Statistical Journal of the IAOS 34 (2018) 271–279
DOI 10.3233/SJI-170358
IOS Press

Variance reduction using a non-informative sampling design

Thomas Zimmermann
Statistisches Bundesamt (Destatis), Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany
Tel.: +49 611 753841; Fax: +49 611 754000; E-mail: Thomas.Zimmermann@destatis.de

Abstract. Official Statistics commonly conducts sample surveys to produce estimates of aggregate statistics with a desired level of precision. For this purpose, design-based methods are used which are suitable for the estimation of finite population quantities such as totals or means. In most cases, however, model-based analyses are applied to the survey data as well. Examples include small area estimation techniques that allow for reliable estimates of finite population quantities in the presence of small sample sizes and socio-econometric models used in academia to test scientific hypotheses. This may cause problems as model-based methods frequently assume a non-informative sampling design and a violation of this assumption can lead to erroneous statistical inferences. We argue in this work that if the application of model-based methods can be anticipated before the sample is drawn, then this knowledge should be incorporated in the survey design. We propose a method called antithetic clustering that enables precise estimates for aggregate figures using design-based estimation methods and does automatically lead to non-informative sampling designs. Our method is compared against other sampling plans designed to achieve precise design-based estimates for aggregates in a simulation study.

Keywords: Variance reduction, design-based inference, model-based analysis, small area estimation

1. Introduction

Traditionally, Official Statistics has adopted a design-based approach to produce estimates using sample data. Thus, the sample is collected by means of probability sampling, i.e. each unit in the population has a known and positive probability of being included in the sample [1, p. 32]. Moreover, an estimator is chosen which possesses certain desirable properties such as design-consistency. Hence, very precise estimates can be obtained provided the sample size is large. While this prerequisite will be met for national statistics or for large subgroups that have been incorporated in the sampling design as strata, it might not be true for some small subgroups. A potential remedy in this case is the use of model-based small area estimation methods [2–4]. A caveat regarding the application of many model-based small area estimation techniques is that they are not design-consistent under general sampling designs. Hence, their design-bias does not vanish as the sample size increases and these methods are consequently not robust against a potential model misspecification. Furthermore, the sampling design can even induce biases of model-based estimates when the model is correctly specified. This phenomenon is known as informative sampling and arises whenever a model that can be validated for the sample differs from the model which holds for the population [5, p. 455]. As a consequence, the sample model cannot be used for inference on the population model without further adjustments. Ignoring this fact may lead to erroneous statistical inferences.

In most applications, estimates for subgroups with small sample sizes as well as estimates on aggregate levels with large sample sizes are needed at the same time. This poses a challenge to the survey planner, as the sampling design has to reflect different and potentially conflicting requirements simultaneously. On the one hand, the sampling design should be built on information related to the variable of interest to enable efficient design-based estimates for aggregate statistics. This could be achieved via stratification [6, p. 450] or

sampling with probabilities proportional to size, where a proportional relationship between the size variable and the variable of interest is desirable [1, p. 88]. On the other hand, these optimised designs may lead to informative sampling and thereby invalidate conclusions drawn from model-based estimation procedures. For those estimators non-informative designs such as simple random sampling (SRS) that do not interfere with the model would be beneficial. However, plain SRS schemes do not use auxiliary information at the design stage and are thus not very suitable for design-based estimation of aggregate figures.

The preceding discussion clearly indicates that the trade-off between design-optimisation and modelling should be already dealt with in the sampling design. Even though both design- and model-based estimates are regularly published by statistical offices [7,8], designs reflecting the needs of both philosophies have rarely been discussed. A notable exception is due to [9], who propose a box-constraint optimal allocation in stratified random sampling (StrRS), where the variance of a national statistic is minimised under an implicit restriction on the range of the sampling weights.

In Section 2, we propose a sampling method that allows for precise design-based estimates and is non-informative by construction. Our approach is based on the technique of antithetic variates, which is a well-known method to reduce the variance in Monte-Carlo simulations [10]. We adapt this approach to the context of survey sampling and derive conditions under which it will yield estimates with a higher precision than SRS.

Section 3 presents the results of a design-based simulation study, where we compare our method against various alternative sampling designs for both design- and model-based estimators.

Finally, concluding remarks are given in Section 4.

