Zins, Stefan; Burgard, Jan Pablo

Working Paper

Planning domain sizes in cluster sampling

Research Papers in Economics, No. 6/20

Provided in Cooperation with:
University of Trier, Department of Economics

Suggested Citation: Zins, Stefan; Burgard, Jan Pablo (2020) : Planning domain sizes in cluster sampling, Research Papers in Economics, No. 6/20, Universität Trier, Fachbereich IV - Volkswirtschaftslehre, Trier

This Version is available at:
https://hdl.handle.net/10419/243473

Universität Trier

Planning Domain Sizes in Cluster Sampling

Stefan Zins
Jan Pablo Burgard

Research Papers in Economics
No. 6/20

Planning Domain Sizes in Cluster Sampling

Stefan Zins and Jan Pablo Burgard

October 1, 2020

Abstract

Multi-stage cluster sampling is a common sampling design of social surveys because populations of interest are often structured by, or partitioned into, disjoint organizational and administrative units. The need to use cluster sampling can conflict with survey planners' goal to select a sample that contains a specific number of elements from certain domains of interest. This can be a complex problem if sampling units, i.e. clusters, cut across the domains of interest, as it is often the case. For example, an analysis require sufficient observations from certain age and gender categories. But the population is clustered within schools, hospitals, establishments, or municipalities and hence age-gender categories cannot be used for stratification.

We propose a quadratic optimization approach to define inclusion probabilities that can be used for drawing balanced cluster samples that comply with predefined sample sizes from domains of interest. Henceforth the clusters may cut across domains. We also provide an application of the proposed solution to the domain size problem for an existing social survey on migration and emigration in Germany.

Keywords: planned domain sizes, balanced sampling, cluster sampling, quadratic optimization

1 Introduction

Cluster sampling is a common sight within sampling designs in social surveys, where instead of sampling individual elements groups of elements are selected. This is because the target populations are often structured according to some hierarchical clustering. For example, residents are clustered within municipalities, students are clustered within schools, and employees are clustered within establishments. Sampling frames for these clustering units can be more easily be compiled and maintained than a collective sampling frame for the entire population. There are also other circumstances that make cluster sampling necessary, like cost and field time constraints. Thus, feasible sampling strategies often involve selecting clusters of units first, then surveying all, or parts, of the elements they contain.

Cluster sampling, however, comes with some disadvantages. In many practical applications it will reduce the efficiency of estimation strategies, because the similarity of elements within the same cluster is likely to increase the sampling variance of estimators (Lohr, 2009, Ch. 9.5). Another drawback of cluster sampling is the difficulty it poses to the planning of domain sizes, which is the focus of this paper. For instance, a survey planner may want to realize a certain allocation of the sample size over predefined age-gender groups of a population of persons. Implementing such an allocation would be straightforward, if a collective sampling frame existed that could be stratified according to these age-gender groups.

Unfortunately, if persons cannot be selected under a single-stage sampling design, realizing planned domain sizes in the sample is complicated. Especially, if clusters cut across domains of interest. Typically clusters are intrinsic to the target population and cannot be designed freely, hence their cutting across domains can be seen as the standard setting. The survey planner has to find a sampling design for the first sampling stage and subsequent sampling stages that controls for the domain sizes in the sample. Solving the problem of finding a multi-stage or cluster sampling design that controls for domain sizes in the sample (i.e., The Domain Size Problem of Cluster

Sampling) is only complicated, if inclusion probabilities of elements are to remain computable in closed form. If inclusion probabilities do not need to be calculable, the problem can be solved by adding a second sampling stage. Sampling at the second-stage (or any subsequent stage) is simply made conditional on the selected cluster sample (or any prior selection of sampling units). But knowing the inclusion probabilities is a prerequisite to construct widely used classes of unbiased or asymptotically unbiased estimators (see e.g. Gelman, 2007; Kott, 2018).

