fractions among hotspots are much smaller than the prescribed sampling fractions. For other values of the sampling fractions no reduction in the number of sampled hotspots is obtained by increasing the fractions, but there is a risk of worsening the spread of the peaking response burden. Balancing and using substrata do not reduce the number of sampled hotspots but realize a more even spread of the peaking response burden without introducing selectivity in the drawn samples. Balanced sampling and using substrata lead to similar results in this simulation, where balanced sampling results in a more even spread in all considered situations.

5. Discussion and Conclusion

In this paper two simulation studies are carried out comparing sampling methods to better spread the peaking response burden for small businesses at Statistics Netherlands (SN). The reason for SN to focus on businesses with a peaking response burden was a letter of complaint from the business community (Commissie van Straalen, Strategische commissie betere regelgeving bedrijven, 2020). Although SN operates a coordinated sampling system (EDS) (Smeets and Boonstra, 2018) that allows to spread the response burden evenly among the businesses, each year several businesses are still heavily sampled. In a pilot study, small and medium-sized businesses with a peaking response burden were labelled as hotspots and a number of sampling methods were evaluated to lighten the peaking response burden either by reducing or by spreading the number of sampled hotspots as much as possible (Gommans et al., 2022, Ippel et al, 2023). In Matei et al. (2023) a sample coordination method has been developed based on balanced sampling by which the spread of the peaking response burden for small businesses at Statistics Netherlands can be improved. This method amongst others is analyzed in simulation studies presented in this paper.

In the simulation studies we focus on spreading the peaking response burden for small businesses and compare the method of balanced sampling with the method used in the pilot, where hotspots are replaced by other units by first removing hotspots from the sampling frame and then by drawing a sample with the prescribed sample size from the remaining (non-hotspot) units. In this method the sampling fractions of the non-hotspot units are increased during the sampling process. The simulation is applied to the Dutch Research & Development (R&D) Survey, which uses a stratified sample design with equal inclusion probabilities within the strata. Because of this, in addition to balanced sampling, we also consider the use of substrata, where separate substrata are created for the hotspots and non-hotspots and random samples are drawn in both substrata with equal sampling fractions. Since most business surveys at SN use stratified sample

designs, it is useful to investigate the use of substrata as an alternative sampling method.

The first simulation considers a practical situation at SN and applies the considered sampling methods to a given R&D sample that is drawn with the EDS, but without taking into account the hotspots. The considered sampling methods are then applied with the aim of replacing the hotspots in the given sample by other (non-hotspot) units as much as possible. The simulation shows that in strata with enough non-hotspot units in the population, that is, when the sampling fraction does not exceed the population fraction of non-hotspot units, the largest number of hotspots can be replaced by increasing the fractions, compared to balancing and using substrata. In that case, a large reduction of the number of sampled hotspots can be achieved when the number of hotspots in the given sample is (relatively) large.

The second simulation investigates the longer-term effects of the sampling methods on the number of sampled hotspots. A series of monthly draws is simulated with a length of 25 years, where the hotspot status is monthly updated for all the businesses in the population. Here the same definition is used as in the pilot. This simulation shows that for a specific range of small sampling fractions a reduction in the number of sampled hotspots can be achieved by increasing the fractions, but with a risk of introducing selectivity in the drawn samples. The realized sampling fractions among hotspots are smaller than the prescribed sampling fractions. For other values of the sampling fractions no reduction in the number of sampled hotspots is obtained by increasing the fractions and there is a risk of worsening the spread of the peaking response burden. This is due to the fact that by replacing hotspots by non-hotspot units, more units in the population are becoming hotspots. At a certain point, there is no room left to replace the sampled hotspots by non-hotspots in the population. Balancing and using substrata do not reduce the expected number of sampled hotspots, but realize a more even spread of the peaking response burden without introducing selectivity in the drawn samples. In this simulation, balanced sampling and using substrata lead to similar results regarding the number of sampled hotspots, while balanced sampling results in the best spread in all considered situations.