2. Antithetic clustering

2.1. Notation

Following [11], we consider a fixed and finite population $U = \{1, \ldots, k, \ldots, N\}$, with values $(y_1, \ldots, y_k, \ldots, y_N)'$ of the variable of interest y. The values y_k are only observed for the elements included in the sample $S \subset U$ of size n, which is drawn using a probability sampling mechanism. The population can be further partitioned into D mutually exclusive domains (or areas) $U_d \subset U, d = 1, \ldots, D$ with domain sizes N_d. Thus, the part of the sample which is taken from domain d is given by $S_d = S \cap U_d$ and the resulting sample size in the domain d is denoted as n_d. Note that depending on the sampling design used, the sample sizes in domains may be random. In this article, we focus on the estimation of the national mean $\mu_Y = (\sum_{k \in U} y_k / N)$ and the domain means $\mu_{Y,d} = (\sum_{k \in U_d} y_k / N_d)$. Estimators of μ_Y and $\mu_{Y,d}$ will be denoted as $\hat{\mu}_Y^*$ and $\hat{\mu}_{Y,d}^*$ where the "*"-symbol refers to the estimation method used. Moreover, we assume that the sampling frame comprises information about a size variable z, whose values z_k are known for all units in the population. Hence, this size variable can be used by the survey planner when constructing the sampling mechanism. Frequently, the survey collects a $(p+1)$-dimensional vector of auxiliary information $\boldsymbol{x}_k = (1, x_{1k}, \ldots, x_{pk})'$ as well. If the corresponding population totals of the vector of auxiliary information $\tau_{\boldsymbol{X}} = \sum_{k \in U} \boldsymbol{x}_k$ are known at the estimation stage, the $\boldsymbol{x}_k$ can be incorporated as covariates in model-based and model-assisted estimation procedures [11, p. 220]. It should be noted that as z_k is assumed to be known for the entire population, it can easily be included among the covariates $\boldsymbol{x}_k$.

Table 1
Algorithm for antithetic clustering

1. Order the elements according to the values of z_k.
2. Set $i = 0$ and assign units 1 and N from the ordered vector to the first cluster.
3. Increase i by 1 and assign units $1 + i$ and $(N - i)$ from the ordered vector to the next cluster.
4. Repeat step 3 until all units have been assigned to a cluster. The procedure yields $L = \lceil N/2 \rceil$ clusters, where the last cluster may either comprise one unit (odd N) or two units (even N).
5. Draw $l > 1$ out of L clusters by means of a simple random sample.

2.2. Our approach

Our aim is to construct a sampling design, which enables precise design-based estimates but does not distort the properties of statistical models. To do so, we combine single stage cluster sampling with the idea of antithetic variates, where pairs of negatively correlated random variables are drawn to reduce the variance in Monte-Carlo simulations [10, Chapter 5]. We call our proposed method antithetic clustering (ATC). It is summarised in Table 1.

Since the sample is drawn using a simple random sample of clusters, all clusters and hence all units $k = 1, \ldots, N$ of the population have the same probability of being included in the sample. Consequently, the sample selection mechanism does not depend on

the values of the variable of interest such that the proposed method is non-informative by construction. Note that the idea to base the sampling design on the sorted values of an auxiliary variable is not new. A systematic sampling approach based on the ordered values of z_k is discussed in [12, Section 3.4.2] and references therein. Technically, systematic sampling can be considered a special case of a single stage cluster sampling design where only one cluster is selected. Thus, unbiased variance estimation under a design-based approach is infeasible as the second-order inclusion probabilities $\pi_{kl} = 0$ for elements which do not belong to the same cluster [1, p. 75]. This limitation does not occur with our approach, as we draw a simple random sample of clusters, where more than one cluster is sampled.

The question that remains is whether our sampling mechanism is suitable for design-based estimation. Therefore, we study the properties of the sample mean under single stage cluster sampling.

2.3. The efficiency of single stage cluster sampling

A sampling design yields estimates with a higher precision than simple random sampling provided its design effect is less than one. This design effect (DEFF) under single stage cluster sampling is closely related to the intraclass correlation coefficient (ICC) in the case of evenly sized clusters where $N_h = \overline{N}_L$ for all $h = 1, \ldots, L$ [13, Section 5.2.2]:

$$DEFF = \frac{\overline{N}_L L - 1}{\overline{N}_L (L-1)} (1 + (\overline{N}_L - 1) ICC) \tag{1}$$

From Eq. (1), it follows that $ICC < \frac{-1}{L\overline{N}_L - 1}$ is required to obtain better estimates under single stage cluster sampling compared to SRS. Note that the ICC is defined as [13, Section 5.2.2]

$$ICC = 1 - \frac{\overline{N}_L}{\overline{N}_L - 1} \frac{SSW_Y}{SSW_Y + SSB_Y} \tag{2}$$

where

$$\begin{aligned} SSB_Y &= \sum_{h=1}^{L} N_h \overline{Y}_h^2 - N \overline{Y}^2 \\ SSW_Y &= \sum_{k=1}^{N} y_k^2 - \sum_{h=1}^{L} N_h \overline{Y}_h^2 \end{aligned} \tag{3}$$

denote the sum of squares between clusters and the sum of squares within clusters, respectively. Using Eqs (2) and (3), we get the following condition for a variance reduction compared to SRS:

$$SSW_Y > \frac{L(\overline{N}_L - 1)}{L - 1} SSB_Y. \tag{4}$$

An implication of Eq. (4) is to create clusters such that most of the variation of the dependent variable is due to variation within the clusters, not between clusters. What does this imply for our ATC approach? Intuitively, the clusters will have a large ratio of the within versus the between variation for the size variable. It can be shown that our approach is optimal among all possible combinations of PSUs, which are exhaustive, mutually exclusive and where one unit with an above-median value of the size variable is clustered with a unit with a below-median value. This follows from applying the rearrangement equality, which is given in [14, p. 261]. Having established a certain optimality of ATC for the size variable, we need to examine the implications for our variable of interest. To do so, we consider models specifying the data generating process.