An example of a two-stage design would be to pool all selected clusters, stratify the pooled elements by domains of interest, and select a stratified simple random sample. If domain sizes are constant across all clusters - and the number of sampled clusters is fixed - then such a design would still have inclusion probabilities that have a straightforward closed form. However, if domain sizes are not constant, then the stratum sizes of the second sampling stage will depend on the set of sampled clusters, thus inclusion probabilities have to be approximated. First order Taylor Series approximations could be used in such a case. Furthermore, variance estimation under such a design would be difficult, as the second order inclusion probabilities also have to be approximated. This is even more complex, especially if the domain sizes of the pooled samples are random or not controlled for. One could use Monte Carlo methods to approximate the second order inclusion probabilities. But this can be very time consuming and impractical, because the process of calculating the first and second order probabilities has to be repeated with every adjustment survey planners do to the pooled design. Because of uncertainties with the required power of statistical test or field work parameters such as response rates, recalculations can be expected to happen frequently during the planning stage. Due to the necessary approximations of the inclusion probabilities, and the complex survey design, the variance estimation is complex as well. These are impractical features making the pooling of clusters a cumbersome sampling strategy for planning and inference.

it is even more complicated to solve the domain size problem with a cluster sampling design that has calculable inclusion probabilities if the distribution of domain sizes is highly unequal across

clusters. To better control for domain sizes in the sample the survey planner could split clusters into multiple strata, where the strata contain clusters of similar domain size. The drawback of this approach is that it increases the complexity of the sampling if different sampling designs within the strata have to be used. For example, if some clusters have prohibitively large domain sizes and thus only parts of them can be surveyed, whereas within other strata domain sizes are much smaller and all elements in a cluster can be surveyed. Such mixed designs require alternative methods for data analysis, typically leading to more complex procedures. For example, the standard formula for the design effect (Kish, 1965, p. 162) can no longer be used (Gabler et al., 1999). Also a stratification of primary sampling units (PSUs) that produce such homogeneous strata with respect to domain sizes has to be found, which might be difficult without the help of optimization methods.

We propose an approach to obtain a sampling design that *ex-ante* controls for the size of domains within the sample and does not need a mixture of sampling designs. The basic idea is to use balanced sampling with inclusion probability weighted domain sizes of clusters as balancing variables. We show how to compute the optimal inclusion probabilities to plan domain sizes that cut across clusters in cluster sampling.

The approach can also be applied to multi-stage sampling. For this, the domain specific sample sizes from each PSU need to be both non-random and known. If more than a single element is surveyed from at least one PSU we have a cluster of units from this PSU in the sample, i.e. the PSUs are our clusters in case of multi-stage sampling.

In Section 2 we show how a balanced cluster sampling design can be found that has fixed domain sizes, predefined by the survey planner, and known inclusion probabilities for the elements populating the clusters.

We show that it is possible to solve the problem of finding inclusion probabilities that fulfill the balancing requirements by treating it as a QP-optimization problem. In Section 3 we apply our

method to solving the problem of developing a sampling design with fixed domain sizes for the German Emigration and Remigration Panel Study. Furthermore, we compare the actual sampling design that has been used for the surveys with our approach. This demonstrates the advantages of balancing the cluster sample directly on the desired size of domains of interest in practice.

Finally, in Section 4 we give a summary of our method and point out which sampling design planning problems our methods might be best suited.

2 Drawing Cluster Samples with Planned Domain Sizes

First we introduce some notation for cluster sampling and domain size planning. Let $\mathcal{U} = \{k|k = 1, \ldots, N\}$, $N > 2$ and $N \in \mathbb{N}$, be the index set of our target population of size N. Further, let $C = \{c|c = 1, \ldots, C\}$, with $C > 2$ and $C \in \mathbb{N}$ be the index set of clusters in the population of interest, (e.g. our PSUs), and $\bigcup_{c \in C} \mathcal{U}_c = \mathcal{U}$, where $\mathcal{U}_c \subset \mathcal{U}$, i.e. $\mathcal{U}_c$ is the index set of our target population within the c-th cluster. We define a cluster sampling design as a discreet probability distribution over set $\mathcal{S}_n \subset \mathbb{P}(C)$, where $\mathcal{S}_n$ is the set of of all n sized subsets of the power set $\mathbb{P}(C)$ of C. Function $p(.)$ is the discreet probability mass function of our sampling design and the set of all $s \in \mathcal{S}_n$ with $p(s) > 0$ is the support of our sampling design. For the sake of simplicity we denote $\mathcal{S}_n$ as the support of our cluster sampling design with fixed samples sizes of n clusters. Then the inclusion probability of the $c - th$ cluster is given by

$$\pi_c = \sum_{s \in \mathcal{S}_n} I_c(s) p(s) \ ,$$

where $I_c(s)$ is an indicator function, assuming a value of one if $c \in s$ and zero otherwise.

