University of Southampton

Faculty of Social Sciences
Department of Social Statistics and Demography

Balanced two-stage equal probability sampling

by

Shoaib Ali

M.Phil.

ORCiD: 0000-0001-6221-9962

A thesis for the degree of
Doctor of Philosophy

November 2022

University of Southampton

Abstract

Faculty of Social Sciences
Department of Social Statistics and Demography

Doctor of Philosophy

Balanced two-stage equal probability sampling

by Shoaib Ali

For two-stage sampling, equal probability sampling method (epsem) by πps-SRS is commonly used in practice. It also provides practical means for controlling cost and fieldwork allocation. In two-stage epsem, same number of elements are selected by SRS from each sampled PSU. As an alternative to this convention for two-stage epsem, one can also select same number of equal-sized sub-clusters by SRS from each sampled PSU. Furthermore, although HT-estimator is design unbiased, when a set of auxiliary variables is known, generalized regression (or GREG) estimator is also commonly used in practice. A comparison of sampling strategies involving two-stage epsem by πps-SRS may provide a useful insights from practical viewpoint. Therefore, four sampling strategies involving two-stage epsem are compared under a two-level regression model which is a intuitive choice for two-stage sampling. A simulation study is also conducted to support theoretical comparison of the sampling strategies.

Cube method for balanced sampling was proposed for selection of PSU's. In two-stage sampling design, cube method can be used when auxiliary variables are know at either PSU-level or at element level. Cube method aims to selected balanced samples with fixed first-order inclusion probabilities. It consists of two-phases: flight- and landing-phase. When its landing-phase is invoked, samples are not exactly balanced. A sampling procedure is proposed which aims to improve landing-phase of the cube method when it is not exactly balanced. In addition, a methodology for the estimation of sampling variance under balanced sampling is also proposed which found to be better than a variance estimator in literature. Simulation studies are conducted to investigate the performance of proposed sampling procedure and variance estimators.

When location data of sampling units is available, it is emphasized in literature to select spatially balanced samples as study variables are expected to have positive spatial autocorrelation. Since there are many spatially balanced sampling methods available, a comparative study of different spatially balanced sampling methods is conducted under a spatial super-population model with varying level of spatial autocorrelation. When both auxiliary and spatial variables are known, doubly balanced sampling is advocated in literature. Spatial or doubly balanced sampling can be used in two-stage sampling depending on availability of spatial and auxiliary variables. Some variables of the study population may have negative spatial autocorrelation, as two-stage designs are often used for socio-economic surveys which include a variety of study variables. Four spatial sampling schemes are suggested to select spatially balanced samples when there are also some variables with negative spatial autocorrelation in the population. A variance estimation methodology is also suggested under the spatially balanced and doubly balanced sampling methods. Simulation studies are conducted to investigate the performance of proposed spatial sampling schemes and variance estimators.

Contents

List of Figures

List of Tables

Declaration of Authorship

Name: Shoaib Ali
Title of thesis: Balanced two-stage equal probability sampling

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.
I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signature: .... Date: September 16, 2022

Acknowledgements

I would like to express my great gratitude to my supervisor Professor Li-Chun Zhang and co-supervisors Professor Nikos Tzavidis and Doctor Angela Luna Hernandez. They have been very supportive and patient throughout my PhD studies. I would also like to thank Professor Paul Smith and Doctor Yves Berger for their constructive and insightful comments during my first-year review and second-year review for the PhD upgrade.

I am specially thankful to SCDTP (previously know as ESRC-DTC) in University of Southampton for providing PhD studentship funding. A lot of thanks to my PhD and office colleagues for their constant encouragement and support. I am very thankful to my parents, siblings and friends from Pakistan for their prayers and support.

Chapter 1

General introduction and literature review

Sample surveys are conducted to collect data about the finite population, a set of elements under study. In a sample survey, a subset of elements from the finite population is selected and required data is collected from the selected elements. The sample survey is planned such that maximum information about the finite population is elicited using limited resources. A sample survey is deemed to be cheaper than complete enumeration of the finite population (or a census) in terms of cost, time and manpower; and it may even be more accurate. The data collected from the sample surveys are used to estimate unknown quantities of the finite population, for example: totals, averages and proportions etc. Sample surveys are widely used by national statistics offices for data collection about human population, businesses, agriculture, forestry, natural resources, environment and many other fields of study in academic research to produce statistics at local and national levels. These statistics play a vital role in policy and decision making for the related departments which ultimately contribute in addressing problems of the society at local, national and global levels.

According to Hansen and Hurwitz (1943), the history of random sampling dates back to early eighteenth century when Daniel Bernoulli (1700-1782) presented theory of random sampling from a population; a century later, Siméon Denis Poisson (1781-1840) investigated gain of using stratification in sampling; Wilhelm Lexis (1837-1914) introduced sampling of clusters of elements. According to Brewer and Gregoire (2009), the idea of sampling from finite population is reckon to be first presented by Anders Nicolai Kiaer (1838-1919) in 1890s; Kiær (1896) argued that 'partial investigation' (or sampling) instead of complete enumeration of the population can be useful for data collection, and proposed a purposive sampling method, named as 'representative method', which could not get enough support from other statisticians at that time. Supporting Kiaer's work, Bowley (1926) presented a report in ISI (International Statistical Institute) commission which summarized the adaptation of work from Bernoulli and Poisson to the theory of sampling from a finite population (Hansen and Hurwitz, 1943). Neyman (1934) published the key article in survey sampling theory. Neyman was later invited to give lectures in Bureau of the Census (U.S.A) (Brewer and Gregoire, 2009). In 1940s, survey sampling had become popular in a wider community of statisticians. Among many others, some key text books about survey sampling theory are given by Hansen et al. (1953a,b), Kish (1965), Cochran (1977) and Särndal et al. (1992).

This chapter describes some basic concepts in survey sampling and presents a review of literature related to problems considered in this thesis. The arrangement of this chapter is as follow. Section 1.1 describes some basic concepts and notations commonly used in survey sampling. Section 1.2 describes cluster sampling, single-stage and two-stage designs in particular. Sections 1.3 and 1.4 describe use of known auxiliary data in generalized regression estimation and in balanced sampling respectively. Section 1.5 describes spatial dependence in the finite populations and spatially balanced sampling, it also reviews related random sampling methods from literature. Section 1.6 describes problem of variance estimation and reviews some variance approximations under balanced and spatially balanced sampling designs. Section 1.7 describes role of super-population model in the theory of survey sampling where the finite population is considered as a realization of an infinite super-population. Finally, Section 1.8 gives an outline of the problems considered in this thesis.

1.1 Basic concepts and notations

In survey sampling, a *finite population* (or target population) is a set of elements under study about which data is required; for example, people, households or agricultural farms in a city or country, or business enterprises in an industry etc. A *sample* is a subset of the finite population which can be categorised as *random sample* or *non-random sample*.

A *study variable* (or survey variable) is a characteristic or value associated with each element of the finite population; for example, status of persons belonging to labour force, weekly expenditure of households, wheat production of farms, annual revenue of business. In a sample survey, often data about many study variables are collected. A *finite* *population parameter* (or finite population quantity) is a function of study variable; it represents characteristic of the finite population which is required to be estimated; for example, proportion of persons in the labour force, total weekly household expenditures, total wheat production, or total revenue of businesses etc.

A *sampling frame* is any material or device which delimit, identify and provide observational access to or establishes contact with elements of the finite population (Särndal et al., 1992, p. 9). Usually, a sampling frame contains list of elements of the finite population or clusters of elements, called *sampling units*; it may also contain additional data such as auxiliary data to be used at sampling or estimation stages. The sample is selected from the sampling frame.

1.1.1 Probability sampling

Sampling is the process of selecting a sample. In *probability sampling* (or random sampling), a sample is selected following the fundamentals of probability mechanism in which each element of the finite population has a non-zero probability of being selected in the sample. In *non-probability sampling* (or non-random sampling), a sample is selected by a subjective approach. Some common non-probability sampling method are convenient sampling, quota sampling, judgemental sampling etc. Probability sampling eliminates selection biases and it is acceptable to the public due to its objectivity (Särndal et al., 1992, p. 9). In practice, probability sampling is widely preferred because it provides probabilistic properties of the sample estimates such as measures of validity and reliability. Some basic and common probability sampling methods are simple random sampling, systematic random sampling, stratified random sampling and cluster sampling. In a sample survey, random sampling can range from simple to a complex one which might be mixture of more than one basic probability sampling methods.

For selection of a random sample, two types of selection schemes are *draw sequential* and *list sequential*. Draw sequential scheme consists of a series of randomised experiments, called *draws*; each draw selects a sampling units for the sample. List sequential scheme is applied to the list of sampling units; it carries out a randomised experiment for each sampling unit in the list which results into selection or rejection of the sampling units; the sample selection may complete before the end of the list. A random sample can be selected by *with-replacement* (WR) or *without-replacement* (WOR) sampling. In with-replacement sampling, one sampling unit can be selected more than once in the sample, because when a sampling unit is selected in the sample it is replace back in the sampling frame before the next selection is performed. In without-replacement sampling, one sampling unit can be selected in the sample only once, because when a sampling unit is selected in the sample it is not replaced back in the sampling frame. Usually, with-replacement sampling is considered to be less efficient between these two; however, it has simpler formulas for sampling variance of sample estimates (see Section 1.6) which sometimes makes it useful for more complex sampling designs (Cochran, 1977, p. 18).

Suppose that the finite population consists of N elements and $U = \{1, 2, 3, ..., N\}$ is set of labels associated with the population elements. Let y be the survey variable and $y_1, ..., y_N$ denote unknown values of the survey variable associated with the N population elements. Let finite population total of the survey variable is required to be estimated, which is defined as

$$Y = \sum_{i \in U} y_i$$

where y_i denotes value of the survey variable associated with the ith population element.

Let a random sample of n elements, denoted by s, is selected from the finite population U. Let Ω denotes set of all possible samples of size n, called *sample space*. Let $p(s)$ denotes set of selection probabilities associated with each sample in the sample space Ω, where $p(s)$ is called *sampling distribution*. Let $I_{(i \in s)}$ denotes an indicator (or binary) variable which take value 1 if ith element is in the sample s, otherwise 0, where $I_{(i \in s)}$ is a random variable called *sample membership indicator*. The probability of being included (or selected) in the sample for ith element is called *first-order inclusion probability*, defined as

$$\pi_i = \sum_{s \in \Omega} I_{(i \in s)} p(s)$$

In the same way, the probability of being included (or selected) in the sample for (i, j)th pair of elements is called *second-order inclusion probability*, defined as

$$\pi_{ij} = \sum_{s \in \Omega} I_{(i \in s)} I_{(j \in s)} p(s)$$

where $i \neq j \in U$.

1.1.2 Sampling design and sampling strategy

Sampling design assigns selection probability to each sample s in the sample space Ω under a specified sampling method; it consists of sample space and probability distribution, denoted by $\Delta_p = \{\Omega, p(s) | s \in \Omega\}$. In modern literature, it is also common to call the probability distribution $p(s)$ the sampling design (Särndal et al., 1992, p. 29). There are only few sampling methods for which the sampling designs are completely known, for example, simple random sampling and equal probability systematic sampling (see Section 1.1.4). For many advance sampling methods, the implied sampling design is not known, for example, sampling methods for probability proportional to size sampling (see Section 1.1.5). Even when the sampling design is not completely known, knowledge about first- and second-order inclusion probabilities plays the most important role in the estimation of finite population parameters. When first- and second-order inclusion probabilities are strictly positive, i.e. $\pi_i > 0$ and $\pi_{ij} > 0$ for all $i \neq j \in U$, the sampling design is called *measurable* (Särndal et al., 1992, p. 33). A sampling design is *fixed-sized*, when the sample size n is fixed for all the samples under the sampling design (Särndal et al., 1992, p. 38). For some sampling methods, the sample size is not same for all the samples, for example, Bernoulli sampling (Särndal et al., 1992, p. 26) and Poisson sampling (see Section 1.1.4).

After the sample has been selected and observed, the *estimate* of the required finite population parameter is calculated based on the sample values using an *estimator*, which is a function of the study variable. Statistical properties of an estimator are studied under the sampling design (or sampling distribution), for example, expectation and variance of the estimator. Let $\hat{Y}$ denotes an arbitrary estimator for the finite population total Y; some specific estimators for the finite population total will be described later in this chapter.

A *sampling strategy* is combination of a sampling design and an estimator $\{\Delta_p, \hat{Y}\}$ (Särndal et al., 1992, p. 30). Relative efficiency of the sampling strategy (with respect to another sampling strategy) is derived based on properties of the estimator under the sampling design in the strategy, for example, expectation or mean squared error (MSE) of the estimator. An estimator is unbiased or *design-unbiased* if its expectation under the sampling design is equal to the finite population parameter being estimated, that is, $\hat{Y}$ is unbiased estimator of Y if $E_p(\hat{Y}) = Y$, where E_p denotes expectation with respect to sampling distribution $p(s)$. The MSE of the estimator is defined as $\text{MSE}(\hat{Y}) = E_p(\hat{Y} - Y)^2$. When an estimator is unbiased the MSE and variance of the estimator are equivalent, that is, $V_p(\hat{Y}) = E_p\big(\hat{Y} - E_p(\hat{Y})\big)^2 = E_p(\hat{Y} - Y)^2 = \text{MSE}(\hat{Y})$, where V_p denotes variance with respect to sampling distribution $p(s)$. The square root of the sampling variance $[V(\hat{Y})]^{1/2}$ is called *standard error* of the estimator $\hat{Y}$. An estimator with smaller MSE is preferred. An estimator is *consistent* when it converge to its parameter in probability. According to Chebyshev Inequality, it holds when variance of estimator tends to zero (Arnab, 2017, p. 78). These criteria are used in order to investigate multiple sampling strategies for estimation of the required finite population parameters. All these properties assumes selection of all possible samples under the given sampling design and calculation of the estimates for all the samples which is a *hypothetical* process. It is not performed in practice rather the values of MSE or variance are also estimated based on one sample.

In addition to a point estimate, often *confidence interval* (C.I.) is also constructed for a population parameter; it gives a range of two values which expect to contain finite population parameter with some fixed confidence level, as it is often done in standard statistical inference. The C.I.'s are constructed using normal distribution approximation for the estimator based on *central limit theorem* where it is is assumed that estimators of the finite population quantities follow a normal distribution for medium and large sample sizes, see (Cochran, 1977, p. 11), (Kish, 1965, p. 14) and (Särndal et al., 1992, p. 56). The $(1-\alpha)100$ percent confidence limits for the estimator $\hat{Y}$ can be computed as follows

$$(1-\alpha)100\% \text{ C.I. } = \left(\hat{Y} - mse(\hat{Y})z_{(\frac{\alpha}{2})},\ \ \hat{Y} + mse(\hat{Y})z_{(\frac{\alpha}{2})}\right)$$

where $mse(\hat{Y})$ is sample estimate of MSE$(\hat{Y})$ since it is also an unknown population quantity and $z_{(\frac{\alpha}{2})}$ is value of standard normal variate at tail probability $\frac{\alpha}{2}$. Sometimes, approximation to the normal distribution is weak due to small sample size, then confidence limits are computed based on the assumption that estimator follows a Student's t-distribution with $(n-1)$ degrees of freedom, given by

$$(1-\alpha)100\% \text{ C.I. } = \left(\hat{Y} - mse(\hat{Y})t_{(\frac{\alpha}{2},n-1)},\ \ \hat{Y} + mse(\hat{Y})t_{\frac{\alpha}{2},(n-1)}\right)$$

where $t_{(\frac{\alpha}{2},n-1)}$ value of variable under t-distribution at probability $\frac{\alpha}{2}$ with degree of freedom $(n-1)$.

1.1.3 Horvitz-Thompson (HT) estimator

Horvitz and Thompson (1952) proposed HT-estimator for population total Y, which is sum of the values for units in the sample weighted by inverse of corresponding inclusion probabilities of the units. Let HT-estimator of Y is denoted by $\hat{Y}_{HT}$, the mathematical expression for the HT-estimator is given by

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} \tag{1.1}$$

where s is random sample of size n. Provided that first-order inclusion probabilities π_i's are known or can be calculated exactly, the HT-estimator is design-unbiased, i.e.,

$$E_p(\hat{Y}_{HT}) = \sum_{s \in \Omega} \hat{Y}_{HT}(s)p(s) = Y$$

where E_p denotes expectation with respect to sampling distribution $p(s)$ and $\hat{Y}_{HT}(s)$ is the HT-estimator based on sample s. From Horvitz and Thompson (1952), sampling variance of the HT-estimator in Eq. (1.1) is given by

$$V_p(\hat{Y}_{HT}) = \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_i) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \tag{1.2}$$

where V_p denotes sampling variance with respect to sampling distribution $p(s)$. When the sample size is fixed, (Sen, 1953; Yates and Grundy, 1953) gave an alternative mathematical expression for the sampling variance $V_p(\hat{Y}_{HT})$, given by

$$V_p(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_i) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{1.3}$$

The estimation of sampling variance of the HT-estimator will be described later in Section 1.6 of this chapter.

1.1.4 Some basic sampling designs

Simple random sampling (SRS)

In SRS, all the sampling units have equal probability of being included in the sample, and all samples of fixed size n has the equal probability of being selected from the sample space. SRS is one of the basic and simplest methods of random sampling. SRS without-replacement is usually considered as reference point when discussing other sampling designs. In many complex survey designs, SRS is often involved at some stage. For example, SRS can be used in some or all the strata under stratified sampling; it is often used in the second-stage of two-stage sampling (see Section 1.2.2), while probability proportion to size sampling (see Section 1.1.5) is often used at the first-stage.

Sampling design for the SRS without-replacement is known. The sample space consists of $\binom{N}{n}$ distinct samples of fixed sample size n; each sample has equal probability of selection, given by $p(s) = \binom{N}{n}^{-1}$ for all s of size n (Särndal et al., 1992, p. 27). When the sampling design is know, one can compute inclusion probabilities of order one, two, up to order n

(in fact, n-order inclusion probability is probability of selection of the sample of size n). In this case, first- and second-order inclusion probabilities are given by

$$\pi_i \equiv \frac{n}{N} \text{ and } \pi_{ij} \equiv \frac{n(n-1)}{N(N-1)}$$

respectively (Särndal et al., 1992, p. 31). The HT-estimator of population total Y under SRS is given by

$$\hat{Y}_{HT}^{SRS} = \frac{N}{n} \sum_{i \in s} y_i$$

which is equivalent to *expansion estimator* $\hat{Y} = N\bar{y}$, where $\bar{y} = n^{-1} \sum_{i \in s} y_i$ is sample mean of y. The sampling variance of $\hat{Y}_{HT}^{SRS}$ is given by

$$V_{WOR}(\hat{Y}_{HT}^{SRS}) = N^2 \left(\frac{1-f}{n} \right) S_y^2,$$

where $f = \frac{n}{N}$ which is known as *sampling fraction*, $S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{Y})^2$ is population variance and $\bar{Y} = \frac{1}{N} \sum_{i \in U} y_i$ is population mean of variable y (Särndal et al., 1992, p. 46). The factor $(1-f)/n$ in the sampling variance above is known as *finite population correction (fpc)* factor. The *fpc* factor is ignored when sampling fraction f is small; according to (Cochran, 1977, p. 25), *fpc* factor is ignored when f is less than 5 percent.

In case of with-replacement sampling, sample containing n sampling units (including multiple selections of sampling units) is referred as *ordered sample*; and it is distinguished from distinct set of sampling units in the ordered sample, which referred as *set-sample* (Särndal et al., 1992, p. 49). The sample space consists of N^n different ordered samples and the sampling distribution is given by $p(s) = 1/N^n$ for all s (ordered sample) of size n. The selection probability of a sampling unit i ($i \in U$) at each draw (or experiment) is $p_i \equiv 1/N$. Following Hansen and Hurwitz (1943), when $\pi_i = np_i = n/N$, the HT-estimator of Y (using the ordered sample s) becomes Hansen-Hurwitz (HH) estimator and has same expression as under SRS without-replacement; though its sampling sampling variance is different, given by

$$V_{WR}(\hat{Y}_{HT}^{SRS}) = N^2 \left(\frac{N-1}{Nn} \right) S_y^2 = \left(\frac{N-1}{N-n} \right) \times V_{WOR}(\hat{Y}_{HT}^{SRS})$$

where $(N-1)(N-n)^{-1} > 1$ (for $n > 1$), which suggest that SRS with-replacement has $(N-1)(N-n)^{-1}$ times larger variance than SRS without-replacement (Cochran, 1977, p. 30). When sampling fraction f is very small, the two sampling variances (under SRS WR and WOR) are roughly same (Särndal et al., 1992, p. 73).

Systematic sampling (SYS)

In equal probability SYS, a sample of size n is selected such that first sampling unit is selected at random with equal probability from the first k units, where $k = N/n$, and every kth unit thereafter. In this way first unit determines the whole sample. This type of systematic sampling is also called an *every kth* systematic sampling (Cochran, 1946, 1977, p. 205). For this sampling method, sampling design is also known; there are k possible samples of size n in the sample space with uniform sampling distribution $p(s) \equiv 1/k$. First order inclusion probabilities are equal for all sampling units, given by $\pi_i \equiv 1/k$. Second order inclusion probability for pairs of sampling units (i, j) which belongs to the same sample is $\pi_{ij} = 1/k$, rest of the pairs have null joint inclusion probabilities. Under equal probability SYS, the HT-estimator for population total Y is given by

$$\hat{Y}_{HT}^{SYS} = k \sum_{i \in s} y_i$$

and its sampling variance is given by

$$V(\hat{Y}_{HT}^{SYS}) = k \sum_{r=1}^{k} (\hat{Y}_r - \bar{Y}_k)^2$$

where $\hat{Y}_r$ is total of y-values based on rth sample $(r = 1, ..., k)$, and $\bar{Y}_k = Y/k$ (Särndal et al., 1992, p. 76). When k is not an integer, it is rounded off to an integer, and some samples may vary in size by one unit. For such situations, different modifications have been proposed in literature including *circular systematic sampling* (Lahiri, 1951) and *fractional interval method*, see (Cochran, 1977, p. 206) and (Särndal et al., 1992, p. 77) for more details. Unbiased estimation of sampling variance is not possible under systematic sampling, because some second-order inclusion probabilities are zero.

1.1.5 Sampling with probability proportional to size

In probability proportional to size sampling, sampling units are selected with unequal inclusion probabilities which are proportional to a known *size variable*. The size variable often represents actual sizes of the sampling units, for instance, number of population elements in the clusters in cluster sampling (see Section 1.2); it may also be a measure of size which is highly correlated with the study variable (Cochran, 1977, p. 252). If z denotes a size variable which is known for each sampling unit, probability proportional to size sampling is expected to be more efficient than equal probability sampling when values of $\frac{y_i}{z_i}$ are tend to be constant for $i \in U$ (Särndal et al., 1992, p. 87). Usually, probability proportional to size samplings with-replacement and without-replacement are termed as '*pps* sampling' and 'πps sampling' respectively.

Originally, Hansen and Hurwitz (1943) introduced the idea of probability proportional to size sampling for *pps* sampling of clusters with unequal sizes (see Section 1.2). For given size variable z_i $(i \in U)$, the probability of selection for a sampling unit i is given by

$$p_i = \frac{z_i}{Z}$$

where $Z = \sum_{i \in U} z_i$. Under *pps* sampling, Hansen and Hurwitz (1943) also proposed HH-estimator $\hat{Y}_{HH}$ for finite population total Y, which becomes HT-estimator when $\pi_i = np_i$, given by

$$\hat{Y}_{HT} = \frac{1}{n} \sum_{i \in s} \frac{y_i}{p_i} = \hat{Y}_{HH}$$

where p_i is probability of selection for ith sampling unit at each selection. Sampling variance of HT-estimator under *pps* sampling (or HH-estimator) is given by

$$V_{pps}(\hat{Y}_{HT}) = \frac{1}{n} \sum_{i \in U} p_i \left(\frac{y_i}{p_i} - Y \right)^2 \tag{1.4}$$

and its unbiased estimator is given in Eq. (1.14).

Under πps sampling with size variable z_i $(i \in U)$, inclusion probability of a sampling unit i is calculated as

$$\pi_i = \frac{nz_i}{Z}$$

such that $0 < \pi_i \leq 1$. If $\pi_i > 1$ for a value of z_i, its value is fixed to unity (i.e. $\pi_i = 1$) and π_i's are calculated again for rest of the sampling units, this process continues until $0 < \pi_i \leq 1$ for all sampling units, see (Särndal et al., 1992, p. 89). Horvitz and Thompson (1952) proposed a design-unbiased HT-estimator under πps sampling. Sampling variance of HT-estimator for Y under πps sampling is given earlier in Eq. (1.2) and in Eq. (1.3) when sampling design has fixed sample size; the problem of variance estimation under πps sampling is reviewed later in Section 1.6.

Since the sampling with probability proportional to size has been proposed, a large body of literature is associated with sampling methods to achieve this sampling design, πps sampling design in particular; a brief review of some sampling methods follows. Madow (1949) extended equal probability SYS for πps sampling; but absence of unbiased estimator for sampling variance is a typical issue associated with systematic sampling. In the early stage, πps sampling methods were proposed which could only select samples of size one ($n = 1$), for example, *cumulative sum method* and a method given by (Lahiri, 1951), see (Cochran, 1977, p. 251) and (Särndal et al., 1992, p. 91). To select πps samples of size larger than one ($n > 1$), Rao et al. (1962) proposed to divide population into n randomly formed groups, and then select one unit from each group, see (Cochran, 1977, p. 266). Brewer (1963) proposed πps sampling method to select sample of size $n = 2$; second-order inclusion probabilities can be computed for this method which makes possible the unbiased variance estimation, see (Cochran, 1977, p. 261) and (Särndal et al., 1992, p. 92). Hájek (1964) proposed *Poisson sampling* which can select πps samples but with random sample size; a rejective sampling method was also discussed which only accepts samples of fixed sample size; this rejective sampling is also known as *conditional Poisson sampling*. *Rao-Sampford* method (Rao, 1965; Sampford, 1967) selected πps samples of fixed size n and allows calculation of exact second-order inclusion probabilities using a recursive formulae. Chao (1982) propose a πps sampling procedure for πps sampling which also allows calculation of second-order inclusion probabilities. Hanif and Brewer (1980) and Brewer and Hanif (1983) reviewed many πps sampling methods and discussed their properties. Deville and Tillé (1998) proposed a class of πps sampling methods including well-known *pivotal method* for πps sampling with fixed sample size. In the text book by Tillé (2006), more advanced equal and unequal probability sampling methods are reviewed with their different implementations through algorithms. Poisson sampling and pivotal method for πps sampling methods are described bellow which are used later in this chapters to describe a variance estimator (see Section 1.6) and a spatially balanced sampling method (see local pivotal method in Section 1.5.3) respectively.

Poisson sampling (PS)

In Poisson sampling by Hájek (1964), each sampling unit is selected independently with inclusion probability π_i, i.e. $P(I_{(i\in s)} = 1) = \pi_i$ and $P(I_{(i\in s)} = 0) = 1-\pi_i$ and its sampling distribution is given by

$$p(s) = \prod_{i\in s} \pi_i \prod_{i\in U-s} (1-\pi_i)$$

where sample s belongs to the sample space which contains all 2^N subsets of U, see (Särndal et al., 1992, p. 85). In Poisson sampling, sample size, denoted by n_s, is random with mean $E_{PS}(n_s) = \sum_{i\in U} \pi_i$ and variance $V_{PS}(n_s) = \sum_{i\in U} \pi_i(1-\pi_i)$. Since sampling units are selected independently under Poisson sampling, therefore second-order inclusion probabilities are simply product of corresponding first-order inclusion probabilities, given by $\pi_{ij} = \pi_i \pi_j$ for all $i \neq j \in U$. Sampling variance of HT-estimator for Y is given by

$$V_{PS}(\hat{Y}_{HT}) = \sum_{i \in U} \pi_i (1 - \pi_i) \frac{y_i^2}{\pi_i^2}$$

In rejective Poisson sampling or *conditional Poisson sampling* (CPS) (Hájek, 1964), πps samples are selected by Poisson sampling and samples with fixed sample size are accepted only. Therefore, inclusion probabilities induced by CPS are not exactly same as fixed π_i's.

Pivotal method

In pivotal method for πps sampling (Deville and Tillé, 1998), two inclusion probabilities π_i and π_j are selected randomly from the set of N inclusion probabilities and updated according to the following updating rule: if $\pi_i + \pi_j < 1$

$$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) & \text{with probability } \pi_i/(\pi_i + \pi_j) \\ (\pi_i + \pi_j, 0) & \text{with probability } \pi_j/(\pi_i + \pi_j) \end{cases}$$

and if $\pi_i + \pi_j \geq 1$

$$(\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1) & \text{with probability } (1 - \pi_i)/(2 - \pi_i - \pi_j) \\ (\pi_i + \pi_j - 1, 1) & \text{with probability } (1 - \pi_j)/(2 - \pi_i - \pi_j) \end{cases}$$

where π_i', π_j' denoted updated inclusion probabilities. The process of updating of inclusion probabilities is repeated until all the sampling units are finished. At each step, sampling outcome is decided for at least one sampling unit and sample is obtained in at most N steps. Sampling design implied by the pivotal method is not known, therefore π_{ij}'s cannot be computed exactly, see Deville and Tillé (1998) for details.

1.2 Cluster sampling

In sample surveys, situations occurs when list of population elements (or element level sampling frame) is not available and it is prohibitively expensive to construct such a list. In such cases, list of larger sampling units containing more than one population elements is used for sampling. In some situations even when the list of population elements is available, sample of larger sampling units is selected in order to meet the economic and workload limitations. It is also common to use cluster sampling when population elements are scattered over a vast geographical area, for example, in national level surveys of people or households, because sampling of population elements may result into very high travel expenses if personal interviews are required and inefficient supervision of field work.

In cluster sampling, population elements are grouped into clusters, called *primary sampling units* (PSU's) (Särndal et al., 1992, p. 125), and a random sample of PSU's is selected. Usually, the clusters are constructed based on some predefined geographical boundaries known from other sources, for example, administrative boundaries or census tracts. Therefore, clusters constitute of nearby population elements. Often, nearby population elements tend to have similar values of study variables; this similarity is represented by a *measure of homogeneity* (Särndal et al., 1992, p. 130) or *intra-cluster correlation* when PSU's have equal sizes (Cochran, 1977, p. 241). Since there is loss of information in cluster sampling due to selection of nearby population element in the form of clusters (or PSU's). Therefore, SRS of population elements is considered to be more efficient than cluster sampling and its efficiency increases with positive value of intra-cluster correlation. However, when efficiency is balanced against the survey cost then cluster sampling may have advantages over the SRS of elements in the situations mentioned in the previous paragraph. Cluster sampling design may consist of one, two or multi stages, called *multi-stage sampling*. When sampling design has more than one stages, PSU's are further divided into smaller sampling units (or clusters) depending on the stages of the sampling design, and population elements or clusters of population elements are selected in the final stage. Singe-stage cluster sampling and two-stage sampling designs are described bellow.

1.2.1 Single-stage cluster sampling

Let elements in the finite population U can be grouped into G clusters of unequal sizes, denoted by $N_1, ..., N_G$, which are considered as PSUs. Let the population of PSU's is denoted by $U_\mathrm{I} = \{U_1, ..., U_G\}$. In *single-stage cluster sampling* (SCS), a random sample of PSU's is selected and all the population elements are observed in the selected PSU's. Let a random sample of n_I PSU's, denoted by s_I, is selected from the population of clusters U_I. When the sample s_I is selected by SRS without-replacement, the HT-estimator of finite population total Y and its sampling variance are given by

$$\hat{Y}_{HT}^{SCS} = (N_\mathrm{I}/n_\mathrm{I}) \sum_{g \in s_\mathrm{I}} Y_g \text{ and } V(\hat{Y}_{HT}^{SCS}) = N_\mathrm{I}^2 \frac{1 - f_\mathrm{I}}{n_\mathrm{I}} \frac{\sum_{g \in U_\mathrm{I}} (Y_g - \bar{Y}_c)^2}{N_\mathrm{I} - 1}$$

respectively, where $Y_g = \sum_{i \in U_g} y_{gi}$ is total of y variable for gth cluster, $f_{\mathrm{I}} = n_{\mathrm{I}}/N_{\mathrm{I}}$ is sampling fraction of PSU's and $\bar{Y}_c = \sum_{i \in U} y_i/G$ is mean per cluster, see (Särndal et al., 1992, p. 129). Sampling variance $V(\hat{Y}_{HT}^{SCS})$ can also be written as follows

$$V(\hat{Y}_{HT}^{SCS}) = \left(1 + \frac{N - N_{\mathrm{I}}}{N_{\mathrm{I}} - 1}\delta\right) V_{WOR}(\hat{Y}_{HT}^{SRS}) + N_{\mathrm{I}}^2 \frac{1 - f_{\mathrm{I}}}{n} \mathrm{Cov}$$

where $V_{WOR}(\hat{Y}_{HT}^{SRS})$ is sampling variance of HT-estimator under without-replacement SRS of population elements, Cov$= \sum_{g \in U_I} (N_g - \bar{N}) N_g \bar{Y}_g^2/(N_g - 1)$ represents covariance between N_g and $N_g \bar{Y}_g^2$, and δ denotes the measure of homogeneity of population elements within PSU's (or clusters) with respect to study variable.

From (Särndal et al., 1992, p. 130), δ is defined as $\delta = 1 - S_w^2/S_y^2$, where S_w^2 is pooled within-cluster variance, given by $S_w^2 = (N - N_{\mathrm{I}})^{-1} \sum_{g \in U_{\mathrm{I}}} \sum_{i \in U_g} (y_{gi} - \bar{Y}_g)^2$, where $\bar{Y}_g = \sum_{i \in U_g} y_{gi}/N_g$ population mean of the gth cluster. The range of values for δ is given by $[-(N_{\mathrm{I}} - 1)/(N - N_{\mathrm{I}}), 1]$. Small and large values of δ indicate low and high levels of homogeneity of the population elements within clusters with respect to the survey variable. At the extreme values: $\delta = -(N_{\mathrm{I}} - 1)(N - N_{\mathrm{I}})$ means all the cluster means are equal, $\delta = 1$ means variance is zero within all the clusters and $\delta = 0$ means total variance is equal to average of within cluster variances, that is, $S_y^2 = S_w^2$.

The variance formula suggests that SCS is better than SRS only if $\delta < 0$ provided that all the PSU's are of same size which is quite unusual in practice when clusters consists of nearby elements. This means SRS is usually better than SCS. When PSU's are of unequal sizes, the relative efficiency also depends on 'Cov' term, in addition to δ. This means positive value of the term 'Cov' even worsen the efficiency of SCS given that the value of δ is positive. When $\delta = 0$ then variance of SCS increases with 'Cov' term in addition to the factor involving variance of SRS. When δ attains its minimum value, which is $-(N_{\mathrm{I}} - 1)/(N - N_{\mathrm{I}})$, then variance of SCS is function of variance of PSU sizes. In summary, homogeneity of elements within PSU's makes SCS inefficient and variation in the PSU sizes makes it worse.

1.2.2 Two-stage sampling

In two-stage sampling, PSU's are further divided into smaller sampling units, called *secondary sampling units* (SSU's), and a random sample of SSU's is selected from each PSU selected in the first-stage. When SSU's are population elements, it is called *two-stage element sampling* and when SSU's are again clusters of elements then it is called *two-stage cluster sampling*. These terms are taken from (Särndal et al., 1992, p. 125). From Cochran (1939), two-stage sampling is also known as *subsampling* and Mahalanobis and Fisher (1944) called it two-stage sampling while discussing area sampling from the agricultural populations.

Two-stage sampling design is an alternative to SCS which provides better flexibility to balance between efficiency and cost. Increasing sample size of PSU's increases the efficiency of SCS which is usually unacceptable due to cost limitations. Two-stage design allows to select larger sample of PSU's and then selecting a random sample of SSU's instead of observing all the SSU's in the selected PSU's. Subsampling of SSU's also requires estimation of PSU level totals Y_g. If within PSU variation is small, estimator of Y_g have small variance even for modest sample size of elements within selected PSU's (Särndal et al., 1992, p. 134).

For two-stage element sampling, equal probability selection method (epsem) by πps-SRS is common in practice, in which PSU's are selected by πps sampling and equal number of elements is selected by SRS from the PSU's selected in the first-stage sample. This design was first introduced by Hansen and Hurwitz (1943) as *pps*-SRS design. One can achieve a fixed sample size of population elements using this design. Another two-stage epsem is SRS-SRS, where SRS of PSU's is selected and a sample of elements is selected with constant sampling fraction from the PSU sampled in first-stage sample. In practice, PSUs often has unequal sizes, therefore, πps sampling of is preferred because it is more efficient than SRS. Two-stage epsem by πps-SRS is also known as *self-weighted design* as it assign uniform weights to y-values in the sample. It also has administrative advantage; since there are equal number of elements to be observed in each sampled PSU, therefore, field work is roughly equal in each PSU (Särndal et al., 1992, p. 141).

1.3 Generalized regression estimator

When population totals are known for a set of auxiliary variable related with the study variable y, generalized regression estimator (GREG) estimator is commonly used in practice. This estimator lies under the umbrella of model assisted approach (see Section 1.7), which is motivated by general linear regression model and the properties are derived under the sampling distribution. The GREG-estimator was given by Cassel et al. (1976) and further discussed by Särndal et al. (1992). It is also considered as a calibration estimator discussed by Deville and Särndal (1992) where sample weights are calculated such that known auxiliary totals are equal to their corresponding GREG-estimates.

Let $x_1, ..., x_q$ denote q number of auxiliary variables and $\mathbf{x}_i = (x_{1i}, ..., x_{qi})^\top$ denotes column vector of ith elements of qth auxiliary variable. Let $\hat{\mathbf{X}}_{HT} = \sum_{i \in s} \mathbf{x}_i / \pi_i$ is vector of HT-estimators which estimates vector of auxiliary totals given by $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$. Let the relationship of auxiliary variables with the study variable y can be expressed by a general linear regression model, given by

$$y_i = \mu_i + \epsilon_i \tag{1.5}$$

where $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ is linear predictor of the model with $\boldsymbol{\beta}$ representing vector of q regression coefficients, and ϵ_i is random error term which is normally distributed with mean 0 and variance σ_i^2. Let the vector $\mathbf{X}$ is known, the GREG-estimator, denoted by $\hat{Y}_{GR}$, for the population total Y is given by

$$\hat{Y}_{GR} = \hat{Y}_{HT} + (\hat{\mathbf{X}}_{HT} - \mathbf{X})^\top \hat{\boldsymbol{B}} \tag{1.6}$$

from (Särndal et al., 1992, p. 232), where $\hat{\boldsymbol{B}}$ is sample estimate of the vector $\boldsymbol{\beta}$, given by

$$\hat{\boldsymbol{B}} = \left(\sum_{i \in s} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\sigma_i^2 \pi_i} \right)^{-1} \sum_{i \in s} \frac{\mathbf{x}_i y_i}{\sigma_i^2 \pi_i} \tag{1.7}$$

from (Särndal et al., 1992, p. 225), where superscript $\top$ denotes transpose of a vector (or matrix). The GREG-estimator can also be written in terms of calibration weights w_i's, given by

$$\hat{Y}_{GR} = \sum_{i \in s} w_i y_i,$$

with calibration weights given by

$$w_i = \frac{1}{\pi_i} \left[1 + (\hat{\mathbf{X}}_{HT} - \mathbf{X})^\top \left(\sum_{i \in s} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i \sigma_i^2} \right)^{-1} \frac{\mathbf{x}_i}{\sigma_i^2} \right]$$

from (Särndal et al., 1992, p. 232). The GREG-estimator is approximately design unbiased and usually have smaller sampling variance as compared to HT-estimator, because GREG-estimator accounts for the variation in y explained by auxiliary variables. Approximate sampling variance of the GREG-estimator is given by

$$V(\hat{Y}_{GR}) \approx \sum_{i \in U} \frac{e_i^2}{\pi_i} + 2 \sum_{i < j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{e_i e_j}{\pi_i \pi_j}$$

from (Särndal et al., 1992, p. 235), where $e_i = y_i - \mathbf{x}_i^\top \boldsymbol{B}$ is ith finite population residual and $\boldsymbol{B}$ is vector of finite population regression coefficients given by

$$\boldsymbol{B} = \left(\sum_{i\in U} \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\sigma_i^2}\right)^{-1} \sum_{i\in U} \frac{\mathbf{x}_i y_i}{\sigma_i^2} \tag{1.8}$$

from (Särndal et al., 1992, p. 227).

The main problem associated with GREG-estimation is unusual calibration weights, including negative and very large positive values, which can potentially reduce the efficiency of GREG-estimator. It happens when number of auxiliary variables is large.

1.4 Balanced sampling

When a set of auxiliary variables correlated with study variables is known before selection of the sample, then samples balanced with respect to known auxiliary variables are tend to be more efficient than unbalanced samples. A sampling design is said to be balanced if population totals of auxiliary variables are equal to their respective HT-estimators. In other words, any sample s selected under the balanced sampling design satisfies the following equations:

$$\hat{\mathbf{X}}_{HT}(s) = \sum_{i\in s} \frac{\mathbf{x}_i}{\pi_i} = \sum_{i\in U} \mathbf{x}_i = \mathbf{X} \tag{1.9}$$

where $\hat{\mathbf{X}}_{HT}(s)$ denotes vector of HT-estimators based on sample s. In the context of balanced sampling, auxiliary variables are sometimes called *balancing variables* and set of equations in (1.9) are referred as *balancing equations*. SRS design is balanced with respect to population size, that is, for $x_i \equiv 1$, $\hat{X}_{HT} = \sum_{i\in s} x_i/\pi_i = (N/n)\sum_{i\in s} x_i = N$. For πps sampling designs, HT-estimator for population size is given by $\hat{X}_{HT} = \sum_{i\in s} x_i/\pi_i = \sum_{i\in s} 1/\pi_i$ is random, where $x_i \equiv 1$.