2.4. ATC under a single level model

Suppose that the relationship between the dependent variable and the size variable used for clustering is given by the simple linear regression model

$$y_k = \beta_0 + \beta_1 z_k + \varepsilon_k, \quad \varepsilon_k \overset{i.i.d.}{\sim} G(0, \sigma^2) \tag{5}$$

where ε_k denotes the error term, which is assumed to be independently and identically distributed according to a distribution G with mean zero and variance σ^2. The expectation of the SSB and SSW under model Eq. (5) can be calculated using

$$\begin{aligned} E_M(SSB_Y) &= \sum_{h=1}^{L} N_h E_M(\overline{Y}_h^2) - N E_M(\overline{Y}^2), \\ E_M(SSW_Y) &= \sum_{k=1}^{N} E_M(y_k^2) - \sum_{h=1}^{L} N_h E_M(\overline{Y}_h^2), \end{aligned} \tag{6}$$

where $E_M(\cdot)$ denotes the expectation with respect to the model. It can be seen from Eq. (6) that we need expressions for the expected values of $\overline{Y}_h^2$, $\overline{Y}^2$ and y_k^2 under model Eq. (5). They are readily available using the variance identity $E(X^2) = [E(X)]^2 + Var(X)$. This leads to

$$\begin{aligned} E_M(y_k^2) &= \beta_0^2 + 2\beta_0\beta_1 z_k + \beta_1^2 z_k^2 + \sigma^2, \\ E_M(\overline{Y}_h^2) &= \beta_0^2 + 2\beta_0\beta_1 \overline{Z}_h + \beta_1^2 \overline{Z}_h^2 + \sigma^2/N_h, \\ E_M(\overline{Y}^2) &= \beta_0^2 + 2\beta_0\beta_1 \overline{Z} + \beta_1^2 \overline{Z}^2 + \sigma^2/N. \end{aligned} \tag{7}$$

Inserting expressions Eq. (7) in Eq. (6) yields the following equations:

$$
\begin{aligned}
E_M(SSB_Y) &= \beta_1^2 \left(\sum_{h=1}^{L} N_h \overline{Z}_h^2 - N \overline{Z}^2 \right) \\
&\quad + \sigma^2 (L-1), \\
E_M(SSW_Y) &= \beta_1^2 \left(\sum_{k=1}^{N} z_k^2 - \sum_{h=1}^{L} N_h \overline{Z}_h^2 \right) \\
&\quad + \sigma^2 (N-L)
\end{aligned} \tag{8}
$$

Equation (8) and utilising

$$
SSB_Z = \sum_{h=1}^{L} N_h \overline{Z}_h^2 - N \overline{Z}^2,
$$

$$
SSW_Z = \sum_{k=1}^{N} z_k^2 - \sum_{h=1}^{L} N_h \overline{Z}_h^2,
$$

as well as $N - L = L$ for even N allow us to arrive at a simple expression for the condition under which ATC will lead to a variance reduction vis-à-vis SRS. It is given by

$$
\begin{aligned}
\frac{E_M(SSW_Y)}{E_M(SSB_Y)} &= \frac{\beta_1^2 SSW_Z + L\sigma^2}{\beta_1^2 SSB_Z + (L-1)\sigma^2} \\
&> \frac{L}{L-1}.
\end{aligned} \tag{9}
$$

If β_1^2 takes a non-zero value, expression Eq. (9) can be further simplified to

$$
\frac{E_M(SSW_Y)}{E_M(SSB_Y)} = \frac{SSW_Z}{SSB_Z} > \frac{L}{L-1}. \tag{10}
$$

Hence, ATC is expected to perform better than SRS under a linear model provided the correlation between the variable of interest is non-zero and the ratio of the within to the between variation in the size variable is greater than $L/(L-1)$. If $\beta_1 = 0$, the variable of interest and the clustering variable are uncorrelated. It should be further noted that relaxing the i.i.d. assumption on the model error to the case of independence also leads to condition Eq. (10) in the presence of a non-zero correlation. Another interesting question relates to the consequences of applying ATC when the population model is given by

$$
\begin{aligned}
y_k &= c \cdot (z_k - \overline{Z})^2 + \epsilon_k, \\
\epsilon_k &\overset{i.i.d.}{\sim} G(0, \sigma_\epsilon^2).
\end{aligned} \tag{11}
$$

In this case, constructing clusters based on z_k will be inefficient and lead to a loss in precision vis-à-vis SRS as the units within a cluster will have very similar values of y_k such that SSB_Y dominates the total sum of squares. A simple remedy in this situation is to determine the cluster membership by applying the ATC method to the values of $a_k = (z_k - \overline{Z})^2$, as model Eq. (5) holds for the auxiliary variable a_k.