In general, we conclude that the method of increasing the sampling fractions can substantially reduce the number of hotspots in a given sample, especially when the number of sampled hotspots is large and when there are enough non-hotspot units in the population. However, for the longer term there is no guarantee that the number of sampled hotspots will remain at a low level using this method. Moreover, there is a risk of introducing selectivity in the samples which can be amplified by spreading the response burden over time. It has therefore been decided to use this method only for strata with small sampling fractions, so that only a few hotspots are replaced by other units and the risk of introducing selectivity is limited.

The method of balanced sampling minimizes the variance of the number of sampled hotspots and prevents the occurrence of large numbers of sampled hotspots. Using substrata provided similar results. While balanced sampling is

more generally applicable, using substrata is probably easier to implement in an existing coordinated sampling system like the EDS. However, a disadvantage of the use of substrata is that, due to rounding issues, the prescribed inclusion probabilities will be less well realized as the strata become smaller. In a follow-up project the implementation of balanced sampling in the EDS will be investigated.

The results can be generalized to other surveys with a stratified sample design, but also to the situation where successive samples are drawn for surveys with different stratifications. In the latter case a basic stratification could be defined by crossing the strata of the separate surveys. Furthermore, the hotspot status of the businesses is used as measure for the response burden in the simulations. The hotspot status is based on the number of surveys for which a business has been sampled within the last 12 months. The use of other measures for the response burden, e.g., weights based on the completion time of the questionnaire, would most likely not lead to other conclusions (studied in Matei et al., 2023).

The method from the pilot which manually removes hotspots from a drawn sample in collaboration with the statistical division is not addressed in the simulation studies. The reason is, that it does not apply sample coordination but in fact reduces the sampling fractions. Also with this method there is a risk of introducing selectivity. In addition, like the other methods from the pilot, this method introduces employee burden at SN because for each sample manual work is required. It has therefore been decided to use this method also in a limited way and to remove only a limited number of hotspots.

A final conclusion is that if a structural reduction of the number of sampled hotspots is desired, the methods from the pilot do not constitute a long-term solution. A potential solution is adjusting the sampling fractions of the business surveys. This includes formalizing the methods from the pilot in the sample designs as well as harmonizing stratifications across different surveys in order to facilitate a more extensive sample coordination by the EDS.

References

CBS (2022). Onderzoeksomschrijvingen: Research and Development. https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksomschrijvingen/research-development.

Gommans, G., Burgerjon, L., and Paulissen, R. (2022). Hotspots. Internal CBS memo.

Grafström, A. (2012). Spatially correlated Poisson sampling. Journal of Statistical Planning and Inference 142(1), 139–147.

Grafström, A., Lisic, J., and Prentius, W. (2022). BalancedSampling: balanced and spatially balanced sampling. R package version 1.6.3, https://CRAN.R-project.org/package=BalancedSampling.

Ippel, L., Gommans, G., Paulissen, R., Burgerjon, L., Kitzen, R., Klingwort, J., Smeets, M., Houben, L. , and Snijkers, G. (2023). Hotspots project. Evaluatie en geleerde lessen. Internal CBS memo.

Klingwort, J. and Smeets, M. (2023). Hotspot Analyses Using the Research & Development and Community Innovation Survey. Internal CBS report.

Matei, A., Smith, P., Smeets, M., and Klingwort, J. (2023). Targetted double control of burden in multiple surveys [in print, publication expected 2023]. Survey Methodology.

Rosén, B. (1997). On sampling with probability proportional to size. Journal of Statistical Planning and Inference 62, 159–191.

Smeets, M. and Boonstra, H. J. (2018). Sampling coordination of business surveys at Statistics Netherlands. In B. Lorenc, P. Smith, M. Bavdaž, G. Haraldsen, D. Nedyalkova, L.-C. Zhang, en T. Zimmermann (Eds.), The Unit Problem and Other Current Topics in Business Survey Methodology, pp. 127–137. Newcastle-upon-Tyne: CambridgeScholars.