According to Deville and Tillé (2004), the history of balanced sampling dates back to early developments of finite population sampling in the beginning of twentieth century. An early concept of balanced sampling named as 'representative method' was given by Kiær (1896). It was to select a sample such that it matches a know quantity. In the early work for balanced sampling, some purposive sampling methods were proposed. Yates (1946); Thionet (1953) also advocated balanced sampling. Some partial solutions for balanced sampling were given by Ardilly (1991); Deville (1992); Hedayat and Majumdar (1995); Deville et al. (1988). Deville and Tillé (2004) proposed cube method for balanced sampling which is widely used in practice. Chauvet and Tillé (2006) gave a faster implementation of the cube sampling method. Fuller (2009b) studied properties of rejective method (Hájek, 1964, 1981) for balanced sampling. Chauvet et al. (2017) studied some sampling strategies involving the cube method and rejective method for balanced sampling. In recent developments, Benedetti et al. (2022) proposed a balanced sampling method based on a global optimisation algorithm called simulated annealing and Leuenberger et al. (2022) suggested a modification which aims to improve efficiency of the fast implementation of the cube method.

1.4.1 The cube method

Deville and Tillé (2004) gave a random sampling method, called *cube method*, which aims to select samples with fixed first-order inclusion probabilities and balanced with respect to a set of known auxiliary variables. The name of this sampling method is motivated by geometric representation of the sampling design using N-dimensional cube (or N-cube), where N denotes number of sampling units in the population and 2^N vertices of the N-cube represent all possible samples (of any size) from the population. In the geometrical representation of sampling design, vector of first-order inclusion probabilities $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)$ is expressed as a convex combination of the vertices of the N-cube. Sampling design under the cube method assigns selection probability $p(s)$ to each vertex of the N-cube such that $E(s) = \boldsymbol{\pi}$, that is, fixed first-order inclusion probabilities are achieved. The set of balancing equations in Eq. (1.9) can be defined as a hyperplane which intersects the N-cube. Selecting a balanced sample is to choose a vertex of the N-cube that remains in the hyperplane. The balanced sampling algorithm in the cube method, randomly reaches a vertex of the N-cube from the vector $\boldsymbol{\pi}$ in such a way that the balancing equations are satisfied, or approximately so.

Algorithm for the cube method randomly transforms elements of the vector $\boldsymbol{\pi}$ into sample membership indicators $\{0, 1\}$. It consists of two phases: *flight-phase* and *landing-phase*. In the flight-phase, a discrete time stochastic process, called *balancing martingale*, transforms the first-order inclusion probabilities into $\{0, 1\}$ indicator one-by-one such that balancing equations and fixed inclusion probabilities are achieved. It starts with the vector $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$, at time $t = 1, ..., T$, three steps are repeated as follow:

1. Generate any vector $\mathbf{u}(t)$, such that $\mathbf{u}(t)$ is kernel of the matrix $\mathbb{A} = (\mathbf{x}_1/\pi_1, ..., \mathbf{x}_N/\pi_N)$, and $u_i(t) = 0$ if $\pi_i(t-1)$ is an integer.

2. Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values of $\lambda_1(t)$ and $\lambda_2(t)$ such that $0 \leq \boldsymbol{\pi}(t-1) + \lambda_1(t)\mathbf{u}(t) \leq 1$, $0 \leq \boldsymbol{\pi}(t-1) - \lambda_2(t)\mathbf{u}(t) \leq 1$.

3. Select
$$\boldsymbol{\pi}(t) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\mathbf{u}(t) \text{ with probability } q(t) \\ \boldsymbol{\pi}(t) - \lambda_1^*(t)\mathbf{u}(t) \text{ with probabiilty } 1 - q(t) \end{cases}$$
where $q(t) = \lambda_2^*(t)/[\lambda_1^*(t) + \lambda_2^*(t)]$.

Above steps are repeated until it is no longer possible to carry out step 1. At the end of flight-phase, vector $\boldsymbol{\pi}(T)$ is obtained, if all the inclusion probabilities are transformed into $\{0, 1\}$ indicators, then the algorithm completes. Otherwise landing-phase is required to achieve the sample. In the landing-phase, balancing equations are compromised in order to get sample of fixed size such that fixed inclusion probabilities are respected. When balancing equations are not exactly satisfied, it is referred as *rounding problem* (Deville and Tillé, 2004; Tillé, 2011).

Let $\boldsymbol{\pi}(T) = \boldsymbol{\pi}^*$, and sampling design for the remaining units is formulated as optimization problem which minimizes the conditional sampling variance $Var(\hat{\mathbf{X}}|\boldsymbol{\pi}^*)$. The landing-phase can be implemented in two way as follow:

- *Linear programming:* The conditional variance $Var(\hat{\mathbf{X}}|\boldsymbol{\pi}^*)$ is minimized using linear programming,
- *Dropping balancing equations:* At the end of flight-phase if sample is not achieved, last variable from the set of auxiliary (or balancing) variables is dropped and flight-phase is implemented again. This process continue unit a sample is achieve. Therefore it is advised to put the auxiliary variables in the order of their importance in the algorithm with this version of landing-phase.

(Deville and Tillé, 2004) advocated sampling strategy of balanced sampling by cube method and GREG-estimator as balanced sampling helps avoiding extreme weights for the GREG-estimator. In their simulation study, it was reported that percentage of samples with negative calibration weights reduced from 32% to 0.1%.

In the implementation of above algorithm for cube method, number of computational operations increase with square of the population size N^2. Chauvet and Tillé (2006) proposed a fast implementation of cube method for which number of computational operations increases with size of the population N. In the fast implementation, flight-phase of the cube method was applied to a subset of population which consisted of $q + 1$ sampling units instead of whole population. Application of flight-phase continues until it provides a vector with at most q components which are not rounded to zero or one. Landing-phase of the cube method is implemented in usual way. An important aspect of fast implementation is that order of population units may have an effect on efficiency of the cube method. Therefore, it was suggested to randomly mix the sampling units in order to maximise randomness of the design or rearrange the sampling units such that rounding problem in the landing-phase is reduced.

Leuenberger et al. (2022) proposed to rearrange the sampling units in the population with respect to their distances from centre of the auxiliary space. This rearrangement aims to reduce the rounding problem of fast implementation of the cube method. The idea was to deal with distant or atypical units first in the flight-phase as much as possible and typical or central units in the landing-phase. At each step of the flight-phase of fast cube method, at least one inclusion probability is rounded to 0 or 1 among the first $q + 1$ units with non-integer inclusion probabilities. Rearrangement of the population sampling units such that atypical units are in the beginning of file provides more chance to the atypical units to be submitted in the flight-phase. Therefore, most of the typical units left to be processed in the landing-phase which is expected to reduce the rounding problem.

1.5 Spatially balanced sampling

When a study population can be mapped over a geographical area using some location data, e.g. geographic coordinates or maps, it is referred as *spatial population*, which can be categorised as discrete or continuous. In discrete spatial populations, response variable is measured at discrete points while in continuous spatial population, response variable can be measured at infinite points (Stevens Jr and Olsen, 2004). A random sample from the finite spatial population is said to be *spatially balanced* when it is well-spread over the population area (Grafström et al., 2012). Spatially balanced sampling methods aim to select well-spread samples using location data of the spatial population. This implies, location data for all the sampling units in the population are supposed to be known in advance of sample selection. Most spatial populations are not homogeneous (Stevens Jr and Olsen, 2004), so that SRS fails to provide a spatially balanced sample. This has become an established fact that spatially balanced samples tend to be more efficient than SRS when there exist some spatial dependence in the spatial population under study (Stevens Jr and Olsen, 2004; Grafström et al., 2012; Benedetti et al., 2017c).

Almost all the spatially balanced sampling methods achieve spatial balance in the sample by manipulating second-order inclusion probabilities. Some basic methods do this directly by assigning zero joint probabilities to adjacent pairs of sampling units and more advanced methods do this implicitly by using other criteria, for example, spatial stratification of the population area, or using pairwise distances of the sampling units (e.g. euclidean distance). A typical problem associated with spatially balanced sampling methods is that they often produce zero or close to zero second-order inclusion probabilities which precludes unbiased estimation of sampling variance of HT-estimator.

In the following subsections, spatial dependence in finite populations and spatial balance of the sample are briefly described followed by the subsection which describes some spatially balanced sampling methods which are commonly used in practice or often appeared in the literature.

1.5.1 Spatial dependence

Spatial dependence or autocorrelation referred as the relationship among values of variable that is a result of geographical arrangement of their locations (Benedetti et al., 2015, p 17). A measure of spatial autocorrelation quantifies the dependence of variable values on the relative positioning of the population units in the space. There is positive spatial autocorrelation, if similar values of the variable are nearby in the space. There is negative spatial autocorrelation, if larger values are surrounded by small values (or small values surrounded by large values) in the space. There is zero spatial autocorrelation if similar values are randomly scattered in the space. Some commonly used measures for quantify the spatial dependence are described bellow.

Spatial weight (or connectivity) matrix

Spatial weight matrix (SWM) or connectivity matrix, denoted by $\mathbf{W}$, describes the observations which are neighbours of each location (Benedetti et al., 2015, p. 18). It is a square matrix with dimension equal to number of observations; if it is computed for a spatial population of size N, it would have dimension $N \times N$. Each entry of the matrix corresponds to (i, j) pair, denoted by w_{ij}, which represents proximity of the locations i, j and it is often measured by a distance measure, for instance, euclidean distance. It can also take binary $(0, 1)$ values, where $w_{ij} = 1$ if (i, j) location are neighbours, $w_{ij} = 0$ otherwise. Diagonal elements of the matrix $\mathbf{W}$ are set to zero, because they represents distance (or proximity) of a location to itself.

Moran's I index for spatial autocorrelation

Moran's I index given by Moran (1948, 1950) is commonly used measure of spatial autocorrelation for quantitative (interval and ratio measurement scale) study variables. Cliff and Ord (1973, 1981) presented a comprehensive work on spatial autocorrelation and suggested a formula to calculate the Moran's I index and joint count statistics for qualitative (nominal measurement scale) variables.

Let x_i denotes a study variable and w_{ij} is measure of connectivity between units i and j, where $w_{ij} = d_{ij}^{-1}$ for $i \neq j$, $w_{ij} = 0$ else, and d_{ij} denotes euclidean distance between locations of units i and j. The formula for Moran's I index is given by

$$I = \frac{N}{W} \frac{\sum_{i\in U} \sum_{j\in U} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i\in U}(x_i - \bar{x})^2} \tag{1.10}$$

where $W = \sum_{i\in U} \sum_{j\in U} w_{ij}$ and $\bar{x}$ is mean of variable x_i. A value of index $I = -(N-1)^{-1}$ means no spatial autocorrelation, and the values less and greater than $-(N-1)^{-1}$ represent negative and positive spatial autocorrelation respectively. Significance of an observed value of index I can be tested based on normal approximation.

Semi-variogram

Semi-variogram measures variability between pair of units at different lags of distance, given by

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i,j\in N(h)} (y_i - y_j)^2 \tag{1.11}$$

where $N(h)$ is number of pairs separated by distance $d(i,j) = h$, where $i, j \in U$. For spatially correlation variable, as the distance between pair of units increase, values of sami-variogram are expected to increase.

1.5.2 Spatial balance of samples

Voronoi polygons for spatial balance

Stevens Jr and Olsen (2004) introduced a measure of spatial balance based on Voronio polygons. It also called Thiessen polygons method (Lister and Scott, 2009). In the sample s, the Voronoi polygon α_i for the unit $i \in s$ includes all population units closer to i than to any other sample unit $k \in s$. Let ν_i be the sum of inclusion probabilities in the ith Voronoi polygon. If a population unit has equal distance to two or more sample units, then it is included in more than one polygons. The inclusion probability of that unit is then divided equally to each polygon, that is, $\nu_i = \sum_{j \in \alpha_i} \pi_j / m_j$, where $m_j = \sum_{i \in s} I_{(j \in \alpha_i)}$. It follows that $\sum_{i \in s} \nu_i = n$. A sample is spatially balanced if $\nu_i = 1$ for all $i \in s$. Thus, the variance

$$\text{SB} = \frac{1}{n} \sum_{i \in s} (v_i - 1)^2 \tag{1.12}$$

can be used as measure of spatial balance for a sample; spatial balance under a sampling design can be computed as average of spatial balances over many samples (Grafström et al., 2012).

Moran's I index for spatial balance

Tillé et al. (2018) proposed to calculate Moran's index for the sample indicator variable, $I_{(i \in s)}$, as measure of spatial balance. Two measures of spatial balance were given, one based on usual Moran's index and other based on its normalized version along with modified spatial weights. The normalization was introduced to restrict the values in range of $[-1, 1]$. Let $\mathbf{I}$ be the vector denoting $I_{(i \in s)}$ variable, $\mathbf{1}_{N \times 1}$ be a vector of ones and $\bar{\mathbf{I}} = \frac{1}{N} \mathbf{1}\mathbf{1}^\top \mathbf{I}$, the measure of spatial balanced based on usual Moran's I index was given by

$$\text{SB}_I = \frac{(\mathbf{I} - \bar{\mathbf{I}})^\top \mathbf{W} (\mathbf{I} - \bar{\mathbf{I}})}{(\mathbf{I} - \bar{\mathbf{I}})^\top (\mathbf{I} - \bar{\mathbf{I}}) \mathbf{1}^\top \mathbf{W} \mathbf{1}}$$

where $\mathbf{W}$ is non-negative spatial weight matrix based on a distance measure, e.g. element w_{ij} indicates how close is unit j to i, a large value means j is neighbour of i and $w_{ii} = 0$ for all $i \in U$. The measure of spatial balance based on normalized version of Moran's index was given by

$$\mathrm{SB}_I^* = \frac{(\mathbf{I}-\bar{\mathbf{I}}_w)^\top\tilde{\mathbf{W}}(\mathbf{I}-\bar{\mathbf{I}}_w)}{\sqrt{(\mathbf{I}-\bar{\mathbf{I}}_w)^\top\mathbf{D}(\mathbf{I}-\bar{\mathbf{I}}_w)(\mathbf{I}-\bar{\mathbf{I}}_w)^\top\mathbf{B}(\mathbf{I}-\bar{\mathbf{I}}_w)}}$$

where elements of the weight matrix $\tilde{\mathbf{W}}$ were defined as

$$\tilde{w}_{ij} = \begin{cases} 1 & \text{if unit } j \in N_{\lfloor k_i \rfloor}, \\ k_i - \lfloor k_i \rfloor & \text{if unit } j \text{ is the } \lceil k_i \rceil\text{th nearest neighbour of } i, \\ 0 & \text{otherwise.} \end{cases}$$

where $k_i = (1/\pi_i - 1)$, $\lfloor a \rfloor$ and $\lceil a \rceil$ are roof and ceiling functions of a which return largest integer smaller than a and $a+1$ respectively.

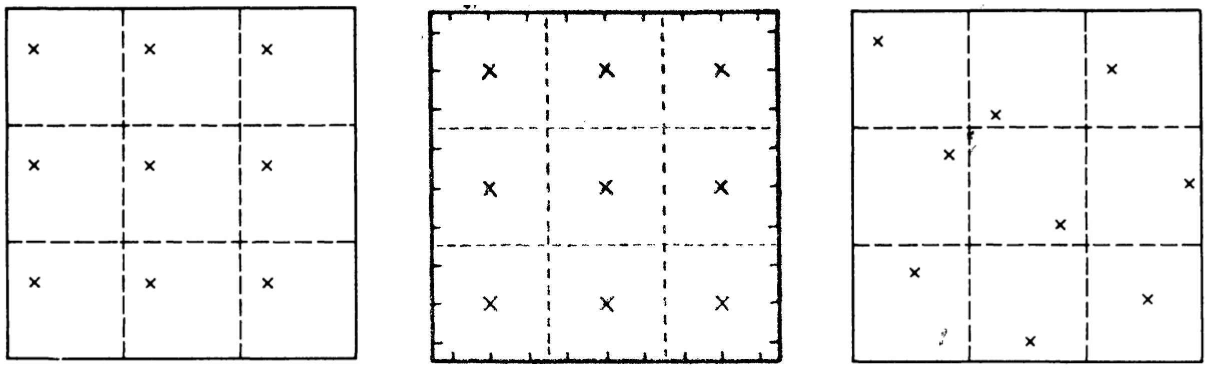
1.5.3 Spatially balanced sampling methods

This subsection describes some spatially balanced sampling methods. The cube method for balanced sampling (with respect to auxiliary variables), described in Section 1.4.1, can also be used for spatially balanced sampling by considering spatial variables (or coordinates) as balancing variables (Benedetti et al., 2017b,c).

Two-dimensional systematic sampling and maximal stratification

When sampling from an areal (or two-dimensional) population, *two-dimensional systematic sampling* (2D-SYS) is an extension of the systematic sampling. For rectangular two-dimensional (or areal) population, Quenouille (1949) proposed 2D-SYS and its different variants along with other sampling methods including two-dimensional random sampling and stratified random sampling with one unit per stratum; later discussed by Das (1950), Bellhouse (1977) and (Cochran, 1977, p. 227). In maximal stratification, population area is divided into spatial strata and one per stratum is selected. In maximal stratification, the methodologies for achieving a good spatial stratification are often criticized.

For 2D-SYS, population area is divided into $k \times k (= n)$ grid cells and one unit is selection from each grid cell. The most basic version of 2D-SYS is *square grid* pattern or aligned 2D-SYS (Cochran, 1977, p. 227) in which a pair of random coordinates determines the sample. In the *central* square grid pattern centre of the grid cells is chosen as sample units. In these basic versions, sampling units are selected equally distant from each other which ensure the spread of sampled units in the population area. When the populations show some kind of cyclic or periodic pattern, these two methods may show potential bias. *Unaligned* 2D-SYS sampling method aims to overcome this problem by selecting the units which are equidistant but random within the rows or columns of grid cells. See Figure 1.1 for a demonstration of these three variants of 2D-SYS. The 2D-SYS sample is more efficient than simple random sample of the coordinates when correlation between two units is monotone decreasing function of distance between two units (Cochran, 1977, p. 227).



(a) 2D-SYS aligned sample: "square grid" pattern.

(b) 2D-SYS aligned sample: "central square grid".

(c) 2D-SYS unaligned sample.

Figure 1.1: Variants of two-dimensional systematic sampling – *Source:* (Cochran, 1977, p. 228).

Balanced sampling excluding contiguous units (BSEC)

Hedayat et al. (1988) gave a sampling method for list of population units arranged in particular order, known as balanced sampling excluding contiguous (BSEC) units. This is an equal probability sampling method. It selects a sample by assigning zero joint inclusion probabilities to the contiguous sampling units in the list of sampling units and equal joint inclusion probability to non-contiguous sampling units. Here the aim is same, to select a sample which consists of the sampling units that are not too close. Efficiency of this sampling method depends upon the serial correlation in the list of population elements, like systematic sampling, in comparison with simple random sampling.

Generalized random-tessellation stratified (GRTS) design

Stevens Jr and Olsen (2004) proposed GRTS design, the most granulated among spatially balanced sampling methods based on spatial stratification, which selects spatially balanced samples with fixed inclusion probabilities. Assuming that sampling frame consists of N points located within a geographic region, the GRTS design stratify the region into small quadrants, each containing at most one population point, and place them in the line of length N using a random process called *hierarchical randomization*, for details see Stevens Jr and Olsen (2004). For equal probability sampling, each quadrant is given unit length and a systematic random sample of size n is selected with random start between $(0, N/n)$. For πps systematic sampling, the quadrants are assigned length proportional to the corresponding inclusion probabilities. The hierarchical randomization process preserves the spatial positioning of the quadrants as much as possible, therefore a systematic sample is well-spread over the population area. The GRTS design is applicable for both discrete and continuous spatial populations. It is able to accommodate additional sample points in case of non-responses while maintaining the spatial balance of the sample. It also have ability to perform inverse sampling.

Stevens Jr and Olsen (2004) also proposed a measure of spatial balance based on Voronoi polygons, see Eq. (1.12). Spatial balance of samples under GRTS design was compared with independent random sampling (IRS) (or *pss* sampling) and spatially stratified sampling (SSS) under different scenarios of non-responding or denial areas and different sample sizes. Numerical results showed that, GRTS design was the most spatially balanced followed by SSS and IRS was the least balanced design.

Stevens Jr and Olsen (2003) gave a variance estimator under GRTS design called local mean (or local neighbourhood) variance estimator, see Eq. (1.17). Since spatially balanced designs have some very small or zero second-order inclusion probabilities, use of usual unbiased variance estimator may not be possible; if possible in some cases it can be unstable. Local neighbourhood estimator is perceived to be more stable and approximately unbiased (Stevens Jr and Olsen, 2004). In 1997, Indiana Department of Environmental Management (IDEM) implemented GRTS design in a survey for biological assessment of streams and rivers in Indiana. For different response variables from this survey, variance estimates were calculated using IRS variance estimator and local neighbourhood variance estimator. The results showed that local neighbourhood variance estimates were smaller than IRS variance estimates.

Spatially balanced sampling using space-filling curves (SFC)

Lister and Scott (2009) proposed a GIS-assisted spatially balanced sampling method which transform the two-dimensional space into one-dimensional line using SFC called Peano curve. The implementation of this method was claimed to be more transparent and simpler than GRTS design, and similar performance in terms of efficiency and spatial balance. In this sampling method, the location of population units in two-dimension is translated into one-dimensional spatial address which are grouped into contiguous groups. Sampling units are selected from these groups randomly.

Spatially correlated Poisson sampling (SCPS)

Grafström (2012) proposed SCPS method to select spatially balanced samples with fixed first-order inclusion probabilities. It is a special case of correlated Poisson sampling, a list sequential sampling method given by Bondesson and Thorburn (2008). Correlated Poisson sampling method first decides sampling outcome for the first unit in the list, then for second unit and so on, up to last unit. After each sampling decision, it updates the inclusion probabilities of the remaining units in list using a set of weights. The weights are chosen within a specified range to achieve required design properties. In SCPS, spatially balanced samples are selected by assigning large positive weights to the population units which are close with respect to distance.

In correlated Poisson sampling method, first unit is selected in the sample with probability $\pi_1^{(0)} = \pi_1$. If unit was included then $I_1 = 1$, otherwise $I_1 = 0$. Generally at step j, when the values for $I_1, ..., I_{j-1}$ have been recorded, unit j is included with probability $\pi_j^{(j-1)}$. Then inclusion probabilities are updated for the units $i = j+1, ..., N$ as follow

$$\pi_i^{(j)} = \pi_i^{(j-1)} - (I_j - \pi_i^{(j-1)})w_j^{(i)}$$

where $w_j^{(i)}$ are weights given by unit j to the units $i = j+1, j+2, ..., N$ and $\pi_i^{(0)} = \pi_i$. For $0 \leq \pi_i^{(j)} \leq 0$, the weights can be chosen within the following range:

$$-\min\left(\frac{1-\pi_i^{(j-1)}}{1-\pi_j^{(j-1)}}, \frac{\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right) \leq w_j^{(i)} \leq \min\left(\frac{\pi_i^{(j-1)}}{1-\pi_j^{(j-1)}}, \frac{1-\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right).$$

A fixed sized sampling design is obtained if weights sum to one, given that inclusion probabilities sum to an integer. In order to achieve spatially balanced sampling design under the SCPS, two strategies were proposed to choose the weights, as described in the following.

- *Maximal Weights:* The unit j gives as much weight as possible to the nearest unit with respect to distance (not in the list) among units $i = j+1, j+2, ..., N$, then as much weight as possible to the second nearest unit and so on, such that the weights sum to 1. If distances are equal for two or more units, then the weight is distributed equally on those units.

- *Gaussian preliminary weights:* The weights follows Gaussian distribution centred at the position of unit j, which were chosen as: $w_j^{(i)*} \propto \exp\{-(d(i,j)/\sigma)^2\}$, where $i = j+1, j+2, ..., N$ and σ is a parameter which control the spread of the weights. Average of distance between unit j and its nearest neighbour was proposed as one option for the choice of σ. Furthermore, weights may not be within the required range and need to be truncated which may create very small variation in the sample size.

For both types of weighting strategies, different order of the units gives a different design, but overall property of the design is same, that is, samples are spatially balanced.

Local pivotal method (LPM)

Grafström et al. (2012) proposed two local pivotal methods (LPMs) to select spatially balanced samples with fixed first-order inclusion probabilities. These methods are based on another sampling method, called pivotal method (Deville and Tillé, 1998) (also see Section 1.1.5), and pairwise euclidean distance of population units. The algorithm for pivotal method selects two units randomly and update their inclusion probabilities such that one of the two units is included in or excluded from the sample. This process is repeated for all units in the population. In this way, a sample with fixed inclusion probabilities is obtained. Local pivotal methods update inclusion probability in same way for two nearby population units, not the randomly selected units. In this way, LPMs avoids selecting nearby units in the sample. Two criteria for choosing nearby units were given. First criterion consisted of four steps and is described in the following

1. Randomly choose one unit i;

2. Choose unit j which is nearest neighbour of i. If there are multiple units having the same distance to i, then randomly choose among them;

3. If j has i as its nearest neighbour, then update the inclusion probabilities using the algorithm for pivotal method;

4. Repeat step 1 to 3 until all units are finished.

In second criterion, step 1, 2 and 4 of first criterion are used. This allows directly updating inclusion probabilities instead of searching for nearest neighbour of unit j. Local pivotal methods based on these two criteria were named as LPM1 and LPM2 respectively. It was shown analytically that both methods ensure local balance of sample size at least for clustered populations.

For variance estimation under LPMs, three estimators were mentioned: variance estimator under independent random sampling or *pss* sampling (Hansen and Hurwitz, 1943), variance estimator under conditional Poisson sampling (CPS) (Hájek, 1964) and "local-mean" variance estimator (Stevens Jr and Olsen, 2003), also see Section 1.6. Local-mean variance estimator was recommended for LPMs. By simulation studies based on real and artificial data sets, it was shown that LPMs are more spatially balanced in the sense of Voronoi polygon measure and more efficient than CPS and GRTS. It was also demonstrated that "local-mean" variance estimator is generally better than other two competitors for the LMPs.

Spatially balanced sampling using product of within-sample distances (PWD)

Benedetti and Piersimoni (2017) proposed a spatially balanced sampling method which aims to minimize within-sample distance using a Markov Chain Monte Carlo (MCMC) algorithm. It does not consider unequal fixed inclusion probabilities. First and second-order inclusion probabilities were derived for a simplified case.

This method aims to select a sample s with selection probability proportional to a distance measure $M(D_s)$ where D_s represents the distance matrix of units in the sample s. In the sampling algorithm, initial sample $s^{(0)}$ is selected randomly. At iteration t the elements of sample $s^{(t)}$ are updated according to the following steps:

1. Select two units at random, one from the sample $s^{(t)}$ and other from outside the sample $s^{(t)}$,

2. In the new sample, denoted by $s_e^{(t)}$, two units are exchanged and the new sample is selected in $s^{(t+1)}$ with probability: $\min\{1, [M(D_{s_e^{(t)}})/M(D_{s^{(t)}})]^{\beta}\}$, where β is a tuning parameter.

3. Repeat first and second steps $q \times N$ times, where q is the maximum number of

iteration each consisting of N attempts. The algorithm stops when no attempt is accepted within qth iteration.

Two distance measures were proposed given by: $M_1(D_s) = \sum_{i \in s} \sum_{j \in s} d_{ij}$ and $M_0(D_s) = \prod_{i \in s} \prod_{j \in s} d_{ij}$, where d_{ij} is euclidean distance between units i and j. The tuning parameter β controls the proportionally between selection probability of a sample and its distance measure. A large value of β is required when relation is more than proportional, although it requires more iterations. The distance measure $M_0(D_s) = \prod_{i \in s} \prod_{j \in s} d_{ij}$, named PWD (proportional to within sample distance) was often recommended during the simulation studies.

Spatially balanced sampling using weakly associated vectors (WAVE)

Jauslin and Tillé (2020) proposed a sampling method which selects spatially balanced samples with fixed first-order inclusion probabilities. In this method, first, a stratification matrix of size $N \times N$ is constructed where ith row of the matrix represents a stratum which contains unit i and its nearest neighbours such that sum of the inclusion probabilities is greater than or equal to one by only one unit. Rows of the matrix sum to one which might be viewed as spatial constraints to achieve a spatially balanced sample. Second, following the idea of cube method (Deville and Tillé, 2004), vector of inclusion probabilities is modified in a random manner and transformed into a selection indicator vector such that the sample achieves spatial constraints and respects the fixed first-order inclusion probabilities. The idea of stratification matrix seems to target at the definition of spatial balance based on Voronoi polygons (Stevens Jr and Olsen, 2004).

A simulation study was conducted using real data set known as `Meuse` which contained variables related to metal concentrations. The proposed method was compared with SRS, GRTS, LPM1, SCPS and HIP under equal probability and with GRTS, LPM1, SCPS and Maxent under πps sampling. Cadmium and zinc were considered as response and size variables respectively. Comparison was made based on MSE of HT-estimator and three measures of spatial balance: SB, SB$_I$ and SB$_I^*$. The WAVE method was more efficient than others only for small sample sizes considered in the study. For large samples SCPS was the most efficient under πps sampling. The two measures of spatial balance based on Moran index showed that WAVE is the most balanced method in all the cases. The measure based on Voronoi polygon indicated that WAVE or LPM were the most balanced under equal probability sampling while SPCS was the most balanced under unequal probability sampling.

1.5.4 Doubly balanced sampling by local cube method

A sampling design is doubly balance if it achieve the balancing equations, see Eq. (1.9), with respect to auxiliary variables and at the same time balanced with respect to spatial coordinates. Grafström and Tillé (2013) gave a sampling method, called *local cube method*, which aims to select samples with fixed first-order inclusion probabilities which are balanced with respect to auxiliary variables and are well-spread over the finite space at the same time. This method is based on cube method for balanced sampling and local pivotal method for spatially balanced sampling, both are described earlier in this chapter. In the algorithm for local cube method, a cluster of $q+1$ nearby units is selected and flight-phase of cube method is applied (where q is number of auxiliary variables). Application of flight-phase on a nearby units selects at least one units from the cluster and reduce the selection chance of other nearby units in the same sample, this is how it achieves local balance. When undecided units are less than $q+1$, landing-phase of the cube method is applied.

1.6 Variance estimation

The HT-estimator is an unbiased estimator under any sampling design when first-order inclusion probabilities are known and positive, i.e. $\pi_i > 0$ for all $i \in U$. An unbiased estimator for sampling variance of the HT-estimator is always desirable. An unbiased variance estimator from Horvitz and Thompson (1952) based on variance formula in Eq. (1.2) is given by

$$\hat{V}_{HT}(\hat{Y}_{HT}) = \sum_{i,j\in s} \left(\frac{\pi_{ij} - \pi_i\pi_i}{\pi_{ij}}\right) \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

provided that $\pi_{ij} > 0$ for all $i,j \in U$ (Särndal et al., 1992, p. 43). When sample size is fixed, an unbiased variance estimator from Sen (1953) and Yates and Grundy (1953) (SYG) based on variance formula in Eq. (1.3), is given by

$$\hat{V}_{SYG}(\hat{Y}_{HT}) = -\frac{1}{2}\sum_{i\in s}\sum_{i<j\in s} \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{1.13}$$

where s is a random sample of fixed size n and $\pi_{ij} > 0$ for all $i,j \in U$ (Särndal et al., 1992, p. 45). For the calculation of above variance estimators, values of both first- and second-order inclusion probabilities are required and all of them are required to be strictly positive. In most cases, first-order inclusion probability are prefixed or easy to compute, whereas second-order inclusion probabilities are unknown under most of the sampling designs including balanced and spatially balanced sampling designs. Even if calculation of second-order inclusion probabilities is possible through some recursive methods, it becomes computationally expensive for large populations. Therefore, sampling variance of HT-estimator is often estimated through approximations. Since the theory of HT-estimator under πps sampling is introduced, many approximations have been proposed in literature to estimate its sampling variance, most of them involve only first-order inclusion probabilities because they are often known.

Variance estimator under *pps* sampling (Hansen and Hurwitz, 1943) is often used for variance estimation under πps sampling, however it usually overestimates the sampling variance since πps sampling tends to be more efficient that *pps* sampling. Hartley and Rao (1962) proposed an approximation for variance estimation under randomized systematic πps sampling under the assumption of $N \to \infty$ for fixed n. While Hájek (1964) proposed a variance approximation under conditional Poisson sampling using assumption of $N \to \infty$ and $(N - n) \to \infty$. Rosén (1991) considered variance estimation under *pps* systematic sampling. Berger (1998a) extended Hajék's approximation for some other sampling designs. Berger (1998b) proposed a variance estimator under Chao's sampling scheme for πps sampling (Chao, 1982). Deville (1999) proposed a variance approximation based on maximum entropy. Based on (Hájek, 1964)'s approximation, Berger (2005) proposed a variance estimator under πps systematic sampling. Haziza et al. (2004) and Haziza et al. (2008) compared 12 estimators for sampling variance of HT-estimator under Rao-Sampford πps sampling procedure (Rao, 1965; Sampford, 1967). There are many other methodologies which are used for variance estimation including jack-knife and bootstrap methods, see Wolter (2007) for details. Those variance approximations which are used or discussed in later chapters of this thesis are described bellow.

Variance approximation based on *pps* sampling

A simple approximation of sampling variance under πps sampling is to use the sampling variance under *pps* sampling (Hansen and Hurwitz, 1943), discussed by (Durbin, 1953), (Cochran, 1977, p. 252), (Särndal et al., 1992, p. 99,422) and (Wolter, 2007, p. 12) among others. It does not require computation of second-order inclusion probabilities. Variance estimator based on this approximation often overestimate the sampling variance because πps sampling tends to have smaller variance than *pps* sampling. Sampling variance of HT-estimator under *pps* sampling is given in Eq. (1.4) and its unbiased estimator is given by

$$\hat{V}_{pps}(\hat{Y}_{HT}) = \frac{1}{n(n-1)} \sum_{i \in s} \left(\frac{y_i}{p_i} - \frac{1}{n} \sum_{i \in s} \frac{y_i}{p_i} \right)^2 \tag{1.14}$$

where $\pi_i = np_i$.

(Deville and Tillé, 2005)'s variance approximation under balanced sampling

Deville and Tillé (2005) suggested that sampling variance of the HT-estimator under cube method can be approximated by the sampling variance of the GREG-estimator under Poisson sampling design. This approximation based on two arguments: first, sampling design under Poisson sampling does not requires computation of second order inclusion probabilities; second, Poisson sampling has maximum entropy and balanced sampling design is conditional of Poisson sampling design; see Deville and Tillé (2005) for details.

Let $\tilde{\pi}_i$ denotes first-order inclusion probabilities under Poisson sampling design, the sampling variance of HT-estimator under poisson sampling (Hájek, 1964) is given by

$$V_{PS}(\hat{Y}_{HT}) = \sum_{i \in U} \frac{y_i^2}{\pi_i^2} \tilde{\pi}_i(1-\tilde{\pi}_i) = \mathbf{z}^T \tilde{\boldsymbol{\Delta}} \mathbf{z}$$

where $\mathbf{z} = (y_1/\pi_1, ..., y_N/\pi_N)^T$ and $\tilde{\boldsymbol{\Delta}} = \text{Diag}[\tilde{\pi}_i(1-\tilde{\pi}_i)]_{i \in U}$ is a diagonal matrix. Note that $\tilde{\pi}_i$'s are unknown and Deville and Tillé (2005) given four approximations for $\tilde{\pi}(1-\tilde{\pi})$.

Following (Hájek, 1964, 1981)'s residual technique, Deville and Tillé (2005) proposed variance approximation under balanced sampling, given by

$$V_{PS}(\hat{Y}_{HT}|\hat{\mathbf{X}}_{HT} = \mathbf{X}) \approx V_{PS}(\hat{Y}_{HT} + (\mathbf{X} - \hat{\mathbf{X}}_{HT})^T \boldsymbol{\beta})$$

where

$$\begin{aligned} \boldsymbol{\beta} &= V_{PS}(\hat{\mathbf{X}}_{HT})^{-1} \text{Cov}_{PS}(\hat{\mathbf{X}}_{HT}, \hat{Y}_{HT}), \\ V_{PS}(\hat{\mathbf{X}}_{HT}) &= \sum_{i \in U} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i^2} \tilde{\pi}_i(1-\tilde{\pi}_i) \\ \text{Cov}_{PS}(\hat{\mathbf{X}}_{HT}, \hat{Y}_{HT}) &= \sum_{i \in U} \frac{\mathbf{x}_i y_i}{\pi_i^2} \tilde{\pi}_i(1-\tilde{\pi}_i) \end{aligned}$$

When the term $\tilde{\pi}(1-\tilde{\pi})$ is approximated by $N\pi_i(1-\pi_i)/(N-q)$ (Hájek, 1981; Deville and Tillé, 2005), the variance approximation under balanced sampling can be written as

$$V(\hat{Y}_{HT}) \approx \frac{N}{N-q}\sum_{i\in U}\frac{\tilde{e}_i^2}{\pi_i^2}\pi_i(1-\pi_i) \tag{1.15}$$

where $\tilde{e}_i = z_i - \tilde{z}_i$, $\tilde{z}_i = \mathbf{A}^T(X\tilde{\mathbf{\Delta}}X^T)^{-1}X\tilde{\mathbf{\Delta}}\mathbf{z}$, $\mathbf{A} = (\mathbf{x}_1/\pi_1, ..., \mathbf{x}_N/\pi_N)$ and $X = (\mathbf{x}_1, ..., \mathbf{x}_N)$. The corresponding variance estimator is given by

$$\hat{V}(\hat{Y}_{HT})_{DT} = \frac{n}{n-q}\sum_{i\in s}\frac{\hat{\tilde{e}}_i^2}{\pi_i}(1-\pi_i) \tag{1.16}$$

where $\hat{\tilde{e}}_i = z_i - \hat{\tilde{z}}_i$, $\hat{\tilde{z}}_i = \mathbf{A}_s^T(X_s\tilde{\mathbf{\Delta}}_sX_s^T)^{-1}X_s\tilde{\mathbf{\Delta}}_s\mathbf{z}_s$, the subscript s denotes values corresponding to sample s. The subscript DT for variance estimator means (Deville and Tillé, 2005)'s variance estimator.

In (Deville and Tillé, 2005)'s variance approximation, the assumption of exact balancing may not be always true. Therefore, the variance estimator based on this approximation can be biased when sampling design is not exactly balanced. For cube method, assuming exact balancing of the design means that this approximation only aims the flight-phase of the cube method, the bias can increase as the sampling variance due to flight-phase decreases (or sampling variance due to landing-phase increases). Breidt and Chauvet (2011) also proposed simulation-based approximation for balanced sampling using cube method. In a simulation study, the variance estimator based on the simulation-based approximation was compared with (Deville and Tillé, 2005)'s variance estimator in Eq. (1.16). The simulation-based variance estimator was approximately unbiased but less efficient as compared to (Deville and Tillé, 2005)'s estimator.

Under spatially balance sampling, second-order inclusion probabilities of nearby units are likely to be zero or very close to zero. For example, in one- and two-dimensional systematic sampling second-order inclusion probabilities are non-zero only for the units which belongs to the same sample. Similarly in BSEC, second-order inclusion probabilities are non-zero only of the non-contiguous units in the list. Therefore, unbiased estimation of variance using Sen-Yates-Grundy estimation is not possible, as second-order inclusion probabilities appears in the denominator. According to Stevens Jr (1997), expression for second-order inclusion probabilities can be produced under GRTS design for continuous populations, although they are not known for finite (or discrete) populations. Due to near-zero second order inclusion probability, these expression may not give a stable variance estimator (Stevens Jr and Olsen, 2004). In another instance, Benedetti et al. (2017a) proposed a model-based variance estimator for two-dimensional systematic sampling, one-per-stratum (or maximal stratification) sampling. This estimator required second-order inclusion probabilities to be know which is true for the two considered design but not for more advanced designs, for instance, LPMs and SCPS. Some variance estimators which are commonly used in practice or often appeared in literature are described in the following. Grafström and Lundström (2013) also proposed a variance estimator when qualitative balancing variables are used for spatial or auxiliary balancing.

Local-mean (or local neighbourhood) variance estimator

Stevens Jr and Olsen (2003) proposed a variance estimator based on local neighbourhood (NBH) for GRTS design, it is also know as local-mean variance estimator (Grafström et al., 2012; Grafström and Lundström, 2013). The expression for the local-mean estimator is given by

$$\hat{V}_{\mathrm{NBH}}(\hat{Y}_{HT}) = \sum_{i \in s} \sum_{j \in D_i} w_{ij} \left(\frac{y_i}{\pi_i} - \bar{y}_{D_i} \right)^2 \tag{1.17}$$

where D_i is a neighbourhood to unit i, containing at least four units, and w_{ij} are weights that decrease as the distance between unit i and j increases. The weights satisfy $\sum_j w_{ij} = 1$ and $\bar{y}_{D_i}$ is a neighbourhood total (Grafström et al., 2012). Local-mean variance estimator is often recommended for spatially balanced sampling, unless a better estimator is available (Stevens Jr and Olsen, 2004; Grafström, 2012; Grafström et al., 2012; Robertson et al., 2013; Benedetti and Piersimoni, 2017).