2.5. ATC under a model with domain effects

While the developments from the previous sections are based on a simple linear regression model, the condition applies as well to a model with domain-specific effects v_d

$$
\begin{aligned}
y_k &= \beta_0 + \beta_1 z_k + v_d + \varepsilon_k, \quad k \in U_d, \\
\varepsilon_k &\overset{i.i.d.}{\sim} G(0, \sigma^2)
\end{aligned} \tag{12}
$$

provided that the sampling design is a two stage design with the domains as strata on the first stage (planned domains) and within domains the ATC procedure is applied. The reason why this holds is that within a domain d, the domain-specific effect v_d in Eq. (12) is a constant and thus absorbed by the intercept β_0. Thus, model Eq. (12) reduces to model Eq. (5) within domains. Since the national mean is a convex combination of the stratum means under StrRS, applying ATC within domains will lead to a variance reduction for the national mean compared to SRS.

Now suppose that the model governing the population is indeed given by Eq. (12), but ATC is applied on the population level directly. This leads to changes for the relevant expectations needed to compute the sum of squares between and within as the cluster can be composed of units from different domains. Hence, the expected values are given by

$$
\begin{aligned}
E_M(y_k^2) &= \beta_0^2 + 2\beta_0\beta_1 z_k + 2\beta_0 v_k + \beta_1^2 z_k^2 \\
&\quad + 2\beta_1 z_k v_k + v_k^2 + \sigma^2, \\
E_M(\overline{Y}_h^2) &= \beta_0^2 + 2\beta_0\beta_1 \overline{Z}_h + 2\beta_0 \overline{V}_h + \beta_1^2 \overline{Z}_h^2 \\
&\quad + 2\beta_1 \overline{Z}_h \overline{V}_h + \overline{V}_h^2 + \sigma^2/N_h, \\
E_M(\overline{Y}^2) &= \beta_0^2 + 2\beta_0\beta_1 \overline{Z} + 2\beta_0 \overline{V} + \beta_1^2 \overline{Z}^2 \\
&\quad + 2\beta_1 \overline{ZV} + \overline{V}^2 + \sigma^2/N,
\end{aligned} \tag{13}
$$

where v_k denotes the domain-specific effect relevant for unit k and $\overline{V}$ and $\overline{V}_h$ refer to the population and cluster means of the domain-specific effects, respectively. Using Eq. (13) in connection with expressions for $E_M(SSB_Y)$ and $E_M(SSW_Y)$ yields:

$$
\begin{aligned}
E_M(SSB_Y) &\approx \beta_1^2 SSB_Z + (L-1)\sigma^2 \\
&\quad + \sum_{h=1}^{L} N_h \overline{V}_h^2 - N \cdot \overline{V}^2, \\
E_M(SSW_Y) &\approx \beta_1^2 SSW_z + L\sigma^2 \\
&\quad + \sum_{d=1}^{D} N_d v_d^2 - \sum_{h=1}^{L} N_h \overline{V}_h^2.
\end{aligned} \tag{14}
$$

Note that expressions Eq. (14) are approximations, since cross-product terms between the domain-specific effects and the clustering variable as well as those between the domain-specific effects and the individual error terms are ignored. These approximations can be motivated as in many applications the cross-product terms are negligible compared to the terms present in Eq. (14). Thus, ATC will be more precise than SRS if

$$\begin{aligned}\frac{E_M(SSW_Y)}{E_M(SSB_Y)} &\approx \frac{\beta_1^2 SSW_Z + SSW_V}{\beta_1^2 SSB_Z + SSB_V} \\ &> \frac{L}{L-1},\end{aligned} \tag{15}$$

where we use

$$SSW_V = \sum_{d=1}^{D} N_d v_d^2 - \sum_{h=1}^{L} N_h \overline{V}_h^2$$

and

$$SSB_V = \sum_{h=1}^{L} N_h \overline{V}_h^2 - N \cdot \overline{V}^2.$$

Hence, the domain-specific effects v_d play a similar role to the values of the size variable z_k.

3. Simulation study

3.1. Simulation set-up

In this section, we present results from a simulation study that compares the proposed ATC approach with other sampling designs which are known to be suitable for design-based estimation. In addition to studying the impact of the sampling designs on aggregate design-based estimates, we also analyse the influence of the designs on design- and model-based estimates for small domains. We consider a fixed and finite population comprising $N = 12000$ units from $D = 30$ domains. In our design-based simulation study, we repeat the process of drawing samples from a fixed population according to various sampling designs $R = 10000$-times. The population is chosen as one realisation of the following nested-error regression model

$$\begin{aligned}&y_k = 10 + x_{1k} + x_{2k} + v_d + \varepsilon_k, \quad k \in U_d, \\ &v_d \overset{i.i.d.}{\sim} N(0, 3^2), \quad \varepsilon_k \overset{i.i.d.}{\sim} N(0, 5^2).\end{aligned} \tag{16}$$