To define domains and domain sizes, let $d = \{1, \ldots, D\}$ be the index set of our D domains of interest and $m_{cd} \in \mathbb{N}$ a non-random integer denoting the elements *surveyed* from the d-th domain of the c-th cluster, if $I_c(s) = 1$. We assume, the survey planner seeks to obtain a sample of clusters

$s \subset \mathcal{S}_n$ such that

$$\sum_{c \in s} m_{cd} = \tau_d, \quad \forall d \in \{1, \ldots, D\}, \tag{1}$$

where, $\tau_d,\ d = 1, \ldots, D$, are the planned domain sizes for the survey. That is, s needs to be selected in such a way that the aggregated domain sizes of the selected clusters equal the desired domain sizes in the sample.

The sampling design that allows us to impose constraints in Equation (1) is balanced sampling, proposed by Royall and Pfeffermann (1982); Deville and Tillé (2004). Balanced sampling is a method of selecting samples with equal or unequal inclusion probabilities, under the condition that the Horvitz-Thompson estimators, (i.e. the inclusion probability weighted sample total), for the known totals of the balancing variables have, in case of perfect balancing, a sampling variance of zero (Tillé and Favre, 2004). The balancing variables can be chosen freely by the survey planner. Thus, using the size of domains of interest per cluster multiplied by the inclusion probabilities of the clusters as a balancing variable will result in fixed sample sizes for domains of interest.

For sampling design $p(.)$ to be (perfectly) balanced on known totals $\tau_d,\ d \in \{1, \ldots, D\}$, the following conditions need to be met

$$\sum_{s \in \mathcal{S}_n} \frac{I_c(s) m_{cd}}{\pi_c} = \tau_d, \quad \forall d = 1, \ldots, D, \quad \forall c \in C. \tag{2}$$

The two constraint in Equations (1) and (2) are conceptually very different. Where Equation (1) constrains the domain sizes in the sample, Equation (2) constrains the estimated totals of the domains on the level of the target population. We could use as balancing constraints in (2) $\bar{\tau}_d = \sum_{c \in C} m_{cd}$, i.e. we balance on the totals of the survey elements form domains $d = 1, \ldots, D$. However, this would not directly satisfy Equation (1). In order to achieve the constraints in Equation (1) with a balanced sample it is necessary to find a suitable balancing variable a_{cd}, such

that

$$\sum_{c \in C} \frac{I_c(s) a_{cd}}{\pi_c} = \tau_d, \qquad \forall d = 1, \ldots, D \quad \forall s \in \mathcal{S}_n. \tag{3}$$

Which is achieved by defining $a_{cd} = \pi_c m_{cd}$ as the balancing variable and consequently $\tau_j = \sum_{c \in C} a_{cd}$. Our aim is to find a vector of inclusion probabilities $\pi \in (0,1)^C$ that satisfies $\tau_d = \sum_{c \in C} a_{cd}$ as close as possible. At the same time it is desirable to deviate as little as possible from some inclusion probabilities π^*, that are initially set by the survey planner. For example probabilities proportional to the total elements surveyed from a cluster, i.e. $\pi_c^* = m_c/\tau$, where $m_c = \sum_{d=1}^{D} m_{cd}$ and $\tau = \sum_{d=1}^{D} \tau_d$. We choose the sum of squared deviations as distance measure between π and π^*, since we would like to penalize more the greater the deviations from the initial vector π^* is. Other distance function can be used, however this results in adopting the following optimization problems.

To find a_{cd}, $c = 1, \ldots, C$, $d = 1, \ldots, D$ that satisfy Equation 3 we formulate the following optimization problem

$$\begin{aligned} \underset{\pi_c \in (0;1),\, \epsilon_c \in \mathbb{R}}{argmin} \;\; & \alpha \frac{\sum\limits_{c \in C} (\pi_c - \pi_c^*)^2}{2} + (1-\alpha) \frac{\sum\limits_{d \in \{1,\ldots,D\}} \epsilon_D^2}{2} \\ & s.t.\, E[\sum_{c \in s} m_{cd}] = \tau_d + \epsilon_d, \;\; \forall d \in \{1, \ldots, D\} \end{aligned} \tag{OP$_1$}$$

The $\epsilon_d \in \mathbb{R}$, $d = 1, \ldots, D$, are some slackness variables, that allow the optimization procedure to slightly deviate from the strict equality constraints, which ensures a non empty solution space. Parameter $\alpha \in \mathbb{R}_{>0}$ can be set by the survey planner to put either more importance on a small deviation of π from π^* (higher $\alpha/(1-\alpha)$) or on the tightness of the constraints (lower $\alpha/(1-\alpha)$) in Problem OP$_1$.