(Grafström and Tillé, 2013)'s variance estimator under doubly balanced sampling

For doubly balanced sampling by local cube method, Grafström and Tillé (2013) introduced variance estimator by combining local-mean variance estimator (Stevens Jr and Olsen, 2003) and variance estimator for balanced sampling (Deville and Tillé, 2005), given by

$$\hat{V}_{\mathrm{DBS}}(\hat{Y}_{HT}) = \frac{n}{n-p} \frac{p+1}{p} \sum_{i \in s} (1 - \pi_i) \left(\frac{e_i}{\pi_i} - \bar{e}_i \right)^2 \tag{1.18}$$

where

$$
\begin{aligned}
e_i &= y_i - \mathbf{x}^\top \hat{\boldsymbol{\beta}} \\
\hat{\boldsymbol{\beta}} &= \left[\sum_{i \in s} (1 - \pi_i) \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i^2} \right]^{-1} \sum_{i \in s} (1 - \pi_i) \frac{\mathbf{x}_i y_i}{\pi_i^2} \\
\bar{e} &= \sum_{i \in G_i} (1 - \pi_i) \frac{e_i}{\pi} \left[\sum_{i \in G_i} (1 - \pi_i) \right]^{-1}
\end{aligned}
$$

and G_i is the set of the $p + 1$ closest units of i in the sample (including i itself).

1.7 Supper-population model

Often it is assumed that the given finite population is a realization of an *infinite* supper-population which is usually expressed by a probability model. Therefore, variable of interest are random variables under the super-population model. When the inference regarding finite population parameters is based on the sampling distribution under the sampling design (or induced by the sampling method), it is usually known as *design-based* (or randomisation) approach. Ordinarily, survey sampling theory is based on design-based approach. If the inference is based on probability distribution assumed under the super-population model, it is known as *model-based* (or prediction) approach, see Valliant et al. (2000). Whereas, when super-population model is used in order to develop a sampling design or an estimation scheme but inference is based on sampling distribution, it is known as *model-assisted* approach, see Särndal et al. (1992) for model-assisted survey sampling.

There is also some literature regarding controversy between design- and model-based inference approaches. An objection on the design-based approach is based on the fact that it disregard the probability distribution under the super-population model of the study variable rather it is based on probability distribution induced by the sampling design. Godambe (1955) demonstrated that no unbiased estimator with least variance exist in the class of linear estimators under the sampling design (or design-based approach). This article stimulated the work regarding consideration of probability models in the theory of survey sampling. Godambe and Joshi (1965) extend the theory of non-existence of least variance for class of non-linear estimators under the design-based approach. Later, the work related to using probability models in estimation of finite population parameters resulted into model-based approach.

In theory, when assessing some sampling strategies for a given finite population, it is useful to envisage some assumptions about the finite population, i.e specify a super-population model. Anticipated mean squared error (AMSE) is model expectation of the sampling variance of an estimator (in a sampling strategy). A general expression for the AMSE of an estimator $\hat{Y}$ of Y is given by

$$\text{AMSE}(\hat{Y}) = E_m\{E_p(\hat{Y} - Y)^2|U\} = E_p\{E_m(\hat{Y} - Y)^2|s\}$$

where E_m and E_p denote expectation functions with respect to super-population model and sampling distribution respectively. In most numerical studies of this thesis, AMSE approach is considered for the comparison of different sampling strategies.

A super-population model with constant mean and homogeneous error variance is given by $y_i = \mu + \epsilon_i$, where ϵ_i random error term identically independently distributed (iid) under normal distribution $N(0, \sigma^2)$. For this model, AMSE of HT-estimator for population total Y under SRS is minimum, see Fuller (2009a). A linear regression model with independent errors is given in Eq. (1.5). For this model, AMSE of HT-estimator for Y under a sampling design $p(s)$ can be written as

$$\text{AMSE}(\hat{Y}_{HT}) = E_p\left[\left(\sum_{i\in s}\frac{\mathbf{x}_i}{\pi_i} - \sum_{i\in U}\mathbf{x}_i\right)^\top \boldsymbol{\beta}\right]^2 + \sum_{i\in U}\left(\frac{1}{\pi_i} - 1\right)\sigma_i^2 \qquad (1.19)$$

For the above AMSE, Godambe and Joshi (1965) gave a lower bound given by

$$\sum_{i\in U}\left(\frac{1}{\pi_i} - 1\right)\sigma_i^2$$

and suggested that HT-estimator with πps sampling such that $\pi_i \propto \mu_i$ has minimum variance in the class of linear estimators. Under the same model, (Särndal et al., 1992, p. 452) suggest that GREG-estimator with πps sampling such that $\pi_i \propto \sigma_i$ is an optimal strategy. Isaki and Fuller (1982) has also studied properties of some sampling strategies under the linear regression super-population model.

Under the linear regression model with correlated errors, $V_m(\epsilon_i) = \sigma_i^2$ and $\text{Cov}_m(\epsilon_i, \epsilon_j) = \sigma_{ij}$ where V_m and Cov_m denote variance and covariance function under super-population model, the AMSE of HT-estimator under sampling distribution $p(s)$ and linear regression model is given by

$$\mathrm{AMSE}(\hat{Y}_{HT}) = E_p\left[\left(\left(\sum_{i\in s}\frac{\mathbf{x}_i}{\pi_i} - \sum_{i\in U}\mathbf{x}_i\right)^{\top}\boldsymbol{\beta}\right]^2 + 2\sum_{i<j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)\sigma_{ij} \tag{1.20}$$

from (Grafström and Tillé, 2013). For an optimal sampling strategy, one would like to minimize the AMSE of the HT-estimator given above, which has two term: first term represent the lack of auxiliary balance and second term involve second-order inclusion probabilities. First term vanishes when sampling design is balance with respect to auxiliary variables. Second term can be minimized be choosing a set of second-order inclusion probabilities. However, it is a hard problem to find a set of second-order inclusion probabilities which corresponds to a sampling design (Gabler and Schweigkoffer, 1990). Spatially balanced sampling methods aims to control second-order inclusion probabilities implicitly with the aim to minimize the AMSE.

1.8 Thesis outline

For two-stage sampling design, equal probability sampling method (epsem) by πps-SRS is commonly used in practice. For two-stag epsem, PSU's are selected by πps sampling at the first-stage, and usually equal number of elements is selected by SRS from the sampled PSU's at the second-stage. Alternatively, a sample of equal-sized clusters can be selected, at the second-stage. Although HT-estimator is unbiased, GREG-estimator is commonly used when population total of some auxiliary variables are known. In this, way four sampling strategies can be formulated based on two-stage epsem and two estimators: HT- and GREG-estimator. Following problems are considered about two-stage epsem by πps-SRS:

1. A comparison of four sampling sampling is conducted under a two-level regression model which aims to provide insights about two-stage epsem from practical viewpoint.
2. Some exploration is done about formulating custom sub-cluster for two-stage espem using auxiliary variable. This formulation of sub-clusters aims to improve the efficiency of two-stage epsem.

As mentioned in earlier sections that balanced sampling with respect to known auxiliary variables in combination with GREG-estimator is advocated in literature. Cube method for balanced sampling is often used in practice. It was originally proposed for selection of PSU's. Regarding balanced sampling of PSU's in two-stag epsem, following two problems are considered.

1. Cube method is not always exactly balanced. A sampling procedure is proposed which aims to improve the cube method when it is not exactly balanced.
2. Second order inclusion probabilities are unknown under cube method, therefore problem of unbiased variance estimation under cube method is not fully addressed. A methodology of variance estimation under balanced sampling is also proposed which naturally follows the proposed procedure for balanced sampling.

The introduction of spatially balanced sampling in social surveys is a recent phenomena. It has mainly been used in natural resource, ecology and environmental surveys. In the context of two-stage sampling design in social surveys, one can select spatially balanced samples either at one or both stages depending on availability of location data. When auxiliary and spatial variables are known, it is also proposed to select doubly balanced sample with respect to auxiliary and spatial variables (or spatial coordinates). In this thesis, following aspects of spatially balanced sampling methods are considered.

1. A variety of spatially balanced sampling methods can be found in literature with varying ability to achieve spatial balanced and efficiency. A comparative study of spatially balanced sampling methods (whose implementation is available in R statistical software) is conducted under spatial super-population model.
2. There might be situation when a survey contain some variables with negative spatial autocorrelation in addition to those with positive spatial autocorrelation. Some spatial sampling strategies are suggested for such situations.
3. Estimation of sampling variance is a common challenge associated with spatially balanced sampling. A variance estimation methodology is proposed for spatially and doubly balanced sampling.

Chapter 2

Sampling strategies related to two-stage equal probability sampling

2.1 Introduction

Two-stage sampling is commonly used in large scale surveys, for instance, in national level household surveys. For two-stage sampling, equal probability sample method (epsem) by πps-SRS is a common practice in which PSU's are selected by πps sampling and equal number of elements are selected by SRS from the sampled PSU's. This is a two-stage element sampling design; a shorthand notation for this design 2Se shall be used in this chapter. An alternative two-stage epsem can be defined as two-stage cluster sampling design, denoted by 2Sc, in which PSU's are selected by πps sampling (same as in 2Se) and a simple random sample of sub-clusters (in stead of elements) is selected from each sampled PSU, where all the sub-clusters have same size.

The HT-estimator of finite population total is unbiased under the two-stage epsem. GREG-estimator, an approximately unbiased estimator of finite population total, is also commonly used in practice when a set of auxiliary totals is known. For socio-economic surveys, these auxiliary totals are often available from last census data or from a previous sample survey. Two sampling designs for two-stage epsem and two estimators for finite population total constitute four sampling strategies given bellow:

1. (2Se, HT)
2. (2Sc, HT)

3. (2Se, GREG)

4. (2Sc, GREG)

Two-stage epsem is an intuitive choice when survey variable can be expressed by a super-population model which assumes constant variance of y-values across the finite population. In this chapter, above four sampling strategies involving two-stage epsem are compared in a systematic way. The comparison is based on AMSE's under the two-level regression model rather than their MSE's for a given finite population; comparison of these sampling strategies with respect to AMSE's is not seen in literature. AMSE represent average performance of a sampling strategy for the finite population parameters under a give super-population model. The aim of this comparison is to single out a preferred sampling strategy which might be useful from practical viewpoint.

Usually cluster sampling is less efficient than element sampling, therefore, two-stage epsem base on two-stage cluster sampling (i.e. 2Sc) is expected to be less efficient than two-stage element sampling (i.e. 2Se) when sub-clusters have positive intra-cluster correlation. Also, equal-sized sub-clusters are rarely found in practice. This may motivate custom formulation of equal-sized sub-clusters. A simplified case is when equal-sized sub-clusters are randomly formed within each PSU. When a set of auxiliary variables is available, one may also look to formulate even better sub-clusters using the known values of auxiliary variables. In the following, a toy example demonstrates that there exist ways of cluster sampling by which sampling variance is zero.

Example 2.1. Let response values are $y_i = 1, ..., 16$ and sample size is $n = 4$. Consider the following arrangement of y_i's with rows as clusters:

$$\begin{bmatrix} 1 & 8 & 9 & 16 \\ 2 & 7 & 10 & 15 \\ 3 & 6 & 11 & 14 \\ 4 & 5 & 12 & 13 \end{bmatrix}$$

where rows of the matrix are four samples which have the same sample mean is 8.5 (and total 34), but different sample variance. The sampling variance of the row-sample mean is 0. This particular cluster sampling is the most efficient design for this population, including compared against SRS of elements.

□

To get cluster sampling like in the above example, unknown y-values are required which is not possible in a sample survey situation. However, auxiliary variables related to y-variable can be used for this purpose like other model-assisted approaches in survey sampling. In this chapter, one way of formulating sub-clusters is explored which aims to obtain sub-clusters such that they are equal-sized and have equal means of auxiliary variables. Assuming that such sub-clusters exists and can be achieved, a preliminary analysis indicates that SRS of these sub-clusters is actually same as equal-probability balanced sampling (with respect to auxiliary variables) of elements in term of AMSE. Furthermore, these sub-clusters might be more difficult to achieve in practice as compared to balanced sampling. However, it may have an advantage over balanced sampling, which is unbiased variance estimation because it is SRS of sub-clusters.

Rest of the chapter is arranged as follows. In Section 2.2, formulation of sub-clusters using auxiliary variables is briefly described which happen to be same as balanced equal probability sampling of elements. In Section 2.3, four sampling strategies involving two-stage epsem are compared with respect to their AMSE's under a two-level super-population model. In Section 2.4, a simulation study is conducted which provides further insights of the comparison. In the last, Section 2.5 gives some conclusions of this chapter. In the following two subsections, formulas of HT-estimator and GREG-estimator under two-stage sampling designs are given.

2.1.1 HT-estimators under two-stage sampling

Let a random sample s_{I} of n_{I} PSU's is selected from the population U_{I} (U_{I} is defined in Section 1.2.1), there are N_{I} PSU's in the population. Let π_g and π_{gh} denotes first- and second-order inclusion probabilities for the gth and (g,h)th PSU's respectively, where $g \in U_{\mathrm{I}}$ and $(g \neq h) \in U_{\mathrm{I}}$. In two-stage element sampling, a random sample s_g of n_g elements is selected from the PSU's selected in the first-stage sample, i.e. $g \in s_{\mathrm{I}}$. Let $\pi_{i|g}$ and $\pi_{ij|g}$ denote first- and second-order inclusion probabilities of ith and (i,j)th elements in the gth PSU, i.e. $i \in U_g$ and $i \neq j \in U_g$. The HT-estimator of population total Y and its sampling variance under two-stage element sampling are given by

$$\hat{Y}_{HT} = \sum_{g=1}^{n_{\mathrm{I}}} \frac{\hat{Y}_{gHT}}{\pi_g} \tag{2.1}$$

$$V(\hat{Y}_{HT}) = \sum_{g,h=1}^{N_{\mathrm{I}}} \left(\frac{\pi_{gh}}{\pi_g \pi_h} - 1\right) Y_g Y_h + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \sum_{i,j=1}^{N_g} \left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1\right) y_{gi} y_{gj} \tag{2.2}$$

where $\hat{Y}_{gHT} = \sum_{i=1}^{m_g} y_i/\pi_{i|g}$ is HT estimator of gth PSU total $Y_g = \sum_{i \in U_g} y_{gi}$, see (Särndal et al., 1992, p. 137).

Under two-stage epsem by πps-SRS (i.e. 2Se), sample size of elements is equal within all sampled PSU's, i.e. $n_g \equiv n_0$. Sampling variance under 2Se can be written as

$$V(\hat{Y}_{HT}^{2Se}) = \sum_{g,h=1}^{N_{\mathrm{I}}} \left(\frac{\pi_{gh}}{\pi_g \pi_h} - 1 \right) Y_g Y_h + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \left(\frac{1 - f_g}{n_0} S_g^2 \right) \tag{2.3}$$

where $f_g = n_0/N_g$, $S_g^2 = \sum_{i \in U_g} (y_{gi} - \bar{Y}_g)^2$ and $\bar{Y}_g = Y_g/N_g$.

For two-stage cluster sampling, let each PSU can be further divided into $N_{\mathrm{II}g}$ sub-clusters of unequal sizes N_{gk} where $g \in U_{\mathrm{I}}$ and $k = 1, ..., N_{\mathrm{II}g}$. Let $U_{g1}, ..., U_{N_{\mathrm{II}g}}$ denote second-stag clusters in gth PSU. Let a random sample $s_{\mathrm{II}g}$ of $n_{\mathrm{II}g}$ second-stage clusters (or sub-clusters) is selected from the gth PSU selected in first-stage sample, i.e. $g \in s_{\mathrm{I}}$. Let $\pi_{k|g}$ and $\pi_{kl|g}$ denotes first- and second-order inclusion probabilities for kth and (k, l)th sub-clusters in the gth PSU, respectively. The HT-estimator of population total Y and its sampling variance under two-stage cluster sampling can be written as

$$\hat{Y}_{HT} = \sum_{g=1}^{n_{\mathrm{I}}} \frac{1}{\pi_g} \sum_{k=1}^{n_{\mathrm{II}g}} \frac{Y_{gk}}{\pi_{k|g}} \tag{2.4}$$

$$V(\hat{Y}_{HT}) = \sum_{g,h=1}^{N_{\mathrm{I}}} \left(\frac{\pi_{gh}}{\pi_g \pi_h} - 1 \right) Y_g Y_h + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \sum_{k,l=1}^{N_{\mathrm{II}g}} \left(\frac{\pi_{kl|g}}{\pi_{k|g} \pi_{l|g}} - 1 \right) Y_{gk} Y_{gl} \tag{2.5}$$

where $Y_{gk} = \sum_{i \in U_k} y_{gi}$ is population total for kth sub-cluster in gth PSU.

Under two-stage epsem by πps-SRS (2Sc), assume that all the sub-clusters are equal sized, i.e. $N_{gk} \equiv N_0$, and equal number of sub-clusters are sampled from each selected PSU, i.e. $n_{\mathrm{II}g} \equiv n_{\mathrm{II}0}$. Sampling variance under 2Sc can be written as

$$V(\hat{Y}_{HT}^{2Sc}) = \sum_{g,h=1}^{N_{\mathrm{I}}} \left(\frac{\pi_{gh}}{\pi_g \pi_h} - 1 \right) Y_g Y_h + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \left(\frac{1 - f_{\mathrm{II}g}}{n_{\mathrm{II}0}} S_{\mathrm{II}g}^2 \right) \tag{2.6}$$

where $f_{\mathrm{II}g} = n_{\mathrm{II}0}/N_{\mathrm{II}g}$, $S_{\mathrm{II}g}^2 = \sum_{k \in U_g} (Y_{gk} - \bar{Y}_{\mathrm{II}g})^2$ and $\bar{Y}_{\mathrm{II}g} = Y_g/N_{\mathrm{II}g}$ is mean per sub-cluster.

2.1.2 GREG-estimators under two-stage sampling

For GREG-estimator, (Särndal et al., 1992, p. 304) described three scenarios for availability of auxiliary data in the two-stage sampling design, given in the following.

Case A: (PSU Auxiliaries). The auxiliary totals are available for all PSU's in the population.

Case B: (Complete Element Auxiliaries). The auxiliary values are available for all elements in the entire population.

Case C: (Limited Element Auxiliaries). The auxiliary values are available for all the elements in selected PSU's only.

Case A leads to regression model at PSU level, Cases B and C lead to regression modelling of the element values. Similarly, above three cases of auxiliaries leads to three different GREG-estimators under two-stage design. Here, we considered the Case B for availability of complete auxiliary data.

Let $\mathbf{x}_{gi}$ denotes a column vector of q-values corresponding to the ith element in the gth PSU. Let relationship of y and the auxiliary variables is expressed by an element level general linear model, given by $y_{gi} = \mu_{gi} + \epsilon_{gi}$, where $\mu_{gi} = \mu(\mathbf{x}_{gi}) = \mathbf{x}_{gi}^{\top}\boldsymbol{\beta}$ is linear predictor, $\boldsymbol{\beta}$ is vector of q regression coefficients and ϵ_{gi} is random error term associated with the value of ith element in the gth cluster, which is normally distributed with mean zero and variance σ_{gi}^2. A population based estimate $\boldsymbol{B}$ of unknown vector $\boldsymbol{\beta}$ is given in Eq. (1.8) which also an unknown finite population quantity. Let $s = \{s_1, ..., s_{n_{\mathrm{I}}}\}$ denotes all the elements in the two-stage sample. A sample estimate of $\boldsymbol{B}$ is given in Eq. (1.7) where $\pi_i = \pi_g \pi_{i|g}$ under 2Se, under 2Sc $\pi_i = \pi_g \pi_{k|g}$ for all $i \in U_k$, and it is assumed $\sigma_i^2 = 1$. GREG-estimator for finite population total Y under two-stage element sampling, from (Särndal et al., 1992, p. 323), is given by

$$\hat{Y}_{GR}^{2Se} = \sum_{i\in U} \hat{y}_i + \sum_{i\in s} \frac{\hat{e}_i}{\pi_i} = \sum_{g=1}^{N_{\mathrm{I}}} \sum_{i=1}^{N_g} \hat{y}_{gi} + \sum_{g=1}^{n_{\mathrm{I}}} \frac{1}{\pi_g} \sum_{i=1}^{m_g} \frac{\hat{e}_{gi}}{\pi_{i|g}} \tag{2.7}$$

where $\hat{y}_{gi} = \mathbf{x}_{gi}^{\top}\hat{\boldsymbol{B}}$ and $\hat{e}_{gi} = y_{gi} - \hat{y}_{gi}$. An approximate variance of GREG-estimator in Eq. (2.7) is given by

$$V(\hat{Y}_{GR}^{2Se}) \approx \sum_{g=1}^{N_{\mathrm{I}}} \sum_{h=1}^{N_{\mathrm{I}}} \left(\frac{\pi_{gh}}{\pi_g \pi_h} - 1 \right) \mathbf{e}_g \mathbf{e}_h + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \sum_{i,j=1}^{N_g} \left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}} - 1 \right) e_{gi} e_{gj}$$

where $\mathbf{e}_g = \sum_{i \in U_g} e_{gi}$, and $e_{gi} = y_{gi} - \mathbf{x}_{gi}^{\top}\boldsymbol{B}$, see (Särndal et al., 1992, p. 325). Under two-stage epsem (2Se), when $n_g \equiv n_0$, sampling variance can be written as

$$V(\hat{Y}_{GR}^{2Se}) \approx \sum_{g=1}^{N_{\mathrm{I}}} \sum_{h=1}^{N_{\mathrm{I}}} \left(\frac{\pi_{gh}}{\pi_g \pi_h} - 1 \right) \mathbf{e}_g \mathbf{e}_h + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \left(\frac{1 - f_g}{n_0} S_{eg}^2 \right)$$

where $S_{eg}^2 = \sum_{i \in U_g} (e_{gi} - \bar{\mathbf{e}}_g)^2$ and $\bar{\mathbf{e}}_g = \mathbf{e}_g / N_g$. Following the formulation of GREG-estimator in Eq. (2.7), GREG-estimator of Y under two-stage cluster sampling can be written as

$$\hat{Y}_{GR} = \sum_{i \in U} \hat{y}_i + \sum_{i \in s} \frac{\hat{e}_i}{\pi_i} = \sum_{g=1}^{N_{\mathrm{I}}} \sum_{k=1}^{N_{\mathrm{II}g}} \sum_{i=1}^{N_{gk}} \hat{y}_{gki} + \sum_{g=1}^{n_{\mathrm{I}}} \frac{1}{\pi_g} \sum_{k=1}^{n_{\mathrm{II}g}} \frac{1}{\pi_{k|g}} \sum_{i=1}^{N_{gk}} \hat{e}_{gki} \tag{2.8}$$

where $\hat{e}_{gki} = y_{gki} - \mathbf{x}_{gki}^{\top}\hat{\boldsymbol{B}}$. Similarly, an approximate variance of GREG-estimator in Eq. (2.8) can be written as

$$V(\hat{Y}_{GR}) \approx \sum_{g=1}^{N_{\mathrm{I}}} \sum_{h=1}^{N_{\mathrm{I}}} \left(\frac{\pi_{gh}}{\pi_g \pi_h} - 1 \right) \mathbf{e}_g \mathbf{e}_h + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \sum_{k=1}^{N_{\mathrm{II}g}} \sum_{l=1}^{N_{\mathrm{II}g}} \left(\frac{\pi_{kl|g}}{\pi_{k|g} \pi_{l|g}} - 1 \right) \mathbf{e}_{gk} \mathbf{e}_{gl}$$

where $\mathbf{e}_{gk} = \sum_{i \in U_k} e_{gki}$, $\mathbf{e}_g = \sum_{k \in U_g} \mathbf{e}_{gk}$, and $e_{gki} = y_{gki} - \mathbf{x}_{gki}^{\top}\boldsymbol{B}$. Under two-stage epsem (2Sc), when all the sub-clusters are equal-sized and $n_{\mathrm{II}g} \equiv n_{\mathrm{II}0}$, sampling variance can be written as

$$V(\hat{Y}_{GR}^{2Sc}) \approx \sum_{g=1}^{N_{\mathrm{I}}} \sum_{h=1}^{N_{\mathrm{I}}} \left(\frac{\pi_{gh}}{\pi_g \pi_h} - 1 \right) \mathbf{e}_g \mathbf{e}_h + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \sum_{k=1}^{N_{\mathrm{II}g}} \sum_{l=1}^{N_{\mathrm{II}g}} \left(\frac{1 - f_{\mathrm{II}g}}{n_{\mathrm{II}0}} S_{e\mathrm{II}g}^2 \right)$$

where $S_{e\mathrm{II}g}^2 = \sum_{k \in U_g} (\mathbf{e}_{gk} - \bar{\mathbf{e}}_{\mathrm{II}g})^2$ and $\bar{\mathbf{e}}_{\mathrm{II}g} = \mathbf{e}_g / N_{\mathrm{II}g}$ is mean per sub-cluster.

2.2 Formulation of sub-clusters using auxiliary variables

For simplicity, assume that a PSU is the finite population (of size N) from which a random sample of n elements is required, and values of one auxiliary variable x_i are known for all the population elements. Assume that this finite population can be expressed by a linear model given by $y_i = x_i + \epsilon_i$ where ϵ_i denotes random error term independently identically distributed under normal distribution given by $N(0, \sigma^2)$.

Now, assumed that M clusters of size $n/2$ can be formed from the population such that cluster means of the auxiliary variable are equal i.e. $\bar{X}_g \equiv c$ for $g = 1, ..., M$, where c is constant. These clusters are named as dynamic clusters because they are not based on fixed geographical boundaries. When a simple random sample of two dynamic clusters is selected, the sampling design is named as *dynamic cluster sampling* (DCS).

Consider following four sampling strategies along with their AMSE's under the super-population model defined above.

I. (SRS, HT): SRS of n elements and HT-estimator for finite population total Y. AMSE of HT-estimator under SRS and above population model is given by

$$\text{AMSE}_{SRS}(\hat{Y}_{HT}) = E_{SRS}(\hat{X}_{HT} - X)^2 + N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \sigma^2 \tag{2.9}$$

where E_{SRS} is expectation function under SRS of elements.

II. (DCS, HT): DCS (defined earlier) and HT-estimator for finite population total Y. The HT-estimator and its sampling variance under DCS can be written as

$$\hat{Y}_{HT} = \sum_{g=1}^{2} \frac{Y_g}{\pi_g} = \frac{M}{2} \sum_{g=1}^{2} Y_g \text{ and } V_{DCS}(\hat{Y}_{HT}) = M^2 \left(\frac{1}{2} - \frac{1}{M}\right) \frac{\sum_{g=1}^{M} (Y_g - \bar{Y}_c)^2}{M-1}$$

respectively, where $\bar{Y}_c$ is population mean per cluster. Let E_{DCS} and E_m denote expectation functions under DCS and super-population model defined above, $\hat{\epsilon}_{HT}$ is HT-estimator of finite population total of errors given by $\epsilon = \sum_{i \in U} \epsilon_i$, the AMSE of HT-estimator under DCS and super-population model (defined above) is given by

$$\begin{aligned}
\text{AMSE}_{DCS}(\hat{Y}_{HT}) &= E_m[E_{DCS}(\hat{Y}_{HT} - Y)^2] \\
&= E_m[E_{DCS}(\hat{X}_{HT} - X)^2] + E_m[E_{DCS}(\hat{\epsilon}_{HT} - \epsilon)^2] \\
&= E_{DCS}(\hat{X}_{HT} - X)^2 + M^2 \left(\frac{1}{2} - \frac{1}{M}\right) \frac{1}{M-1} E_m \left[\sum_{g=1}^{M} (\epsilon_g - \bar{\epsilon}_c)^2\right] \\
&= E_{DCS} \left(\frac{M}{2} \sum_{g=1}^{2} X_g - X\right)^2 + M^2 \left(\frac{1}{2} - \frac{1}{M}\right) \frac{1}{M-1} E_m \left[\sum_{g=1}^{M} (\epsilon_g - \bar{\epsilon}_c)^2\right] \\
&= E_{DCS} \left(\frac{M}{2} \frac{n}{2} \sum_{g=1}^{2} \bar{X}_g - N\bar{X}\right)^2 + M^2 \left(\frac{1}{2} - \frac{1}{M}\right) \frac{1}{M-1} \frac{n}{2} (M-1)\sigma^2
\end{aligned}$$

$$\begin{aligned}\text{AMSE}_{DCS}(\hat{Y}_{HT}) &= E_{DCS}\left(\frac{N}{2}\sum_{g=1}^{2}\bar{X}_g - N\bar{X}\right)^2 + \left(\frac{2N}{n}\right)^2\left(\frac{1}{2} - \frac{n}{2N}\right)\frac{n}{2}\sigma^2 \\ &= N^2 E_{DCS}\left(\frac{1}{2}\sum_{g=1}^{2}\bar{X}_g - \bar{X}\right)^2 + N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2\end{aligned}$$

since x-means $\bar{X}_g$ of all clusters are equal under DCS, therefore first term vanishes and AMSE is given by

$$\text{AMSE}_{DCS}(\hat{Y}_{HT}) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2$$

III. (EBS, HT): Equal-probability balanced sampling (EBS) of n elements (with respect to auxiliary variable x_i) and HT-estimator for finite population total Y. Since, $\hat{X}_{HT} = X$ under balanced sampling, therefore, AMSE of HT-estimator is given by

$$\text{AMSE}_{EBS}(\hat{Y}_{HT}) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2$$

IV. (SRS, GREG): SRS of n elements and GREG-estimator $\hat{Y}_{GR}$ for finite population total Y.

$$\text{AMSE}_{SRS}(\hat{Y}_{GR}) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2$$

From Godambe and Joshi (1965), lower bound of AMSE under the aforementioned population model is given by

$$N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2$$

which is achieved by sampling strategy I., II. and III. For general linear regression model, only approximate expression for variance of GREG-estimator can be obtained, therefore, it achieve the above lower bound approximately.

For DCS, the lower bound of AMSE is only achieve in the ideal situation when dynamic clusters exist and can be formulated. In reality, it might be difficult or not possible to achieve such clusters exactly. Therefore, additional variation is added when cluster means of auxiliary variables or sizes of clusters are not exactly equal. Balanced sampling methods already exist which can achieve the same lower bound of AMSE, although exact balanced sampling is also not always possible. One possible advantage of DCS over

EBS might be unbiased variance estimation because variance estimators for HT-estimator under balanced sampling and for GREG-estimator under SRS may involve some bias which becomes negligible for large sample size. Whereas, in two-stage epsem for large scale surveys, second-stage sample is often relatively small.

2.3 Two-stage epsem by πps-SRS under two-level regression model

A two-level model for the population under two-stage design is a natural choice. In two-level model, there is an additional PSU-level (or cluster level) random error term, also called random effects, therefore it is also known as random effects model. When PSU- and element-level covariates, also called fixed effects, are present in the model, it is usually called mixed effect model. A general terminology of linear mixed models (LMM) is used when random and fixed effects are linearly related with the response variable. A LMM or random effects model assumes a special kind of intra-cluster (or intra-class) correlation which only takes positive values, which is commonly observed in practice. When more levels of random effects are added in the model it is called multi-level regression models.

In this section, first, a two way interaction between multi-stage sampling and multi-level modelling is described in the following paragraphs. After that, in the following subsection, four sampling strategies (mentioned in Section 2.1) involving two-stage epsem, HT- and GREG estimators are compared with respect to their AMSE's under a homoscedastic two-level regression model. Under a homoscedastic model, variance of error terms in the model is constant across the population elements.

Sampling techniques, for instance, generalized regression estimation and balanced sampling are motivated by an underlying super-population model which assume that relationship of a study variable with a set of auxiliary variables can be expressed by a linear regression model. There is not much literature regarding use of two-level (or multi-level) models in sampling design. One particular article which has used LMM at design and estimation stage is described as follows. Breidt and Chauvet (2012) proposed penalized balanced sampling which is motivated by LMM, that is, underlying model for the finite population is LMM. In this sampling methodology, samples are selected using the cube method (see Section 1.4.1) which are balanced with respect to a penalized set of fixed effects (or auxiliary variables) and random effects in the model. In addition, expression for LMM-assisted regression estimator is also given which is calibrated with respect to totals of a set of random effects in addition to fixed effects (or auxiliary variables).

There is a large body of literature associated with incorporating characteristics of multi-stage sample design in the standard estimation techniques for multi-level models. Usually, multilevel models are used to account for the hierarchical structure which exist in the data regardless of the sample design used to collect the data (Goldstein, 1991). Multi-stage sampling design is often motivated by low cost of the sample survey, its stages are are also based on natural hierarchy of the study population. Therefore, it is often of interest to the data analysts analysing multi-stage survey data using multi-level models (Pfeffermann et al., 1998; Jones, 1993). Therefore many estimation methods for multilevel model are motivated by the multi-stage sampling design, where hierarchy of both model and sample design is same. For the analysis of data from multi-stage sample, unequal probability sampling at any stage may induce bias in the model estimates, therefore, unequal probabilities should be taken into account in the estimation (Pfeffermann et al., 1998; Rabe-Hesketh and Skrondal, 2006). To address the above problem, Skinner (1989) gave pseudo likelihood method for single-level models. Pfeffermann et al. (1998) suggested weighting procedures for the estimation of model parameter for two-level models. Rabe-Hesketh and Skrondal (2006) address the problem for linear mixed models in the similar context. Pfeffermann et al. (2006) considered this problem under informative sampling, where inclusion probabilities are related to the response variable conditioned on model covariates. Other contributions in this area also include Skinner and de Toledo Vieira (2007) and Rao et al. (2013).

2.3.1 Comparison of sampling strategies

Let the relationship of response variable y and auxiliary variables $(x_1, ..., x_q)$ is expressed by a two-level random effect model, given by

$$y_{gi} = \mu_{gi} + v_g + e_{gi} \tag{2.10}$$

where $\mu_{gi} = \mathbf{x}_{gi}^\top \boldsymbol{\beta}$ is linear predictor, v_g is random effect associated with gth PSU such that $v_g \sim N(0, \sigma_v^2)$ and e_{gi} is random error associated with ith response value in gth PSU such that $e_{gi} \sim N(0, \sigma_e^2)$. This implies, we have $Var(y_{gi}|\mathbf{x}_{gi}) = \sigma^2 = \sigma_v^2 + \sigma_e^2$ for all $i \in U_g$ and $g \in U_{\mathrm{I}}$; $Cov(y_{gi}, y_{gj}|\mathbf{x}_{gi}\mathbf{x}_{gj}) = \sigma_v^2$ for $i \neq j \in U_g$ and $Cov(y_{gi}, y_{gj}|\mathbf{x}_{gi}\mathbf{x}_{hj}) = 0$ for $g \neq h \in U_{\mathrm{I}}$. The intra-cluster correlation is given by $\rho = \sigma_v^2/\sigma^2$. AMSE's of the four sampling strategies under this two-level regression model are given below.

1. (2Se, HT): From Eq. (A.2) in Appendix A.1, AMSE of HT-estimator under two-stage elements sampling by πps-SRS and population model in Eq. (2.10) is given by

$$\begin{aligned}\text{AMSE}(\hat{Y}_{HT}) = {} & V_1\left(\sum_{g=1}^{n_{\text{I}}}\frac{\mu_g}{\pi_g}\right) + \sum_g \frac{1}{\pi_g}V_2\left(\frac{N_g}{n_g}\sum_{i\in s_g}\mu_{gi}\right) \\ & + \sum_{g=1}^{N_{\text{I}}}\left(\frac{1}{\pi_g}-1\right)N_g\left(1+(N_g-1)\rho\right)\sigma^2 + \sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\left(\frac{N_g}{n_g}-1\right)N_g(1-\rho)\sigma^2\end{aligned}$$

where $\mu_g = \sum_{i\in U_g}\mu_{gi}$, V_1 and V_2 denotes variance functions under the first- and second-stage sampling designs respectively. Under two-stage epsem by πps-SRS (2Se), when $n_g \equiv n_0$, above ASME can be written as

$$\begin{aligned}\text{AMSE}(\hat{Y}_{HT}^{2Se}) = {} & V_1\left(\sum_{g=1}^{n_{\text{I}}}\frac{\mu_g}{\pi_g}\right) + \sum_g \frac{1}{\pi_g}V_2\left(\frac{N_g}{n_0}\sum_{i\in s_g}\mu_{gi}\right) \\ & + \sum_{g=1}^{N_{\text{I}}}\left(\frac{1}{\pi_g}-1\right)N_g\left(1+(N_g-1)\rho\right)\sigma^2 + \sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\left(\frac{N_g}{n_0}-1\right)N_g(1-\rho)\sigma^2\end{aligned} \tag{2.11}$$

2. (2Sc, HT): From Eq. (A.4) in Appendix A.1, AMSE of HT-estimator under two-stage cluster sampling by πps-SRS and population model in Eq. (2.10) is given by

$$\begin{aligned}\text{AMSE}(\hat{Y}_{HT}) = {} & V_1\left(\sum_{g=1}^{n_{\text{I}}}\frac{\mu_g}{\pi_g}\right) + \sum_g \frac{1}{\pi_g}V_2\left(\frac{N_{\text{II}g}}{n_{\text{II}g}}\sum_{k=1}^{n_{\text{II}g}}\mu_{gk}\right) \\ & + \sum_{g=1}^{N_{\text{I}}}\left(\frac{1}{\pi_g}-1\right)N_g\left(1+(N_g-1)\rho\right)\sigma^2 + \sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\left(\frac{N_{\text{II}g}}{n_{\text{II}g}}-1\right)\{N_g(1-\rho) \\ & + \left(\frac{N_{\text{II}g}}{N_{\text{II}g}-1}\right)S^2_{N_{gk}}\rho\Big\}\,\sigma^2\end{aligned}$$

where $\mu_{gk} = \sum_{i\in U_k}\mu_{gki}$ and $S^2_{N_{gk}}$ is variance of second-stage cluster sizes within gth PSU. Under two-stage epsem by πps-SRS (2Sc), when $n_{\text{II}g} \equiv n_{\text{II}0}$ and sub-clusters have same size $N_{gk} = N_0$, which implies $S^2_{N_{gk}} = 0$ and $N_{\text{II}g}N_0 = N_g$, above AMSE can be written as

$$\begin{aligned}\text{AMSE}(\hat{Y}_{HT}^{2Sc}) = {} & V_1\left(\sum_{g=1}^{n_{\text{I}}}\frac{\mu_g}{\pi_g}\right) + \sum_g \frac{1}{\pi_g}V_2\left(\frac{N_g}{n_{\text{II}0}N_0}\sum_{k=1}^{n_{\text{II}0}}\mu_{gk}\right) \\ & + \sum_{g=1}^{N_{\text{I}}}\left(\frac{1}{\pi_g}-1\right)N_g\left(1+(N_g-1)\rho\right)\sigma^2 + \sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\left(\frac{N_g}{n_{\text{II}0}N_0}-1\right)N_g(1-\rho)\sigma^2\end{aligned} \tag{2.12}$$

3. (2Se, GREG): From Eq. (A.3) in Appendix A.1, approximate AMSE's of GREG-estimator under two-stage elements sampling by πps-SRS and two-level population model is given by

$$\mathrm{AMSE}(\hat{Y}_{GR}) \approx \sum_{g=1}^{N_{\mathrm{I}}} \left(\frac{1}{\pi_g} - 1\right) N_g \left(1 + (N_g - 1)\rho\right) \sigma^2 + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \left(\frac{N_g}{n_g} - 1\right) N_g (1-\rho)\sigma^2$$

Under two-stage epsem by πps-SRS (2Se), above AMSE can be written as

$$\mathrm{AMSE}(\hat{Y}_{GR}^{2Se}) \approx \sum_{g=1}^{N_{\mathrm{I}}} \left(\frac{1}{\pi_g} - 1\right) N_g \left(1 + (N_g - 1)\rho\right) \sigma^2 + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \left(\frac{N_g}{n_0} - 1\right) N_g (1-\rho)\sigma^2 \tag{2.13}$$

4. (2Sc, GREG): From Eq. (A.5) in Appendix A.1, approximate AMSE's of GREG-estimator under two-stage cluster sampling by πps-SRS and two-level population model is given by

$$\mathrm{AMSE}(\hat{Y}_{GR}) \approx \sum_{g=1}^{N_{\mathrm{I}}} \left(\frac{1}{\pi_g} - 1\right) N_g \left(1 + (N_g - 1)\rho\right) \sigma^2 + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}} - 1\right) \Big\{ N_g(1-\rho) + \left(\frac{N_{\mathrm{II}g}}{N_{\mathrm{II}g} - 1}\right) V(N_{gk})\rho \Big\} \sigma^2$$

Under two-stage epsem by πps-SRS (2Sc), above AMSE can be written as

$$\mathrm{AMSE}(\hat{Y}_{GR}^{2Sc}) \approx \sum_{g=1}^{N_{\mathrm{I}}} \left(\frac{1}{\pi_g} - 1\right) N_g \left(1 + (N_g - 1)\rho\right) \sigma^2 + \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g} \left(\frac{N_g}{n_{\mathrm{II}0} N_0} - 1\right) N_g (1-\rho)\sigma^2 \tag{2.14}$$

First, let us compare AMSE's of sampling strategies (2Se, HT) and (2Sc, HT) from Eq. (2.11) and Eq. (2.12) respectively. Generally, (2Se, HT) has smaller AMSE than that of (2Sc, HT), because sampling of sub-clusters under 2Sc is generally less efficient than sampling of elements at the second-stage. When size of sub-clusters is such that N_0 is factor of n_0, i.e. $n_{\mathrm{II}0} N_0 = n_0$, then all the corresponding terms in both AMSE's are same except third term; for AMSE under (2Sc, HT), it is given

$$V_2\left(\frac{N_g}{n_{\mathrm{II}0} N_0} \sum_{k=1}^{n_{\mathrm{II}0}} \mu_{gk}\right)$$

which is expected to be larger than third term under (2Se, HT). It involves variance of sub-cluster totals μ_{gk} (or means $\bar{\mu}_{gk} = \mu_{gk}/N_0$), which is zero when sub-cluster means $\bar{\mu}_{gk}$ are same. In Section 2.2, it is demonstrated that it can be achieved if sub-clusters are formulated such that sub-cluster means of auxiliary variables are same. In that case, the sampling strategy (2Sc, HT) would have smaller AMSE than that under (2Se, HT). In another case, if mean of the super-population model, i.e. μ_{gi}, is constant within PSU's, then the two sampling strategies would have the same AMSE's.