Following [11, Section 5.2], the values of the explanatory variables were generated as $x_{1k} \overset{i.i.d.}{\sim} U(1, 11)$ and $x_{2k} \overset{i.i.d.}{\sim} U(-5, 5)$, respectively. As estimators for the small area means we consider the Horvitz-Thompson (HT) estimator of the mean, a modified generalised regression (GREG) estimator, and the Battese-Harter-Fuller (BHF) estimator due to [15]. The HT estimator of a domain mean is given by $\hat{\mu}_{Y,d}^{HT} = N_d^{-1} \sum_{k \in S_d} w_k \cdot y_k$, where $w_k = \pi_k^{-1}$ denotes the design weight given by the inverse inclusion probability. We deliberately introduce model misspecification as the models fitted for both the GREG and the BHF estimator do not include x_{1k} among the covariates. The modified GREG estimator is defined as [4, Section 2.5]:

$$\hat{\mu}_{Y,d}^{GREG} = \hat{\mu}_{Y,d}^{HT} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_d^{HT})' \hat{\beta}_{WLS} \tag{17}$$

where $\bar{\boldsymbol{X}}_d$ and $\bar{\boldsymbol{x}}_d^{HT}$ denote the vector of the population mean and the HT estimator of the sample mean of $\boldsymbol{x}_k = (1, x_{2k})'$ in area d. Moreover, $\hat{\beta}_{WLS}$ refers to vector of regression coefficients obtained from regressing y_k on x_{2k} using weighted least squares with weights w_k. The BHF estimator is given by [4, Section 7.1]:

$$\begin{aligned}\hat{\mu}_{Y,d}^{BHF} = \hat{\gamma}_d (\bar{y}_d^{SRS} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_d^{SRS})' \hat{\beta}_{GLS}) \\ + (1 - \hat{\gamma}_d) \bar{\boldsymbol{X}}_d' \hat{\beta}_{GLS}\end{aligned} \tag{18}$$

where $\bar{y}_d^{SRS}$ and $\bar{\boldsymbol{x}}_d^{SRS}$ denote unweighted estimates of the mean and mean vectors of y_k and $\boldsymbol{x}_k$, respectively. Furthermore, $\hat{\beta}_{GLS}$ denotes the estimated regression vector that is obtained from regressing y_k on x_{2k} with random effects for the areas using generalised least squares. It can be seen that the modified GREG estimator corrects the HT estimator by an adjustment that depends on the difference between the vector of the population mean and the corresponding HT estimate of the sample mean for the auxiliary information. The model-based BHF estimator Eq. (18) can be considered as a weighted average between a survey regression estimator, $\bar{y}_d^{SRS} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_d^{SRS})' \hat{\beta}_{GLS}$, and the regression-synthetic component $\bar{\boldsymbol{X}}_d' \hat{\beta}_{GLS}$, where the weight $\hat{\gamma}_d$ increases with the sample size n_d [4, Section 7.1].

As estimators for the national mean we focus on the HT and GREG estimators, as the sample design is typically constructed to enable design-based estimates on the national level. The HT estimator for the national mean is given by

$$\hat{\mu}_Y^{HT} = N^{-1} \sum_{k \in S} w_k \cdot y_k,$$

while the GREG estimator follows as

$$\hat{\mu}_Y^{GREG} = \hat{\mu}_Y^{HT} + (\bar{\boldsymbol{X}} - \bar{\boldsymbol{x}}^{HT})' \hat{\beta}_{WLS},$$

where $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{x}}^{HT}$ denote the vector of the population mean and the HT estimator of the sample mean of $\boldsymbol{x}_k = (1, x_{2k})'$ respectively.

Table 2
Average absolute relative bias of the domain estimates

Estimator	$E(n_d)$	ATC	Cube-πps	Cube-SRS	Pivotal	Rejective	SRS	StrRS
HT	< 10	0.003	0.005	0.004	0.005	0.002	0.002	0.003
	10–30	0.001	0.003	0.002	0.003	0.002	0.001	0.002
	> 30	0.001	0.002	0.002	0.002	0.002	0.001	0.001
GREG	< 10	0.019	0.019	0.018	0.019	0.018	0.018	0.019
	10–30	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	> 30	0.001	0.001	0.001	0.002	0.001	0.001	0.001
BHF	< 10	0.061	0.132	0.061	0.132	0.062	0.061	0.061
	10–30	0.034	0.090	0.034	0.090	0.034	0.034	0.034
	> 30	0.013	0.085	0.012	0.086	0.013	0.013	0.012

To introduce variation in the domain sizes, we proceed in a similar fashion to [11, Section 5.2] and allocate the units to domains with probabilities proportional to $exp(q_d)$ where q_d is drawn as an independently and identically distributed random variable with a uniform distribution over the interval between 0 and 2.9. This results in domain sizes N_d varying between 57 and 1167 units. The sampling schemes are applied to the population as a whole, irrespective of the domain membership. Hence, the domain-specific sample sizes n_d are random variables and even sample sizes of zero are possible. We fix the total sample size to $n = 500$, such that the expected sample size for domain d follows as $E(n_d) = n \cdot N_d/N = 500 \cdot N_d/12000$. Thus our approach reflects the situation where the sample is designed to obtain national estimates of adequate precision, but where the need to produce reliable domain estimates is not addressed at the design stage. We focus on this unplanned domain case, because it is a situation frequently encountered in practice.