Strategische commissie betere regelgeving bedrijven (2020). Onevenredige enquêtebelasting van specifieke ondernemers in het MKB. https://open.overheid.nl/repository/ronl-7973b111-ee24-4d3e-84d5-01d3dfdcb9f1/1/pdf/advies%20strat%20cie%20200605.pdf.

Appendix A. R&D population

stratum h	N_h	n_h	f_h	$N_{\mathrm{HS},h}$	$F_{\mathrm{HS},h}$	$\mathrm{E}(n_{\mathrm{HS},h})$	$\mathrm{V}(n_{\mathrm{HS},h})$
STP:GK4:SBI1	1300	26	0.02	1	0.00	0.02	0.00
STP:GK4:SBI10	67	8	0.12	4	0.06	0.48	0.40
STP:GK4:SBI100	21	8	0.38	0	0.00	0.00	0.00
STP:GK4:SBI11	29	14	0.48	8	0.28	3.84	1.51
STP:GK4:SBI12	10	8	0.80	6	0.60	4.80	0.43
STP:GK4:SBI13	127	8	0.06	4	0.03	0.24	0.22
STP:GK4:SBI14	36	8	0.22	3	0.08	0.66	0.47
STP:GK4:SBI15	166	8	0.05	3	0.02	0.15	0.15
STP:GK4:SBI17	72	8	0.11	14	0.19	1.54	1.11
STP:GK4:SBI19	124	8	0.06	3	0.02	0.18	0.15
STP:GK4:SBI2	13	8	0.62	0	0.00	0.00	0.00
STP:GK4:SBI20	89	8	0.09	12	0.13	1.08	0.83
STP:GK4:SBI21	28	8	0.29	3	0.11	0.87	0.58
STP:GK4:SBI22	267	8	0.03	0	0.00	0.00	0.00
STP:GK4:SBI24	458	8	0.02	0	0.00	0.00	0.00
STP:GK4:SBI25	24	8	0.33	0	0.00	0.00	0.00
STP:GK4:SBI26	13	8	0.62	2	0.15	1.24	0.42
STP:GK4:SBI29	40	8	0.20	1	0.03	0.20	0.19
STP:GK4:SBI3	9	8	0.89	0	0.00	0.00	0.00
STP:GK4:SBI33	101	10	0.10	1	0.01	0.10	0.09
STP:GK4:SBI34	320	8	0.03	2	0.01	0.06	0.08
STP:GK4:SBI35	53	8	0.15	4	0.08	0.60	0.51
STP:GK4:SBI36	38	8	0.21	8	0.21	1.68	1.08
STP:GK4:SBI40	10	8	0.80	3	0.30	2.40	0.37
STP:GK4:SBI41	217	8	0.04	5	0.02	0.20	0.15
STP:GK4:SBI42	19	8	0.42	4	0.21	1.68	0.81
STP:GK4:SBI43	96	8	0.08	3	0.03	0.24	0.22
STP:GK4:SBI44	31	8	0.26	4	0.13	1.04	0.69
STP:GK4:SBI45	309	8	0.03	2	0.01	0.06	0.08
STP:GK4:SBI46	20	14	0.70	17	0.85	11.90	0.56
STP:GK4:SBI49	60	8	0.13	3	0.05	0.39	0.33
STP:GK4:SBI5	9	8	0.89	7	0.78	6.23	0.17
STP:GK4:SBI50	21	8	0.38	0	0.00	0.00	0.00
STP:GK4:SBI51	786	16	0.02	0	0.00	0.00	0.00
STP:GK4:SBI52	164	8	0.05	0	0.00	0.00	0.00
STP:GK4:SBI53	1811	37	0.02	3	0.00	0.06	0.00
STP:GK4:SBI54	947	10	0.01	0	0.00	0.00	0.00
STP:GK4:SBI55	3600	36	0.01	22	0.01	0.22	0.35
STP:GK4:SBI56	2527	25	0.01	1	0.00	0.01	0.00
STP:GK4:SBI57	890	8	0.01	0	0.00	0.00	0.00
STP:GK4:SBI58	88	8	0.09	1	0.01	0.09	0.07