Now, let us compare AMSE's of sampling strategies (2Se, GREG) and (2Sc, GREG) from Eq. (2.13) and Eq. (2.14). Under the same condition of second-stage sample size considered, that is $n_{\text{II}0}N_0 = n_0$, the two strategies have same AMSE's. Here, the key fact for this is that mean of the model residuals is constant, which is zero. Therefore, no need to formulate sub-clusters which has same means of auxiliary variables. If mean of the residuals is not constant due to some model miss-specification, then the AMSE's may not be same any more.

The comparison of sampling strategies (2Se, HT) and (2Se, GREG) is obvious, because GREG-estimator is generally more efficient than HT-estimator for given sampling design, provided that auxiliary variables has some correlation with study variable. Same applied for the comparison of sampling strategies (2Sc, HT) and (2Sc, GREG).

In the comparison of sampling strategies above, two conditions were imposed either on the sampling design or population model, given by

- Sub-clusters are equal-sized under two-stage epsem by πps-SRS (2Sc).
- Mean of residuals under the population model is constant (which is zero).

The sensitively of these two conditions is explored by a simulation study in the following section.

2.4 Simulation study

A simulation study is conducted to empirically illustrate the comparison of sampling strategies in the previous section. There are two assumed conditions: constant second-stage sub-cluster sizes, constant mean of regression errors. In addition: the presence of error correlation ρ (or intra-cluster correlation) under the population model may have an 'interaction' if any of the two condition fails.

Finite population: For the simulation study, an example population of households is simulated for Southampton (a postcode area in England) using UK census data for year 2011. This data is published by Office for National Statistics (2011) and available from the website `www.nomisweb.co.uk`. The population consist of $N = 276203$ households with at least one resident. Two-stage sampling is used in many household surveys conducted by ONS (Office for National Statistics), for example, two-stage sample of postal addresses is selected in Family Resource survey (FRS) and Living Cost and Food (LCF) surveys. In these household sruveys, usually postcode sectors are considered as PSU's. Therefore, same PSU's are considered in this simulation study. According to UK census data for year 2011, there are $N_{\mathrm{I}} = 108$ postcode sectors (or PSU's) in the population of households in Southampton. Five auxiliary variables (`hh.compstn`, `hh.size`, `hh.tenure`, `hrp.gender`, `hrp.SeC`) and two study variables (`log.hh.income`, `hh.IntCon`) are considered in this population. The description for these variables is given in Table 2.1 where NS-SeC stands for national statistics socio-economic classification.

Table 2.1: Description of variables in the example finite population of Southampton

Notation	Variable	Type
`log.hh.income`	Natural logarithm of household weekly gross income	Continuous
`hh.IntCon`	High-speed internet connection in the household	Binary
`hh.comp`	Household composition	Categorical
`hh.size`	Household size	Discrete
`hh.tenure`	Household tenure type	Categorical
`hrp.gender`	Gender of household reference person (HRP)	Binary
`hrp.SeC`	NS-SeC of household reference person (HRP)	Categorical

Auxiliary variables are generated using marginal and joint distributions at postcode sector (or PSU) level. The distributions are obtained from UK census 2011 data. Description of categories of the auxiliary variables and population level proportions are given in Appendix A.2.

Study variables are generated using a two-level regression model. Generating variable from the model requires to specify some realistic values of regression coefficients for the model and variance components for PSU-level (σ_v) and element-level (σ_{gi}) error terms. For this purpose, sample survey data from LCF 2017-18 is used which contains all the variables considered in this simulation study. The data set for LCF sample survey 2017-18 was published by Office for National Statistics (2019), and it is available from UK Data Service website given by `ukdataservice.ac.uk`. The categories of covariates in the LCF survey data were not exactly same as in the census data. Approximately similar categories are constructed using the existing categories of the survey data. Regression coefficients are computed by fitting a linear regression model for `log.hh.income` and a logistic regression model for `hh.IntCon` using LCF survey data set. A summary of both fitted models is given in Appendix A.3.

To simulate study variables `log.hh.income` and `hh.IntCon` in the finite population, two-level linear and logistic regression models are used respectively. Fixed parts of the models, μ_{gi}, are specified using regression coefficients of the fitted models based on LCF survey data (as described earlier). The residual variance of the fitted regression models is used as total error variance for the two-level regression models (which are used to generate the study variables). Total variance is partitioned into two components: PSU-level and element-level error variances by a percentage. Different percentages of variance components are used such that five different values of intra-cluster correlation are obtained given by $\rho = (0.01, 0.05, 0.10, 0.15, 0.20)$, see Table 2.3 for values of the total error variance σ^2 and PSU-level error variance component σ_v^2. The choice of percentages is somehow arbitrary but is not completely unrealistic for the population considered in this simulation study. The value of intra-cluster correlation very much depend on the survey variable and sizes of PSU's. There is not much literature which reports values of intra-cluster correlation for the variables and PSU sizes considered in this simulation. However, we found an article by Valliant et al. (2015) which report values of intra-cluster correlation for some study variables collected in household surveys in United States. The values of intra-cluster correlation for these variable range from 0.002 to 0.148 for the PSU's of average size 4253, and from 0.003 to 0.191 for the PSU's of average size 1316. In our simulation, the average size of PSU's is 2557, not very far from them. Therefore, we assumed two-level model with intra-cluster correlation ranges from 0.01 to 0.20.

After generating a finite population, we refitted the regression models in order to obtain coefficient of determination R^2. For logistic regression model, R^2 was calculated as: $1-$ Deviance of fitted model/Deviance of Null model. The average values of R^2 and other descriptives of the survey variables based on many realized finite populations and refitting of the models are reported in Table 2.2.

Table 2.2: Descriptives of two survey variables based on many realizations of the finite populations under the model.

	$P(y=1)$	R^2	min(y)	Q_1	mean(y)	Q_3	Var(y)
`log.hh.income`		0.5450	3.2074	5.9488	6.4612	6.9899	0.5656
`hh.IntCon`	0.9315	0.2428					

Now in order to investigate the sensitivity of GREG-estimator against constant mean of the errors, we generated two more survey variables in addition to the original one such that a fixed effect is added to the model with two level of variation. That is, $y_{gi} = \mu_{gi} + w_{gi} + v_g + e_{gi}$, where w_{gi} is fixed effect with two values of variance σ^2_{w1} and σ^2_{w2}, which are 15% and 30% of variance of model mean, i.e $Var(\mu_{gi})$.

Table 2.3: For linear and logistic regression models, values of total error variance σ^2 and PSU-level error variance component σ^2_v against different ρ-values; σ^2_{w1} and σ^2_{w2} are two values of variance of fixed effect added in the regression models.

		ρ	0.01	0.05	0.10	0.15	0.20		
Variables	Model	σ^2			σ^2_v			σ^2_{w1}	σ^2_{w2}
`log.hh.income`	Linear	0.2577	0.0026	0.0129	0.0258	0.0387	0.0515	0.0462	0.0923
`hh.IntCon`	Logistic	0.4340	0.0044	0.0220	0.0440	0.0660	0.0880	0.9239	1.8478

Two-stage sampling: Samples are selected by two two-stage sampling design: two-stage element sampling by πps-SRS (2Se) which is epsem, and two-stage cluster sampling by πps-SRS (2Sc) which is only epsem if sub-clusters have same size. These designs are described bellow:

- **2Se:** For two-stage element sampling, first- and second-stage sample sizes are approximately same as in FRS survey 2016-17 for England (where, 1417 PSU's were selected from over 12000 and 25 addresses were selected per PSU). In this simulation study, first-stage sampling fraction of PSU's is $f_{\text{I}} = 0.10$ ($n_{\text{I}} = 11$) and $n_g \equiv n_0 = 30$ households are selected from each sampled PSU. Size of the final sample is $n = 330$ households.

- **2Sc:** For two-stage cluster sampling, no natural sub-clusters exit in the population. Therefore, artificial sub-clusters of same size $N_{gk} \equiv N_0 = 5$ are generated such that required sample of 30 households is achieved by selecting $n_{\text{II}g} \equiv n_{\text{II}0} = 6$ SSU's within each sampled PSU. The sizes of 87 PSU's (out of 108) are not multiple of 5. Therefore, each of these PSU's contains one sub-cluster of size not equal to 5 households. This formulation of sub-clusters is denoted by 2Sc1, see Table 2.4.

Formulation of the sub-clusters: In order to investigate the sensitivity against unequal sizes of sub-clusters, we generate two more clustering variables in addition to 2Sc1, denoted by 2Sc2 and 2Sc3. These variables are generated such that average size of SSU's (or sub-clusters) is approximately 5, i.e. $\bar{N}_{gk} \approx 5$, and coefficient of variation, $cv(N_{gk})$ varies from small to large value. In Table 2.4, N_{II} denotes total number of sub-clusters in the population and $\{N_{\text{II}}|N_{gk} \neq 5\}$ denotes number of sub-clusters whose size is not equal to 5 households. Sub-clusters are formulated within each PSU as follow:

1. select a random value c from the discrete uniform distribution $U(a, b)$ as sizes of sub-cluster N_{gk} , where $(a, b) = (5, 5)$ for 2Sc1, $(a, b) = (4, 6)$ for 2Sc2, and $(a, b) = (3, 7)$ for 2Sc3,

2. assign SSU indicator to the c number of elements, starting from the first element of the PSU, subtract the value c from PSU size N_g.

3. repeat steps 1. and 2. until the left over PSU size is less than or equal to 7 (the maximum size a sub-cluster can have). The remaining number of elements are assigned to the last sub-cluster within the PSU.

Above formulations of sub-clusters do not take into account the proximity of population elements. We can say that these are randomly formed clusters, since population data is randomly generated. In terms of AMSE, we expect sampling sub-clusters to be same as element sampling at the second-stage. However, variation in the cluster sizes makes cluster sampling less efficient even when they are randomly formed.

Table 2.4: Descriptives for second-stage clusters in 2Sc design

Formulation	$cv(N_{gk})$	$\min(N_{gk})$	$\max(N_{gk})$	$\text{avg}(N_{gk})$	N_{II}	$\{N_{\text{II}}\|N_{gk} \neq 5\}$
2Sc1	0.014	2	6	4.998	55260	87
2Sc2	0.164	1	6	4.995	55290	36697
2Sc3	0.282	1	7	5.002	55211	44127

Computation of AMSE: Under the given two-level population model $B_1 = 10000$ finite populations are generated. From each population $B_2 = 1$ sample is selected by two-stage element sampling design (2Se) and two-stage cluster sampling designs: 2Sc1, 2Sc2 and 2Sc3. The AMSE's of HT- and GREG-estimator are calculated as $\sum_{B_1} \sum_{B_2} (\hat{Y} - Y)^2/B_1$, where $\hat{Y}$ represent HT- and GREG-estimator. In principle, B_2 should also be a large value, however one sample from each finite population can also give an approximate value of the AMSE's and avoids heavy numeric computations. We treated (2Se, GREG) as baseline strategy, and calculated percent relative efficiency of this strategy based on the AMSE. Percent relative efficiencies of baseline strategy with respect other sampling strategies for different values of ρ and two survey variables `log.hh.income` and `hh.IntCon` are given in Tables 2.5 and 2.6 respectively.

Table 2.5: Percent relative value of AMSE for the strategy (2Se, GREG) with respect to other three strategies for five different values of ρ: for survey variable `log.hh.income`.

ρ	0.01		0.05		0.10		.15		0.20	
Design	HT	GR	HT	GR	HT	GR	HT	GR	HT	GR
`log.hh.income`										
2Se	373.5	100.0	250.7	100.0	198.0	100.0	174.0	100.0	158.5	100.0
2Sc1	382.5	100.4	259.2	98.4	203.5	100.4	180.4	101.0	160.0	98.1
2Sc2	1952.1	102.3	1153.3	100.4	789.7	99.1	614.0	100.3	495.5	98.8
2Sc3	4998.3	102.5	2865.4	101.2	1902.7	102.1	1450.0	100.9	1137.5	100.9
`log.hh.income` with σ^2_{w1}										
2Se	371.1	100.0	250.5	100.0	197.9	100.0	174.0	100.0	158.4	100.0
2Sc1	379.8	100.0	259.2	98.7	203.4	100.4	180.5	101.0	159.9	98.1
2Sc2	1934.3	102.0	1150.6	100.5	788.2	99.2	613.8	100.6	494.4	98.6
2Sc3	4951.1	102.1	2858.0	101.3	1898.6	102.1	1449.1	101.1	1134.8	100.8
`log.hh.income` with σ^2_{w2}										
2Se	364.9	100.0	248.2	100.0	196.8	100.0	173.3	100.0	158.0	100.0
2Sc1	373.1	100.6	255.9	98.3	202.4	100.8	179.7	101.2	159.6	98.2
2Sc2	1899.7	102.5	1138.4	100.4	783.8	99.1	611.7	101.0	494.1	99.0
2Sc3	4859.4	101.8	2826.3	101.2	1887.7	102.4	1442.8	101.3	1133.3	101.0

Results for study variable `log.hh.income`, from Table 2.5, shows that baseline sampling strategy (2Se, GREG) is the most efficient in almost all the scenarios considered in this simulation study. However, as the value of intra-cluster correlation increases, efficiency of baseline strategy tend to decrease with respect to all the other sampling strategies. When size of sub-clusters is equal (i.e. under 2Sc1), there is not much difference between subsampling of sub-clusters and subsampling of elements i.e. sampling strategy (2Se, HT) and (2Sc1, HT) are approximation same. Same applies for (2Se, GREG) and (2Sc1, GREG). As the sizes of sub-clusters differ, subsampling of sub-clusters (i.e. 2Sc2, 2Sc3) loses its efficiency as compared to subsampling of elements (i.e. 2Se); and loss of efficient is much more quick when using HT-estimator (i.e. for (2Sc2, HT) and (2Sc3, HT)) as compared to using GREG-estimator. Adding a fixed w_{gi} in the model (which aims to make error mean non-constant) has no substantial effect on the efficiency of GREG-estimator.

Results for study variable `hh.IntCon`, from Table 2.6, shows that efficiency of baseline sampling strategy (2Se, GREG) with respect to (2Se, HT) is unchanged for different values of intra-cluster correlation. While, its efficiency with respect to (2Sc1, HT) decrease as value of ρ increase; and this decrease is quicker when variation in sizes of sub-clusters increases (i.e. for (2Sc2, HT) and (2Sc3, HT)). Unequal sizes of sub-clusters has not substantial impact on efficiency of baseline sampling strategy. Adding a fixed w_{gi} in the model has some effect on the efficiency of baseline strategy for this study variable, this effect becomes prominent for σ_{w2} and large values of ρ.

Table 2.6: Percent relative value of AMSE for the strategy (2Se, GREG) with respect to other three strategies for five different values of ρ: for survey variable `hh.IntCon`.

ρ	0.01		0.05		0.10		.15		0.20	
Design	HT	GR	HT	GR	HT	GR	HT	GR	HT	GR
`hh.IntCon`										
2Se	121.4	100.0	122.3	100.0	123.5	100.0	123.6	100.0	121.2	100.0
2Sc1	125.6	99.7	123.1	99.4	122.9	100.7	122.7	100.6	124.0	102.0
2Sc2	307.5	101.3	300.5	99.8	295.9	101.7	288.4	98.3	290.0	98.8
2Sc3	650.5	102.0	628.7	96.8	624.1	99.2	605.4	100.7	601.2	98.9
`hh.IntCon` with σ^2_{w1}										
2Se	113.4	100.0	114.1	100.0	114.2	100.0	111.7	100.0	112.2	100.0
2Sc1	115.8	97.7	115.1	98.3	109.6	94.7	113.4	98.0	112.5	99.8
2Sc2	323.3	98.6	324.9	98.6	321.6	98.5	318.8	100.3	311.0	101.5
2Sc3	725.9	98.1	714.9	99.2	708.9	99.9	702.7	99.5	700.2	103.4
`hh.IntCon` with σ^2_{w2}										
2Se	115.3	100.0	115.2	100.0	113.5	100.0	113.7	100.0	115.4	100.0
2Sc1	117.2	99.8	114.3	98.3	114.7	98.0	112.9	96.1	116.3	98.8
2Sc2	254.7	99.2	248.6	96.8	241.8	96.3	241.1	98.8	246.4	98.2
2Sc3	518.7	98.7	503.7	98.7	499.2	96.4	487.8	99.9	493.7	99.9

Results for the both study variables shows that relative values of AMSE under (2Se, HT) with respect to AMSE under (2Sc1, HT) is much smaller than that under sampling strategies (2Sc2, HT) and (2Sc3). Even for some cases, AMSE's under (2Se, HT) and (2Sc1, HT) are approximately same. This is because sub-clusters are randomly formed and they are equal-sized under 2Sc1. Only few PSU's has one sub-cluster with different size under 2Sc1. This demonstrate the efficiency of custom formulation of sub-clusters.

2.5 Conclusions

Four sampling strategies involving two-stage epsem by πps-SRS are explored with respect to their AMSE's. Subsampling of elements in two-stage epsem is commonly used. Sampling strategy involving two-stage epsem and GREG-estimator is preferred strategy. Subsampling of equal-sized sub-clusters is also considered as an alternative sampling design. Some preliminary analysis shows that if one can formulate sub-clusters with certain properties, i.e. equal size and equal means of auxiliary variable, two-stage espem with subsampling of sub-clusters may have an advantage. However, further work is required to achieve such sub-clusters.

In the preferred strategy, GREG-estimator assumed that study variable and auxiliary variables are related through a linear regression model. Sensitivity of GREG-estimator for one type of model miss-specification, that is model errors have non-constant mean, is analysed by a simulation study. The amount of misspecification considered here has not much effect on the efficiency of GREG-estimator. In future, more sophisticated study might be done to find out the threshold of model miss-specification where GREG-estimator can be inefficient as compared to HT-estimator. Furthermore, other types of model miss-specifications can also be analysed, for example, finding out pair of values (R^2, n) for which GREG-estimator can go wrong as compared to HT-estimator, where R^2 is coefficient of determination for the underlying regression model and n is sample size.

Chapter 3

Improving the cube method

3.1 Introduction

When a set of auxiliary variables related to response variables is available before sample selection, then *balanced* samples with respect to the known auxiliary variables tend to have smaller sampling variance than that of unbalanced samples. A sampling design is balanced with respect to auxiliary variables when HT-estimators of auxiliary totals are equal to their respective known auxiliary totals for any balanced sample. *Cube method*, given by Deville and Tillé (2004), is a well known sampling method which aims to select balanced samples with fixed first-order inclusion probabilities. The sample selection procedure under the cube method consists of two phases: flight-phase and landing-phase. In the flight-phase, a stochastic process called balancing martingale transforms the first-order inclusion probabilities into 0 or 1 one-by-one in order to get the sample which is balanced with respect to auxiliary variables. If the flight-phase is able to transform all the inclusion probabilities into 0 or 1, then the sample selection completes. Otherwise, the landing-phase is required which compromises the balancing equations in order to achieve the sample. Invoking the landing-phase of the cube method generally implies that the selected sample is not exactly balanced.

In this chapter, a practical way of improving the cube method is proposed when it is not exactly balanced. The proposed procedure, called *two-step cube* method, is based on repeated sampling from the cube method and then minimizing the average lack of balance in the realized cube samples. Solution for one of the technical components of 'two-step cube method' is tentative and has scalability issue which might be improved in future. A variance approximation under balanced sampling is also proposed which naturally follows the proposed sampling procedure.

Consider the problem of estimating finite population total Y of response variable using HT-estimator $\hat{Y}_{HT}$. For now, usual HT subscript for the HT-estimator is dropped in order to avoid notational complexity. This will be adopted again from Section 3.5 when notation for another estimator is introduced. Assuming that the relationship of response variable and set of auxiliary variables can be expressed using the linear regression model in Eq. (1.5) which has independent normally distributed random errors with mean 0 and variance σ_i^2, the AMSE of HT-estimator $\hat{Y}$ under this linear model and the sampling distribution $p(s)$ is given by

$$\text{AMSE}(\hat{Y}) = E_p\left[\left(\sum_{i\in s}\frac{\mathbf{x}_i}{\pi_i} - \sum_{i\in U}\mathbf{x}_i\right)^{\top}\boldsymbol{\beta}\right]^2 + \sum_{i\in U}\sigma_i^2\left(\frac{1}{\pi_i}-1\right)$$

where E_p denotes expectation with respect to the sampling distribution $p(s)$. In vector notation, the vector $\hat{\mathbf{X}} = \sum_{i\in s}\mathbf{x}_i/\pi_i$ estimates $\mathbf{X} = \sum_{i\in U}\mathbf{x}_i$ with no bias and AMSE$(\hat{Y})$ can be written as

$$\text{AMSE}(\hat{Y}) = \boldsymbol{\beta}^{\top}\boldsymbol{\Lambda}_p\boldsymbol{\beta} + \sum_{i\in U}\sigma_i^2\left(\frac{1}{\pi_i}-1\right) \tag{3.1}$$

where

$$\boldsymbol{\Lambda}_p = E_p\left\{(\hat{\mathbf{X}}-\mathbf{X})^{\top}(\hat{\mathbf{X}}-\mathbf{X})\right\} \tag{3.2}$$

which is variance-covariance matrix of the q-vector $\hat{\mathbf{X}}$ and it represents the *imbalance* with respect to the q auxiliary variables under sampling distribution $p(s)$. The lth diagonal element $\Lambda_{ll}(p)$ of the matrix $\boldsymbol{\Lambda}_p$ represents the sampling variance of the HT-estimator $\hat{X}_l = \sum_{i\in s} x_{il}/\pi_i$ of the auxiliary total $X_l = \sum_{i\in U} x_{il}$ under the sampling distribution $p(s)$, where $l = 1, ..., q$.

The quadratic term $\boldsymbol{\beta}^T\boldsymbol{\Lambda}_p\boldsymbol{\beta}$ represents the contribution of imbalance in the AMSE in Eq. (3.1) with respect to q auxiliary variables under sampling distribution $p(s)$. As the coefficient of determination under the linear model increases the contribution of imbalance in the AMSE becomes more important. When sampling design is exactly balanced the imbalance matrix $\boldsymbol{\Lambda}_p$ is zero and the quadratic term also reduces to zero. The AMSE is minimum when sampling design is exactly balanced and first-order inclusion probabilities are proportional to standard deviations of the error term in the linear model, that is,

$\pi_i \propto \sigma_i$ for all $i \in U$. In practice, error variances under the model are unknown, therefore inclusion probabilities are usually chosen to be proportional to a known size variable.

To be practical, it is difficult to work with the variance-covariance matrix $\mathbf{\Lambda}_p$ as measure of imbalance because of its multidimensionality. Therefore, a scaler measure which represents the variance-covariance matrix is needed. Chen et al. (2016) discussed three types of scaler measures for the multidimensional variability, two of them represented a variance covariance matrix: *trace* and *determinant* of the variance-covariance matrix, and third measure was entropy. Trace and determinant of the variance-covariance matrix are sum and product of its eigenvalues respectively.

In general, entropy measures are directly related to multivariate data rather than variance-covariance matrix of the data. They aims to capture higher order variability in the multivariate data, while variance-covariate matrix is based on only second-order variability. Entropy can be used for non-normal data. Most of the entropy measures are related to variance-covariance matrix when data follow multivariate normal distribution. A common measure of multivariate variability is *joint entropy*. For a N-dimensional random vector $X = (X_1, ..., X_N)$ in real space R^d with probability density function $p_X(x)$, *Shannon's joint entropy* of X is defined as $H(X) = -\int_{R^2} p_X(x) \log p_X(x) d_x$. If X is jointly normal $H(X)$ becomes function of determinant of variance-covariance matrix Σ of the normal distribution (Chen et al., 2016).

In this study, trace of the variance-covariance matrix $\mathbf{\Lambda}_p$ is used as scalar measure for imbalance, named as *total imbalance*, which can be written as

$$tr(\mathbf{\Lambda}_p) = \sum_{l=1}^{q} \Lambda_{ll}(p) = \sum_{l=1}^{q} E_p(\hat{X}_l - X_l)^2 = \sum_{l=1}^{q} \sum_{s \in \Omega} p(s)(\hat{X}_l(s) - X_l)^2 \tag{3.3}$$

where $\hat{X}_l(s)$ denotes the HT-estimate for X_l based on the sample s. The trace is preferred over determinant as scalar measure of the matrix $\mathbf{\Lambda}_p$ because it has advantages regarding the problem under study here. Trace is well defined even when the variance-covariance matrix is singular. That is, when one or more eigenvalues are zero, the determinant is always zero but trace has a value depending on other non-zero eigenvalues. It also shows that trace provides better control over the quadratic term in the AMSE as compared to determinant of the variance-covariance matrix $\mathbf{\Lambda}_p$. Since the aim is to minimize the quadratic term, therefore a better control of the term is helpful.

The basic idea underling the proposed procedure is as follow. When the cube method is not exactly balanced, the relative squared difference is given by $D_l = (\hat{X}_l - X_l)^2/N$.

It becomes negligible for large n and fixed value of q (Deville and Tillé, 2004). In the landing-phase, cube method does not control the total imbalance explicitly. Moreover, when the observed $D_l(s)$ from a sample s selected by cube method is non-negligible, it is not clear whether the given sample was selected unluckily or actually the total imbalance under cube method is large. One can get an empirical estimate of total imbalance by repeated sampling under the cube method. Having selected many samples by the cube method, each sample has different observed value of $\sum_{l=1}^{q} D_l$. A sample with the smallest value of $\sum_{l=1}^{q} D_l$ would be a good choice, but it does not achieve a sampling design with fixed first-order inclusion probabilities. Selecting a random sample out of these samples shall again result into sampling by cube method. A sensible choice would be to select a random sample out of the many cube samples with a different sampling distribution such that the total imbalance is smaller than that of the cube method. Implementation of this idea is given in the next section.

Rest of the chapter is arranged as follow. In the Section 3.2, the proposed sampling procedure is given. Its theoretical properties are discussed in Section 3.3. In the Section 3.4, a simulation study is conducted which compares the propose sampling procedure with the cube method. A methodology for the estimation of sampling variance under balanced sampling is proposed in Section 3.5. A simulation study in Section 3.6 assesses the performance of proposed variance estimators and compares with an estimator from the literature. Section 3.7 gives conclusions and some future work directions.

3.2 Proposed procedure: two-step cube

The proposed procedure consists of two steps: first, a finite number of samples are selected using cube method and empirical value of total imbalance is computed; second, a sampling distribution for the selected cube samples is determined by minimizing the empirical total imbalance. However, for the minimization problem in the second step only a tentative solution is given. The finite set of samples selected using the cube method is named as *realized cube sample space*, and *empirical distribution* over the realized cube sample space is calculated. First-order inclusion probabilities under the new sampling distribution, determined by minimization, are same as those calculated under the empirical distribution implied by the cube method. When a sample is selected from the realised cube sample space using the sampling distribution which minimises the empirical total imbalance, then total imbalance is expected to be equal or smaller than than of cube method.

Let $c(s)$ is set of selection probabilities associated with samples under the cube method, and denotes the sampling distribution implied by the cube method. Let $\mathbf{\Lambda}_c$ and $tr(\mathbf{\Lambda}_c)$ denote the variance-covariance matrix of vector $\hat{\mathbf{X}}$ and the total imbalance under the cube method respectively. Expressions for $\mathbf{\Lambda}_c$ and $tr(\mathbf{\Lambda}_c)$ can be obtained by substituting $c(s)$ for $p(s)$ in Eq. (3.2) and Eq. (3.3) respectively. The two steps for the proposed procedure are given in the following:

Step 1: Let K samples of fixed size n are selected from U by using cube method and $\Omega_K = \{s_k; k = 1, ..., K\}$ denotes the realized cube sample space of size K, Ω_K may contain a sample multiple times, therefore it not a set of distinct samples. The minimum size of Ω_K (or minimum value of K) is such that all the population units are contained in it. The impact of small and large values of K would become more clear in the next section where theoretical properties of the proposed procedure are discussed. Let the vector $\boldsymbol{\lambda} = [\lambda_k = 1/K]$ denotes the empirical distribution over Ω_K which assigns a probability mass of $1/K$ to each sample s_k in Ω_K. An empirical estimate for the variance-covariance matrix $\mathbf{\Lambda}_c$ based on $\boldsymbol{\lambda}$ is given by

$$\hat{\mathbf{\Lambda}}_{\boldsymbol{\lambda}} = E_{\boldsymbol{\lambda}}\left\{(\hat{\mathbf{X}} - \mathbf{X})^\top(\hat{\mathbf{X}} - \mathbf{X})\right\}$$

where $E_{\boldsymbol{\lambda}}$ denotes expectation with respect to the empirical distribution $\boldsymbol{\lambda}$. Similarly, an empirical estimate of total imbalance under the cube method is given by

$$tr(\hat{\mathbf{\Lambda}}_{\boldsymbol{\lambda}}) = \sum_{l=1}^{q} \hat{\Lambda}_{ll}(\boldsymbol{\lambda}) = \sum_{l=1}^{q}\sum_{k=1}^{K} \lambda_k(\hat{X}_l(s_k) - X_l)^2 = \sum_{l=1}^{q} \frac{1}{K}\sum_{k=1}^{K}(\hat{X}_l(s_k) - X_l)^2$$

where $\hat{X}_l(s_k)$ is HT-estimate of the auxiliary total X_l based on the sample s_k, and $\hat{\Lambda}_{ll}(\boldsymbol{\lambda})$ is lth diagonal element of $\hat{\mathbf{\Lambda}}_{\boldsymbol{\lambda}}$ which simply represent empirical estimate of sampling variance of HT-estimator $\hat{X}_l$ under the cube method. First-order inclusion probabilities implied by the empirical distribution $\boldsymbol{\lambda}$ are given by

$$\pi_i(\boldsymbol{\lambda}) = \sum_{k=1}^{K} I_{i\in s_k}\lambda_k = \frac{1}{K}\sum_{k=1}^{K} I_{i\in s_k}$$

where $I_{i\in s_k}$ is indicator variable which takes value 1 if $i \in s_k$, 0 otherwise, and $i \in U$.

Step 2: Let $\boldsymbol{\lambda}^*(\neq \boldsymbol{\lambda})$ denotes another sampling distribution (which needs to be determined here) over the realized cube sample space Ω_K such that $tr(\hat{\mathbf{\Lambda}}_{\boldsymbol{\lambda}^*}) \leq tr(\hat{\mathbf{\Lambda}}_{\boldsymbol{\lambda}})$ and $\pi_i(\boldsymbol{\lambda}^*) = \pi_i(\boldsymbol{\lambda})$ for all $i \in U$, where $\pi_i(\boldsymbol{\lambda}^*)$'s are first-order inclusion probabilities and

$tr(\hat{\mathbf{\Lambda}}_{\boldsymbol{\lambda}^*})$ is total imbalance implied by the sampling distribution $\boldsymbol{\lambda}^*$ for the given realized cube sample space Ω_K. Expression for $tr(\hat{\mathbf{\Lambda}}_{\boldsymbol{\lambda}^*})$ is given by

$$tr(\hat{\mathbf{\Lambda}}_{\boldsymbol{\lambda}^*}) = \sum_{l=1}^{q} \hat{\Lambda}_{ll}(\boldsymbol{\lambda}^*) = \sum_{l=1}^{q} \sum_{k=1}^{K} \lambda_k^* (\hat{X}_l(s_k) - X_l)^2$$

where matrix $\hat{\mathbf{\Lambda}}_{\boldsymbol{\lambda}^*}$ is the variance-covariance matrix of vector $\hat{\mathbf{X}}$ under the sampling distribution $\boldsymbol{\lambda}^*$. In order to obtain the sampling distribution $\boldsymbol{\lambda}^*$ over Ω_K, an optimization problem is formulated in the following:

Minimize:

$$C(\boldsymbol{\lambda}^*) = tr(\hat{\mathbf{\Lambda}}_{\boldsymbol{\lambda}^*}) = \sum_{l=1}^{q} \sum_{k=1}^{K} \lambda_k^* (\hat{X}_l(s_k) - X_l)^2$$

Subjected to:

$$(i).\ \ 0 < \lambda_k^* < 1,\ \forall k,$$
$$(ii).\ \ \sum_{k=1}^{K} \lambda_k^* = 1,$$
$$(iii).\ \ \pi_i(\boldsymbol{\lambda}^*) = \pi_i(\boldsymbol{\lambda}),\ \forall i \in U.$$

where $C(\boldsymbol{\lambda}^*)$ denotes cost function for the solution vector $\boldsymbol{\lambda}^*$ in the above optimization problem. In the optimization problem above, first and second constraints ensure that λ_k^* is a sampling distribution while third constraint ensures that first-order inclusion probabilities under the proposed procedure are same as those computed under the cube method.

For the above optimization problem, a stochastic global optimization algorithm known as *simulated annealing* algorithm is used. Mullen (2014) discussed implementations for different optimization algorithms including simulated annealing algorithm which are available in two different packages of `R` statistical software (R Core Team, 2022). First, `R`-function `optim(method="SANN")` in `R`-package `stats`; second, `R`-function and package `GenSA` (Xiang et al., 2013). These implementations for the simulated annealing algorithm rarely produce a sensible solution for this particular problem because of its complexity. Simulated annealing algorithm is easy to program and manipulate manually in `R`. An `R`-code is created for the simulated annealing algorithm which is giving a reasonable solution, but has scalability problem as the population size becomes large. If a better algorithm is available, one can plug-in but the idea of the proposed procedure does not change. A brief description of simulated annealing and algorithm used for this problem are given in Appendix A.4.

3.3 Theoretical properties of the proposed procedure

In this section, it is shown that the proposed sampling procedure achieves the same what cube achieves in terms of first-order inclusion probabilities, but has smaller or equal total imbalance. The two-step cube method achieve these results for any value of K. For a given application of the two-step cube method, an empirical estimate of total imbalance is minimised and empirical estimates of fixed inclusion probabilities are achieved base on K samples. Therefore, relative gain with respect to cube method is just an empirical estimate of the true gain. As value of K increases it brings more confidence in the achieved gain. Furthermore, one can assess the potential gain in terms of realised total imbalance under the two-step cube method for a given application.

3.3.1 First-order properties

The first-order inclusion probability implied by the empirical distribution $\boldsymbol{\lambda}$ can be written as

$$\pi_i(\boldsymbol{\lambda}) = E_{\boldsymbol{\lambda}}(I_{(i\in s)}|\Omega_K) = \frac{1}{K}\sum_{s\in\Omega_K} I_{(i\in s)} = \frac{1}{K}\sum_{k=1}^{K} I_{(i\in s_k)}$$

which is expectation of $I_{(i\in s)}$ with respect to $\boldsymbol{\lambda}$. It may, therefore, be referred to as an empirical estimator of the sample inclusion probability induced by the cube method, denoted by

$$\pi_i(c) = E_c(I_{(i\in s)}) = \pi_i$$

see Deville and Tillé (2004), where E_c denotes expectation with respect to sampling distribution $c(s)$ and π_i is ith design inclusion probability. Under the two-step cube method $\boldsymbol{\lambda}^*$ is probability mass function over Ω_K such that $\pi_i(\boldsymbol{\lambda}^*) = \pi_i(\boldsymbol{\lambda})$. It can also be written as

$$\pi_i(\boldsymbol{\lambda}^*) = E_{\boldsymbol{\lambda}^*}(I_{(i\in s)}|\Omega_K) = \sum_{s\in\Omega_K} I_{i\in s}\lambda^*(s) = \frac{1}{K}\sum_{s\in\Omega_K} I_{(i\in s)} = E_{\boldsymbol{\lambda}}(I_{(i\in s)}|\Omega_K) = \pi_i(\boldsymbol{\lambda}) \quad (3.4)$$

for all $i \in U$, where Eq. (3.4) holds by definition under two-step cube method. It also implies that

$$
\begin{aligned}
E_{\boldsymbol{\lambda}^*}[\hat{X}_l(s)|\Omega_K] = E_{\boldsymbol{\lambda}^*}\left(\sum_{i\in U}\frac{x_{il}}{\pi_i}I_{i\in s}|\Omega_K\right) &= \sum_{i\in U}\frac{x_{il}}{\pi_i}E_{\boldsymbol{\lambda}^*}\left(I_{i\in s}|\Omega_K\right) \\
&= \sum_{i\in U}\frac{x_{il}}{\pi_i}E_{\boldsymbol{\lambda}}\left(I_{i\in s}|\Omega_K\right) = E_{\boldsymbol{\lambda}}[\hat{X}_l(s)|\Omega_K] = \hat{\hat{X}}_l \qquad (3.5)
\end{aligned}
$$

where

$$\hat{\hat{X}}_l = \frac{1}{K}\sum_{s\in\Omega_K}\hat{X}_l(s)$$

Result 3.1. The two-step cube sample inclusion probability is π_i, for $i \in U$.
Proof: Let $c^*(s)$ denotes the sampling distribution implied by the proposed two-step cube method. Denote by E_{c^*} the expectation with respect to the sampling distribution $c^*(s)$ under the two-step cube method. Expectation of $I_{(i\in s)}$ random variable under two-step cube is given by

$$E_{c^*}(I_{(i\in s)}) = E_{\Omega_K}[E_{\boldsymbol{\lambda}^*}(I_{(i\in s)}|\Omega_K)]$$

Using Eq. (3.4)

$$E_{c^*}(I_{(i\in s)}) = E_{\Omega_K}[E_{\boldsymbol{\lambda}}(I_{(i\in s)}|\Omega_K)] = E_{\Omega_K}\left(\frac{1}{K}\sum_{s\in\Omega_K}I_{(i\in s_k)}\right) = \frac{1}{K}\sum_{k=1}^{K}E_{\Omega_K}(I_{(i\in s_k)})$$

where E_{Ω_K} is expectation over many Ω_K's under the cube method. Theoretically, an infinite number of Ω_K's can be selected which results into an infinite number of cube samples of size n . Computation of expectation based on infinite number of cube samples results into true expectation under the cube method, that is, E_{Ω_K} is equivalent to E_c. This does not depend on the value of K. For example, when $K = 1$, selecting an infinite number of E_{Ω_K}'s is same as selecting an infinite number of cube samples. Therefore, it follows

$$E_{c^*}(I_{(i\in s)}) = \frac{1}{K}\sum_{k=1}^{K}E_c(I_{(i\in s_k)}) = \frac{1}{K}\sum_{k=1}^{K}\pi_i = \pi_i$$

□

Result 3.2. Under the two-step cube method $E_{c^*}(\hat{X}_l) = X_l$, where $l = 1, ..., q$.
Proof. Expectation of $\hat{X}_l$ under two-step cube is given by

$$E_{c^*}(\hat{X}_l) = E_{\Omega_K}[E_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K)]$$

Using result from Eq. (3.5)

$$\begin{aligned} E_{c^*}(\hat{X}_l) &= E_{\Omega_K}(\hat{\hat{X}}_l) \\ &= E_{\Omega_K}\left(\frac{1}{K}\sum_{s\in\Omega_K}\hat{X}_l(s)\right) = \frac{1}{K}\sum_{s\in\Omega_K}E_{\Omega_K}\left(\hat{X}_l(s)\right) \\ &= \frac{1}{K}\sum_{s\in\Omega_K}E_c(\hat{X}_l) = \frac{1}{K}\sum_{s\in\Omega_K}X_l = X_l \end{aligned}$$

□

3.3.2 Second-order properties

The conditional imbalance $\hat{\Lambda}_l(\boldsymbol{\lambda}^*)|\Omega_K$ is the mean squared error (MSE) of $\hat{X}_l$ under the empirical distribution $\boldsymbol{\lambda^*}$ over Ω_K, whereas the conditional imbalance $\hat{\Lambda}_l(\boldsymbol{\lambda})|\Omega_K$ is that with respect to $\boldsymbol{\lambda}$, i.e.

$$\begin{aligned} \hat{\Lambda}_l(\boldsymbol{\lambda}^*) &= \sum_{k=1}^{K}\lambda_k^*(\hat{X}_{lk}-X_l)^2 = \mathrm{MSE}_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K) = V_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K) + E^2_{\boldsymbol{\lambda}^*}(\hat{X}_l - X_l|\Omega_K) \\ \hat{\Lambda}_l(\boldsymbol{\lambda}) &= \sum_{k=1}^{K}\lambda_k(\hat{X}_{lk}-X_l)^2 = \mathrm{MSE}_{\boldsymbol{\lambda}}(\hat{X}_l|\Omega_K) = V_{\boldsymbol{\lambda}}(\hat{X}_l|\Omega_K) + E^2_{\boldsymbol{\lambda}}(\hat{X}_l - X_l|\Omega_K) \end{aligned}$$

The two-step cube method minimizes the total conditional imbalance, which implies the following

$$tr(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\lambda}^*}) = \sum_{l=1}^{q}\hat{\Lambda}_l(\boldsymbol{\lambda}^*) \le \sum_{l=1}^{q}\hat{\Lambda}_l(\boldsymbol{\lambda}) = tr(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\lambda}}) \tag{3.6}$$

From Eq. (3.4), $\pi_i(\boldsymbol{\lambda}^*) = \pi_i(\boldsymbol{\lambda})$ for all $i \in U$, which implies

$$\begin{aligned} E_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K) &= E_{\boldsymbol{\lambda}}(\hat{X}_l|\Omega_K) \\ E^2_{\boldsymbol{\lambda}^*}(\hat{X}_l - X_l|\Omega_K) &= E^2_{\boldsymbol{\lambda}}(\hat{X}_l - X_l|\Omega_K) \end{aligned}$$

Therefore, Eq. (3.6) becomes

$$\sum_{l=1}^{q}V_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K) \le \sum_{l=1}^{q}V_{\boldsymbol{\lambda}}(\hat{X}_l|\Omega_K) \tag{3.7}$$

Now consider unconditional imbalance (or MSE of $\hat{X}_l$). Since $E(\hat{X}_l) = X_l$ under the both cube and two-step cube methods, therefore MSE and variance are equivalent under both methods, given by

$$\text{MSE}_{c^*}(\hat{X}_l) = V_{c^*}(\hat{X}_l) = V[E_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K)] + E[V_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K)]$$
$$\text{MSE}_c(\hat{X}_l) = V_c(\hat{X}_l) = V[E_{\boldsymbol{\lambda}}(\hat{X}_l|\Omega_K)] + E[V_{\boldsymbol{\lambda}}(\hat{X}_l|\Omega_K)]$$

By minimising the total condition imbalance, the two-step cube method aims to reduce the unconditional total imbalance (or combined unconditional MSE of $\hat{X}_l$'s), i.e.

$$\sum_{l=1}^{q} \text{MSE}_{c^*}(\hat{X}_l) \leq \sum_{l=1}^{q} \text{MSE}_c(\hat{X}_l) \tag{3.8}$$

which will be proved in Result 3.4 later in this section.