We apply a variety of sampling designs in order to compare our propose method with other sampling designs that are frequently used to obtain precise design-based estimates. The first two designs that are used in our study are SRS and the ATC approach described in Section 2.2, where the latter uses $z_k = x_{1k}$ as the auxiliary information at the design stage. Furthermore, we consider StrRS with 10 strata, where the stratum membership is determined by the deciles of x_{1k} and 50 units are taken from each stratum. It should be noted that this allocation of sample sizes to the strata is very close to the optimal Neyman-Tschuprow allocation for the population constructed by model Eq. (16). Additionally, we apply a rejective sampling procedure (Rejective) as described by [16] using SRS as the initial sampling procedure. In our simulation study, the SRS sample was rejected as long as $(\bar{x_1}^{SRS} - \bar{X}_1)^2 \cdot Var(\bar{x_1}^{SRS})^{-1} < 0.5^2$, where $\bar{X}_1$ and $\bar{x_1}^{SRS}$ denote the population mean and SRS estimate of x_{1k}, respectively and $Var(\bar{x_1}^{SRS})$ refers to the variance of $\bar{x_1}^{SRS}$. This corresponds to an empirical rejection rate in our simulation study of 79.16%. Furthermore, we also study pivotal sampling (Pivotal) introduced by [17], where the inclusion probabilities are chosen to be proportional to x_{1k}, i.e. $\pi_k \propto x_{1k} \forall k$. Finally, we apply the cube method due to [18] with balancing constraints on the population size and the total of x_{1k}. In our study, we consider two variants of the cube method: using (i) equal inclusion probabilities, i.e. $\pi_k = nN^{-1} \forall k$ (Cube-EPSEM) and (ii) $\pi_k \propto x_{1k} \forall k$ (Cube-πps).

It can be seen that x_{1k} is incorporated as auxiliary information at the design stage for all sampling designs except SRS, i.e. we set $z_k = x_{1k}$ Doing so permits potential variance reductions for design-based and model-assisted estimators. Hence, the simulation study facilitates a comparison of the ATC approach with other popular sampling designs that are known to be suitable in a design-based framework. Moreover, the consequences of applying a particular design on model-based small area estimates can be studied. Note that due to the construction of the size variable $z_k = x_{1k}$, similar domain-specific sample sizes n_d result for the different sampling designs. Nevertheless there are important differences among the sampling designs as the Cube-πps and the pivotal methods draw samples with probabilities proportional to size. Since the size variable influences the variable of interest but is not included among the covariates, the issue of informative sampling arises for these two sampling designs.

3.2. Results

The simulation results for domain estimates in terms of the average absolute relative bias (AARB) are summarised in Table 2. We average the results according to the expected sample size in the domains, $E(n_d)$. We clearly see that the Monte-Carlo biases of the HT estimator are very close to zero under any sampling de-

Table 3
Average relative root mean squared error of the domain estimates

Estimator	$E(n_d)$	ATC	Cube-πps	Cube-SRS	Pivotal	Rejective	SRS	StrRS
HT	< 10	0.470	0.509	0.468	0.507	0.472	0.470	0.470
	10–30	0.261	0.281	0.261	0.282	0.260	0.260	0.260
	> 30	0.163	0.177	0.163	0.177	0.163	0.163	0.163
GREG	< 10	0.214	0.265	0.212	0.267	0.214	0.214	0.213
	10–30	0.098	0.122	0.098	0.123	0.097	0.098	0.098
	> 30	0.060	0.075	0.060	0.076	0.060	0.061	0.060
BHF	< 10	0.118	0.166	0.118	0.167	0.118	0.118	0.118
	10–30	0.081	0.116	0.081	0.116	0.081	0.081	0.081
	> 30	0.054	0.099	0.053	0.100	0.054	0.054	0.054

sign and sample size. This finding is expected as the HT estimator is known to be design-unbiased. For the modified GREG estimator, we observe small Monte-Carlo biases in the group of very small domains with $E(n_d) < 10$. These biases vanish as the sample size increases. This finding is also consistent with the theory as the modified GREG estimator is asymptotically unbiased. In the case of the model-based BHF estimator, we observe highly different results depending on the sampling mechanism used. For designs with equal inclusion probabilities, the absolute biases decrease as the sample size increases and reach values of 1.2 to 1.3 per cent in the group of the largest domains. The reason for this is that the BHF estimator is biased when conditioning on the random effects v_d, which is precisely what is done in a fixed finite population setting. This conditional bias is supposed to decrease as the sample size increases, because the weight $\hat{\gamma}_d$ on the survey regression component approaches 1. It should be noted, however, that much larger values of the AARB for the BHF estimator result under the Cube-πps and the pivotal sampling designs, respectively. Under both designs the sampling mechanism is informative, which causes biased estimates when using the BHF method.