STP:GK4:SBI59	11	8	0.73	5	0.45	3.65	0.59
STP:GK4:SBI60	352	8	0.02	7	0.02	0.14	0.15
STP:GK4:SBI61	121	8	0.07	1	0.01	0.07	0.07
STP:GK4:SBI62	479	8	0.02	0	0.00	0.00	0.00
STP:GK4:SBI63	3432	34	0.01	0	0.00	0.00	0.00
STP:GK4:SBI64	101	8	0.08	1	0.01	0.08	0.07
STP:GK4:SBI65	86	8	0.09	1	0.01	0.09	0.07
STP:GK4:SBI67	43	8	0.19	6	0.14	1.14	0.80
STP:GK4:SBI68	1269	25	0.02	5	0.00	0.10	0.00
STP:GK4:SBI69	134	8	0.06	3	0.02	0.18	0.15
STP:GK4:SBI7	511	8	0.02	7	0.01	0.14	0.08
STP:GK4:SBI70	286	8	0.03	0	0.00	0.00	0.00
STP:GK4:SBI71	34	8	0.24	0	0.00	0.00	0.00
STP:GK4:SBI72	301	8	0.03	0	0.00	0.00	0.00
STP:GK4:SBI73	369	8	0.02	1	0.00	0.02	0.00
STP:GK4:SBI74	916	9	0.01	0	0.00	0.00	0.00
STP:GK4:SBI75	807	16	0.02	2	0.00	0.04	0.00
STP:GK4:SBI76	940	19	0.02	1	0.00	0.02	0.00
STP:GK4:SBI77	139	70	0.50	3	0.02	1.50	0.69
STP:GK4:SBI78	19	10	0.53	0	0.00	0.00	0.00
STP:GK4:SBI79	473	10	0.02	1	0.00	0.02	0.00
STP:GK4:SBI8	12	8	0.67	6	0.50	4.02	0.73
STP:GK4:SBI80	233	8	0.03	0	0.00	0.00	0.00
STP:GK4:SBI81	95	8	0.08	0	0.00	0.00	0.00
STP:GK4:SBI82	229	8	0.03	0	0.00	0.00	0.00
STP:GK4:SBI83	764	16	0.02	1	0.00	0.02	0.00
STP:GK4:SBI84	85	8	0.09	0	0.00	0.00	0.00
STP:GK4:SBI85	101	8	0.08	0	0.00	0.00	0.00
STP:GK4:SBI86	644	8	0.01	0	0.00	0.00	0.00
STP:GK4:SBI87	197	8	0.04	1	0.01	0.04	0.08
STP:GK4:SBI88	34	8	0.24	5	0.15	1.20	0.80
STP:GK4:SBI89	532	8	0.02	0	0.00	0.00	0.00
STP:GK4:SBI90	1581	16	0.01	0	0.00	0.00	0.00
STP:GK4:SBI91	221	8	0.04	0	0.00	0.00	0.00
STP:GK4:SBI92	903	9	0.01	0	0.00	0.00	0.00
STP:GK4:SBI93	180	8	0.04	0	0.00	0.00	0.00
STP:GK4:SBI94	114	8	0.07	0	0.00	0.00	0.00
STP:GK4:SBI95	14	8	0.57	0	0.00	0.00	0.00
STP:GK4:SBI96	453	8	0.02	0	0.00	0.00	0.00
STP:GK4:SBI97	405	8	0.02	0	0.00	0.00	0.00
STP:GK4:SBI98	42	8	0.19	0	0.00	0.00	0.00
STP:GK4:SBI99	372	8	0.02	0	0.00	0.00	0.00

Table A.1: population U of R&D survey 2020.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2017–2018	2017 to 2018 inclusive
2017/2018	Average for 2017 to 2018 inclusive
2017/'18	Crop year, financial year, school year, etc., beginning in 2017 and ending in 2018
2013/'14–2017/'18	Crop year, financial year, etc., 2015/'16 to 2017/'18 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbs.nl/information