The inequality (3.7) represents the estimated gain of the two-step cube regarding (3.8). The two-step cube is expected to have total imbalanced equal or smaller than cube regardless of the value of K. Whereas, with larger K, the gain is estimated more precisely, because of smaller $V(V_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K))$; indeed, if $K \to \infty$, then the gain is evaluated exactly.

Result 3.3. The expectation of $\frac{K}{K-1}V_{\boldsymbol{\lambda}}(\hat{X}_l|\Omega_K)$ over repetitions of Ω_K is equal to $\Lambda_l(c)$, that is,

$$E_{\Omega_K}\left(\frac{K}{K-1}V_{\boldsymbol{\lambda}}(\hat{X}_l|\Omega_K)\right) = \Lambda_l(c)$$

Proof: Given any $l = 1, ..., q$, let $Z_{l,k} = \hat{X}_l(s_k)$, for $k = 1, ..., K$. By construction, $Z_{l,1}, ..., Z_{l,K}$ is an IID sample of Z_l, where $Z_l = \hat{X}_l$ is the HT-estimator of X_l based on a cube sample, such that by definition

$$V_c(Z_l) = \Lambda_l(c)$$

is the corresponding cube sampling variance of $\hat{X}_l$. It follows that

$$E_{\Omega_K}\left(\frac{1}{K-1}\sum_{k=1}^{K}(Z_{l,k} - \bar{Z}_l)^2\right) = \Lambda_l(c) \text{ where } \bar{Z}_l = \sum_{k=1}^{K}\frac{Z_{l,k}}{K} = \hat{\bar{X}}_l$$

Hence,

$$E_{\Omega_K}[V_{\boldsymbol{\lambda}}(\hat{X}_l|\Omega_K)] = \frac{K-1}{K}E_{\Omega_K}\left(\frac{1}{K-1}\sum_{k=1}^{K}[\hat{X}_l(s_k) - \hat{\bar{X}}_l]^2\right) = \frac{K-1}{K}\Lambda_l(c)$$

□

Result 3.4. The total imbalance under the two-step cube sampling is smaller or equal to that under the cube method, that is,

$$tr(\boldsymbol{\Lambda}_{c^*}) \leq tr(\boldsymbol{\Lambda}_c)$$

Proof: Let the $\boldsymbol{\Lambda}_{c^*}$ and $tr(\boldsymbol{\Lambda}_{c^*})$ be the variance-covariance matrix and total imbalance of vector $\hat{\mathbf{X}}$ under the two-step cube method respectively. Given any Ω_K, by construction of the two-step cube we have Eq. (3.7) so that

$$E_{\Omega_K}\left(\sum_{l=1}^{q} V_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K)\right) \leq E_{\Omega_K}\left(\sum_{l=1}^{q} V_{\boldsymbol{\lambda}}(\hat{X}_l|\Omega_K)\right) = \frac{K-1}{K}\sum_{l=1}^{q}\Lambda_l(c)$$

where the right-hand side follows from Result 3.3. For the left hand side, we notice that lth diagonal element of the imbalance matrix under two-step cube sampling can be written as

$$\begin{aligned} V_{c^*}(\hat{X}_l) &= E_{\Omega_K}[V_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K)] + V_{\Omega_K}[E_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K)] = E_{\Omega_K}[V_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K)] + V_{\Omega_K}(\hat{\hat{X}}_l) \\ &= E_{\Omega_K}[V_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K)] + \frac{1}{K}V_c(\hat{X}_l) = E_{\Omega_K}[V_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K)] + \frac{1}{K}\Lambda_l(c) \end{aligned}$$

It follows that

$$E_{\Omega_K}\left(\sum_{l=1}^{q} V_{\boldsymbol{\lambda}^*}(\hat{X}_l|\Omega_K)\right) = \sum_{l=1}^{q} V_{c^*}(\hat{X}_l) - \frac{1}{K}\sum_{l=1}^{q}\Lambda_l(c) = \sum_{l=1}^{q}\Lambda_l(c^*) - \frac{1}{K}\sum_{l=1}^{q}\Lambda_l(c)$$

$$\sum_{l=1}^{q}\Lambda_l(c^*) - \frac{1}{K}\sum_{l=1}^{q}\Lambda_l(c) \leq \frac{K-1}{K}\sum_{l=1}^{q}\Lambda_l(c)$$

$$\sum_{l=1}^{q}\Lambda_l(c^*) \leq \sum_{l=1}^{q}\Lambda_l(c)$$

$$tr(\boldsymbol{\Lambda}_{c^*}) \leq tr(\boldsymbol{\Lambda}_c)$$

□

3.3.3 Assessment of potential gains in practice

When it is assumed that response variables and auxiliary variables are related under a linear model with independent errors, response variable can be expressed as a linear function of auxiliary variables and finite population residuals, given by $y_i = \mathbf{x}_i^\top \boldsymbol{B} + e_i$, where $i \in U$. Sampling variance of HT-estimator can be partitioned into two terms as follows

$$V_p(\hat{Y}) = \boldsymbol{B}^\top \boldsymbol{\Lambda}_p \boldsymbol{B} + V_p(\hat{e})$$

where $\boldsymbol{B}$ is vector of finite population regression coefficients and $\hat{e}$ is HT-estimator of finite population total of population residuals. In the above equation, first term is a quadratic form which represent sampling variance explained by auxiliary variables, and second term is sampling variance of HT-estimator for finite population residual total. Before selection of the sample, total imbalance can be estimated empirically under both two-step cube and cube methods. Under the two-step cube, reducing total imbalance aims to reduce the first term. However, the second term under two-step cube method might be different from that under the cube method because this term can vary from one finite population to other under the same model. It is hard to compare second term under two methods unless some historic or proxy values of the residuals are available.

In order to say something about gain in terms of sampling variance of $\hat{Y}$, the magnitude of gain in terms of estimated total imbalance is important for a given application of the two-step cube. When gain in terms of total imbalance is large, the difference of first terms under two sampling methods is much larger than difference of the second terms, therefore, comparison of second terms under the two methods becomes relatively unimportant. In that case, it is worth making an assessment for potential gain in terms of total imbalance. For a given population, one can make a comparison of the following quantities.

- **Cube:** An empirical distribution of estimated total imbalance can be obtained under the cube method. It is easy to obtain many realized cube sample spaces Ω_K's and calculate total imbalance for each of them. In this way, an empirical distribution of total imbalance under the cube method is obtained.
- **Two-step cube:** In practice, it is hard to replicate two-step cube method due to time limitation, since the algorithm is time consuming. Therefore, total imbalance can be calculated for one realization of two-step cube method. In this way, we have only one estimate of the total imbalance under the two-step cube.
- **Comparison:** Compare the one estimated value under two-step cube against the

empirical distribution under cube. Intersection of the left tail area of the empirical distribution (under cube method) and the point estimate (under two-step cube method) gives an idea about how likely the two-step cube is better than the cube method. This will be demonstrated in simulation study for two-step cube. For this comparison, a large value of K brings more confidence in terms of gain under the two-step cube.

3.4 Simulation study for two-step cube method

A simulation study is conducted in order to compare the efficiency of cube method and proposed procedure, the two-step cube method. Four types of comparison are made as described bellow

- First, following the simulation study from (Deville and Tillé, 2004), cube and two-step methods are compared based mean squared errors (MSE) relative to that under unbalanced sampling;
- Second, a comparison is drawn between two-step cube and cube sampling from ordered population (Leuenberger et al., 2022);
- Third, potential gain of two-step cube method with respect to total imbalance is analysed;
- Fourth, cube and two-step cube methods are compared with respect to AMSE of HT-estimator.

3.4.1 Comparison based on mean squared errors

Deville and Tillé (2004) conducted a simulation study to compare cube method with πps sampling based on values of MSE of HT- and GREG-estimators under the cube method relative to those under πps sampling. Same criteria is followed in this simulation study. Therefore, samples are selected under three sampling designs give by

- unbalanced sampling by probability proportion to size sampling without replacement (πps sampling),
- balanced sampling by cube method, and

- balanced sampling by two-step cube method.

In this simulation study, a data set MU284 is used from (Särndal et al., 1992, p. 652-9) which is related to 284 municipalities in Sweden. Description of variables and a correlation matrix for the data set are given in Tables A.6 and A.7 of Appendix A.5 respectively. The data set MU284 is available from the `R`-package `sampling` (Tillé and Matei, 2021). Four largest municipalities are removed from the data set because these municipalities have one inclusion probability when selecting sample with probability proportion to size variable. The remaining 280 municipalities are grouped into $N = 50$ clusters based on clustering variable `CL` already given in the data set. This becomes a modified version of the 'Clustered MU284' from (Särndal et al., 1992, p. 600-1). This modification comes from the simulation study in Deville and Tillé (2004) that was carried out with two additional variables `SOC82` and `R85` which are not available in the current version of data set MU284. The sample size is $n = 20$ and first-order inclusion probabilities are proportion to the variable `P75`, total population in 1975, for all the three sampling design mentioned above. The simulation study is based on $B = 5000$ repeated samples and is performed as follow:

πps sampling: Select B samples under unbalanced sampling which is performed using cube algorithm with inclusion probabilities as the only balancing variable which is equivalent to πps sampling.

Cube sampling: Select B samples using the cube method with three balancing variables `P75, RMT85` and `ME84`. (Note: there were four balancing variables in (Deville and Tillé, 2004)'s simulation, `P75, RMT85, SOC82` and `ME84`). Cube samples are selected using fast implementation of cube method available in `R`-package `BalancedSampling` (Grafström and Lisic, 2019).

Two-step cube sampling: To select B samples under the two-step cube method, following three steps are repeated B times:

(i). Select $K = 1000$ samples using cube method which are considered as realized cube sample space Ω_K.

(ii). Obtain a sampling distribution $\boldsymbol{\lambda}^*$ over Ω_K under two-step cube method using simulated annealing algorithm.

(iii). Select a sample from Ω_K based on $\boldsymbol{\lambda}^*$ sampling distribution.

An explanation for above three step is as follow. Theoretically the sampling distribution $\boldsymbol{\lambda}^*$ for a given Ω_K is one realization of the two-step cube method. In order to make a fair comparison, B realizations of two-step cube are required. Therefore, we obtained B realized cube sample spaces $\Omega_{K,1}, ..., \Omega_{K,B}$ each of size K. For each $\Omega_{K,b}$, we obtain a sampling distribution $\boldsymbol{\lambda}^*{}_b$ under the two-step cube method using simulated annealing algorithm, where $b = 1, ..., B$. From each $\Omega_{K,b}$, a sample is selected based on its respective sampling distribution $\boldsymbol{\lambda}^*{}_b$, where $b = 1, ..., B$. In this way, we get B samples under the two-step cube method.

Based on the B samples selected above using three sampling methods, following quantities are computed:

- Empirical MSE of HT- and GREG-estimators of the finite population totals of all the variables (both response and auxiliary). The values of empirical MSE's under the cube method (Cube) and two-step cube (2Cube) method relative to those under πps sampling are given in Table 3.1.

- For a convenient comparison of cube and two-step cube methods, percent relative efficiency (PRE) of two-step cube method is also calculated, see Table 3.1.

- Furthermore, estimates of Monte Carlo errors (MCE) for empirically estimated MSE's are also calculated. Relative values of MCE's are given in Table 3.2. Values for MCE are calculate as follow. Let $g = (\hat{X}_{HT} - X)^2$ which can be calculated for each of B samples selected independently under the three designs mentioned above. As an Mote Carlo estimate we calculated expectation of g, given by $\bar{g} = \sum_{b=1}^{B} (\hat{X}_{HT} - X)^2$ which is MSE($\hat{X}_{HT}$). A standard error of $\bar{g}$ is given by $\sigma_{\bar{g}} = \sigma_g/\sqrt{B}$ which represents the Monte Carlo error of MSE($\hat{X}_{HT}$). This quantity can be estimated by

$$\hat{\sigma}_{\bar{g}} = \frac{SD(g)}{\sqrt{B}} = \frac{1}{\sqrt{B}(B-1)} \sum_{b=1}^{B} (g_b - \bar{g})^2 = \text{MCE}$$

  and Relative MCE is computed as

$$\frac{\text{MCE}}{\text{Empirical MSE}}.$$

Table 3.1: Empirical MSE's under cube (Cube) and two-step cube (2Cube) relative to the values under πps sampling; Percent relative efficiency (PRE) of two-step cube compared to the cube method.

	HT-estimator				GREG-estimator			
	πps	Cube	2Cube	PRE	πps	Cube	2Cube	PRE
Auxiliary variables:								
P75	-	-	-	-	-	-	-	-
RMT85	1.00	0.12	0.05	245.49	0.00	0.00	0.00	175.35
ME84	1.00	0.12	0.03	397.51	0.00	0.00	0.00	183.65
Survey variables:								
P85	1.00	0.71	0.68	105.08	0.79	0.67	0.65	101.84
CS82	1.00	0.91	0.89	102.57	1.05	0.91	0.89	102.39
S82	1.00	0.79	0.75	105.72	0.81	0.76	0.74	103.34
REV84	1.00	1.09	1.08	101.63	0.95	1.07	1.05	101.30
SIZE	1.00	0.85	0.81	105.02	0.85	0.82	0.80	103.22
S82-CS82-SS82	1.00	0.76	0.72	106.11	0.76	0.73	0.71	103.05
CS82-SS82	1.00	0.89	0.83	106.69	0.84	0.86	0.82	105.29

In Table 3.1, first row is omitted because values of MSE's of two estimators (HT and GREG) are zero (or near zero) as P75 is size variable under the three designs and is exactly balanced. While comparing the relative values of the MSE's for other variables, the two-step cube method with GREG-estimator has the smallest MSE's among all the six strategies. For HT-estimator, the imbalance (or MSE) with respect two auxiliary variables under two-step cube is smaller than that under the cube method, it is reduced by 245% for RMT85 and by 397% for ME84.

As expected, MSE's of the GREG-estimators for three auxiliary totals are approximately zero because GREG-estimators are calibrated for auxiliary totals. PRE reported in the table shows that two-step cube have smaller imbalances with respect to auxiliary variables as compared to cube method, although in this case, values of imbalances are very small and of less importance to compare. Similarly, MSE's of GREG-estimators for totals of survey variables under two-step cube method are smaller than that of cube method. Therefore, results of this simulation study clearly suggest that two-step cube have ability to improve efficiency of cube method.

Since values reported in Table 3.1 are empirical estimates of the true MSE's, therefore, an estimate of Monte Carlo error is also computed for each empirical estimate of MSE. In Table 3.2, values of relative MCE are not very different for cube and two-step cube methods, however GREG-estimator tends to have larger error values as compared to HT-estimator. It might be due to additional variation of calibration weights in the GREG-estimator. Furthermore, small values of relative MCE (ranging from 2% to 4.6%) indicate that choice of value for $K = 1000$ is good enough. A smaller value of K would have shown larger variability in the empirically estimated MSE's and resulted into reduced confidence in the estimated gain under the two-step cube.

Table 3.2: Relative Monte Carlo error (MCE) of empirically estimated MSE's of HT- and GREG-estimators under πps sampling, cube and two-step cube methods.

	HT-estimator			GREG-estimator		
	πps	Cube	2Cube	πps	Cube	2Cube
Auxiliary variables:						
P75	0.00	0.00	0.00	0.00	0.00	0.00
RMT85	0.019	0.021	0.025	0.041	0.041	0.040
ME84	0.020	0.020	0.032	0.045	0.046	0.036
Survey variables:						
P85	0.020	0.020	0.020	0.020	0.020	0.020
CS82	0.022	0.022	0.021	0.020	0.022	0.021
S82	020	0.021	0.021	0.019	0.021	0.021
REV84	0.022	0.022	0.022	0.020	0.022	0.022
SIZE	0.021	0.022	0.021	0.020	0.021	0.021
S82-CS82-SS82	0.021	0.022	0.022	0.019	0.022	0.021
CS82-SS82	0.020	0.021	0.020	0.020	0.021	0.020

3.4.2 Comparison with cube sampling from ordered finite population

Leuenberger et al. (2022) suggested to rearrange the population units in decreasing order with respect a multivariate distance measure from centre of the auxiliary space. This aims to improve the landing-phase of fast implementation of the cube method. Centre of the auxiliary space is computed as mean of expanded values of the auxiliary variables given by

$$\bar{\mathbf{Z}} = \frac{1}{N} \sum_{i \in U} \mathbf{z}_i$$

where $\mathbf{z}_i = \mathbf{x}_i / \pi_i$. Three different measures of multivariate distance from the centre are computed including Mahalanobis distance, Projection depth and Tukey depth. First was computed using `R`-function `mahalanobis()` and other two were computed using `R`-package `DepthProc` (Kosiorowski and Zawadzki, 2022). Population was rearranged in four ways: randomly ordered and decreasing order of three different measures of distance.

Again, $B = 5000$ samples of size $n = 20$ are selected by πps sampling and cube method from the population without any rearrangement, and cube sampling from the populations with four different rearrangements which are denoted as follow

- Cube-R: cube sampling from randomly ordered population,
- Cube-M: cube sampling from population rearranged with respect to Mahalanobis distance,
- Cube-P: cube sampling from population rearranged with respect to Projection depth,
- Cube-T: cube sampling from population rearranged with respect to Tukey depth.

Using B samples, MSE's of HT-estimators of totals are computed under all six designs. Relative values of MSE's with respect to that under πps sampling are given in Table 3.3. Relative values of MSE's for cube method (Cube) are slightly different as compared to those in Table 3.1 because a different set of $B = 5000$ sample is used in this simulation. Furthermore, values are shown up to three digits after the decimal point as most values are same up to second digit.

Table 3.3: Relative MSE's of HT-estimators under cube sampling from order populations.

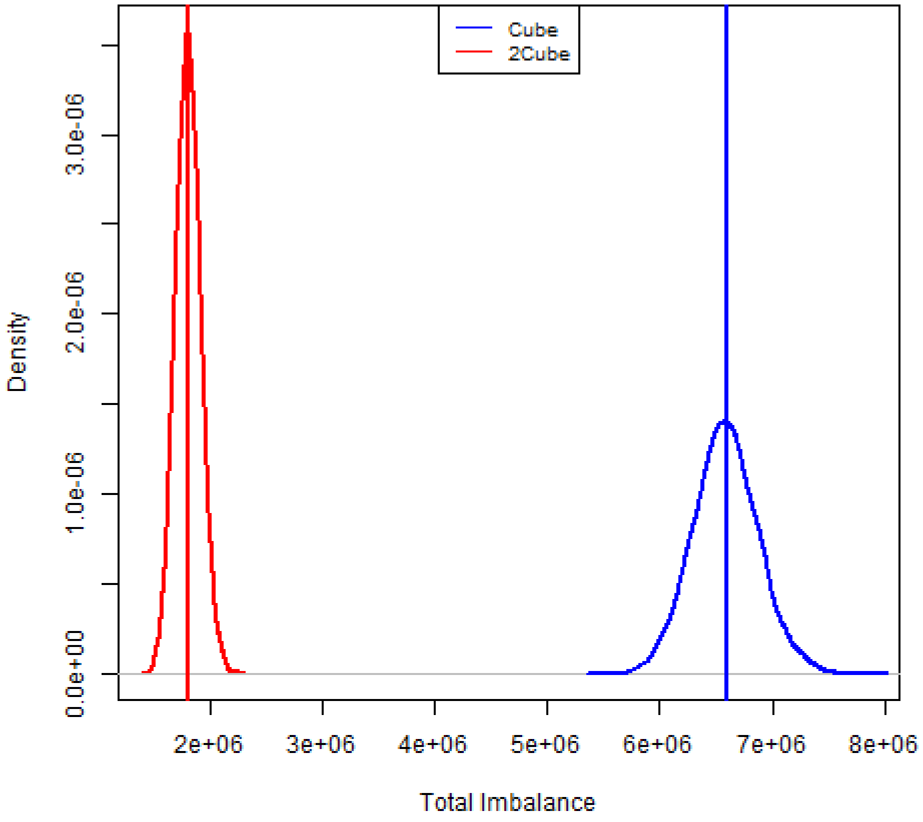
	πps	Cube	Cube-R	Cube-M	Cube-P	Cube-T
Auxiliary variables:						
P75	-	-	-	-	-	-
RMT85	1.000	0.114	0.113	0.112	0.115	0.109
ME84	1.000	0.116	0.114	0.115	0.116	0.110
Survey variables:						
P85	1.000	0.714	0.674	0.694	0.689	0.698
CS82	1.000	0.925	0.855	0.925	0.918	0.930
S82	1.000	0.759	0.725	0.746	0.769	0.769
REV84	1.000	1.041	1.048	0.981	1.065	1.061
SIZE	1.000	0.811	0.780	0.811	0.832	0.821
S82-CS82-SS82	1.000	0.725	0.689	0.710	0.748	0.731
CS82-SS82	1.000	0.844	0.813	0.833	0.814	0.844

In Table 3.3, first thing to look at is reduction in terms of imbalance with respect to auxiliary variables. Rearrangement of the given population slightly reduce the imbalance, particularly for the case of Tukey depth. In comparison with two-step cube method, this reduction is far less which suggest that the two-step cube method is not much affect by the rearrangement of the population. One reason could be that the two-step cube explicitly targets the imbalance with respect to auxiliary variables.

3.4.3 Assessment of potential gain in terms of total imbalance

Empirically estimated total imbalanced under the three designs and estimated percent gain (in terms of PRE) of two-step cube over cube are given in Table 3.4. Empirical distribution of total imbalance under cube and two-step cube are computed based $B = 5000$ Ω_K's. These densities are plotted in Figure 3.1. Vertical lines in the middle of each of two densities denotes averages based on $B = 5000$ values.

Figure 3.1: Empirical density of MSE of HT-estimator under cube and two-step cube.



It can be seen from Figure 3.1 that there is big distance between two densities and no common area between them. This suggests that the two-step cube is expected to be

Table 3.4: Total Imbalance = $tr(\hat{\mathbf{\Lambda}})$; Percent relative efficiency (PRE) of two-step cube as compared to cube method.

	πps	Cube	2Cube	PRE
$tr(\hat{\mathbf{\Lambda}})$	56828867.05	6851831.59	1739187.58	393.97

always better than the cube method in terms of total imbalance with respect to auxiliary variables in the given finite population. If there was some common area between the two densities which would have indicated likelihood of the two-step cub method being less or equally efficient as compared to cube method in terms of total imbalance. Since two-step cube minimises the total imbalance, therefore it not expected to be less efficient than the cube method in terms of total imbalance. As mentioned earlier in Section 3.3.3 when computation of empirical density of total imbalance under two-step cube is time consuming, one can compare one value of empirical total imbalance under the two-step cube with the density of total imbalance under cube method.

3.4.4 Comparison of cube and two-step cube methods based on AMSE

In this part of the simulation, two-step cube is compared with cube method based on empirically computed AMSE's of HT- and GREG-estimators. In order to calculate empirical AMSE's, the data set modified 'Clustered MU284' described earlier is used as follows. First, population based regression coefficients are computed by fitting linear regression models for each response variable using the three auxiliary variables. A summary of these fitted models is given Table 3.5.

Table 3.5: Summary of regression models fitted for seven survey variables given three auxiliary variables using modified 'Clustered MU284' data set.

	B_0	P75(B_1)	RMT85(B_2)	ME84(B_3)	$\hat{\sigma}$	Adj.R^2
P85	3.93	0.67	0.03	0.00	6.42	0.99
CS82	24.57	0.15	0.01	-0.00	16.44	0.43
S82	147.14	2.49	-0.24	0.00	40.35	0.74
REV84	2059.31	45.56	0.33	0.84	2991.16	0.87
SIZE	4.15	0.04	-0.00	0.00	0.87	0.50
S82-CS82-SS82	59.70	1.20	-0.15	0.00	19.29	0.54
CS82-SS82	-38.31	-0.98	0.10	-0.00	36.49	0.22

Then coefficients (B_0, B_1, B_2, B_3) are used in place of true model coefficients and response variables are regenerated by adding a an error term to the linear predictor as follows

$$y_i = B_0 + B_1\texttt{P75} + B_2\texttt{RMT85} + B_3\texttt{ME84} + \epsilon_i$$

where ϵ_i is generate from normal distribution with mean zero and variance σ^2. Three different values of $\sigma^2 = (1, 10, 20)$ are used, averaged values of coefficient of determination based on 5000 realisations for each value of σ^2 are shown in Table 3.6.

Table 3.6: Values of coefficient of determination (averaged over 5000 realisations) for regenerated populations from linear regression models using B_0, B_1, B_2, B_3 coefficients and different values of σ^2.

	$\sigma^2 = 1$	$\sigma^2 = 10$	$\sigma^2 = 20$
P85	1.000	0.998	0.997
CS82	0.996	0.957	0.917
S82	1.000	0.998	0.996
REV84	1.000	1.000	1.000
SIZE	0.439	0.071	0.036
S82-CS82-SS82	0.998	0.978	0.957
CS82-SS82	0.998	0.979	0.957

From each of the seven regression models, $M = 5000$ finite populations are generated. From each finite population, $B = 5000$ samples are selected using πps sampling, cube and two-step cube method. Empirical AMSE's of HT- and GREG-estimators are computed based on $M \times B$ samples under the three sampling methods. Percent relative efficiency of cube and two-step cube methods with respect to πps sampling is calculated using empirical AMSE's of HT- and GREG estimators which are given Table 3.7.

Table 3.7: Percent relative efficiency of two-step cube method (2Cube) and cube method (Cube) with respect to πps sampling based on empirical AMSE's of HT- and GREG-estimators.

	$\sigma^2 = 1$				$\sigma^2 = 10$				$\sigma^2 = 20$			
	HT-estimator		GREG-estimator		HT-estimator		GREG-estimator		HT-estimator		GREG-estimator	
	Cube	2Cube	Cube	2Cube	Cube	2Cube	Cube	2Cube	Cube	2Cube	Cube	2Cube
SIZE	119.34	126.79	122.28	128.30	106.46	110.42	107.87	111.19	102.97	106.18	104.04	106.82
P85	172.16	193.93	194.16	203.61	127.76	135.55	134.34	138.37	115.68	120.91	119.51	122.63
CS82	110.94	117.70	112.29	118.51	110.33	116.81	111.64	117.60	109.63	115.86	110.90	116.62
S82	129.98	139.85	134.76	142.25	129.95	139.81	134.72	142.20	129.88	139.72	134.64	142.10
REV84	92.78	97.94	92.95	98.12	92.78	97.94	92.95	98.12	92.78	97.94	92.95	98.12
S82-CS82-SS82	140.01	151.75	147.15	155.32	139.70	151.32	146.77	154.87	139.26	150.75	146.23	154.25
CS82-SS82	145.84	158.76	154.34	162.90	144.81	157.44	153.07	161.46	143.86	156.19	151.92	160.11

In Table 3.7, results show that two-step cube method has smallest AMSE's for all the cases. The efficiency of both cube and two-step cube decreases as value of error variance increase (i.e. coefficient of determination decreases). This indicates that choice of auxiliary variables have great impact on the efficiency of balanced sampling designs. Relative efficiency for survey variable `REV84` is less than 100 which means that balance sampling design used in this simulation is not useful for this variable as compared to πps sampling. By looking at intercept term of the fitted model for this variable in Table 3.5, which is very large, one can imagine that balancing with respect to intercept (or population size) is more important than other variables. This fact is also evident from Table 3.1.

3.5 Variance estimation

In this section, a methodology for variance estimation under balanced sampling is proposed which is motivated by natural decomposition of sampling variance of HT-estimator. When relationship of response variable and auxiliary variables can be expressed using a linear regression model, then response variable can be written as linear function of auxiliary variables and population residuals for a given finite population. The sampling variance of HT-estimator can be partitioned into two terms: first term is quadratic term which captures the effect of imbalance on the sampling variance, and second term is sampling variance of HT-estimator when sampling design is exactly balanced which can be approximated by the sampling variance of GREG-estimator. Berger (2005) also used similar kind of partition for variance estimation under πps systematic sampling, although it was motivated by (Hájek, 1964)'s residual technique for variance estimation. In this case, the quadratic term was known because second-order inclusion probabilities under systematic sampling were known and second term was approximated by (Hájek, 1964)'s variance approximation. Later, Berger et al. (2009) used this estimator for variance estimation of GREG-estimator.

Let $y_i = \mathbf{x}_i^\top \boldsymbol{B} + e_i$, where e_i is finite population residual associated with ith element and $\boldsymbol{B}$ is vector of finite population regression coefficients. The HT-estimator of population total can be written as $\hat{Y}_{HT} = \hat{\mathbf{X}}_{HT}^\top \boldsymbol{B} + \hat{e}_{HT}$, where $\hat{e}_{HT}$ is HT-estimator for finite population total of residuals given by $e = \sum_{i \in U} e_i$. The sampling variance of the HT-estimator $\hat{Y}_{HT}$ under the sampling distribution $p(s)$ can be written as

$$V_p(\hat{Y}_{HT}) = V_p\left(\hat{\mathbf{X}}_{HT}^\top \boldsymbol{B}\right) + V_p(\hat{e}_{HT}) = \boldsymbol{B}^\top \boldsymbol{\Lambda}_p \boldsymbol{B} + V_p(\hat{e}_{HT})$$

where $\boldsymbol{\Lambda}_p$ is defined in Eq. (3.2). Second term in the right side of above equation denotes sampling variance of HT-estimator under exactly balanced sampling design which is approximated by the sampling variance of GREG-estimator. Therefore, the proposed approximation for the sampling variance of HT-estimator is given by

$$V_p(\hat{Y}_{HT}) \approx \boldsymbol{B}^{\top}\boldsymbol{\Lambda}_p\boldsymbol{B} + V_p(\hat{Y}_{GR})$$

where $\hat{Y}_{GR}$ is GREG-estimator of the population total given in Eq. (1.6). Above approximation under two-step cube method can be written as

$$V_{c^*}(\hat{Y}_{HT}) \approx \boldsymbol{B}^{\top}\boldsymbol{\Lambda}_{c^*}\boldsymbol{B} + V_{c^*}(\hat{Y}_{GR}) \tag{3.9}$$

where vector of regression coefficients $\boldsymbol{B}$ in the first term is estimated by its corresponding sample estimate $\hat{\boldsymbol{B}}$ given in Eq. (1.7), and matrix $\boldsymbol{\Lambda}_{c^*}$ is estimated using its empirical estimator given by

$$\hat{\boldsymbol{\Lambda}}_{c^*} = E_{\boldsymbol{\lambda}^*}\left\{(\hat{\mathbf{X}}_{HT} - \mathbf{X})^{\top}(\hat{\mathbf{X}}_{HT} - \mathbf{X})\right\}$$

where $E_{\boldsymbol{\lambda}^*}$ is expectation with respect to sampling distribution $\boldsymbol{\lambda}^*$ over Ω_K. Based on the proposed approximation in Eq. (3.9), a variance estimator for HT-estimator of finite population total under two-step cube method is given by

$$\hat{V}_{c^*}(\hat{Y}_{HT}) = \hat{\boldsymbol{B}}^{\top}\hat{\boldsymbol{\Lambda}}_{c^*}\hat{\boldsymbol{B}} + \hat{V}_{c^*}(\hat{Y}_{GR}) \tag{3.10}$$

where $\hat{V}_{c^*}(\hat{Y}_{GR})$ is estimator for the sampling variance of GREG-estimator under the two-step cube method.

Like HT-estimator, sampling variance of the GREG-estimator and its variance estimator also involve second order inclusion probabilities. To avoid the calculation of second-order probabilities, a couple of approximations for the sampling variance of GREG-estimator are adopted from literature. Using these approximations, two proposed variance estimators for HT-estimator under two-step cube method are given bellow.

1. First, approximation based on *pps* sampling is used which is already described in Section 1.6. Sampling variance of HT-estimator under *pps* is stated again as follows

$$V_{WR}(\hat{Y}_{HT}) = \frac{1}{n}\sum_{i\in U} p_i \left(\frac{y_i}{p_i} - \sum_{i\in U} y_i\right)^2$$

where $np_i = \pi_i$. Corresponding estimator for the sampling variance based on above approximation is given by

$$\hat{V}_{WR}(\hat{Y}_{HT}) = \frac{1}{n(n-1)}\sum_{i\in s}\left(\frac{\hat{y}_i}{p_i} - \sum_{i\in s}\frac{\hat{y}_i}{\pi_i}\right)^2$$

An estimator for sampling variance of GREG-estimator based on above approximation can be written as

$$\hat{V}_{WR}(\hat{Y}_{GR}) = \frac{1}{n(n-1)}\sum_{i\in s}\left(\frac{\hat{e}_i}{p_i} - \sum_{i\in s}\frac{\hat{e}_i}{\pi_i}\right)^2$$

where $\hat{e}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{B}}$. By substituting $\hat{V}_{WR}(\hat{Y}_{GR})$ in the second term of the proposed variance estimator in Eq. (3.10), it gives the following

$$\hat{V}_{c^*}(\hat{Y}_{HT})_{WR} = \hat{\boldsymbol{B}}^\top\hat{\boldsymbol{\Lambda}}_{c^*}\hat{\boldsymbol{B}} + \frac{1}{n}\sum_{i\in U} p_i\left(\frac{e_i}{p_i} - \sum_{i\in U} e_i\right)^2 \tag{3.11}$$

□

2. Second, Deville and Tillé (2005) suggested that sampling variance of HT-estimator under exactly balanced sampling design can be approximated by sampling variance of GREG-estimator under Poisson sampling design based on (Hájek, 1964)'s residual technique for variance approximation, also see Section 1.6. This variance approximation suffers from bias because an exactly balanced sampling design is not always achieved (Breidt and Chauvet, 2011). In the proposed variance approximation, (Deville and Tillé, 2005)'s variance approximation is used for sampling variance of GREG-estimator.

Using Eq. (1.16), (Deville and Tillé, 2005)'s variance estimator of GREG-estimator can be written as

$$\hat{V}(\hat{Y}_{GR})_{DT} = \frac{n}{n-q}\sum_{i\in s}\frac{\hat{\hat{e}}_i^2}{\pi_i}(1-\pi_i) \tag{3.12}$$

where $\hat{\hat{e}}_i$ is defined with Eq. (1.16) and q is number of auxiliary variables. By substituting $\hat{V}(\hat{Y}_{GR})_{DT}$ in the second term of proposed variance estimator under two-step cube method in Eq. (3.10), it gives the following

$$\hat{V}_{c^*}(\hat{Y}_{HT})_{PS} = \hat{\boldsymbol{B}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\lambda}^*}\hat{\boldsymbol{B}} + \frac{n}{n-q}\sum_{i\in s}\frac{\hat{\hat{e}}_i^2}{\pi_i}(1-\pi_i) \tag{3.13}$$

where subscript PS denotes variance estimator based on approximation under Poisson sampling (Hájek, 1964).

□

The proposed methodology for variance estimation can also be used for the cube method, and proposed variance approximation for HT-estimator under cube method is given by

$$V_c(\hat{Y}_{HT}) \approx \boldsymbol{B}^{\top}\boldsymbol{\Lambda}_c\boldsymbol{B} + V_c(\hat{Y}_{GR}) \tag{3.14}$$

where $\boldsymbol{\Lambda}_c$ is variance-covariance matrix of $\mathbf{X}_{HT}$ under cube method and $V_c(\hat{Y}_{GR})$ is sampling variance of GREG estimator under cube method. Similarly, two proposed variance estimators under cube method can be written as

$$\hat{V}_c(\hat{Y}_{HT})_{WR} = \hat{\boldsymbol{B}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\lambda}}\hat{\boldsymbol{B}} + \frac{1}{n(n-1)}\sum_{i\in s}\left(\frac{\hat{e}_i}{p_i} - \sum_{i\in s}\frac{\hat{e}_i}{\pi_i}\right)^2 \tag{3.15}$$

and

$$\hat{V}_c(\hat{Y}_{HT})_{PS} = \hat{\boldsymbol{B}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\lambda}}\hat{\boldsymbol{B}} + \frac{n}{n-q}\sum_{i\in s}\frac{\hat{e}_i^2}{\pi_i}(1-\pi_i) \tag{3.16}$$

respectively, where $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\lambda}}$ is empirical estimate of $\boldsymbol{\Lambda}_{\boldsymbol{c}}$.

In the following section, performance of the two proposed variance estimators is assessed under two-step cube method, they are compared with (Deville and Tillé, 2005)'s variance estimator under the cube method. It would also be useful to make a comparison with (Breidt and Chauvet, 2011)'s variance estimator. It might not be possible at the moment because algorithm used to compute the results for this variance estimator is not available for `R` statistical package.

3.6 Simulation study for variance estimation

A simulation study is carried out in order to explore the performance of two proposed variance estimators under cube method and two-step cube method. The two proposed variance estimator are also compared with one existing estimator from Deville and Tillé (2005) under the cube method. The simulation set up is adapted from Breidt and Chauvet (2011) where different variance estimators were compared under the cube method. In the simulation study, there are three sampling designs, each design has a finite population consisting of a response variable and three auxiliary variables which are related under a linear regression model. In the first and second sampling designs, samples are selected with probabilities proportional to a size variable generated form the uniform distribution, while in the third sampling design samples are selected with equal probabilities. The three sampling designs are further described in the following.

Sampling design 1: The population U_1 is of size $N = 40$. The sample size is $n = 15$ and samples are selected with probability proportional to a variable generated from uniform distribution. For $i \in U_1$, define $z_{i2} = i, z_{i3} = 1/i, z_{i4} = 1/i^2$ and let $\bar{z}_j = N^{-1} \sum_{i \in U} z_{ij}$ and $s^2_{zj} = (N-1)^{-1} \sum_{i \in U} (z_{ij} - \bar{z}_j)^2$ denotes the usual empirical mean and variance respectively, for $j = 2, 3, 4$. Then $\mathbf{x}_i^T = (x_{1i}, x_{2i}, x_{3i}, x_{4i})$, where $x_{1i} = \pi_i$ and $x_{ij} = (z_{ij} - \bar{z}_j)/s_j$ for $j = 2, 3, 4$.

Sampling design 2: The population U_2 is of size $N = 30$. The sample size is $n = 10$ and samples are selected with probability proportional to a variable generated from uniform distribution. Define $x_{1i} = 1$, $x_{2i} = 1$ if $i \in \{1, ..., 15\}$ and 0 otherwise, $x_{3i} = 1$ if $i \in \{11, ..., 25\}$ and 0 otherwise, and $x_{4i} = 1$ if $i \in \{1, ..., 5\} \cup \{21, ..., 30\}$ and 0 otherwise.

Sampling design 3: The population U_3 is of size $N = 45$. The sample size is $n = 15$ and samples are selected with equal probability sampling. Define $x_{1i} = 1$, $x_{2i} = 1$ if $i \in \{1, ..., 15\}$ and 0 otherwise, $x_{3i} = 1$ if $i \in \{16, ..., 30\}$ and 0 otherwise, and $x_{4i} = 1$ if $i \in \{1, ..., 9\} \cup \{16, ..., 21\} \cup \{31, ..., 39\}$ and 0 otherwise.

In each population, response variable y is generated according to the linear regression model, given by

$$y_i = \beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + \sigma\epsilon_i$$

where ϵ_i are generated from the standard normal distribution. For the populations U_1 and U_3, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$. For population U_2, $\beta_1 = 0$, $\beta_2 = \beta_3 = \beta_4 = 1$. In each population, the coefficient σ is chosen such that the model R^2 (coefficient of determination) is approximately equal to 0.5. Pairwise correlations between variables in the three populations are given in Table 3.8.

In the simulation study, biases and variances of five variance estimators are computed by Monte Carlo simulation method based on 5000 samples from each population. The five estimators are as follows:

- under two-step cube method, two proposed estimators given by $\hat{V}_{c^*}(\hat{Y}_{HT})_{WR}$ and $\hat{V}_{c^*}(\hat{Y}_{HT})_{PS}$ from Eq. (3.11) and Eq. (3.13) respectively,

- under cube method, two proposed estimators given by $\hat{V}_c(\hat{Y}_{HT})_{WR}$ and $\hat{V}_c(\hat{Y}_{HT})_{PS}$ from Eq. (3.15) and Eq. (3.16) respectively, and one estimator from literature (Deville and Tillé, 2005) given by $\hat{V}_c(\hat{Y}_{HT})_{DT}$ from Eq. (1.16).

Table 3.8: Pairwise correlations between variables in the populations of three sampling designs in the simulation for variance estimation.

	Sampling design 1.				*Sampling design* 2.				*Sampling design* 3.			
	y	x_2	x_3	x_4	y	x_2	x_3	x_4	y	x_2	x_3	x_4
y	1.00	-0.24	0.56	0.65	1.00	0.16	0.38	0.16	1.00	0.19	0.18	0.55
x_2	-0.24	1.00	-0.61	-0.39	0.16	1.00	-0.33	-0.33	0.19	1.00	-0.50	0.09
x_3	0.56	-0.61	1.00	0.95	0.38	-0.33	1.00	-0.33	0.18	-0.50	1.00	-0.19
x_4	0.65	-0.39	0.95	1.00	0.16	-0.33	-0.33	1.00	0.55	0.09	-0.19	1.00

As a measure of bias and variance of a point estimator $\hat{\theta}$ of a parameter θ, Monte Carlo (MC) precent relative bias (RB) and percent relative stability (RS) are calculated, given by

$$RB_{MC}(\hat{\theta}) = 100 \times \frac{1}{\theta}\left(\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b - \theta\right) \text{ and } RS_{MC}(\hat{\theta}) = 100 \times \frac{1}{\theta}\left(\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \theta)^2\right)^{\frac{1}{2}} \tag{3.17}$$

respectively, where $B = 5000$ are number of samples in the Monte Carlo simulation. For the computation of these measures, simulations for cube and two-step methods are performed as follow.