For any sampling design with inclusion probabilities π that solve the optimization problem in (OP$_1$), $E(\sum_{c \in s} m_{cd})$ is equal to $\sum_{c \in C} \pi_c m_{cd}$. Therefore, the optimization Problem (OP$_1$) can be

rewritten as:

$$\underset{\pi_c \in (0;1),\, \epsilon_c \in \mathbb{R}}{argmin} \; \alpha \frac{\sum\limits_{c \in C} (\pi_c - \pi_c^*)^2}{2} + \beta \frac{\sum\limits_{d \in 1,\ldots,D} \epsilon_D^2}{2} \qquad s.t. \sum_{c \in C} \pi_c m_{cd} = \tilde{\tau}_d + \epsilon_d, \forall d \in \{1, \ldots, D\} \tag{OP$_2$}$$

Optimization Problem (OP$_2$) can then be formulated as a quadratic programming problem as follows:

$$\begin{aligned} \underset{x=(\pi\ \epsilon),\, \pi \in (0;1),\, \epsilon \in \mathbb{R}^D}{argmin} \quad & x'Gx - g'tx \\ s.t.\ & A_1'x = \tau_x \\ & A_2'x > 0 \\ & A_3'x > -1 \end{aligned} \tag{OP$_3$}$$

with

$$G = diag(diag((\alpha)_{1\times C}), diag((\beta)_{1\times D})$$

$$g = (\alpha \cdot \pi^*, (0)_{1\times D})$$

$$A_1 = \begin{pmatrix} X \\ diag((-1)_{1\times D}) \end{pmatrix}, \quad A_2 = \begin{pmatrix} diag((1)_{1\times C}) \\ 0_{D\times C} \end{pmatrix}, \quad A_3 = \begin{pmatrix} diag((-1)_{1\times C}) \\ 0_{D\times C} \end{pmatrix} \quad .$$

Problem (OP$_2$) can be solved with standard quadratic programming problem solvers such as quadprog in R by Turlach and Weingessel (2019). The resulting vector of inclusion probabilities π can then be used in combination with a balanced sampling design, which balances on variable $a_{cd} = \pi_c m_{cd}$, to obtain a sampling design that has a support that fulfills the condition in Equation (1).

3 Applications

To illustrate the benefit of our proposed method we compare it with the sampling design of a survey that required samples with planned domain sizes, but had no collective sampling frame available.

3.1 GERPS Sampling Design

The *German Emigration and Remigration Panel Study (GERPS)* includes a sample survey of persons that have either emigrated out of, or migrated back to, Germany. The aim of the survey is to assess the effects this event has on different aspects of peoples lives, like education and professional careers (Ette, 2020). The study had two domains of interest when planning the sampling design, emigrants and remigrates. It was planned to interview 3000 emigrants and remigrates each. Assumed response and contact rates for remigrates and emigrants lead to a gross sample that was required to contain 21467 emigrants and 20202 remigrates. That is, within the notation of Section 2 we have, $D = 2$, and $\tau_1 = 21467$, and $\tau_2 = 20202$.

Contact information for emigrants and remigrates is available at the level of municipalities in Germany. Emigrants are asked to notify the municipality in which they reside prior to their departure about their intent to move abroad and leave an address under which they can be contacted. Remigrates have to register in municipalities to which they move, as all residents in Germany are required to. To gain access to the information on emigrants and remigrates municipalities have to be contacted individually and cooperate voluntarily with the survey planner.

The major problem with selecting the gross samples was the high concentration of emigrants and remigrates within a relatively small number of municipalities. There were, in total, 11054, municipalities in Germany at the time the survey was designed. See Bundesamt (2017) for a detailed description of available municipalities in 2017 in Germany. However most of them report

no emigrants or remigrates at all or they have less than 3 emigrants and/or remigrates, in which case no information on the number of emigrants or remigrates was given. Data on the number emigrants and remigrates was then available form 4638 municipalities.

Figure 3.1 shows the Lorenz curves for the distribution of emigrants and remigrates over these 4638 municipalities in the sampling frame. The black and the red curve correspond to emigrants and remigrates, respectively.