In order to assess the precision of the domain estimates, we consider the average relative root mean squared error (ARRMSE) over domains reported in Table 3. The results show an interesting pattern for any estimation method and domain size. On the one hand, there is the group of equal probability sampling designs that yield very similar results for a particular choice of an estimator and domain size. On the other hand, the Cube-πps and the pivotal designs also yield very similar results for given combinations of an estimator and domain size, but their ARRMSEs are larger than for the equal probability sampling designs. In case of the HT and modified GREG estimators, this finding corresponds to a larger variance as compared to the other designs, as these estimators did not suffer from biases (Table 2). An explanation for this behaviour of the HT estimator is the presence of the intercept term in the data generating process Eq. (16), which causes the inefficiencies of the HT estimator based on inclusion probabilities $\pi_k \propto x_{1k}$ [11, p. 227]. In order to explain the performance of the modified GREG estimator, we may note that the regression vector $\hat{\beta}_{WLS}$ estimated by weighted least squares with weights w_k under an equal probability sampling design is identical to the solution that would have been obtained by ordinary least squares. When the sampling design uses unequal inclusion probabilities, however, using weighted least squares will lead to an increase of the variance vis-à-vis ordinary least squares. Indeed, we observe the largest Monte-Carlo variances of $\hat{\beta}_{WLS}$ when we use Cube-πps and the pivotal designs (not reported here). Regarding the model-based BHF estimator, the higher values for the ARRMSE under Cube-πps and the pivotal designs are due to informative sampling. Furthermore, a comparison of the three different estimation methods in terms of the ARRMSE clearly indicates advantages for the BHF estimator, which yields the best results for all domain sizes. This finding does not come as a surprise, since our simulation study contains small unplanned domains, where design-based estimation methods are not suitable.

The results for the national estimates are shown in Table 4, where RBias refers to the relative bias of the national estimates, while RRMSE indicates the relative root mean squared error and ACR denotes the average confidence interval coverage rate. All numerical entries in Table 4 are rounded to three decimal places. Regarding the biases, we see that all combinations of an estimator and a design yield unbiased estimates. A closer look at the precision of the national estimates reveals that the equal probability sampling designs which use auxiliary information at the design stage yield the best results for both estimators. The RRMSE under these designs is about 10 per cent

Table 4
Results for the national estimates

Estimator	Design	RBias	RRMSE	ACR
HT	SRS	0	0.019	0.951
	Rejective	0	0.017	–
	Pivotal	−0.002	0.028	–
	StrRS	0	0.017	0.950
	Cube-πps	0	0.021	0.942
	Cube-SRS	0	0.017	0.948
	ATC	0	0.017	0.947
GREG	SRS	0	0.017	0.949
	Rejective	0	0.015	–
	Pivotal	0.001	0.021	–
	StrRS	0	0.015	0.951
	Cube-πps	0	0.018	–
	Cube-SRS	0	0.015	–
	ATC	0	0.015	0.945

smaller than the RRMSE under SRS for a given estimator. Hence, incorporating auxiliary information at the design stage helps to achieve a variance reduction as compared to SRS. Furthermore, we see that designs using inclusion probabilities proportional to x_{1k} produce estimates with a larger RRMSE for both estimators. Unlike the results for the domain estimates, however, the Cube-πps method yields estimates that are much more precise than pivotal sampling. An explanation for this finding is that the Cube-πps method includes a calibration constraint for the total size of the population. Moreover, the results show advantages of the GREG estimator as compared to HT estimator for a given sampling design. This is simply due to the fact that the GREG estimator incorporates additional information about x_{2k} which could not be used at the design stage. Finally, the average confidence interval coverage rates are very close to the nominal rate of 95 per cent for all combinations of estimator and designs. Note that we did not compute variance estimates under the rejective and pivotal sampling procedures. Moreover, when the sample was selected by the cube method, we only produced variance estimates for the HT estimator using the residual technique developed by [19].

4. Concluding remarks

We have proposed a novel allocation mechanism of ultimate sampling units to clusters, which in connection with single stage cluster sampling allows realising variance reductions for design-based estimation methods versus SRS. Moreover, this allocation mechanism yields equal inclusion probabilities and therefore avoids the issue of informative sampling. Thus, our approach does not distort the properties of model-based estimation procedures. Therefore, our method is well-suited for modern surveys, where design-based estimates are produced at aggregate levels and at the same time model-based estimates are published for domains with small sample sizes. Further advantages of our proposal are that it is both very simple to implement and, perhaps even more importantly, also very easy to communicate to the public.