(i) **Empirical true sampling variance under the cube method:** Select $C = 100000$ samples using the cube method. Empirical true sampling variance is calculated as

$$EV_c(\hat{Y}) = \frac{1}{C}\sum_{m=1}^{C}(\hat{Y}(s_m) - Y)^2 \tag{3.18}$$

where $m = 1, ..., C$ and $EV_c(\hat{Y})$ will be used in Eq. (3.17) as θ in order to calculate $RB_{MC}(\hat{\theta})$ and $RS_{MC}(\hat{\theta})$, when $\hat{\theta} = \left(\hat{V}_c(\hat{Y})_{WR}, \hat{V}_c(\hat{Y})_{PS}, \hat{V}_c(\hat{Y})_{DT}\right)$. Using C cube samples, percentage relative flight-phase efficiency (FE) of the cube method is also computed which is given by

$$FE = \frac{EV_F(\hat{Y}_{HT})}{EV_c(\hat{Y}_{HT})} \times 100$$

where $EV_F(\hat{Y}_{HT})$ is empirical true variance due to flight-phase of the cube method. As value of FE increases negative bias of the variance estimator $\hat{V}_c(\hat{Y})_{DT}$ decreases (Breidt and Chauvet, 2011).

(ii) **(Deville and Tillé, 2005)'s estimator under the cube method:** Select $B = 5000$ samples using the cube method, and calculate $RB_{MC}(\hat{\theta})$ and $RS_{MC}(\hat{\theta})$ for $\hat{V}_c(\hat{Y})_{DT}$.

(iii) **Two proposed estimators under the cube method:** First, calculate $B = 5000$ estimates of regression coefficient vector $\hat{\boldsymbol{B}}$ based on samples selected in (ii). Then, select B realized cube samples spaces of size $K = 1000$ independently, denoted by $\Omega_{K_1}, ..., \Omega_{K_B}$. Calculate B matrices $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\lambda}_1}, ..., \hat{\boldsymbol{\Lambda}}_{\boldsymbol{\lambda}_B}$, where $B = 5000$. Compute 5000 values of $\hat{V}_c(\hat{Y})_{WR}$ and $\hat{V}_c(\hat{Y})_{PS}$. Now $RB_{MC}(\hat{\theta})$ and $RS_{MC}(\hat{\theta})$ can be calculated for $\hat{V}_c(\hat{Y})_{WR}$ and $\hat{V}_c(\hat{Y})_{PS}$.

(iv) **Empirical true sampling variance under the two-step cube method:** In (iii), B realised cube sample spaces $\Omega_{K_1}, ..., \Omega_{K_B}$ are already obtained. For all these realised cube sample space, sampling distributions $\boldsymbol{\lambda^*}_1, ..., \boldsymbol{\lambda^*}_B$ under two-step cube method are obtained using simulated annealing algorithm, as described in Step 2 of the proposed sampling procedure. In this way, B estimates $\boldsymbol{\Lambda}_{\boldsymbol{\lambda^*}_1}, ..., \boldsymbol{\Lambda}_{\boldsymbol{\lambda^*}_B}$ can be obtained. Now, select B samples under two-step cube, i.e. one sample from each of Ω_{K_b} using $\boldsymbol{\lambda^*}_b$ sampling distribution, where $b = 1, ..., B$.

In principal a larger number of samples should have been used to calculated empirical true variance, for example $C = 100000$ samples are used for the cube method. This amount of sampling from two-step cube is computationally expensive. Therefore, *bootstrap method* is used which avoids additional iterations of two-step cube method. Calculate $\hat{Y}$ (HT-estimator of Y) estimate for each of B samples under two-step cube. Select 1000 bootstrap samples of size B from the initial vector of $\hat{Y}$-estimates of size B. For each of bootstrap samples, variance of $\hat{Y}$ is computed. Then an average of 1000 bootstrap sampling variances is computed which is considered as empirical true sampling variance under two-step cube method.

(v) **Two proposed variance estimators under the two-step cube method:** In (iv), B samples under two-step cube method are already selected. Based on these samples under two-step cube method, compute B estimates of $\hat{\boldsymbol{B}}$. Now B values for $\hat{V}_{c^*}(\hat{Y})_{WR}$ and $\hat{V}_{c^*}(\hat{Y})_{PS}$ can be calculated. Based on these B values, computed $RB_{MC}(\hat{\theta})$ and $RS_{MC}(\hat{\theta})$ for $\hat{V}_{c^*}(\hat{Y})_{WR}$ and $\hat{V}_{c^*}(\hat{Y})_{PS}$.

Table 3.9 shows measures of relative bias and stability for two proposed variance estimators under the two-step cube method (2Cube), two proposed and an existing variance estimators under the cube method (Cube). It also shows measures of relative bias and stability for the estimates of regression coefficients used in the computation of variance estimators. Smaller values of these measures indicate better performance of the estimators. Furthermore, Table 3.11 shows confidence intervals and percent coverage rates (PCR) for HT-estimator of population total using two proposed and an existing variance estimators under both cube and two-step cube methods.

From Table 3.9, results show that the two proposed estimators are highly biased and unstable for *Sampling design* 1 under both cube and two-step cube methods. Estimates of regression coefficients also have the same issue. One possible reason which can be noticed from Table 3.8 that two auxiliary variables x_3 and x_4 are highly correlated with correlation coefficient $r^2 = 0.95$. The problem of multicollinearity resulted into the highly biased regression coefficients with high stand error. Consequently, the proposed variance estimators are also highly biased and inefficient.

As a simple measure of avoiding the problem of multicollinearity, variable x_4 is excluded from the process of variance estimation. Results are shown in Table 3.10. This modification seems to reduce the problem. Bias and variability of the two proposed variance estimators have reduced under both cube and two-step cube methods, even it has improved the performance of the existing estimator under the cube method. Now, proposed estimator based on Poisson approximation is less biased and more stable than the existing estimator under the cube method.

For *Sampling design* 2, both proposed variance estimators $\hat{V}_c(\hat{Y})_{WR}$ and $\hat{V}_c(\hat{Y})_{PS}$ have smaller percent bias than that of the existing estimator $\hat{V}_c(\hat{Y})_{DT}$ under the cube method. While, the proposed estimator $\hat{V}_c(\hat{Y})_{PS}$ has the smallest percent of bias and highest stability among these three estimators. Under the two-step cube method, the estimator based on Poisson approximation $\hat{V}_{c^*}(\hat{Y})_{PS}$ performs way better than the estimator based on *pps* sampling approximation $\hat{V}_{c^*}(\hat{Y})_{WR}$ in terms of both bias and stability.

For *Sampling design* 3, the proposed variance estimator based on Poisson approximation $\hat{V}_c(\hat{Y})_{PS}$ performs slightly better than the existing estimator in terms of bias and stability under the cube method. Under the two-step cube method, again the estimator based on Poisson approximation $\hat{V}_{c^*}(\hat{Y})_{PS}$ is better than the estimator based on *pps* sampling approximation in terms of both bias and stability.

Base on the results from Table 3.11, the two proposed estimators always have higher coverage rate than the existing estimator under cube method. The propose estimator

Table 3.9: Measures of relative bias (RB) and relative stability (RS), from Eq. (3.17), for the variance estimators and estimates of regression coefficients under cube (Cube) and two-step cube (2Cube) methods.

		Sampling design 1		*Sampling design* 2		*Sampling design* 3	
	FE	44.8		69.0		86.8	
	Estimators	$RB_{MC}(\hat{\theta})$	$RS_{MC}(\hat{\theta})$	$RB_{MC}(\hat{\theta})$	$RS_{MC}(\hat{\theta})$	$RB_{MC}(\hat{\theta})$	$RS_{MC}(\hat{\theta})$
2Cube	$\hat{V}_{c^*}(\hat{Y})_{WR}$	122.77	1380.59	12.86	63.73	15.82	41.39
	$\hat{V}_{c^*}(\hat{Y})_{PS}$	49.85	1113.42	4.45	60.71	-0.69	32.40
	$\hat{B}_0$	-25.11	146.52	-	-	0.39	51.20
	$\hat{B}_2$	88.19	-149.22	-9.63	25.95	-0.97	43.00
	$\hat{B}_3$	47.38	-146.03	-2.75	14.39	-0.25	44.85
	$\hat{B}_4$	9.40	221.60	2.56	28.12	0.03	31.67
Cube	$\hat{V}_c(\hat{Y})_{WR}$	605.25	7130.64	8.96	55.53	20.71	44.12
	$\hat{V}_c(\hat{Y})_{PS}$	531.84	7110.24	1.60	53.33	4.07	33.27
	$\hat{V}_c(\hat{Y})_{DT}$	-64.71	68.56	-14.51	55.20	-6.81	34.38
	$\hat{B}_0$	-51.68	330.75	-	-	0.55	51.30
	$\hat{B}_2$	83.74	-152.28	-9.09	25.63	-1.41	44.29
	$\hat{B}_3$	34.09	-196.29	-2.87	14.88	0.04	44.48
	$\hat{B}_4$	-25.63	438.54	2.71	28.49	0.29	31.91

Table 3.10: Excluding variable x_4, measures of relative bias (RB) and relative stability (RS) for the variance estimators and estimates of regression coefficients under cube (Cube) and two-step cube (2Cube) methods.

		Sampling design 1	
	Estimators	$RB_{MC}(\hat{\theta})$	$RS_{MC}(\hat{\theta})$
2Cube	$\hat{V}_{c^*}(\hat{Y})_{WR}$	53.25	234.54
	$\hat{V}_{c^*}(\hat{Y})_{PS}$	-36.07	47.05
	$\hat{B}_0$	-25.33	61.18
	$\hat{B}_2$	-99.70	-145.25
	$\hat{B}_3$	-123.71	-138.96
Cube	$\hat{V}_c(\hat{Y})_{WR}$	74.69	242.59
	$\hat{V}_c(\hat{Y})_{PS}$	-17.74	50.64
	$\hat{V}_c(\hat{Y})_{DT}$	-51.87	61.97
	$\hat{B}_0$	-29.12	64.67
	$\hat{B}_2$	-96.04	-148.49
	$\hat{B}_3$	-120.03	-138.94

based on Poisson approximation has highest coverage rate in all case. As expected, the proposed estimator based on with-replacement approximation overestimate the sampling variance under the cube and two-step cube method. In summary, the proposed estimator based on Poisson approximation performs better than existing estimator and other competitive estimator based on with-replacement sampling approximation, provided that the regression model is not misspecified.

Table 3.11: Average of 95 percent confidence limits (L=lower, U=upper) of HT-estimator based on 5000 samples, and percent coverage rate (PCR).

		Sampling design 1			*Sampling design* 2			*Sampling design* 3		
	Estimators	L	U	PCR	L	U	PCR	U	L	PCR
2Cube	$\hat{V}_{c^*}(\hat{Y})_{WR}$	-33.42	113.66	97.86	35.53	54.02	94.40	81.45	116.67	92.78
	$\hat{V}_{c^*}(\hat{Y})_{PS}$	-20.41	100.65	92.26	35.91	53.65	93.56	82.74	115.37	91.12
	$\hat{V}_{c}(\hat{Y})_{DT}$	9.30	70.94	78.50	36.34	53.07	88.46	83.85	114.72	90.70
Cube	$\hat{V}_{c}(\hat{Y})_{WR}$	-42.52	122.98	98.06	35.03	54.37	93.64	81.63	116.94	94.32
	$\hat{V}_{c}(\hat{Y})_{PS}$	-29.32	109.78	92.48	35.38	54.01	92.52	82.88	115.69	92.92
	$\hat{V}_{c}(\hat{Y})_{DT}$	9.38	71.08	77.70	36.27	53.13	88.62	83.85	114.72	90.70

Generally, the proposed estimator based on Poisson approximation is better than estimator based on *pps* sampling approximation. As the imbalance (or rounding problem) of the cube method increases bias and instability of the existing estimator $\hat{V}_c(\hat{Y})_{DT}$ increases. Therefore, performance of the proposed estimator (based on Poisson approximation) in terms of bias and stability increases as compared ot the existing estimator. The proposed estimators are more sensitive to the fitness of the model, therefore, an assessment of model fitness might be useful before using the proposed estimators.

3.7 Conclusions and future work

Balanced sampling with respect to known auxiliary variables tends to improve the efficiency of estimates when response variable is linearly related with known auxiliary variables. Cube method aims to select balanced samples using an algorithm consisting of two phases, flight-phase and landing-phase. Whenever landing-phase is invoked, auxiliary balance of the samples is compromised in order to achieve a sampling design with fixed first-order inclusion probabilities. In this chapter, a practical way of improving the landing-phase of cube method is proposed. The proposed procedure, named two-step cube, aims to minimises total imbalance of the implied sampling design while respecting fixed first-order inclusion probabilities. Minimising total imbalance also aims to reduce the AMSE of HT-estimator under a linear super-population model. The cube method is surely exactly balanced when there is only one balancing variable. When there are multiple auxiliary variables, landing-phase is often invoked in order to select a sample. This means, there is often scope to improve the cube method by using two-step cube methodology. It is shown theoretically that total imbalance under two-step cube is equal to or less than that of the cube method.

The proposed two-step cube method depends on an optimisation algorithm in order to minimise realised total imbalance. In the optimization problem, calibration based on an $N \times K$ matrix is used in order to achieve fixed inclusion probabilities, where K is a finite number of samples under the cube method and N is population size. Computational cost of the optimisation algorithm increases with both N and K. For a given application of two-step cube, an estimate of gain in terms of total imbalance can be obtained and a larger value of K brings more confidence in the estimated gain. A global optimisation algorithm known as 'simulated annealing' is used here, which suffers from scalability problem as the size of either quantity N or K becomes large. A further investigation is required to explore the relation of these two quantities and time required for the optimisation algorithm to terminate. In future, further explore of optimisation algorithms may also help to find an optimisation algorithm with no or reduced scalability problem. Whereas, proposed procedure of the two-step cube method does not change.

The proposed methodology for the variance estimation is naturally followed by the proposed sampling procedure. Two variance estimators are proposed under any balanced sampling design. Results from the simulation study suggest that the proposed variance estimator which uses variance approximation based on Poisson sampling from literature (Hájek, 1964; Deville and Tillé, 2005) tend to be better than the other estimator (based on *pps* sampling) under both cube and two-step cube methods. Also, proposed variance estimator tends to perform better than an existing variance variance from Deville and Tillé (2005) under the cube method. From the simulation study, it is also noticed that the proposed variance estimation methodology is sensitive to model miss-specification as it depends on sample estimates of the regression coefficients. Whenever an underlying assumption of the model is violated which effects the estimation of regression coefficients, performance of the proposed variance estimators is expected to be affected too. For example, problem of multicollinearity was seen in the simulation study for variance estimation.

The proposed variance approximation can also be seen as bias correction for the existing variance approximation from Deville and Tillé (2005). When correction term goes wrong due to model miss-specification, the bias correction term can be dropped and proposed estimator shall be reduced to the Deville and Tillé (2005)'s estimator. Alternatively, there is possibility to use estimators of regression coefficients which are robust against problem of multicollinearity. This requires further exploration in this area.

Chapter 4

Spatially balanced two-stage equal probability sampling

4.1 Introduction

When study population exhibit positive spatial autocorrelation, spatially balanced sampling tends to improve the efficiency of sample estimates. It aims to select samples which are well-spread over the finite population area. It is widely used in environmental, ecology, forestry, agriculture and natural resource surveys (Stevens Jr and Olsen, 2004; Grafström et al., 2012; Benedetti et al., 2015; Robertson et al., 2018). National statistics institutes are increasingly geo-referencing the usual list sampling frames, and application of spatially balanced sampling in other areas including socio-economic and business surveys is being considered (Dickson et al., 2014, 2019; Abi, 2019; Filipponi et al., 2019). Two-stage sampling design is commonly used in socio-economic surveys, specifically in large scale (e.g. national level) household surveys. In this chapter, some aspects of spatially balanced sampling are investigated which have been overlooked or not completely addressed in the past, while focusing on use of spatially balanced sampling in two-stage equal probability sampling method (epsem) for socio-economic surveys.

It is now an establish fact that spatially balanced sampling designs are more efficient than those designs which do not account for spatial location data, provided that there exist spatial dependence in the study population, that is, near population units are tend to be similar (Stevens Jr and Olsen, 2004; Grafström et al., 2012; Benedetti et al., 2017c). In the last few years, an increasing number of sampling methods are proposed which has used spatial location data in different ways to select spatially balanced samples. Most of these sampling methods are described in Section 1.5.3. From practical point of view, choosing an appropriate spatially balanced sampling method for the given spatial population is important. Some comparative studies for different such methods are also found in the literature, which compared them with respect to their efficiency (i.e. MSE of the HT-estimator) and a measure of spatial balance from Stevens Jr and Olsen (2004) described in Section 1.5.2. Certainly, these studies supported the fact that spreading the sample over the population area improved the efficiency of estimates, but also indicated that different methods behave differently with respect to spatial structure of the study population (i.e. distribution of population units in the geographic space and nature of the spatial autocorrelation) and other parameters of sampling design (i.e. sampling fraction). In this chapter, some simulation studies are reviewed from literature; a comparative study of spatially balanced sampling methods is also conducted which has extended the comparison to another criterion of assessment, which is AMSE of the HT-estimator. The AMSE assesses performance of a sampling strategy averaged over many spatial populations under a given spatial super-population model.

In practice, GREG-estimator is commonly used in socio-economic studies, since population totals of some auxiliary variables are often known from previous surveys or administrative data sources. In addition to auxiliary population totals, when values of auxiliary variables are known for sampling units, balanced sampling by cube method can be used either at first, second or both stages depending on the availability of variables. Balanced sampling by cube method in combination with GREG-estimator reduces the problem of negative calibration weights (Deville and Tillé, 2004). Similarly, doubly balanced sampling can be used in two-stage sampling when location data of sampling units is also available. Grafström and Lundström (2013) argued that spatially balanced samples are also balanced with respect to auxiliary variable under certain conditions on auxiliary variables. However, two types of balanced designs are not same in general. It would be interesting to investigate how much auxiliary balance is achieved by a spatially balanced designs. Moreover, how both types of balances (auxiliary and spatial) interact under doubly balanced sampling design at varying levels of variation explained by auxiliary variables, and if there are situations when one type of balance is more important than the other.

In contrast with environmental and natural resources surveys, socio-economic survey are considered to be multi-objective and may contain a variety of study variables. Although presence of spatial autocorrelation is assumed, there might be situations when some variables exhibit complete randomness or negative spatial autocorrelation. Negative spatial autocorrelation is more likely to materialise for aggregated data, therefore, population of PSU's may exhibit negative spatial autocorrelation. Furthermore, such surveys often contains variables measured at variety of measurement scales, for example, quantitative and categorical variables. For such situations, it is also important to investigate that how bad spatially balanced samples are for those variables which are not spatially correlated. Altieri and Cocchi (2021) conducted a simulation study to investigate performance of some spatially balanced sampling methods when the study population exhibit negative spatial autocorrelation, and proposed an alternative spatial sampling design. This study suggested that spatially balanced sampling method can under-perform in comparison to SRS. In this chapter, some sampling schemes are proposed based on cube and local cube methods. Simulation studies are also conducted to investigate performance of the proposed sampling schemes.

Estimation of sampling variance with no bias under spatially balanced sampling designs is often a challenge. Because, most of these designs induce zero or near-zero second-order inclusion probabilities which preclude unbiased estimation of the sampling variance. Local-mean variance estimator from Stevens Jr and Olsen (2003) is often used in practice and recommended for many spatially balanced sampling methods. It was originally proposed for GRTS design. An extension of this estimator for doubly balanced sampling design by local cube method was given by Grafström and Tillé (2013). Both estimators are described in Section 1.6. In this chapter, proposed variance estimation methodology for balanced sampling in Section 3.5 is extended for both spatially and doubly balanced sampling designs. This extension uses the idea of eigenvector spatial filtering from spatial modelling literature Griffith (2003).

This chapter is arranged as follow. In Section 4.2, a simulation study is conducted for comparison of spatially balanced sampling methods based on AMSE of HT-estimator under equal probability sampling, assuming different spatial structure of population units and study variables under a spatial super-population model with positive spatial autocorrelation. In Section 4.3, the simulation study is extend for the comparison of balanced sampling by cube method, spatially balanced sampling methods and doubly balanced sampling by local cube method, under a simple common mean model and a linear regression model with one auxiliary variable and spatially correlated random errors. In Section 4.4, problem of sampling from populations with negative spatial autocorrelation is considered and some spatial sampling schemes are proposed. In Section 4.5, variance estimation methodology for balanced sampling in Section 3.5 is extended for spatially balanced and doubly balanced sampling designs. In the last, Section 4.6 presents some conclusions from this chapter.

4.2 Comparative study of spatially balanced sampling methods under the spatial super-population model

A variety of sampling methods have been proposed in literature to select spatially balanced samples (see Section 1.5.3). For these sampling methods, numerical studies based on real and artificial spatial populations were also conducted to make comparison of them with respect to their efficiency (based on MSE) and ability to achieve spatial balance in the samples. In most of these studies, measure of spatial balance based on Voronoi polygons, given by Stevens Jr and Olsen (2004), was used. Recently, other measures of spatial balance based on Moran's I index are also introduced by Tillé et al. (2018), also see Jauslin and Tillé (2020). From these numerical studies, it can be noticed that the most spatially balanced method may not be the most efficient for a given population; relative efficiency of a spatially balanced sampling method with respect to others might be different for different spatial populations, it may also change with the sampling fraction. It is important from the practical perspective to choose a suitable spatially balanced sampling method for given spatial population when there are multiple sampling methods available to select a spatially balanced sample. One may make a choice based on measure of spatial balance, but the commonly used measure of spatial balance may not lead to most efficient sampling method. Other criteria might be to make an assessment based on average performance of spatially balanced sampling methods over many spatial populations. Therefore, comparison of spatially balanced sampling methods based on AMSE under super-population model is done in this section.

In the following, some comparative simulation studies for spatially balanced sampling methods are reviewed from the literature and facts mentioned in previous paragraph are highlighted. Thereafter, a simulation study is conducted which compares spatially balanced sampling methods with respect to AMSE of HT-estimator under super-populations with different spatial structures.

4.2.1 A review of comparative simulation studies from literature

With the aim to individualise the most proper spatial sampling design for a sample survey of businesses, Dickson et al. (2014) conducted simulation experiments and compared spatially balanced sampling methods including BSS ("balanced spatial sampling" using cube method where spatial coordinates are used as balancing variables), local pivotal methods (LPM1, LPM2), SCPS and DBSS ("doubly balanced spatial sampling" using local cube method where spatial coordinates are used as balancing variables) with SRS under equal probability sampling and with random pivotal method (RPM) under πps sampling. A geo-referenced data set of ($N = 822$) retail stores located in the province of Trento (Italy), for the year of 2009, was used as finite spatial population and problem of estimating total sales of stores was considered. An empirical semi-variogram of the response variable was presented which suggested a short-term spatial trend in the population. Equal probability and πps samples were selected with respect to sampling fractions $f = (0.06, 0.09, 0.12)$. For πps sampling, number of employees was used as size variable. For LPM1, LPM2 and SCPS, calibration weights were also used in the HT-estimator based on linear and quadratic forms of spatial coordinates. In the linear form, horizontal and vertical coordinates were used while in the quadratic form, square and cross products of spatial coordinates were used. For BSS and DBSS, linear and quadratic forms of spatial coordinates were used as balancing variables.

Under equal probability sampling, DBSS was the most spatially balanced design, though there was not big difference as compared to LPM1. If the measure of spatial balance was rounded up to two decimal points, they were equally spatially balanced. DBSS was the most efficient design using balancing variables in linear form for $f = (0.06, 0.09)$ and in quadratic form for $f = 0.12$ under equal probability. LPMs and SCPS gained their maximum efficiency using calibration weights based on quadratic form of spatial coordinates. Under πps sampling, LPM2 was the most spatially balanced design, whereas all the spatially balanced sampling methods were approximately equally efficient (with difference of less than 1% in terms of efficiency gain); they attained their maximum efficiency using no calibration weights and linear form of spatial coordinates as balancing variables. While comparing LPMs and SCPS methods only, LPMs were more spatially balanced than SCPS across all scenarios and they tend to be more efficient than SCPS. In this study, spatial population has spatial trend; DBSS is the most spatially balanced and efficient under equal probability sampling; under πps sampling, LPMs are the most spatially balanced but not more efficient than others.

In another simulation study, Dickson et al. (2019) used a larger data set of ($N = 4592$) businesses located in Torento (Italy), for year 2009, to demonstrate that using LPMs (LPM1, LPM2) and SCPS for balanced sampling with respect to non-spatial (auxiliary) variables is better than stratified sampling (with proportional allocation) and cube method. Problem of estimating total turnover of businesses was considered. Two non-geographical variables, “sector of activity” and “number of employees” were used as balancing variables in LPMs, SCPS and cube method; the population was stratified with respect to these two variables (by converting them in categorical variables) for stratified sampling. Equal probability samples were selected with respect to sampling fractions $f = (0.11, 0.22, 0.33)$. LPM1 and SCPS were more spatially balanced and efficient than stratified sampling and the cube method. LPM1 was the most spatially balanced, but SPCS was the most efficient method. Furthermore, the relative efficiency of SCPS with respect to LPM1 was increasing with the sample size.

Benedetti et al. (2017b) conducted a simulation study to compare spatially balanced sampling methods GRTS, LPM1, SCPS, BSS, DBSS, and some alternative approaches based on spatial stratification. Here, this study is used to draw a comparison of spatially balanced sampling methods only, as authors have also concluded against the spatial stratification in terms of efficiency. In this study, artificial spatial populations were generated with three spatial frames of the population units: highly clustered, clustered and sparse. For each spatial frame, six response variables (or spatial populations) were simulated: without and with a linear spatial trend, and three levels of spatial dependence: low, medium and high. Equal probability samples of different sizes were selected with respect to sampling fractions given by $f = (0.01, 0.05)$. The comparison was made in terms of root mean squared error (RMSE) and spatial balance measure.

The most spatially balanced design were LPM1 for highly clustered and clustered spatial frames when $f = 0.01$, while it was BDSS for rest of the cases. First consider the populations with linear trend. DBSS and BSS were often the most efficient designs, DBSS for medium and high spatial correlation and BSS for low spatial correlation across all the spatial frames and sampling fractions. LPM1 and SCPS were approximately equally efficient for all the scenarios with one exception when LPM1 was 3% more efficient than SCPS for highly clustered spatial frame when $f = 0.01$. Now consider populations with no spatial trend. For the sparse spatial frame, all the five sampling methods tended to be approximately equally efficient with low and medium spatial correlation for both sampling fractions, while LPM1, SCPS and DBSS were approximately same and more efficient than GRTS and BSS for high spatial correlation. For highly clustered and clustered spatial frames, the results were same as those for sparse spatial frame for low and medium spatial correlations; for high spatial correlation, LPM1 and SCPS were approximately same and more efficient than DBSS when $f = 0.01$, while they were less efficient than DBSS when $f = 0.05$. In this study, doubly balancing with respect to spatial coordinates was beneficial when there was spatial linear trend with at least medium spatial correlation in the population. It was also beneficial for populations with no spatial trend and clustered spatial frame but only for larger sample size, this supports previously reviewed study from Dickson et al. (2014). However, it tends to loose efficiency as compared to spatial balancing (e.g. LPM1 and SCPS) when sample size is small, e.g. $f = 0.01$ in this study.

For the same comparison above, Benedetti et al. (2017b) also used two data sets as finite spatial populations, known as Mercer-Hall and Baltimore, which are available from R-packages spData (Bivand et al., 2021) and agridat (Wright, 2021) respectively. The first data set is 20×25 regular grid with $N = 500$ observations, and grain yield in pounds was considered as the response variable. The second data set consists of spatial points with $N = 211$ observations, and the sales price of the houses in Baltimore (Maryland, USA) was considered as response variable. Both the variables are known to have spatial trend (Benedetti et al., 2017b), and Moran indices for 'gain yield' and 'houses price' are also calculated, given by 0.0666 and 0.1198 respectively. Equal probability samples of two different sizes were selected with respect to sampling fractions $f = (0.02, 0.10)$ for Mercer-Hall and $f = (0.05, 0.24)$ for Baltimore data sets. For the population based on Mercer-Hall data set, LPM1, SCPS and DBSS were approximately equal efficient for both sampling fractions, but SCPS and DBSS were slightly more spatially balanced than LPM1. For the population base on Baltimore data set, DBSS was the most efficient sampling design, and LPM1 and SCPS were approximately equally efficient for $f = 0.05$. Whereas, SCPS was the most efficient design for $f = 0.24$, but less spatially balance than BDSS and LPM1. In this study, doubly balanced spatial sampling has not added much to the efficiency except one (Baltimore, $f = 0.05$) out of four cases; SCPS achieved higher efficient (despite being less spatially balanced) as compared to LPM1 for larger sampling fraction $f = 0.24$.

Benedetti et al. (2017c) used the same two data sets for the comparison of spatially balanced sampling methods DUST, GRTS, BSS, LPMs, SCPS, DBSS under equal probability sampling. An additional feature of this study was to select spatially balanced samples with respect to squared values of the spatial coordinates, which added nothing to the efficiency of designs.

Using same simulation design as in Benedetti et al. (2017b), Benedetti and Piersimoni (2017) compared a newly proposed sampling method, called PWD (see Section 1.5.3), with GRTS, BSS, SCPS and LPM with respect to both efficiency and spatial balance. In this study, three additional response variables were generated having quadratic spatial trend (with low, medium and high spatial correlation) for each of three spatial frames for the populations units (highly clustered, clustered and sparse). Equal probability samples were selected with respect to sampling fractions $f = (0.05, 0.10)$. PWD was the most spatially balanced and the most efficient design for all the scenarios considered in this study. LPM and SCPS were approximately equally efficient and spatially balanced for almost all the scenarios with an exception that LPM was slightly more spatially balanced for highly clustered spatial frame.

Filipponi et al. (2019) conducted a simulation study to compare spatially balanced methods, DBSS with linear and quadratic spatial coordinates, LPM1, LPM2 for the estimation of under-coverage rate and population size under capture-recapture set up. The finite population was consisted of ($N = 35585$) census tracts in Emilia Romagna region of Italy. Equal probability and πps samples were selected with different sampling fractions $f = (0.01, 0.05, 0.10)$. Around 10% gain in efficiency by spatially balanced methods was observed under equal probability sampling and there was not any considerable efficiency gain under πps sampling. Same comparison was performed while sampling from nine smaller regions of Emilia Romagna which produced results with larger efficiencies of spatially balanced sampling methods. DBSS with quadratic spatial coordinates was the most efficient design followed by LPMs and SCPS respectively.

Jauslin and Tillé (2020) conducted a simulation study using a real data set known as `Meuse`. It contained variables about different metal concentrations. A newly proposed method WAVE was compared with SRS, GRTS, LPM1, SCPS and HIP (an spatially balanced sampling method considered by Jauslin and Tillé (2020)) under equal probability sampling; and with GRTS, LPM1, SCPS under πps sampling. Samples were selected with respect to sampling fractions $f = (0.10, 0.19, 0.32)$. The comparison was made based on values of MSE of HT-estimator and three measures of spatial balance, here comparison is made only based on commonly used measure of spatial balanced from Stevens Jr and Olsen (2004). The VAWE was the most efficient method under equal and πps sampling only for the $f = 0.10$ and it was most spatially balanced for $f = 0.19$ under equal probability sampling and for $f = (0.10, 0.19)$ under πps sampling. The HIP was the most efficient under equal probability sampling for $f = (0.19, 0.32)$ while it was not the most spatially balanced. The SCPS was the most efficient under πps sampling for $f = (0.19, 0.32)$ and it was the most spatially balanced only for $f = 0.19$. The LPM was the most spatially balanced for $f = 0.32$ under both equal probability and πps sampling but it was not the most efficient for any of the cases considered in the study.

Based on above numerical studies, PWD design is the most efficient and spatially balanced design as compared to BSS, GRTS, LPMs, SCPS and DBSS under equal probability sampling, however it may not perform the same for πps sampling. DBSS tend to be more efficient than BSS, GRTS, LPMs and SCPS when population has spatial trend and strong spatial correlation, it may not be beneficial when population has no spatial trend and sample fraction is small (e.g. $f = 0.01$). It is seen in most cases that LPM1 and SCPS are more efficient than GRTS and approximately equally efficient, however for the cases when sampling fractions are large (e.g. $f = 0.24$ from Benedetti et al. (2017b) and $f = (0.19, 0.32)$ from Jauslin and Tillé (2020)) SCPS is more efficient than LPM1.

There were also cases where measure of spatial balance did not lead to the most efficient spatially balanced design. In order to investigate these results with respect to another aspect of sampling design called anticipated mean squared error (AMSE), a simulation study is conducted in the following section.

4.2.2 Simulation study for comparison of spatially balanced sampling methods

Following the simulation studies from Benedetti et al. (2017b) and Benedetti and Piersimoni (2017), three spatial frames (or spatial configurations of population units) are considered in this simulation study: highly clustered, clustered and sparse, each of size $N = 400$. In fact, these spatial frames with 1000 spatial points are available from R-package `Spbsampling` (Pantalone et al., 2019) and a random sample of 400 points is used for this simulation study. The reason of using smaller size of spatial frames is to reduce the computation time for calculation of AMSE's. The three spatial frames are shown in Figure 4.1.

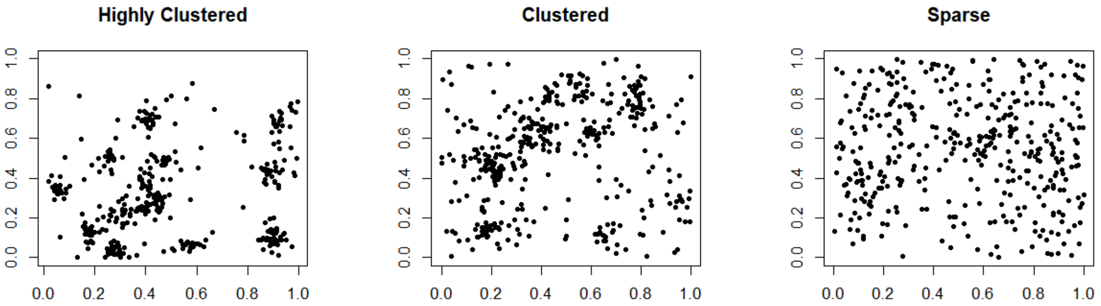


Figure 4.1: Three spatial frames for the spatial populations.

For each spatial frame, nine spatial populations (or response variables) are generated, six of them are simulated using two-dimensional Gaussian random field as follow. Using R-package `geoR` (Ribeiro Jr et al., 2020), three random error terms $\epsilon_{i1}, \epsilon_{i2}$ and ϵ_{i3} are generated from stationary Gaussian random field with three levels of positive spatial autocorrelation low, medium and high respectively, where $i \in U$. Exponential covariance function is used, given by $\text{cov}(d_{ij}) = \sigma \exp(-d_{ij}/\phi)$, where d_{ij} is Euclidean distance between pair of (i, j) units, and $i, j \in U$. Three different levels of positive spatial autocorrelation are achieved by choosing three values of the range parameter $\phi = (0.001, 0.01, 0.1)$ and value of scale parameter $\sigma = 2$ is fixed for all scenarios. Three variables with no spatial trend, $y_{ij} = \epsilon_{ij}$, $j = 1, 2, 3$, and three variables with a linear spatial trend $y_{ij} = \beta(cx_i + cy_i) + \epsilon_{ij}$, $j = 4, 5, 6$, are obtained, where cx and cy denotes spatial x- and y-coordinates respectively, and $\beta = 6.5$ is chosen which explain 75% to 85% variation in the response variables. All the of variables are standardized with mean $\mu = 5$ and standard deviation $\sigma = 1$, i.e. $z_{ij} = 5 + (y_{ij} - \bar{y}_j)/s_j$, $j = 1, ..., 6$, where $\bar{y}_j$ and s_j are mean and standard deviation of variables y_j respectively. Three binary variables $y_{(0.1)}, y_{(0.3)}$ and $y_{(0.5)}$ are computed by assign value 1 to 10%, 30% and 50% largest values of responses variables z_4, z_6 and z_7 respectively. For all the spatial variables, values of Moran's index averaged of 5000 realisations are reported in Table 4.1.

Table 4.1: Values of Moran's index I for nine populations and three spatial frames averaged over 5000 realizations of populations.

	No spatial trend			Linear spatial trend			Binary responses		
	z_1	z_2	z_3	z_4	z_5	z_6	$y_{(0.1)}$	$y_{(0.3)}$	$y_{(0.5)}$
Highly clustered	0.006	0.079	0.304	0.401	0.420	0.474	0.176	0.225	0.237
Clustered	-0.001	0.037	0.210	0.397	0.406	0.445	0.112	0.152	0.161
Sparse	-0.002	0.016	0.141	0.299	0.304	0.333	0.073	0.097	0.102

Equal probability samples of different sizes $n = (4, 8, 20, 32, 44)$ are selected which correspond to sampling fractions $f = (0.01, 0.03, 0.05, 0.08, 0.11)$. Random samples are selected using SRS as benchmark method and eight spatially balanced sampling methods: GRTS, BSS, LPM1, LPM2, SCPS, DBSS and PWD. R-packages `spsurvey` (Dumelle et al., 2021), `BalancedSampling` (Grafström and Lisic, 2019), and `Spbsampling` (Pantalone et al., 2019) are used for the selection of random samples.

For each of the six scenario of spatial super-population model (3 spatial autocorrelations × 2 spatial trends), $nrp = 5000$ finite spatial populations were generated. From each realization of the finite spatial population, $nrs = 5000$ samples are selected under equal probability; mean squared error (MSE) of HT-estimator for population total was calculated using 5000 samples, as given in Eq. (4.1). In order to obtain AMSE of the HT-estimator, MSEs are averaged over 5000 realizations of the spatial populations, as given in Eq. (4.2).

$$\text{MSE}(\hat{Y}_{HT}) = \frac{1}{nrs} \sum_{s=1}^{nrs} \left(\hat{Y}_{HT}(s) - Y \right)^2 \tag{4.1}$$

$$\text{AMSE}(\hat{Y}_{HT}) = \frac{1}{nrp}\sum_{p=1}^{nrp}\text{MSE}(\hat{Y}_{HT})(p) \tag{4.2}$$

where $\text{MSE}(\hat{Y}_{HT})(p)$ is MSE of HT-estimator for pth spatial population out of 5000. Let AMSE_{srs} and AMSE_{sp} denote AMSE's of HT-estimator under SRS and spatially balanced sampling designs respectively. Relative values of AMSE's, given by

$$\frac{AMSE_{sp}}{AMSE_{srs}}$$

are reported in Tables 4.2, 4.3 and 4.4 for highly clustered, clustered and spars spatial frames respectively. Measure of spatial balance for spatially balanced sampling methods are not reported here, they are expected be same as in simulation study of Benedetti et al. (2017b), because it only depends on the spatial configuration of population units i.e. spatial frame.

From Tables 4.2, 4.3 and 4.4, results show that PWD design is the most efficient design in terms of AMSE of HT-estimator under all the scenarios of equal probability sampling considered in this simulation study. For comparison of other sampling methods, let us consider first the populations with linear spatial trend. GRTS is always least efficient among GRTS, LPMs, SCPS and DBSS. It gains efficiency with sample size and spatial correlation and it is more efficient than BSS for some cases with large values of sample size and spatial correlation in this study. For the highly clustered spatial frame with populations having medium and high spatial autocorrelation, DBSS is the second most efficient design (after PWD) followed by SCPS design, while BSS is the second most efficient for low spatial autocorrelation followed by DBSS. LPMs tend to be less efficient than SCPS, they gain efficiency relative to SCPS as sample size increases. Moving from highly clustered to clustered spatial frame, BSS takes over the place of DBSS as second most efficient design for populations with medium (in addition to low) spatial autocorrelation for smaller sample sizes, and for all the sample sizes when moving further from clustered to sparse spatial frame. Similarly, when spatial clustering of population units decreases, LPMs gain efficiency against SCPS more quickly; efficiency of LPM2 against LM1 increases, even LPM2 is more efficient than LPM1 for smaller sample sizes ($n = 4, 12, 20$) under sparse spatial frame. Whereas LPM1 and LPM2 tend to be equally efficient for larger sample size.

Now, let us consider populations with no spatial trend. For highly clustered spatial frame, LPM1 tend to be the second most efficient design (after PWD) for smaller sample sizes

($n = 4, 12$), whereas DBSS takes over this place as sample size and spatial autocorrelation become larger. LPM1 tends to be better than SCPS for medium values of spatial autocorrelation and less efficient for high values. The three designs, LPMs, SCPS and DBSS tend to be equally efficient for low values of spatial autocorrelation, regardless of samples size. When moving from highly clustered to clustered and sparse spatial frames, GRTS, LPMs, SCPS and DBSS are approximately equally efficient for populations with low and medium spatial autocorrelation, whereas for populations with high spatial autocorrelation SCPS design tend to be second most efficient design (after PWD) and it becomes approximately equally efficient to DBSS as the sample size increases.

For binary responses, the three methods LPM1, SCPS and DBSS are more efficient than GRTS and BSS, these three methods are approximately equally efficient. However as sample size gets larger, SCPS and LPM1 tend to loose their efficiencies as compared to DBSS for highly clustered and clustered spatial frames and SCPS looses its efficiency more quickly as compared to LPM1. For sparse spatial frame, only LPM1 tends to loose its efficiency as compared to both SCPS and DBSS as sample size increases.