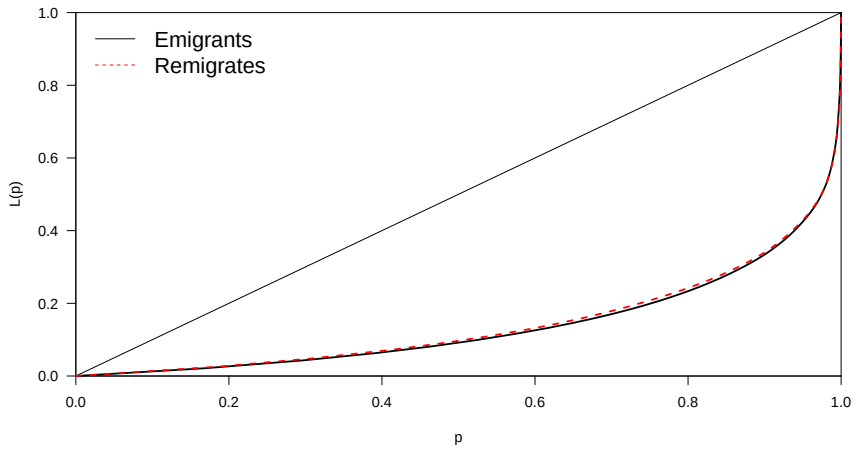

Figure 1: Lorenz curve of the number of remigrates and emigrants in municipalities in Germany

As apparent in the Lorenz curves, both distributions have a relatively high concentration, with emigrants having a slightly higher concentration. The Gini coefficients, i.e. double the area between bisector and these Lorenz curves, are 0.7237 and 0.7148 for emigrants and remigrates respectively.

Additionally to the required number of emigrants and remigrates in the sample the condition was imposed that sampled sizes for municipalities must not exceed 70. The mentioned requirements and the high concentration of emigrants and remigrates made lead to the decision not to use a single stage cluster sampling or a uniform two-stage sampling. The former was done to reduce the concentration of remigrates and emigrants from the largest municipalities in the sample. The later, because finding an appropriate first stage sampling design for municipalities and second stage

sampling designs for emigrants and remigrates from each municipality in the frame was too complex problem to solve manually. Implementing a design that fulfills all the above conditions is difficult, because of the large heterogeneity in the number of emigrants and remigrates across the municipalities. It involves the complex task of finding the number of emigrants and remigrates that have to be surveyed in each municipality, such that the desired domain sizes are met with a sample of 70 municipalities. However, this problem can be solved by finding the solution to optimization problem OP_2 or OP_3 in Section 2.

To fulfill the stated requirements for the sample design it was decided to split the sampling frame into two strata, where different sampling designs could be applied independently within those strata. The first stratum was comprised of the 10 municipalities with the highest sum of emigrants and remigrates, while the second stratum included all the others. Within the first stratum the ten municipalities formed strata again. Then a stratified sample with simple random sampling within municipalities was used for each domain within the first stratum of municipalities. The sample sizes for emigrants and remigrates within the first stratum of municipalities was allocated proportional to the domain sizes of the municipalities.

For the second stratum a single stage cluster sample was used, where all emigrants and remigrates of each sampled municipalities were selected. We selected the sample of municipalities from the second domain using a balanced sampling design (Deville and Tillé, 2004), with inclusion probability proportional to the sum of both domains with each municipality and a balancing variables equal to the number of emigrants and remigrates within each of the 16 federal states of Germany were used. In order to reduce the range of inclusions probabilities all municipalities with less than 5 emigrants or remigrates were excluded from the sampling. This left 2133 municipalities in the sampling frame of stratum 2. The first stratum of the sampling frame then contained around 26% of the emigrant and 30% of the migrant population, while consisting only of around 0.47% of municipalities in the frame.

Deciding on the number of emigrants and remigrates to be selected from the two municipality

strata involved an interactive procedure using a simulation to approximate the expected domain sizes in the second stratum, which is random. Their variances are limited by the used of the above described balance sampling design, however.

3.2 Alternative Sampling Design

The method presented in Section 2 allows for a straightforward and transparent way of planning samples sizes for domains of interest in the presence of clustered populations. Using the method presented in Section 2 we can construct a sampling design, that does not require a predefined cluster stratification, in order to define a survey sample satisfying domain size targets τ_1 and τ_2.

We apply our method to four scenarios that differ by the sampling frames used.