We compared our proposed method against a number of alternative sampling designs aiming at variance reduction for design-based estimation methods in a simulation study under a misspecified model. The results of this study showed very similar results of the ATC method, the cube method with equal inclusion probabilities, StrRS where the strata are defined by the deciles of the auxiliary variable and a rejective sampling procedure. All of these methods make use of the auxiliary information at the design stage and use equal (initial) inclusion probabilities. Sampling designs based on sampling with probabilities proportional to size were shown to be less efficient for the estimation of national estimates and led to biased model-based small domain estimates due to informative sampling.

In comparison to the rejective sampling procedure, our approach allows fixing the inclusion probabilities in advance and it permits the use of simple unbiased design-based variance estimators. Furthermore, our sampling procedure is a SRS of clusters and, thus, very fast even for large populations. This is a distinct advantage over the cube method, which can be time-consuming for large populations. Moreover, using ATC we avoid the need for approximations to second-order inclusion probabilities.

In contrast to sampling with probabilities proportional to size, our proposal is more robust with respect to a misspecification of the implicitly assumed model. This is highlighted by the results of the simulation study, where designs based on sampling with probabilities proportional to size led to inefficiencies owing to the presence of an intercept term in the population model. Additionally, sampling with probabilities proportional to size is clearly suboptimal for HT estimation in situations where the size variable is negatively correlated with the variable of interest.

Alternatively, one could consider StrRS approaches towards optimal model-based stratification for the GREG estimator, which have been discussed in Section 12.4 of [1]. However, they require knowledge about the error structure of the assisting regression model and a rule to determine the stratum member-

ship. Thus, the survey planner needs a comprehensive knowledge about the model, which is by far more demanding than knowing the values of some size variable.

Future research may focus on a generalization of the ATC approach to account for multiple auxiliary variables simultaneously when constructing antithetic clusters. One option in this regard could be to apply a principal component analysis to the standardized matrix of auxiliary information and to base the clustering on the values of an appropriate distance of the principal components from their origin.

Acknowledgments

The author is very grateful to the associate editor and two anonymous referees for their comments and suggestions, which helped to improve the paper substantially.

References

[1] Särndal CE, Swensson B, Wretman J. Model assisted survey sampling. New York: Springer; 1992.
[2] Pfeffermann D. New important developments in small area estimation. Statistical Science. 2013; 28(1): 40-68.
[3] Jiang J, Lahiri P. Mixed model prediction and small area estimation. Test. 2006; 15(1): 1-96.
[4] Rao JNK, Molina I. Small area estimation. Hoboken: John Wiley & Sons, Inc; 2015.
[5] Pfeffermann D, Sverchkov M. Inference under informative sampling. In: Pfeffermann D, Rao CR, eds. Handbook of statistics vol 29B: Sample Surveys: Inference and Analysis. New York: Elsevier; 2009. p. 455-487.
[6] Hidiroglou MA, Lavallee P. Sampling and estimation in business surveys. In: Pfeffermann D, Rao CR, eds. Handbook of statistics vol 29A: Sample Surveys: Design, Methods, and Applications. New York: Elsevier; 2009. p. 441-470.
[7] Little RJ. Calibrated Bayes, an alternative inferential paradigm for official statistics. Journal of Official Statistics. 2012; 28(3): 309-334.
[8] Little RJ. Calibrated Bayes, an alternative inferential paradigm for official statistics in the era of big data. Statistical Journal of the IAOS. 2015; 31(4): 555-563.
[9] Gabler S, Ganninger M, Münnich R. Optimal allocation of the sample size to strata under box constraints. Metrika. 2012; 75(2): 151-161.
[10] Rizzo ML. Statistical computing with R. Boca Raton: CRC Press; 2007.
[11] Lehtonen R, Veijanen A. Design-based methods of estimation for domains and small areas. In: Pfeffermann D, Rao CR, eds. Handbook of statistics vol 29B: Sample Surveys: Inference and Analysis. New York: Elsevier; 2009. p. 219-249.
[12] Valliant R, Dorfman AH, Royall RM. Finite population sampling and inference: a prediction approach. New York: Wiley; 2000.
[13] Lohr S. Sampling: Design and Analysis. Pacific Grove: Duxbury Press; 1999.
[14] Hardy GH, Littlewood JE, Polya G. Inequalities. Cambridge: Cambridge university press; 1952.
[15] Battese GE, Harter RM, Fuller WA. An error component model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association. 1988; 83(401): 28-36.
[16] Fuller WA. Some design properties of a rejective sampling procedure. Biometrika. 2009; 96(4): 933-944.
[17] Deville JC, Tillé Y. Unequal probability sampling without replacement through a splitting method. Biometrika. 1998; 85(1): 89-101.
[18] Deville JC, Tillé Y. Efficient balanced sampling: The cube method. Biometrika. 2004; 91(4): 893-912.
[19] Deville JC, Tillé Y. Variance approximation under balanced sampling. Journal of Statistical Planning and Inference. 2005; 128(2): 569-591.