To summarise, spatially balanced sampling is better choice against SRS when there is knowledges about presence of positive spatial autocorrelation in study population. Further information about spatial structure of population units and study variable can be helpful in choosing an appropriate spatially balanced design for the given study population. In case of equal probability sampling, PWD is the most efficient design which is not applicable when πps sampling is required. For very large populations, this design may lead to computational disadvantage over other designs, see Benedetti and Piersimoni (2017) for details. Apart from PWD, other spatially balanced sampling methods considered here are able to select samples with unequal fixed inclusion probabilities. While comparing these methods under equal probability sampling, DBSS outperforms others when there exists spatial trend with strong spatial correlation in the population, whereas BSS is better than DBSS when spatial correlation is weak. For populations with no spatial trend, when there exist strong spatial correlation and sample size is large (e.g. such that $f \geq 0.05$) DBSS is better than others; when sample size is not large LPM1 and SCPS are more efficient than DBSS for clustered and sparse popuations respectively.

Table 4.2: Relative values of AMSE's of HT-estimator under different sampling designs with respect to SRS, for different sample sizes (n) from highly clustered populations with low, medium and high spatial autocorrelations.

		No spatial trend			Linear spatial trend			Binary responses		
n		z_1	z_2	z_3	z_4	z_5	z_6	$y_{(0.1)}$	$y_{(0.3)}$	$y_{(0.5)}$
4	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.990	0.874	0.652	0.650	0.619	0.942	0.915	0.909
	BSS	1.000	0.996	0.926	0.412	0.411	0.388	0.964	0.954	0.951
	LPM1	1.000	0.988	0.816	0.506	0.504	0.457	0.919	0.876	0.867
	LPM2	1.000	0.989	0.829	0.493	0.491	0.447	0.923	0.884	0.876
	SCPS	0.999	0.989	0.819	0.446	0.444	0.397	0.920	0.879	0.870
	DBSS	1.000	0.988	0.815	0.412	0.409	0.361	0.917	0.877	0.867
	PWD	0.999	0.967	0.704	0.322	0.314	0.243	0.858	0.817	0.802
12	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	0.999	0.970	0.706	0.448	0.440	0.372	0.852	0.797	0.786
	BSS	1.000	0.996	0.908	0.297	0.296	0.267	0.957	0.940	0.937
	LPM1	1.000	0.959	0.555	0.315	0.304	0.201	0.774	0.692	0.677
	LPM2	1.000	0.960	0.576	0.317	0.307	0.209	0.785	0.707	0.692
	SCPS	1.000	0.961	0.562	0.307	0.297	0.194	0.778	0.698	0.683
	DBSS	0.999	0.960	0.562	0.302	0.292	0.190	0.778	0.698	0.683
	PWD	0.999	0.883	0.346	0.245	0.217	0.083	0.627	0.539	0.524
20	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.952	0.611	0.376	0.364	0.275	0.797	0.729	0.715
	BSS	1.000	0.994	0.906	0.277	0.276	0.245	0.957	0.939	0.935
	LPM1	0.999	0.930	0.467	0.303	0.285	0.167	0.711	0.624	0.608
	LPM2	0.999	0.934	0.491	0.302	0.285	0.172	0.726	0.643	0.628
	SCPS	0.999	0.936	0.488	0.299	0.284	0.168	0.725	0.641	0.625
	DBSS	0.999	0.931	0.466	0.287	0.270	0.151	0.712	0.626	0.609
	PWD	0.997	0.847	0.323	0.285	0.247	0.116	0.597	0.512	0.496
32	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	0.999	0.932	0.538	0.325	0.308	0.205	0.749	0.673	0.659
	BBS	1.000	0.993	0.895	0.269	0.267	0.235	0.952	0.930	0.927
	LPM1	0.998	0.893	0.404	0.287	0.261	0.135	0.659	0.571	0.554
	LPM2	0.998	0.903	0.426	0.285	0.261	0.137	0.676	0.589	0.573
	SCPS	0.998	0.905	0.419	0.285	0.261	0.136	0.671	0.585	0.569
	DBSS	0.998	0.896	0.395	0.274	0.248	0.119	0.655	0.567	0.550
	PWD	0.993	0.811	0.282	0.259	0.214	0.079	0.559	0.472	0.456
44	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	0.998	0.915	0.499	0.308	0.288	0.183	0.721	0.642	0.627
	BBS	0.999	0.994	0.896	0.255	0.253	0.221	0.953	0.932	0.928
	LPM1	0.997	0.865	0.367	0.273	0.241	0.117	0.625	0.539	0.523
	LPM2	0.997	0.878	0.388	0.271	0.242	0.120	0.643	0.558	0.541
	SCPS	0.997	0.880	0.385	0.268	0.240	0.117	0.641	0.556	0.539
	DBSS	0.997	0.869	0.355	0.258	0.227	0.099	0.620	0.532	0.515
	PWD	0.992	0.788	0.267	0.247	0.198	0.070	0.539	0.455	0.439

Table 4.3: Relative values of AMSE's of HT-estimator under different sampling designs with respect to SRS, for different sample sizes (n) from clustered populations with low, medium and high spatial autocorrelation.

		No spatial trend			Linear spatial trend			Binary responses		
n		z_1	z_2	z_3	z_4	z_5	z_6	$y_{(0.1)}$	$y_{(0.3)}$	$y_{(0.5)}$
4	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.994	0.891	0.539	0.538	0.513	0.954	0.927	0.921
	BBS	1.000	0.997	0.936	0.382	0.381	0.364	0.974	0.959	0.955
	LPM1	1.000	0.994	0.842	0.472	0.471	0.434	0.937	0.896	0.886
	LPM2	1.000	0.994	0.853	0.483	0.482	0.448	0.940	0.903	0.893
	SCPS	1.000	0.994	0.846	0.444	0.442	0.406	0.939	0.899	0.888
	DBSS	1.000	0.994	0.841	0.396	0.395	0.357	0.936	0.895	0.886
	PWD	0.998	0.979	0.709	0.304	0.300	0.233	0.864	0.824	0.811
12	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.982	0.750	0.397	0.393	0.340	0.883	0.831	0.819
	BSS	1.000	0.997	0.919	0.275	0.274	0.253	0.967	0.948	0.943
	LPM1	1.001	0.979	0.647	0.324	0.320	0.244	0.834	0.761	0.745
	LPM2	0.999	0.980	0.667	0.322	0.318	0.246	0.844	0.775	0.760
	SCPS	1.000	0.981	0.656	0.306	0.302	0.226	0.839	0.768	0.753
	DBSS	1.000	0.979	0.641	0.287	0.283	0.204	0.832	0.758	0.742
	PWD	0.999	0.941	0.485	0.248	0.236	0.130	0.731	0.651	0.633
20	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.973	0.679	0.351	0.345	0.274	0.844	0.781	0.768
	BSS	1.000	0.998	0.923	0.264	0.263	0.241	0.969	0.951	0.947
	LPM1	1.000	0.964	0.559	0.302	0.293	0.197	0.778	0.697	0.681
	LPM2	1.000	0.967	0.583	0.300	0.292	0.200	0.792	0.715	0.699
	SCPS	1.000	0.967	0.574	0.291	0.283	0.189	0.788	0.710	0.693
	DBSS	0.999	0.965	0.556	0.275	0.268	0.169	0.779	0.697	0.680
	PWD	0.999	0.921	0.430	0.246	0.228	0.110	0.685	0.602	0.585
32	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.962	0.616	0.311	0.303	0.221	0.805	0.735	0.721
	BSS	1.001	0.999	0.918	0.248	0.248	0.225	0.967	0.948	0.943
	LPM1	1.000	0.945	0.492	0.278	0.265	0.159	0.729	0.645	0.629
	LPM2	1.000	0.949	0.511	0.273	0.262	0.159	0.743	0.661	0.645
	SCPS	1.000	0.953	0.503	0.271	0.261	0.155	0.740	0.657	0.640
	DBSS	1.000	0.946	0.482	0.256	0.244	0.135	0.726	0.641	0.623
	PWD	0.998	0.897	0.382	0.239	0.216	0.096	0.644	0.562	0.543
44	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	0.998	0.951	0.567	0.288	0.277	0.186	0.774	0.699	0.683
	BSS	0.999	0.998	0.911	0.245	0.245	0.220	0.964	0.943	0.938
	LPM1	0.999	0.926	0.443	0.268	0.252	0.138	0.692	0.607	0.590
	LPM2	0.998	0.932	0.463	0.266	0.250	0.140	0.707	0.624	0.607
	SCPS	0.998	0.938	0.452	0.264	0.250	0.136	0.703	0.617	0.600
	DBSS	0.998	0.928	0.432	0.252	0.236	0.119	0.688	0.601	0.584
	PWD	0.996	0.876	0.344	0.239	0.211	0.086	0.612	0.529	0.511

Table 4.4: Relative values of AMSE's of HT-estimator under different sampling designs with respect to SRS, for different sample sizes (n) from sparse populations with low, medium and high spatial autocorrelation.

		No spatial trend			Linear spatial trend			Binary responses		
n		z_1	z_2	z_3	z_4	z_5	z_6	$y_{(0.1)}$	$y_{(0.3)}$	$y_{(0.5)}$
4	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.998	0.915	0.613	0.612	0.593	0.965	0.946	0.942
	BSS	1.000	0.999	0.950	0.408	0.407	0.394	0.980	0.969	0.966
	LPM1	1.000	0.998	0.879	0.550	0.549	0.520	0.952	0.923	0.916
	LPM2	1.000	0.998	0.886	0.527	0.526	0.499	0.954	0.927	0.921
	SCPS	1.000	0.999	0.874	0.486	0.486	0.455	0.950	0.920	0.913
	DBSS	1.000	0.999	0.881	0.444	0.444	0.415	0.953	0.924	0.917
	PWD	1.000	0.996	0.824	0.326	0.325	0.281	0.928	0.893	0.884
12	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.994	0.814	0.453	0.451	0.408	0.916	0.879	0.871
	BSS	1.001	0.999	0.935	0.294	0.293	0.276	0.974	0.959	0.956
	LPM1	1.000	0.992	0.714	0.371	0.369	0.301	0.871	0.814	0.803
	LPM2	1.000	0.993	0.726	0.359	0.357	0.292	0.877	0.823	0.811
	SCPS	1.000	0.994	0.703	0.328	0.326	0.255	0.867	0.808	0.795
	DBSS	1.000	0.993	0.709	0.317	0.315	0.246	0.869	0.812	0.800
	PWD	1.001	0.988	0.635	0.270	0.266	0.181	0.833	0.764	0.751
20	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.989	0.754	0.391	0.388	0.331	0.885	0.838	0.829
	BSS	1.000	0.999	0.938	0.273	0.273	0.255	0.975	0.961	0.957
	LPM1	1.000	0.986	0.631	0.326	0.323	0.236	0.824	0.757	0.744
	LPM2	1.001	0.987	0.644	0.319	0.315	0.232	0.831	0.767	0.753
	SCPS	1.000	0.987	0.617	0.298	0.295	0.204	0.818	0.750	0.735
	DBSS	1.001	0.987	0.623	0.289	0.286	0.197	0.822	0.753	0.739
	PWD	1.000	0.972	0.529	0.259	0.252	0.144	0.768	0.690	0.673
32	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.984	0.689	0.336	0.331	0.260	0.849	0.794	0.783
	BSS	0.999	0.999	0.930	0.259	0.258	0.239	0.971	0.955	0.952
	LPM1	1.000	0.978	0.552	0.297	0.291	0.188	0.773	0.701	0.687
	LPM2	1.000	0.979	0.567	0.294	0.288	0.188	0.782	0.712	0.698
	SCPS	1.000	0.979	0.537	0.277	0.271	0.164	0.767	0.692	0.677
	DBSS	1.000	0.978	0.541	0.269	0.263	0.157	0.771	0.695	0.680
	PWD	0.999	0.952	0.440	0.251	0.240	0.116	0.704	0.623	0.607
44	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.001	0.980	0.648	0.324	0.319	0.238	0.825	0.765	0.753
	BSS	1.000	0.998	0.929	0.255	0.254	0.235	0.971	0.955	0.952
	LPM1	1.001	0.969	0.501	0.283	0.275	0.162	0.738	0.663	0.648
	LPM2	1.001	0.969	0.516	0.281	0.273	0.162	0.748	0.674	0.660
	SCPS	1.001	0.970	0.489	0.270	0.263	0.146	0.734	0.657	0.641
	DBSS	1.000	0.969	0.488	0.262	0.255	0.138	0.735	0.657	0.641
	PWD	0.999	0.937	0.396	0.249	0.234	0.103	0.668	0.586	0.570

4.3 Epsem by auxiliary and spatial balancing

Definitions of both auxiliary and spatial balance (from Eqs. (1.9) and (1.12) respectively) involve fixed first-order inclusion probabilities only. In order to achieve respective balance, both designs manipulate second-order inclusion probabilities. Manipulation by spatial sampling designs is relatively more explicit which aims to assign zero or very small second-order inclusion probabilities to the nearby units in the space. Grafström and Lundström (2013) suggested that spatially balanced samples are also balanced (with respect to auxiliary variables) under certain conditions on auxiliary variables. In general, spatial and auxiliary balancing may not be the same and simply requiring both equations to satisfy may not be feasible. Let consider a toy example below:

Example 4.1. Let $U = \{1, 2, 3\}$. Let $x_i = i$, for $i = 1, 2, 3$, and $X = 6$. Suppose the sample s is of the size 2. For spatial balancing, let $d(x_i, x_j) = 0$ if $x_i = x_j$ and 1 otherwise. Let $s = (1, 3)$. Based on definition of spatial balance in Eq. (1.12), $\alpha_1 = \{1, 2\}$ and $\alpha_3 = \{2, 3\}$, so that $m_1 = m_3 = 1$ and $m_2 = 2$. For the sample to be spatially balanced, we must have $\nu_1 = \pi_1 + \pi_2/2 = \nu_3 = \pi_3 + \pi_2/2$, i.e. $\pi_1 = \pi_3$. Similarly, for $s = (1, 2)$ to be spatially balanced, we must have $\pi_1 = \pi_2$ as well. Therefore, the only spatially balanced design is $\Pr(1, 2) = \Pr(1, 3) = \Pr(2, 3) = 1/3$ with $\pi_i \equiv 2/3$. However, given $\pi_i \equiv 2/3$, there is only one auxiliary balanced sample $s = (1, 3)$, so that auxiliary balance design with $\pi_i \equiv 2/3$ does not exist.

When both types of balanced designs are feasible, achieving one type of balance may compromise the other. In the following this aspect of spatial and auxiliary balancing is investigated under very simple common mean model and a linear regression model (with spatially correlation errors).

4.3.1 Common mean model

A common mean model with constant error variance is given by, $y_i = \mu + \epsilon_i$, where μ is the common mean and ϵ_i's are normally distributed random errors with mean 0 and variance 1. Considering auxiliary variable $x_i \equiv 1$, although epsem is auxiliary balanced, it does not imply epsem because there exist numerous auxiliary balanced designs. For instance, let $\pi_1 = 1$ and $\pi_i = (n-1)/(N-1)$ for all $i \neq 1$, where the first element is selected with probability one and epsem of size $n - 1$ is applied to the rest of the population.

On the other hand, spatial balancing can lead to epsem, given a particular distance metric.

Let $x_i = i$, for $i \in U$, or the real spatial coordinates. Assume distance $d(x_i, x_j) = 0$ if $x_i = x_j$ and 1 otherwise. For any s and $i \in s$: $\alpha_i = \{i\} \cup (U \setminus s)$, such that $m_j = 1$ for any $j \in s$, and $m_j = n$ for any $j \notin s$. For s to be spatial balanced, it must be $\pi_i = \pi_k$ for any $i \neq k \in s$. For a spatially balanced design, this must hold for any s, and it must be epsem.

One may argue that epsem is spatially and auxiliary balanced, therefore it is equally efficient to the SRS in terms of AMSE of HT-estimator under the common mean model with independent errors. According to (Zhang, 2008), model expectation of the sampling variance (i.e. AMSE) of the estimator is defined as first-order Bayes risk, and model variance of the sampling variance is defined as second-order Bayes risk. Any two designs with the same AMSE may still differ in terms of their second-order Bayes risks. The larger the second-order Bayes risk, the less one has control over the actual sampling variance for a given population. In the following it is shown that SRS is the spatial and auxiliary balanced design with minimum first- and second-order Bayes risk under the common mean model with independent errors.

Result 4.1. SRS is the spatial and auxiliary balanced design with minimum first- and second-order Bayes risks under the common mean model with independent errors.

Let $p(s)$ is any spatial and auxiliary balanced design, possibly with unequal π_i's, under the homogeneous model with independent errors, sampling variance of HT-estimator for population total can be written as

$$V_p(\hat{Y}_{HT}) = \sum_{k \in U} a_k e_k^2 + 2 \sum_{i<j \in U} a_{ij} e_i e_j$$

where $a_k = 1/\pi_k - 1$, and $a_{ij} = \pi_{ij}/\pi_i\pi_j - 1$. Let E_m and V_m denote expectation and variance under the model respectively. Second-order Bayes risk of HT-estimator under design $p(s)$ and model can be written as

$$V_m\left(V_p(\hat{Y}_{HT})\right) = E_m\left(V_p(\hat{Y}_{HT})^2\right) + \left[E_m\left(V_p(\hat{Y}_{HT})\right)\right]^2$$

such that

$$E_m\left(V_p(\hat{Y}_{HT})^2\right) = \sum_{k \in U} a_k^2 E_m(e_k^4) + 2 \sum_{k<l \in U} a_k a_l E_m(e_k^2 e_l^2) + 4 \sum_{i<j \in U} a_{ij}^2 E_m(e_i^2 e_j^2)$$

and

$$E_m\left(V_p(\hat{Y}_{HT})\right) = \sum_{k \in U} a_k E_m(e_k^2)$$

where all the other cross-product terms vanishes provided uncorrelated regression errors.

When $\pi_i \equiv \pi$, it implies $\sum_{i \neq j \in U} a_{ij}^2 = \sum_{i \neq j \in U} \pi_{ij}^2/\pi^4 - 2\sum_{i \neq j \in U} \pi_{ij}/\pi^2 + N(N-1)$, where $\sum_{i \neq j \in U} \pi_{ij} = n(n-1)$ under fixed-size sampling without replacement. Let $\mu_{4k} = E_m(e_k^4)$ under the homogeneous model with independent errors. The second-order Bayes risk of a strategy consisting of $\hat{Y}_{HT}$ and epsem is given by

$$V_m\left(V_p(\hat{Y}_{HT})\right) = \sum_{k \in U}(N/n-1)^2(\mu_{4k} - \sigma^4) + 4\sum_{i<j \in U}(\pi_{ij}N^2/n^2 - 1)^2\sigma^4$$

which is minimised at $\pi_{ij} = n(n-1)/N(N-1)$ for fixed-size sampling without replacement.

□

Previously in Section 4.2.2, different spatially balanced sampling methods are compared with SRS with respect to AMSE of HT-estimator under the common mean models ($\mu = 5$) with spatially correlated random errors. The comparison is extended here and these methods are compared with respect to second-order Bayes risk of HT-estimator under the same models. Results are shown in Tables 4.5, 4.6 and 4.7 for highly clustered, clustered and spars spatial frames respectively.

The results shows that SRS has the smallest second-order Bayes risk, this is just a confirmation of Result 4.1, as all the populations are simulated under common mean model with $\mu = 5$ and $\sigma = 1$ and with correlated errors. One noticeable result is that PWD often has much larger values of second-order Bayes risk relative to other methods and it remains larger for low and medium spatial correlations across all the samples sizes and spatial frames. Whereas for high spatial correlation, it decreases with sample size and tend to be same as those for other methods, specially for populations with linear spatial trend. In general, values of second-order Bayes risk (relative to those under SRS) increases with values of spatial correlation; as sample size increases, these values tend to increase for small values of spatial correlation and they tend to decrease for large values of spatial correlation. Another result which does not go with the general pattern of the results that relative values of second-order Bayes risk for BSS decreases (when populations have spatial trend and spatial correlation is high) with much slower rate such that they remain considerably larger than those under others sampling methods.

Table 4.5: Relative values of second-order Bayes risk under different sampling designs with respect to SRS, for different sample sizes (n) from highly clustered populations with low, medium and high spatial autocorrelation.

		No spatial trend			Linear spatial trend			Binary responses		
		z_1	z_2	z_3	z_4	z_5	z_6	$y_{(0.1)}$	$y_{(0.3)}$	$y_{(0.5)}$
4	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.089	1.190	12.636	1.910	3.631	30.918	3.143	8.476	9.701
	BSS	1.028	1.218	21.668	1.355	2.242	24.855	3.439	12.617	14.972
	LPM1	1.014	1.277	21.940	1.620	2.871	25.775	4.782	15.214	16.857
	LPM2	1.050	1.256	18.472	1.503	2.580	22.627	4.470	12.598	13.759
	SCPS	1.084	1.391	23.886	1.663	2.867	23.841	5.075	16.217	17.961
	DBSS	1.050	1.319	24.305	1.533	2.585	21.197	4.934	16.526	18.480
	PWD	16.134	26.741	168.305	4.518	9.215	57.849	132.584	141.400	140.967
12	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.111	1.756	18.986	1.926	3.025	14.621	11.557	14.355	13.350
	BSS	1.044	1.335	24.821	1.661	2.860	24.642	5.675	13.977	14.311
	LPM1	1.262	2.635	24.857	1.877	3.001	8.568	24.706	21.602	18.422
	LPM2	1.244	2.353	22.879	1.762	2.742	7.924	21.885	19.655	16.927
	SCPS	1.309	2.657	23.316	1.882	2.985	7.337	22.781	20.472	17.744
	DBSS	1.225	2.496	23.215	1.768	2.827	7.603	21.801	19.893	17.359
	PWD	36.512	109.916	104.760	7.079	18.801	16.632	123.587	126.016	126.505
20	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.307	2.273	19.952	1.730	2.636	11.081	21.206	16.760	14.481
	BSS	1.068	1.264	24.316	1.601	2.732	30.247	6.782	13.726	14.097
	LPM1	1.557	3.484	21.320	1.809	2.826	7.734	33.626	20.378	16.875
	LPM2	1.474	3.000	19.325	1.673	2.557	6.982	28.498	18.654	15.622
	SCPS	1.767	3.418	19.690	1.917	2.893	7.516	29.738	18.906	16.067
	DBSS	1.518	3.282	19.282	1.704	2.553	6.295	29.855	18.958	15.752
	PWD	28.622	57.126	36.993	5.863	10.923	9.656	80.533	57.175	50.607
32	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.419	3.046	20.400	1.972	2.488	6.905	26.311	17.199	14.827
	BSS	1.091	1.389	25.218	1.754	2.654	22.956	7.271	13.669	14.102
	LPM1	1.740	4.478	17.887	1.945	2.352	3.991	33.490	17.821	14.680
	LPM2	1.622	3.718	16.933	1.855	2.256	3.900	29.824	16.913	14.140
	SCPS	2.149	4.456	17.239	2.008	2.490	3.893	31.142	17.347	14.555
	DBSS	1.724	4.341	16.393	1.871	2.243	3.356	31.322	16.963	14.025
	PWD	23.988	43.498	25.932	5.120	6.883	3.795	61.701	37.249	32.733
44	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.584	3.271	20.985	1.979	2.823	5.889	29.673	17.609	15.372
	BSS	1.035	1.335	27.745	1.730	3.017	23.564	7.820	14.070	15.021
	LPM1	1.946	4.584	16.415	1.927	2.473	3.400	33.818	16.755	14.210
	LPM2	1.767	3.930	16.102	1.860	2.403	3.215	30.690	16.257	13.996
	SCPS	2.397	4.510	16.498	2.010	2.574	3.263	32.311	16.704	14.656
	DBSS	2.008	4.470	15.354	1.888	2.386	2.678	32.343	16.126	13.696
	PWD	18.359	27.764	20.196	4.876	6.607	3.408	51.242	27.653	25.340

Table 4.6: Relative values of second-order Bayes risk under different sampling designs with respect to SRS, for different sample sizes (n) from clustered populations with low, medium and high spatial autocorrelation.

		No spatial trend			Linear spatial trend			Binary responses		
n		z_1	z_2	z_3	z_4	z_5	z_6	$y_{(0.1)}$	$y_{(0.3)}$	$y_{(0.5)}$
4	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.085	1.131	13.333	1.786	2.895	25.680	2.736	8.335	10.182
	BBS	1.070	1.132	15.524	1.256	1.810	22.078	2.482	8.561	11.350
	LPM1	1.090	1.207	18.905	1.549	2.298	23.256	4.072	12.148	14.362
	LPM2	1.074	1.257	15.805	1.333	1.933	19.267	3.598	10.139	11.691
	SCPS	1.167	1.295	18.794	1.641	2.359	22.972	4.225	12.067	13.832
	DBSS	1.147	1.255	20.605	1.510	2.219	21.875	4.038	13.080	15.686
	PWD	15.895	23.163	93.494	4.839	8.441	44.352	90.331	95.028	99.590
12	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.119	1.442	18.275	2.070	3.006	18.309	9.102	12.421	12.575
	BBS	1.041	1.132	17.666	1.606	2.188	23.507	3.729	9.626	11.582
	LPM1	1.251	1.695	23.038	1.853	2.574	13.839	15.083	16.868	15.826
	LPM2	1.204	1.599	20.787	1.715	2.345	12.452	12.961	14.989	14.434
	SCPS	1.373	1.880	21.198	1.922	2.442	12.335	13.502	15.709	15.264
	DBSS	1.213	1.708	22.694	1.726	2.391	11.621	14.414	16.521	15.887
	PWD	29.317	58.466	100.811	7.067	12.820	23.635	96.454	104.998	101.273
20	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.315	1.729	21.677	2.080	2.918	14.594	14.782	14.299	13.956
	BBS	1.043	1.146	19.681	1.785	2.388	24.094	4.442	9.886	11.812
	LPM1	1.446	2.175	22.387	1.946	2.592	10.099	21.607	16.758	15.208
	LPM2	1.338	1.931	20.706	1.823	2.389	9.530	18.453	15.205	14.024
	SCPS	1.656	2.229	21.138	2.063	2.577	9.389	19.419	16.051	14.831
	DBSS	1.483	2.070	22.007	1.876	2.424	8.341	20.360	16.389	15.070
	PWD	28.061	53.459	50.116	5.746	9.679	8.725	70.853	57.803	54.187
32	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.420	2.116	19.739	1.721	2.520	11.617	17.697	15.018	13.675
	BBS.	1.058	1.148	18.317	1.521	2.170	23.540	4.491	9.786	11.451
	LPM1	1.713	2.810	17.485	1.642	2.150	7.205	23.118	15.211	13.151
	LPM2	1.607	2.471	16.807	1.589	2.046	6.889	20.270	14.513	12.864
	SCPS	2.084	2.996	16.841	1.757	2.258	6.975	21.695	15.317	13.298
	DBSS	1.711	2.783	17.288	1.595	2.048	5.780	22.548	15.281	13.259
	PWD	20.827	35.854	28.138	4.390	6.789	6.475	50.635	36.685	31.914
44	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.546	2.466	20.402	1.886	2.563	12.324	19.996	14.871	13.545
	BBS	1.054	1.159	19.551	1.663	2.328	29.752	4.615	9.484	11.307
	LPM1	2.002	3.190	16.234	1.799	2.145	7.103	23.636	13.979	12.201
	LPM2	1.826	2.783	15.920	1.710	2.090	7.058	21.397	13.640	11.978
	SCPS	2.439	3.484	16.050	1.904	2.342	6.959	23.209	14.391	12.470
	DBSS	2.038	3.193	15.893	1.789	2.079	5.792	23.097	13.938	12.181
	PWD	17.963	27.890	22.290	3.999	5.279	5.494	42.239	27.868	24.855

Table 4.7: Relative values of second-order Bayes risk under different sampling designs with respect to SRS, for different sample sizes (n) from sparse populations with low, medium and high spatial autocorrelation.

		No spatial trend			Linear spatial trend			Binary responses		
n		z_1	z_2	z_3	z_4	z_5	z_6	$y_{(0.1)}$	$y_{(0.3)}$	$y_{(0.5)}$
4	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.086	1.169	8.643	2.022	2.437	16.021	2.136	5.752	6.696
	BBS	1.103	1.149	9.308	1.322	1.512	10.958	1.938	5.717	6.942
	LPM1	1.077	1.112	10.128	1.695	1.998	14.618	2.578	6.864	7.862
	LPM2	1.101	1.160	8.899	1.479	1.695	12.168	2.436	6.118	6.814
	SCPS	1.116	1.177	11.573	1.747	2.002	13.770	2.819	7.817	8.938
	DBSS	1.081	1.159	11.500	1.535	1.827	12.723	2.712	7.664	8.851
	PWD	10.376	11.885	46.863	3.803	4.697	21.582	61.247	44.706	40.055
12	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.157	1.272	11.939	2.173	2.585	14.758	5.245	7.575	7.733
	BBS	1.021	1.136	11.910	1.639	1.961	15.716	2.764	6.391	7.180
	LPM1	1.176	1.410	14.631	1.821	2.136	10.602	8.419	10.075	9.579
	LPM2	1.192	1.330	13.761	1.723	1.980	10.010	7.585	9.594	9.168
	SCPS	1.329	1.573	15.541	1.872	2.136	9.264	9.202	11.021	10.612
	DBSS	1.250	1.399	15.837	1.771	2.100	9.535	8.609	10.879	10.492
	PWD	6.344	7.246	24.851	2.726	3.282	9.928	24.415	21.818	20.012
20	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.268	1.354	13.020	2.118	2.480	12.551	7.206	9.066	8.865
	BBS	1.040	1.062	12.169	1.681	1.947	16.301	3.039	6.670	7.502
	LPM1	1.364	1.578	13.765	1.729	2.012	7.818	11.122	10.928	9.967
	LPM2	1.339	1.531	13.203	1.718	1.945	7.311	10.234	10.448	9.681
	SCPS	1.547	1.836	14.188	1.838	2.083	6.723	12.216	11.574	10.648
	DBSS	1.440	1.591	14.951	1.779	2.002	6.863	11.611	11.801	10.854
	PWD	12.592	15.286	25.400	3.701	4.446	7.453	38.118	27.623	24.562
32	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.461	1.648	13.918	1.866	2.128	6.483	9.964	9.747	9.416
	BBS	1.086	1.087	12.165	1.534	1.835	10.888	3.283	6.411	7.167
	LPM1	1.633	2.020	12.551	1.644	1.866	3.688	14.475	10.627	9.456
	LPM2	1.576	1.891	12.317	1.579	1.776	3.555	13.342	10.310	9.257
	SCPS	1.972	2.388	12.644	1.696	1.939	3.102	15.236	11.144	9.961
	DBSS	1.649	2.026	13.255	1.632	1.802	3.160	14.666	11.002	10.024
	PWD	19.509	23.151	21.054	4.451	5.617	3.821	42.097	27.529	23.780
44	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.607	1.823	13.210	2.325	2.613	7.757	11.815	10.431	9.381
	BBS	1.076	1.081	12.454	1.950	2.294	15.146	3.444	6.995	7.417
	LPM1	2.008	2.389	11.240	2.148	2.254	3.949	16.515	11.131	9.388
	LPM2	1.823	2.226	11.103	2.086	2.200	3.982	15.364	10.774	9.119
	SCPS	2.491	2.889	11.188	2.240	2.397	3.545	17.387	11.372	9.661
	DBSS	1.991	2.453	11.665	2.129	2.265	3.467	16.675	11.449	9.753
	PWD	16.971	20.570	15.059	4.661	5.687	3.271	36.802	23.267	18.866

4.3.2 Linear regression model

Now consider the linear regression model given by $y_i = \mu_i + \epsilon_i$, where $\mathbf{x}_i^\top \boldsymbol{\beta}$ is linear predictor and ϵ_i is random error term. When the errors are independent with constant variance, the AMSE of HT-estimator is minimum under balanced equal probability sampling design and spatial balancing in addition to auxiliary balance (i.e. doubly balanced sampling) is expected to add nothing in terms of efficiency. As pointed out in literature and discussed in previous sections, errors can be correlated in reality. Therefore, spatial balancing in addition to auxiliary balance is expect to improve the AMSE in this case. As mentioned earlier, two types of balance (auxiliary and spatial) are not same in general. There might be compromise between the two under a doubly balanced design. In order to investigate this, simulation study from Section 4.2.2 is partly extend (using only sparse spatial frame and sample sizes $n = 20, 44$) for the linear regression model with spatially correlated errors given by $y_i = 1 + x_i\beta + \epsilon_i$, where $x_i = i/N$.

For the spatial frame with sparse population units, twelve spatial populations $z_{ij}(i = 1, 2, 3, 4; j = 1, 2, 3)$ are generated, with three level of spatial autocorrelation low, medium, hight and four values of regression coefficient β to simulate different levels of variation explained $(10\%, 30\%, 50\%, 70\%)$ by the linear predictor of the model. To computer the first- (i.e. AMSE) and second-order Bayes risks of HT-estimator under the given model, 5000 populations are generated and 5000 equal probability samples are selected from each population under the sampling designs including auxiliary, spatially and doubly balanced sampling, given by

- CUBE: Balanced sampling with respect to auxiliary variable using cube method,
- LPM$_X$: Balance with respect to auxiliary variable using LPM1,
- LPM: Spatially balance sampling using LPM1,
- LCUBE: Doubly balanced sampling using local cube method.

Two different sample sizes are considered $n = (20, 44)$. Values of first- and second-order Bayes risks relative to those under SRS are shown in Tables 4.8 and 4.9 respectively. First row in the Table 4.8, denoted by V$(\hat{X}_{HT})$, is auxiliary balance achieved by sampling designs relative that of SRS. The results for the AMSE's shows that the most balanced design is cube method, LPM method (i.e. LPM$_X$) achieves some balance with respect to auxiliary variable but less than cube method. Doubly balanced sampling design compromise some auxiliary balance in order to achieve spatial balance, it may not be suitable when variation explained by the auxiliary variable is relatively high, for instance, for those populations with low spatial autocorrelation i.e. z_{41} and z_{42}. For auxiliary and doubly balanced designs, second-order Bayes risk (from Table 4.9) tend to decrease as the proportion of variation explained by the auxiliary variable increases, while values for LPM do change in the same direction but are not very different for larger sample size.

Table 4.8: Relative values of AMSE's with respect to SRS under auxiliary, spatial and doubly balanced sampling designs, and linear regression model with correlated errors.

	$n = 20$					$n = 44$				
	SRS	CUBE	LPM$_X$	LPM	LCUBE	SRS	CUBE	LPM$_X$	LPM	LCUBE
$\mathrm{V}(\hat{X}_{HT})$	1.000	0.046	0.056	0.994	0.060	1.000	0.022	0.029	1.038	0.031
z_{11}	1.000	0.888	0.889	0.999	0.889	1.000	0.887	0.887	1.004	0.887
z_{12}	1.000	0.887	0.887	0.987	0.884	1.000	0.886	0.887	0.975	0.874
z_{13}	1.000	0.881	0.879	0.678	0.660	1.000	0.878	0.878	0.567	0.550
z_{21}	1.000	0.728	0.731	0.998	0.732	1.000	0.727	0.728	1.010	0.728
z_{22}	1.000	0.727	0.730	0.988	0.728	1.000	0.726	0.728	0.987	0.718
z_{23}	1.000	0.714	0.715	0.740	0.539	1.000	0.712	0.714	0.657	0.448
z_{31}	1.000	0.564	0.568	0.997	0.570	1.000	0.560	0.563	1.017	0.564
z_{32}	1.000	0.563	0.567	0.990	0.567	1.000	0.560	0.563	0.999	0.555
z_{33}	1.000	0.548	0.550	0.803	0.419	1.000	0.544	0.547	0.750	0.345
z_{41}	1.000	0.365	0.372	0.996	0.374	1.000	0.357	0.361	1.025	0.362
z_{42}	1.000	0.365	0.371	0.991	0.372	1.000	0.356	0.361	1.014	0.357
z_{43}	1.000	0.352	0.358	0.877	0.278	1.000	0.344	0.348	0.860	0.224

Table 4.9: Relative values of second-order Bayes risk with respect to SRS under auxiliary, spatial and doubly balanced sampling designs and linear regression model with correlated errors.

	$n = 20$					$n = 44$				
	SRS	CUBE	LPM$_X$	LPM	LCUBE	SRS	CUBE	LPM$_X$	LPM	LCUBE
z_{11}	1.000	2.820	3.062	1.345	3.115	1.000	2.947	3.798	2.052	3.505
z_{12}	1.000	2.750	3.012	1.499	3.109	1.000	2.913	3.712	2.305	3.586
z_{13}	1.000	3.372	3.615	13.075	8.327	1.000	3.867	4.437	13.051	8.712
z_{21}	1.000	3.674	3.829	1.299	3.862	1.000	3.887	4.502	2.040	4.241
z_{22}	1.000	3.574	3.715	1.427	3.774	1.000	3.843	4.428	2.295	4.195
z_{23}	1.000	6.259	6.336	12.124	5.044	1.000	7.126	7.433	13.383	4.737
z_{31}	1.000	3.375	3.483	1.262	3.490	1.000	3.605	4.029	2.038	3.826
z_{32}	1.000	3.289	3.373	1.371	3.392	1.000	3.575	3.977	2.288	3.743
z_{33}	1.000	7.983	7.939	10.684	3.335	1.000	8.967	9.110	12.895	2.706
z_{41}	1.000	2.246	2.322	1.227	2.305	1.000	2.420	2.661	2.041	2.537
z_{42}	1.000	2.205	2.263	1.319	2.251	1.000	2.415	2.639	2.285	2.477
z_{43}	1.000	7.657	7.547	8.286	2.161	1.000	8.518	8.553	11.185	1.477

4.4 Sampling from populations with negative spatial autocorrelation

The phenomenon of negative spatial autocorrelation is not as common as its positive version. Therefore, it has received relatively less attention, and literature is scant about the treatment of negative spatial autocorrelation in spatial statistics. According to Griffith and Arbia (2010), "Negative spatial autocorrelation refers to a geographic distribution of values, or a map pattern, in which the neighbours of locations with large values have small values, the neighbours of locations with intermediate values have intermediate values, and the neighbours of locations with small values have large values". Griffith (2011) and Chun and Griffith (2018) studied impact of positive and negative spatial autocorrelations on distributions of random variables respectively. Griffith and Arbia (2010) and Chun and Griffith (2018) mentioned some examples of negative spatial autocorrelation from literature.

In the context of socio-economic surveys, one may come across the situation when some study variables are spatially correlated in positive sense while some others are suspected to have negative spatial autocorrelation. Selecting a spatially balanced sample may compromise the efficiency of estimates for study variables with negative spatial autocorrelation. For such kind of surveys, one should be looking for a spatial sampling design which achieves spatial balance but also minimize the efficiency loss for those variables with negative spatial autocorrelation.

The efficiency of non-spatial sampling designs can be affected by either version (positive or negative) of the spatial autocorrelation, as it can change the distributions of variables under study. Almost all the spatially balanced sampling designs are motivated by the phenomenon of positive spatial autocorrelation in the study populations. These designs may not perform well or even under-perform as compared to non-spatial designs when spatial populations exhibit negative spatial autocorrelation (Altieri and Cocchi, 2021).

In this section, some sampling schemes are proposed which use cube and local cube algorithms (Deville and Tillé, 2004; Grafström and Tillé, 2013) and Moran eigenvector spatial filtering specification of a spatial regression model (Griffith, 2003, 2019). Before that bellow a spatial sampling design based on spatial entropy from literature is reviewed which aims to take into account both versions of spatial autocorrelation.

Spatial sampling using spatial entropy

Altieri and Cocchi (2021) proposed a weighting criteria, for SCPS spatially balanced sampling design (Grafström et al., 2012), based on measure of *spatial entropy*. In general, *entropy* is a heterogeneity measure for a random variable. In sampling theory, *sampling entropy* is associated with a sampling design which measures the randomness of the samples under that design. Spatial entropy measures the heterogeneity of the random variable in two-dimensional space. It is decomposed into two factors, one factor which is related to spatial configuration of the population units is used to construct weights for the SCPS design, for more details see the original article from Altieri and Cocchi (2021).

Altieri and Cocchi (2021) conducted a simulation study based on different artificial spatial populations to compare spatially balanced sampling methods including LPM, SCPS, PWD and the proposed spatial sampling design named as SPI (SCPS with weights based on spatial entropy) with SRS. In the spatial populations, binary response variables (with proportion of ones $p = 0.25, 0.5$) were simulated with different spatial structures including regular (negative spatial autocorrelation), random (no spatial variation), compact (strong positive spatial autocorrelation) and multi-cluster (weak positive spatial autocorrelation). A regular grid of points was used as spatial configuration of population units. Equal probability samples were selected with respect to three different sampling fractions $f = (0.01, 0.05, 0.10)$ under the sampling designs. The proposed SPI design was the most efficient for populations with negative spatial autocorrelation and with no spatial variation. The PWD was the most spatially balanced and the most efficient for multi-clustered population under all the three sampling fractions. It was the most efficient for the population with compact spatial structure only when the sampling fraction was $f = 0.01$, while LPM and SCPS were more efficient than PWD when sampling fraction was $f = (0.05, 0.10)$, both being equally spatially balanced and equally efficient with slight differences.

4.4.1 Spatial sampling using eigenvectors of modified SWM

For analysis of spatial data, two basic and commonly used regression models are autoregressive (AR) model and simultaneous autoregressive (SAR) model. They are also known as spatial lag model and spatial error lag model respectively, given by

$$\begin{aligned} &\text{SA:} && \mathbf{y} = \rho\mathbf{M}\mathbf{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &\text{SAR:} && \mathbf{y} = X\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{M})^{-1}\boldsymbol{\epsilon} \end{aligned}$$

where $\mathbf{M}$ is row standardised spatial weight matrix (SWM), see Section 1.5.1, ρ is spatial dependence parameter such that $|\rho| < 1$ and $\boldsymbol{\epsilon}$ is vector of independently identically distributed (iid) random errors under the standard normal distribution. Spectral decomposition of the matrix $\mathbf{M}$ gives N orthogonal eigenvector and associated eigenvalues. Each eigenvector represents a unique pattern which characterise the spatial dependence in the response values, i.e. y_i's. Negative or positive sign eigenvalue tells the nature, while value of the eigenvalue represents strength of the spatial autocorrelation associated with the eigenvector.