Scenario F : The sampling frame includes all 4638 municipalities that have 3 or more emigrants and/or remigrates.

Scenario S : The sampling frame includes all 2143 municipalities that have 5 or more emigrants and remigrates.

Scenario R : The sampling frame includes all 4638 municipalities that have 3 or more emigrants and/or remigrates. The domain sizes of the 10 largest municipalities are set to the number of emigrants and remigrates sampled from these municipalities in the GERPS sampling design.

Scenario RS : The same sampling frame as in Scenario R, but only with the 2143 municipalities that have 5 or more migrants and remigrates.

To evaluate how well our proposed method in Section 2 is able to select inclusion probabilities that, in conjunction with a balanced design, fulfill the condition in Equation 3, we conducted a simulation study. First, for each scenario, we compute the inclusion probabilities by solving optimization Problem OP_3 as described in Equation (3). Following the notation introduced in

Section 2 we have m_{c1} and m_{c2} as the number of emigrants and remigrates in the c-th municipality, respectively. As a starting vector π^* for the optimization of the inclusion probabilities we use a "probability proportional to size approach", with

$$\pi_c^* = 70 \cdot \frac{m_{c1} + m_{c2}}{\sum_{c \in C} (m_{c1} + m_{c2})}, \quad \forall c \in C. \tag{4}$$

Table 1: Expected relative Bias of domain and sample sizes

	sample sizes	emigrants	remigrates
F	-0.000311	0.000000	-0.000000
S	-0.000671	0.000000	-0.000000
R	0.019218	0.000006	-0.000009
RS	0.018762	0.000006	-0.000009

As can be seen in Table 1 the optimized inclusion probabilities for all scenarios lead to low expected biases. In Scenarios R and RS the total of the inclusion probabilities is around 71.3 indicating that about 71 municipalities should be drawn for the sample. In Scenarios F and S the total of the inclusion probabilities is almost 70 and hence reaching the targeted values for the sample size almost perfectly. In all Scenarios, the targeted domain sizes for emigrants and remigrates are met precisely. That is $\sum_{c \in C} \pi_c x_{cd} \approx \tau_d$, $d = 1, 2$ is archived with high accuracy.

Given the scenarios and their computed inclusion probabilities we draw 2000 balanced samples using the cube method to evaluate how good the used implementation of the balanced samples method (Grafström and Lisic, 2019), fulfills our balancing conditions, as show in Equation (3). As balancing variables we use the following three:

$$\begin{aligned} a_{c1} &= \pi_c m_{c1} \\ a_{c2} &= \pi_c m_{c2} \\ b_c &= \pi_c \end{aligned} \tag{5}$$

Variables a_{c1} and a_{c2} are used to control for domain sizes τ_1 and τ_2 in the sample and b_c to control for the number of clusters sampled, because $\sum_{c \in C} \pi_c \approx 70$. Inclusion probabilities

π_c, $c = 1, \ldots, C$ are the solution to Problem OP_3.

Figure 2 shows the distribution of the relative bias from the planned domains and sample size target for the four different scenarios. The Scenarios F and S behave similarly as do R and RS.

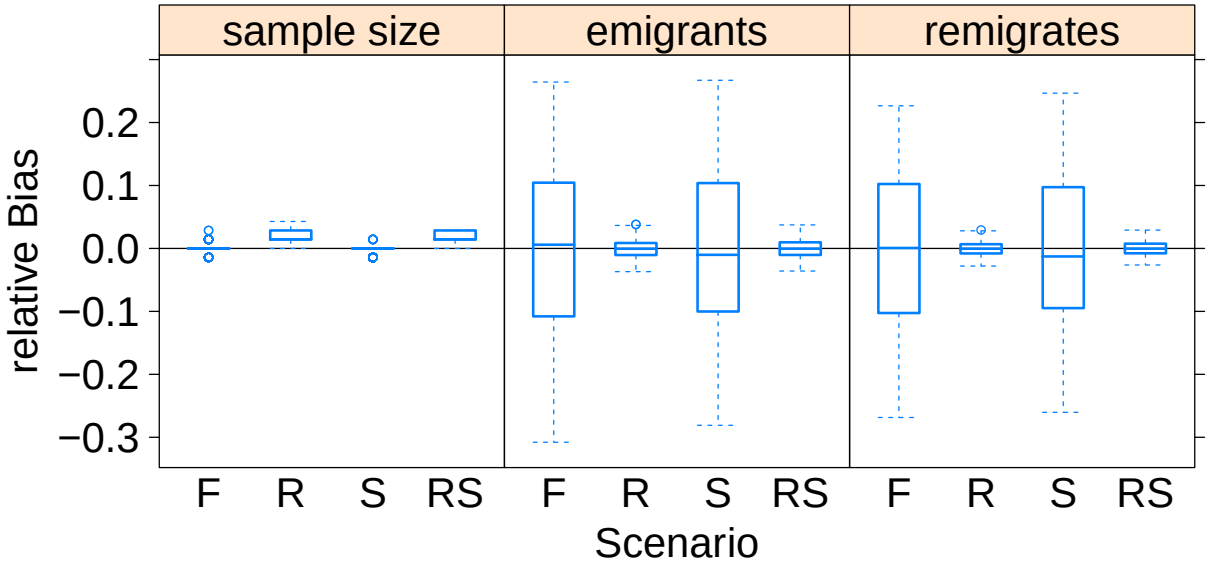

Figure 2: Boxplot of relative Bias of the number of municipalities, emigrants and remigrates in the sample over 2000 repeated samples

The driving factor of differences is here the reduction in the domain sizes of the largest 10 municipalities. This can be explained by the reduced concentration of domain size in the scenarios R and RS, which makes it easier to balanced the samples using the cube method (Deville and Tillé, 2004). This means that at the cost of increasing the municipality sample by only one element, it is possible to find a balanced sampling design that almost perfectly fulfills the required domain sizes without having to combine multiple sampling design as in the original design of the GERPS study.

4 Summary and Outlook

With the methodology presented in Section 2 we address a common problem when planning a sampling design for a clustered population. We show how to control for the size of domains of

interest in the sample, when domains of interest cut across sampling clusters. E.g., planning the size of age and gender categories in a sample of persons when persons are clustered within municipalities or institutions like schools, hospitals or retirement homes.

We propose to solve a quadratic optimization problem to find a vector of inclusion probabilities which can be used, by a balanced sampling design, to achieve the planned domain sizes. To demonstrate our approach we apply it to solve the problem of planning the number of emigrants and remigrates of the GERPS study, which use a more conventional approach, of combining different independent sampling designs to solve the domain problem. We showed that we can produce a sampling design that achieves the same goals while having a much more streamlined work flow, which does not require manually setting the different parameters of multiple sampling designs. As only one sampling design was to be accounted for when using our proposed method, variance estimation and substantial analysis on the sampled data is less complex and in line with the standard approach. Therefore, the presented approach allows to apply available software to estimate sampling variances, such as the R *survey* package (Lumley, 2004).

References

Bundesamt, A. S. (2017), *Statistisches Jahrbuch, 2017: Deutschland und Internationales*, Statistisches Bundesamt.

Deville, J.-C. and Tillé, Y. (2004), "Efficient balanced sampling: the cube method," *Biometrika*, 91, 893–912.

Ette, A. (2020), "German Emigration and Remigration Panel Study (GERPS)," .

Gabler, S., Häder, S., and Lahiri, P. (1999), "A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering," *Survey Methodology*, 25, 105–106.

Gelman, A. (2007), "Struggles with Survey Weighting and Regression Modeling," *Statist. Sci.*, 22, 153–164.

Grafström, A. and Lisic, J. (2019), *BalancedSampling: Balanced and Spatially Balanced Sampling*, r package version 1.5.5.

Kish, L. (1965), *Survey Sampling*, Wiley.

Kott, P. S. (2018), "A design-sensitive approach to fitting regression models with complex survey data," *Statist. Surv.*, 12, 1–17.

Lohr, S. L. (2009), *Sampling: design and analysis*, Nelson Education, 2nd ed.

Lumley, T. (2004), "Analysis of Complex Survey Samples," *Journal of Statistical Software*, 9, 1–19, r package verson 2.2.

Royall, R. M. and Pfeffermann, D. (1982), "Balanced samples and robust Bayesian inference in finite population sampling," *Biometrika*, 69, 401–409.

Tillé, Y. and Favre, A.-C. (2004), "Coordination, Combination and Extension of Balanced Samples," *Biometrika*, 91, 691–709.

Turlach, B. A. and Weingessel, A. (2019), *quadprog: Functions to Solve Quadratic Programming Problems*, r package version 1.5-8.