Under the eigenvector spatial filtering (ESF) model specification (Griffith, 2003), a set of synthetic proxy variables is added in the regression model as control variables in order to account for spatially correlated error in the regression model. The proxy variables are extracted as eigenvectors from spatial weight matrix $\mathbf{M}$. Using ESF, spatial regression model can be written as

$$\mathbf{y} = X\boldsymbol{\beta} + E_k\boldsymbol{\beta}_E + \boldsymbol{\epsilon} \tag{4.3}$$

where $\boldsymbol{\epsilon}$ is vector of iid random errors under the standard normal distribution after spatial autocorrelation being filtered out by adding eigenvectors E_k's, as control variables, in the mean function of regression model. There are N pairs of eigenvalues and eigenvectors which can be large in number for large data set. Therefore, a two-step criteria is used to selected relevant set of eigenvectors: first eigenvectors are selected such that $|\lambda_1/\lambda_N| > 0.25$, where λ_1 and λ_N are the smallest and largest eigenvalues of the matrix $\mathbf{M}$; second, further selection is done by fitting stepwise (backward or forward) linear regression.

Since eigenvectors with large negative eigenvalues characterise negative spatial autocorrelation, therefore balanced sampling with respect to these eigenvectors is expected to given protection against loss of efficiency for variables with negative spatial autocorrelation. The idea for the proposed spatial sampling scheme is to select samples, using local cube algorithm, which is spatially balanced with respect to spatial coordinates and balanced with respect to spatial coordinates and the eigenvectors (associated with negative spatial autocorrelation). Under this sampling scheme, it is expected that some efficiency will be compromise for variables with positive spatial autocorrelation. Following three sampling schemes are explored through a simulation study:

1. LCUBE$_{(sp,E)}$: Selecting samples spatially balanced with respect to spatial coordinates and balanced with respect to eigenvectors associated with negative spatial autocorrelation of the population,

2. LCUBE$_{(E,sp)}$: Selecting samples spatially balanced with respect to spatial coordinates and balanced with respect to eigenvector and spatial coordinates (in the order they are written in the subscript, since landing phase of the local cube methods drops the last variable to complete the sample selection),

3. CUBE$_{(sp,E)}$: Selecting samples balanced with respect to spatial coordinates and eigenvectors using cube method.

4. CUBE$_{(E,sp)}$: Selecting samples balanced with respect to eigenvectors and spatial coordinates using cube method.

4.4.2 Simulation study for the proposed spatial sampling schemes

In this simulation study, three spatial frames: highly clustered, clustered and sparse, are considered. For each spatial frame, nine variables (or spatial populations) are simulated each of size $N = 400$: six with positive $(z_1, ..., z_6)$ and three with negative $(z_7, ..., z_9)$ spatial autocorrelation. Variables with positive spatial autocorrelation are simulated in the similar manner as in simulation study from Section 4.2.2 with spatial trend. In order to simulate variables with negative spatial autocorrelation, first a variable z was simulated from normal distribution with mean 5 and unity standard deviation. Then, spatially balanced samples of size $40, 120$ and 200 were selected from each frame. Values of the z variable for selected units were replaced by a value 10, which gave z_7, z_8, z_9 for selected values $40, 120$ and 200 respectively. One realization for z_7, z_8, z_9 for each of three spatial frames is show in Figure 4.2.

From each spatial population, 5000 equal probability samples are selected with respect to sampling fractions given by $f = (0.05, 0.11)$ i.e. $n = (20, 44)$. AMSE's of HT-estimator of population totals for each of nine study variables are computed in the same manner as in the simulation study from Section 4.2.2. Relative values of AMSE's under spatial sampling design with respect to SRS are shown in Tables 4.10, 4.11 and 4.12 for the spatial frames with highly clustered, clustered and sparse spatial configuration of population units.

Results from the Tables 4.10, 4.11 and 4.12 show that proposed sampling schemes tend to be more efficient than spatial balanced sampling designs for populations with negative spatial autocorrelation, but also compromise some efficiency for populations with positive spatial autocorrelation. By increasing sample size, efficiency of the proposed sampling schemes increases; compromise of efficiency tend to decrease for spatial population with linear spatial trend, but it increase for populations with no spatial trend.

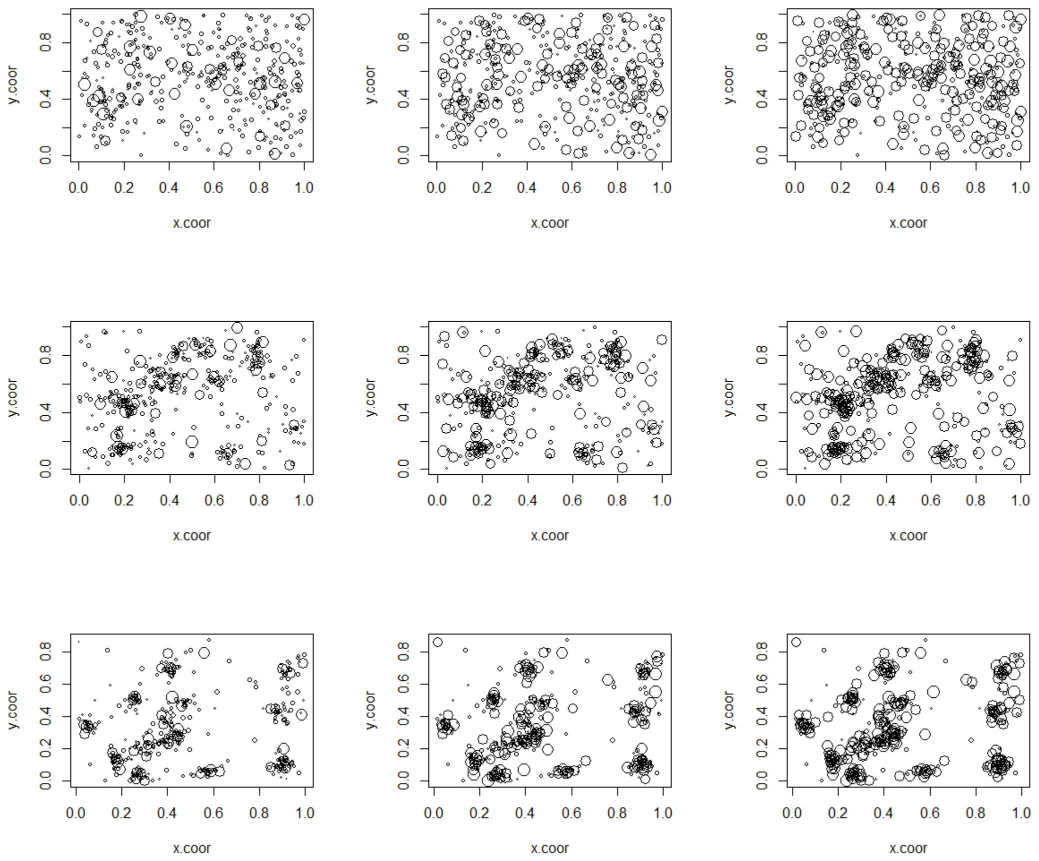


Figure 4.2: Populations with negative spatial autocorrelations: reading from left to right, top layer show z_7, z_8, z_9 for sparse spatial frame, middle layer shows z_7, z_8, z_9 for clustered spatial frame and bottom layer show z_7, z_8, z_9 for highly clustered spatial frame

One issue with the proposed sampling schemes which might not be acceptable that they can be less efficient than SRS for populations with no spatial trend having low and medium spatial autocorrelations. In this regard, two sampling schemes based on local cube method seems to be less problematic as compared to those based on cube method. On other hand, sampling schemes based on local cube method achieve lesser efficiency for the populations with negative spatial autocorrelation.

A direct comparison with the spatial sampling method (denoted by SPI) from Altieri and Cocchi (2021) may not be possible, as its implementation in R software is not available until now, to our best knowledge. An indirect comparison might be possible by computing the values of MSE's relative to SRS from the simulation study conducted in the original article, see Table 4.13. A brief description of the simulation already provided earlier in this section. The results in Table 4.13 shows that SPI method is always better than SRS and spatially balanced sampling methods for populations with negative spatial autocorrelation. However, it compromise more in terms of efficiency for populations with positive spatial autocorrelation as compared to proposed sampling strategies. For instance, percent gain in efficiency as compared to spatially balanced sampling is computed, using Eq. (4.4) for SPI spatial sampling method (where MSE_{sp} denotes minimum MSE achieved by spatially balanced designs) and proposed sampling schemes using Eq. (4.5). Results are shown in Table 4.14. For populations with negative spatial spatial autocorrelation, maximum gain under SPI (as compared to spatially balanced sampling) is 48% and maximum loss is 1449% (for population with positive spatial autocorrelation). For the proposed sampling schemes, maximum gain is 26% while maximum loss is 362%.

$$\text{Percent gain in efficiency} = \left(1 - \frac{\text{MSE}_{SPI}}{\text{MSE}_{sp}}\right) \times 100 \tag{4.4}$$

$$\text{Percent gain in efficiency} = \left(1 - \frac{\text{AMSE}_{proposed}}{\text{AMSE}_{PWD}}\right) \times 100 \tag{4.5}$$

The four proposed strategies are based on a heuristic idea of using eigenvector for spatial sampling, a further exploration about them may provide better results. For instance, selection of appropriate eigenvectors; usually a larger set of eigenvectors characterise negative spatial autocorrelation as compared to those which characterise positive spatial autocorrelation. This was also evident from the simulation studies presented here. An adjustment in the threshold $|\lambda_1/\lambda_N| > 0.25$, or selection of particular eigenvectors based on some known information about spatial pattern of the population units might be useful.

Table 4.10: Relative AMSE's for propose sampling strategies and spatial balanced design with respect to SRS for highly clustered spatial populations

n		z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9
20	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	0.999	0.951	0.607	0.375	0.364	0.275	1.023	1.030	1.032
	BBS	1.000	0.994	0.907	0.277	0.276	0.247	1.004	1.004	1.000
	LPM1	0.999	0.929	0.463	0.303	0.286	0.167	1.037	1.043	1.043
	LPM2	0.999	0.934	0.487	0.302	0.286	0.171	1.033	1.043	1.042
	SCPS	0.999	0.935	0.484	0.299	0.284	0.168	1.035	1.042	1.038
	DBSS	0.999	0.931	0.462	0.288	0.271	0.151	1.036	1.043	1.043
	PWD	0.994	0.848	0.317	0.283	0.248	0.115	1.042	1.047	1.045
	LCUBE$_{(sp,E)}$	1.008	0.988	0.595	0.287	0.282	0.179	1.025	1.016	0.985
	LCUBE$_{(E,sp)}$	1.011	0.998	0.614	0.373	0.370	0.270	1.025	1.010	0.973
	CUBE$_{(sp,E)}$	1.010	1.041	0.957	0.285	0.293	0.264	0.998	0.981	0.946
	CUBE$_{(E,sp)}$	1.012	1.052	0.976	0.332	0.342	0.315	0.997	0.975	0.935
44	SRS.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS.1	0.998	0.914	0.496	0.308	0.288	0.183	1.037	1.057	1.067
	BBS.1	1.000	0.994	0.899	0.255	0.253	0.223	1.004	1.005	1.001
	LPM1.1	0.999	0.864	0.364	0.273	0.241	0.117	1.066	1.096	1.102
	LPM2.1	0.998	0.877	0.385	0.271	0.242	0.120	1.058	1.091	1.100
	SCPS.1	0.998	0.879	0.381	0.269	0.240	0.116	1.058	1.089	1.091
	DBSS.1	0.998	0.868	0.351	0.258	0.227	0.099	1.064	1.095	1.101
	PWD.1	0.992	0.787	0.263	0.247	0.198	0.069	1.073	1.108	1.113
	LCUBE$_{(sp,E)}$	1.009	0.945	0.457	0.262	0.246	0.124	1.049	1.053	1.023
	LCUBE$_{(E,sp)}$	1.012	0.953	0.468	0.310	0.296	0.175	1.048	1.050	1.015
	CUBE$_{(sp,E)}$	1.011	1.045	0.943	0.262	0.270	0.238	1.000	0.979	0.940
	CUBE$_{(E,sp)}$	1.013	1.052	0.950	0.291	0.300	0.269	0.998	0.977	0.936

Table 4.11: Relative AMSE's for propose sampling strategies and spatial balanced design with respect to SRS for clustered spatial populations

n		z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9
20	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.974	0.683	0.351	0.345	0.274	1.024	1.032	1.035
	BBS	1.000	0.998	0.925	0.263	0.263	0.240	1.003	1.003	1.003
	LPM1	0.999	0.965	0.563	0.301	0.293	0.197	1.037	1.044	1.046
	LPM2	0.999	0.966	0.586	0.300	0.292	0.200	1.035	1.043	1.045
	SCPS	1.000	0.969	0.576	0.291	0.283	0.188	1.035	1.043	1.045
	DBSS	1.000	0.965	0.559	0.275	0.267	0.169	1.037	1.044	1.046
	PWD	1.000	0.922	0.428	0.246	0.228	0.109	1.045	1.047	1.048
	LCUBE$_{(sp,E)}$	1.002	1.090	0.861	0.275	0.295	0.235	1.003	0.945	0.863
	LCUBE$_{(E,sp)}$	1.002	1.109	0.934	0.526	0.551	0.505	0.997	0.930	0.837
	CUBE$_{(sp,E)}$	1.002	1.095	1.081	0.272	0.294	0.284	0.985	0.932	0.860
	CUBE$_{(E,sp)}$	1.003	1.111	1.135	0.439	0.464	0.463	0.981	0.919	0.836
44	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.951	0.570	0.288	0.277	0.186	1.040	1.061	1.070
	BBS	1.000	0.998	0.914	0.245	0.244	0.220	1.004	1.005	1.003
	LPM1	1.000	0.926	0.446	0.269	0.251	0.138	1.067	1.099	1.104
	LPM2	1.000	0.932	0.466	0.265	0.250	0.140	1.061	1.094	1.101
	SCPS	1.000	0.937	0.454	0.264	0.250	0.135	1.063	1.092	1.097
	DBSS	1.000	0.929	0.434	0.252	0.236	0.119	1.065	1.097	1.103
	PWD	1.000	0.877	0.343	0.239	0.211	0.086	1.079	1.111	1.114
	LCUBE$_{(sp,E)}$	1.003	1.091	0.712	0.251	0.271	0.178	1.022	0.959	0.857
	LCUBE$_{(E,sp)}$	1.003	1.099	0.746	0.389	0.411	0.324	1.020	0.953	0.844
	CUBE$_{(sp,E)}$	1.004	1.106	1.073	0.249	0.273	0.259	0.985	0.924	0.837
	CUBE$_{(E,sp)}$	1.003	1.114	1.101	0.353	0.378	0.369	0.983	0.919	0.827

Table 4.12: Relative AMSE's for proposed spatial sampling schemes and spatial balanced design with respect to SRS for sparse spatial populations.

	n	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9
20	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	1.000	0.990	0.754	0.391	0.388	0.331	1.021	1.030	1.035
	BBS	1.000	0.999	0.938	0.273	0.272	0.255	1.003	1.004	1.004
	LPM1	1.000	0.985	0.632	0.326	0.322	0.237	1.037	1.044	1.047
	LPM2	1.000	0.985	0.645	0.318	0.315	0.232	1.035	1.043	1.046
	SCPS	0.999	0.987	0.618	0.298	0.295	0.205	1.038	1.043	1.044
	DBSS	0.999	0.986	0.624	0.289	0.285	0.197	1.036	1.044	1.046
	PWD	0.999	0.972	0.530	0.258	0.251	0.144	1.042	1.048	1.048
	LCUBE$_{(sp,E)}$	1.000	1.012	0.699	0.280	0.282	0.204	1.027	1.023	1.005
	LCUBE$_{(E,sp)}$	1.001	1.019	0.711	0.333	0.337	0.261	1.026	1.018	0.995
	CUBE$_{(sp,E)}$	1.002	1.027	0.967	0.273	0.279	0.262	0.999	0.985	0.959
	CUBE$_{(E,sp)}$	1.001	1.031	0.978	0.305	0.311	0.295	0.998	0.983	0.954
44	SRS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRTS	0.999	0.979	0.647	0.323	0.318	0.238	1.038	1.060	1.073
	BBS	1.000	0.999	0.929	0.255	0.254	0.235	1.005	1.005	1.005
	LPM1	1.000	0.967	0.502	0.282	0.274	0.162	1.069	1.099	1.109
	LPM2	1.000	0.968	0.516	0.280	0.272	0.163	1.065	1.095	1.106
	SCPS	0.999	0.970	0.490	0.270	0.263	0.146	1.071	1.097	1.103
	DBSS	0.999	0.968	0.489	0.262	0.254	0.138	1.066	1.097	1.106
	PWD	0.997	0.939	0.397	0.248	0.234	0.103	1.080	1.110	1.116
	LCUBE$_{(sp,E)}$	1.000	1.009	0.581	0.259	0.261	0.156	1.051	1.059	1.043
	LCUBE$_{(E,sp)}$	1.001	1.015	0.593	0.295	0.297	0.194	1.049	1.056	1.035
	CUBE$_{(sp,E)}$	1.001	1.033	0.962	0.257	0.264	0.244	1.001	0.984	0.953
	CUBE$_{(E,sp)}$	1.002	1.036	0.967	0.280	0.287	0.268	0.999	0.982	0.948

Table 4.13: Values of MSE from simulation of Altieri and Cocchi (2021); relative values of MSE with respect to SRS; and percent gain with respect to spatially balanced sampling.

			Value of MSE			MSE/MSE$_{srs}$		
p	n		SRS	SBS	SPI	SBS	SPI	Percent gain
0.5	50	Compact	30963	1944	10035	0.063	0.324	-416
		Multicluster	31792	18390	25615	0.578	0.806	-39
		Regular	29924	31112	26939	1.040	0.900	13
		Random	30040	31016	24363	1.032	0.811	21
	125	Compact	12026	448	3446	0.037	0.287	-669
		Multicluster	11745	4734	4832	0.403	0.411	-2
		Regular	11756	12042	7723	1.024	0.657	36
		Random	12196	11657	9550	0.956	0.783	18
	250	Compact	5659	139	1135	0.025	0.201	-717
		Multicluster	5609	1623	2992	0.289	0.533	-84
		Regular	5657	5439	2830	0.961	0.500	48
		Random	5762	5592	4463	0.970	0.775	20
0.25	50	Compact	22969	1985	8075	0.086	0.352	-307
		Multicluster	23459	12228	20761	0.521	0.885	-70
		Regular	22312	23621	20801	1.059	0.932	12
		Random	23028	21939	19939	0.953	0.866	9
	125	Compact	8864	443	3416	0.050	0.385	-671
		Multicluster	8670	3079	6903	0.355	0.796	-124
		Regular	8693	9200	7234	1.058	0.832	21
		Random	9087	8796	8131	0.968	0.895	8
	250	Compact	4288	147	2277	0.034	0.531	-1449
		Multicluster	4117	1083	2806	0.263	0.682	-159
		Regular	4332	4468	4263	1.031	0.984	5
		Random	4035	4080	3361	1.011	0.833	18

Table 4.14: Percent relative gain of propose spatial sampling schemes with respect PWD spatially balanced sampling design for highly clustered, clustered and sparse spatial frames.

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9
(Highly cluster $n = 20$)									
LCUBE$_{(sp,E)}$	-1	-17	-88	-1	-14	-56	2	3	6
LCUBE$_{(E,sp)}$	-2	-18	-94	-32	-49	-136	2	3	7
CUBE$_{(sp,E)}$	-2	-23	-202	-1	-18	-130	4	6	9
CUBE$_{(E,sp)}$	-2	-24	-208	-17	-38	-175	4	7	11
(Highly cluster $n = 44$)									
LCUBE$_{(sp,E)}$	-2	-20	-73	-6	-24	-79	2	5	8
LCUBE$_{(E,sp)}$	-2	-21	-78	-26	-50	-154	2	5	9
CUBE$_{(sp,E)}$	-2	-33	-258	-6	-36	-245	7	12	16
CUBE$_{(E,sp)}$	-2	-34	-261	-18	-52	-290	7	12	16
(Cluster $n = 20$)									
LCUBE$_{(sp,E)}$	-0	-18	-101	-12	-29	-115	4	10	18
LCUBE$_{(E,sp)}$	-0	-20	-118	-114	-142	-362	5	11	20
CUBE$_{(sp,E)}$	-0	-19	-153	-11	-29	-160	6	11	18
CUBE$_{(E,sp)}$	-0	-20	-165	-78	-103	-324	6	12	20
(Cluster $n = 44$)									
LCUBE$_{(sp,E)}$	-0	-24	-107	-5	-28	-108	5	14	23
LCUBE$_{(E,sp)}$	-0	-25	-117	-62	-94	-279	5	14	24
CUBE$_{(sp,E)}$	-0	-26	-213	-4	-29	-203	9	17	25
CUBE$_{(E,sp)}$	-0	-27	-221	-47	-79	-331	9	17	26
(Sparse $n = 20$)									
LCUBE$_{(sp,E)}$	-0	-4	-32	-8	-12	-42	1	2	4
LCUBE$_{(E,sp)}$	-0	-5	-34	-29	-34	-81	2	3	5
CUBE$_{(sp,E)}$	-0	-6	-82	-6	-11	-82	4	6	8
CUBE$_{(E,sp)}$	-0	-6	-84	-18	-24	-105	4	6	9
(Sparse $n = 44$)									
LCUBE$_{(sp,E)}$	-0	-7	-46	-5	-12	-52	3	5	7
LCUBE$_{(sp,E)}$	-0	-8	-49	-19	-27	-88	3	5	7
CUBE$_{(E,sp)}$	-0	-10	-142	-3	-13	-137	7	11	15
CUBE$_{(sp,E)}$	-0	-10	-144	-13	-23	-160	7	12	15

4.5 Variance estimation

Despite the fact that spatially or doubly balanced sampling designs are more efficient than designs which are not balanced under the assumption of positive spatial autocorrelation, it is still important to estimate the variance (or sampling error) of the estimates which is often a challenge. For example, an unbiased estimate of the sampling variance is desirable for the calculation of confidence intervals. It is also important to know a valid variance estimate when balancing out the precision of the sampling design with other costs involved in the sample survey. For spatially balanced sampling designs, local-mean variance estimator from Stevens Jr and Olsen (2003) is commonly used in practice and often recommended in literature (Stevens Jr and Olsen, 2004; Grafström, 2012; Grafström et al., 2012; Robertson et al., 2013; Benedetti and Piersimoni, 2017). For doubly balanced sampling, Grafström and Tillé (2013) proposed a variance estimator by combining estimators in Eq. (1.17) and Eq. (1.16) from Stevens Jr and Olsen (2003) and Deville and Tillé (2005) respectively. Both the estimators are described in Section 1.6. In this section, methodology of variance estimation for balanced sampling from Section 3.5 is extended for spatially and doubly balanced sampling designs. Again, the variance approximation is based on decomposition of the sampling variance of HT-estimator assuming the Moran eigenvectors spatial filtering formulation of the assumed spatial super-population model.

In the case of doubly balanced sampling using super-population model specification from Eq. (4.3), response values can be written as $y_i = \mathbf{x}_i^\top \boldsymbol{B} + E_k^\top \boldsymbol{B}_E + e_i$, where e_i is finite population residual associated with ith element, $\boldsymbol{B}$ and $\boldsymbol{B}_E$ are vectors of finite population regression coefficients associated with auxiliary variables and eigenvectors used in the working model respectively. The HT-estimator of population total can be written as $\hat{Y}_{HT} = \hat{\mathbf{X}}_{HT}^\top \boldsymbol{B} + \hat{\mathbf{E}}_{HT}^\top \boldsymbol{B}_E + \hat{e}_{HT}$, where $\hat{\mathbf{E}}_{HT}$ and $\hat{e}_{HT}$ estimates vector of totals of eigenvectors and total of finite population residuals. The sampling variance of the HT-estimator $\hat{Y}_{HT}$ under the doubly balanced sampling design can be written as

$$V_{\text{ESF}}(\hat{Y}_{HT}) = \boldsymbol{B}^\top \boldsymbol{\Lambda} \boldsymbol{B} + \boldsymbol{B}_E^\top \boldsymbol{\Lambda}_E \boldsymbol{B}_E + V(\hat{e}_{HT}) \tag{4.6}$$

where the matrix $\boldsymbol{\Lambda}$ is defined in Eq. (3.2) and is estimated empirically (in the same way as for variance estimation under balanced sampling in Chapter 3), $\hat{\boldsymbol{\Lambda}}_E$ is also estimated empirically based on many samples under the given sampling design, $\boldsymbol{B}$ vectors are also estimated in the similar manner (using Eq. (1.7)) based on the given sample, and $V(\hat{e}_{HT})$ is estimated using local-mean variance estimator from Eq. (1.17). Here, the subscript ESF stands for eigenvectors spatial filtering. Based on the proposed approximation, variance estimator under doubly balanced sampling is given by

$$\hat{V}_{\text{ESF}}(\hat{Y}_{HT}) = \hat{\boldsymbol{B}}^\top \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{B}} + \hat{\boldsymbol{B}}_E^\top \hat{\boldsymbol{\Lambda}}_E \hat{\boldsymbol{B}}_E + \hat{V}_{\text{NBH}}(\hat{e}_{HT}) \tag{4.7}$$

where $\hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Lambda}}_E$, $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{B}}_E$ denote the estimators of their respective finite population quantities. In case of spatially balanced, there are no auxiliary variables, therefore, first term in the variance approximation should be dropped and remaining terms are estimated in the same manner as described earlier.

In the following, two Examples 4.2 and 4.3 are presented based on simulation studies from literature (Grafström et al., 2012; Grafström and Tillé, 2013), which compare the proposed variance estimator against the local-mean variance estimator under spatially balanced designs and against an estimator from Grafström and Tillé (2013) under doubly balanced sampling using local cube method, respectively. The comparison is made in terms of percent relative bias and precision using Eq. (3.17).

Example 4.2. To compare the proposed variance estimator with local-mean variance estimator in Eq. (1.17) under spatially balanced sampling, two artificial spatial populations from Example. 5 and 6 of Grafström et al. (2012) are used; they consisted of $N_1 = 400$ and $N_2 = 200$ observations and denoted by U_1 and U_2 here, respectively. The population U_1 has strong spatial trend, while U_2 was generated (with no spatial trend) from Gaussian random field with exponential covariance function. Equal probability samples of different sizes, $n = (16, 32, 48)$ from U_1 and $n = (30, 50, 90)$ from U_2, are selected under GRTS, LPM1, LPM2, SCPS, DBSS and PWD spatially balanced sampling methods. From each population, 1000 samples of each size are selected under the aforementioned sampling designs, and empirical values of sampling variances are calculated for each design. Two variance estimators, proposed $\hat{V}_{\text{ESF}}(\hat{Y}_{HT})$ and local-mean $\hat{V}_{\text{NBH}}(\hat{Y}_{HT})$, are computed for each of 1000 samples. `R`-package `spsurvey` (Dumelle et al., 2021) is used for the calculation of local-mean variance estimator. While computing the proposed variance estimator, the same 1000 samples are used to compute the empirical estimates of the variance-covariance matrices $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_E$. Percent relative bias and stability using Eq. (3.17) are calculated. The results are reported in Tables 4.15 and 4.16 for populations U_1 and U_2 respectively.

Under the proposed variance estimation, eigenvectors and eigenvalues of the row standardised weight matrices are computed for both the spatial populations. Using the selection criterion $\lambda_1/\lambda_N > 0.25$ (where λ_1 and λ_N are smallest and largest eigenvalues respectively), four and seven eigenvectors (i.e. E_k's) were chosen for U_1 and U_2 respectively. Moran's I indices of spatial autocorrelation for response variables in the populations U_1 and U_2 are given by 0.35 and 0.12 respectively. After fitting linear regression model, given by $y_i = \mu + E_{ki}\beta_{E_k} + \epsilon_i$, for both the populations, Moran's I indices of population residuals for U_1 and U_2 are given by 0.16 and 0.04 respectively. These computations have not been used anywhere in calculation of proposed variance estimator, rather they are made to get an idea about amount of spatial correlation accounted by eigenvector spatial filtering under the working spatial model.

From Table 4.15, results for population U_1 show that the proposed variance estimator

Table 4.15: Results for U_1 population in Example 4.2: Relative percent bias and variance of proposed and local-mean variance estimators based on 1000 samples under spatially balanced sampling designs.

		Percent relative bias		Percent relative MSE	
n		$\hat{V}_{\text{NBH}}(\hat{Y}_{HT})$	$\hat{V}_{\text{ESF}}(\hat{Y}_{HT})$	$\hat{V}_{\text{NBH}}(\hat{Y}_{HT})$	$\hat{V}_{\text{ESF}}(\hat{Y}_{HT})$
16	GRTS	162.649	58.530	168.137	62.310
	LPM1	98.448	50.301	104.951	53.844
	LPM2	111.483	56.021	118.498	59.665
	SCPS	203.152	88.370	210.347	92.696
	DBSS	227.085	108.096	233.902	112.141
	PWD	759.583	356.439	768.670	362.270
32	GRTS	65.488	26.518	70.464	29.329
	LPM1	120.784	76.308	124.696	77.940
	LPM2	104.814	66.612	108.337	68.202
	SCPS	239.471	132.457	243.686	134.363
	DBSS	262.832	151.829	266.825	153.630
	PWD	886.202	479.720	891.774	482.175
48	GRTS	50.076	25.312	54.032	28.012
	LPM1	109.167	74.886	111.586	76.223
	LPM2	91.931	62.656	94.488	64.089
	SCPS	248.501	146.058	251.417	147.890
	DBSS	260.532	154.832	263.417	156.640
	PWD	860.644	497.296	865.238	500.111

is better than local-mean variance estimator in terms of both bias and precision under all the spatially balanced sampling designs considered in this study. Biases of the both estimator tend to decrease with sample size under GRTS design LPM2, while decrease under GRTS design is much quicker than LPM2. They behave approximately in similar manner for rest of the sampling designs, that is, their percent relative bias and MSE increase when moving from $n = 16$ to $n = 32$ whereas they become stable when moving from $n = 32$ to $n = 48$. From Table 4.16, results for population U_2 also suggest that the proposed variance estimator is better than the local-mean variance estimator in terms of both bias and stability.

Overall, both the estimators overestimate the true sampling variance under all the sampling designs. Although the magnitude of bias is very large (often more than 100% of MSE) for the population with spatial trend U_1, still the proposed variance estimator reduces the bias. Proposed estimator tend to perform better for population U_1 as compare to population U_2. One reason might be the greater spatial trend in the U_1, and proposed estimator aims at the mean function of the underlying super-population model. The

Table 4.16: Results for U_2 population in Example 4.2: Relative percent bias and variance of proposed and local-mean variance estimators based on 1000 samples under spatially balanced sampling designs.

		Percent relative bias		Percent relative MSE	
n		$\hat{V}_{\text{NBH}}(\hat{Y}_{HT})$	$\hat{V}_{\text{ESF}}(\hat{Y}_{HT})$	$\hat{V}_{\text{NBH}}(\hat{Y}_{HT})$	$\hat{V}_{\text{ESF}}(\hat{Y}_{HT})$
30	GRTS	-0.369	3.660	21.405	25.281
	LPM1	16.334	8.358	31.453	28.198
	LPM2	11.914	3.874	26.592	24.157
	SCPS	16.259	6.227	29.635	24.377
	DBSS	23.231	9.999	35.634	28.020
	PWD	36.251	11.936	44.901	27.202
50	GRTS	25.150	23.708	32.746	30.922
	LPM1	57.424	47.031	62.847	52.592
	LPM2	35.739	27.281	42.021	34.221
	SCPS	43.559	35.466	48.990	41.269
	DBSS	50.238	38.854	55.845	44.675
	PWD	91.014	72.325	95.127	76.537
90	GRTS	59.976	57.870	62.511	60.083
	LPM1	91.369	82.831	93.207	84.571
	LPM2	82.452	75.297	84.482	77.170
	SCPS	73.069	66.738	75.206	68.737
	DBSS	95.749	85.789	97.823	87.706
	PWD	140.718	127.178	142.495	128.900

PWD design often has much larger bias as compared to other designs, since it is the most efficient design among the designs considered in this study.

□

Example 4.3. To compare the proposed variance estimator against the variance estimator in Eq. (1.18) from Grafström and Tillé (2013) under doubly balanced sampling, a simulation study is partially replicated from the same article (Grafström and Tillé, 2013). In this simulation study, a spatial data set, known as `meuse`, consisting of $N = 164$ observations is used as a spatial population. The data set `meuse` was obtained from `R`-package `gstat` (Gräler et al., 2016), and eight variables from this data set are used, described in Table 4.17. Three variables: zinc, lead and cadmium are considered as study populations; other three: copper, elev and om are used as auxiliary (or balancing) variable; x and y are used for spatial balancing.

From the spatial population, 10000 samples of size $n = 50$ are selected with equal probability using local cube method. Using these samples, true empirical sampling variance was

Table 4.17: Description of variables in `meuse` spatial data set.

	variables	description
1	x :	x-coordinates,
2	y :	y-coordinates,
3	cadmium:	topsoil cadmium concentration,
4	copper:	topsoil copper concentration,
5	lead:	topsoil lead concentration,
6	zinc:	topsoil zinc concentration,
7	elev:	relative elevation,
8	om:	organic matter, as percentage.

Table 4.18: Results for Example 4.3: Ratio between average of the estimated variances (using two estimators $\hat{V}_{\text{DBS}}(\hat{Y}_{HT})$ and $\hat{V}_{\text{ESF}}(\hat{Y}_{HT})$) and true empirical variance based on 10000 simulations under equal probability doubly balanced sampling by local cube method.

	$\hat{V}_{\text{DBS}}(\hat{Y}_{HT})$	$\hat{V}_{\text{ESF}}(\hat{Y}_{HT})$	No. of E_k's
zinc	1.06	1.03	5
lead	1.14	1.02	15
cadmium	0.74	0.82	25

computed. For each sample, values for the proposed variance estimator $\hat{V}_{\text{ESF}}(\hat{Y}_{HT})$ are calculated for all the three study variables and then averaged over 10000 values. A ratio is computed between the average of 10000 estimates of proposed variance estimator and true empirical sampling variance. Ideally, the same ratio should have been calculated for the existing variance estimator $\hat{V}_{\text{DBS}}(\hat{Y}_{HT})$ for comparison. Due to complexity of calculations for the existing estimator, the values of ratio for this variance estimator are copied from the original simulation study, see first row in the last section of Table 2 from Grafström and Tillé (2013). The values of ratios for both variance estimators and three study variables are shown in Table 4.18. A value of the ratio less than 1 represents underestimation while a value greater than one indicates that the variance estimator overestimates the true sampling variance; a value close to 1 is desirable.

For the proposed variance estimation in this case, the criteria of choosing eigenvectors given by $\lambda_1/\lambda_N > 0.20$ did not work. Therefore, hit and trial method is used to find out appropriate number of eigenvectors to be used in the variance estimation such that a better estimator can be found. This suggests that a more sophisticated criteria is needed for proposed variance estimation under doubly balanced sampling. For this study, number of eigenvectors used for each study variables are given in the last column of Table 4.18.

From Table 4.18, results shows that the proposed variance estimator $\hat{V}_{\text{ESF}}(\hat{Y}_{HT})$ has smaller bias than that of the existing variance estimator $\hat{V}_{\text{DBS}}(\hat{Y}_{HT})$. However, both the estimators overestimate the sampling variance for study populations zinc and lead, and underestimate for the study population cadmium.

□

Simulation studies in the above examples suggest that the proposed variance estimation methodology is more accurate and more precise than the local-mean (or local-neighbourhood) variance estimation.

4.6 Conclusions

Production and use of geo-referenced (or spatial data) is not restricted to geo-sciences any more; its utilization in different fields is increasing every day. An essential aspect of the geo-reference data is that it reveals impact of spatial location on the data values. A common phenomena is positive spatial autocorrelation, where values which are collected from nearby locations are tend to be similar. Often, sample surveys are conducted to collect data. It is emphasized in literature that sampling designs should take into account the location of the sampling units. Spatially balanced sampling method aims to select units which are well-spread over the study area using location of sampling units and tend to improve the efficiency of sample estimates.

A variety of spatially balanced sampling methods is found in literature which aims to select well-spread sample using different algorithms and techniques. Spatially balanced sampling methods based on the distance measures between pair of units are often more efficient than those based on spatial partitioning of the study area. All the spatially balanced sampling methods assume that the study population has positive spatial autocorrelation including spatial trends and different kinds spatial patterns. Some methods may perform better than others for a particular type of populations. Comparisons of different sampling method were conducted in literature on the basis of MSE and measure of spatial balance. Here, those comparison are extent on the basis of AMSE under a spatial super-population model which assume positive spatial autocorrelation. Under equal probability sampling, spatially balanced sampling design which selects samples with probability proportional to product of within sample distant (PWD) is found to have smallest AMSE among the sampling methods compared here. On the other hand, it has high variability in MSE (i.e. high second-order Bayes risk). It means that performance of this method is better than others on average but its efficiency may vary a lot for different spatial populations.

When a set of geo-referenced auxiliary variables is known, doubly balanced sampling with respect to both auxiliary variables and the location is suggested in literature. Spatially balanced sampling also achieve some balance with respect to auxiliary variable under some conditions, while they are not same in general. In most cases doubly balanced sampling improve the efficiency of estimates (in terms of AMSE) as compared to auxiliary balanced sampling. When variation explained by auxiliary variables is very high relative to spatial correlation in the study variable, doubly balanced sampling should be used cautiously.

In the context of multi-objective surveys, for instance, socio-economic surveys collect data on a variety of variables; some study variables may have negative spatial autocorrelation or complete randomness, whereas phenomenon of positive spatial autocorrelation is common. Spatially balanced samples can be less efficient than SRS for the estimation of parameters for those variables with negative spatial autocorrelation. The proposed spatial sampling schemes aims to minimize the loss of the efficiency for variables with negative spatial autocorrelation. They tend to achieve efficiency (in terms of AMSE) which is comparable with spatially balanced sampling designs and also have smaller loss of efficiency for those population with negative spatial autocorrelation. There are some scenarios, for instance, when positive spatial autocorrelation is low, the proposed sampling schemes may also be less efficient than SRS but not spatially balanced sampling methods.

Two-stage sampling design is often motivated by low survey cost. Spatially or doubly balanced sampling aims to spread the sampling units in the population area. Further work is required to investigate the effect of spatially balanced sampling on survey cost under two-stage sampling design. Furthermore, when PSU's are selected under spatially balanced sampling, distance between centres of PSU's is measured. For areal sampling units (which is the often case for PSU's), there exist other methods of measuring proximity. One may also look to investigate how different ways of measuring distance impact the efficiency of spatially balanced sampling of PSU's.

Variance estimation is a typical challenge associated with spatially balanced sampling design. Local-mean variance estimator is commonly used for most of the spatially balanced sampling designs. Propose variance estimator under spatially balanced sampling tends to perform better than local-mean variance estimator under almost all the spatially balanced sampling designs considered in this study when samples are selected with equal probability. The proposed variance estimator under doubly balanced sampling also perform better than the existing estimator, however, it needs further work for better selection of appropriate eigenvectors.

Here, it is important to mention that proposed spatial sampling schemes and variance estimators based on heuristic idea of using eigenvector spatial filtering (ESF) from spatial data analysis literature. This idea needs further exploration for better results. Since selection of eigenvectors for spatial sampling and for variance estimation is based on the a fixed criteria which is adopted from literature. It was realised during simulation studies that deviating from that fixed criteria might be useful for some situation, for example, in case of variance estimation under doubly balanced sampling different number of eigenvectors are used to get better results.

Appendix A

A.1 AMSE derivations for HT- and GREG-estimator under two-stage sampling by πps-SRS and two-level model

The AMSE of the HT-estimator is given by $\text{AMSE}(\hat{Y}_{HT}) = E_m[E_p(\hat{Y}_{HT} - Y)^2]$, where E_m and E_p denotes expectation under model and sampling design respectively. Since $y_{gi} = \mu_{gi} + v_g + e_{gi} = \mu_{gi} + \epsilon_{gi}$, where $\epsilon_{gi} = v_g + e_{gi}$. Therefore, AMSE can be written as

$$\text{AMSE}(\hat{Y}_{HT}) = E_m[E_p(\hat{\mu}_{HT} - \mu)^2] + E_m[E_p(\hat{\epsilon}_{HT} - \epsilon)^2] \tag{A.1}$$

where μ_{HT} is HT-estimator of total $\mu = \sum_{i\in U} \mu_{gi}$ and ϵ_{HT} is HT-estimator of population total $\epsilon = \sum_{i\in U} \epsilon_{gi}$.

AMSE under two-stage elements sampling: General formulation of sampling variance under two-stage element sampling design, from (Särndal et al., 1992, p. 307), is given by

$$\begin{aligned} V(\hat{Y}_{HT}) &= V_1[E_2(\hat{Y}_{HT})] + E_1[V_2(\hat{Y}_{HT})] \\ &= V_1\left[E_2\left(\sum_{g\in s_\text{I}} \frac{\hat{Y}_{HTg}}{\pi_g}\right)\right] + E_1\left[V_2\left(\sum_{g\in s_\text{I}} \frac{\hat{Y}_{HTg}}{\pi_g}\right)\right] = V_1\left(\sum_{g\in s_\text{I}} \frac{Y_g}{\pi_g}\right) + \sum_{g\in U_\text{I}} \frac{1}{\pi_g} V_2\left(\hat{Y}_{HTg}\right) \end{aligned}$$

Using Eq. (A.1)

$$\text{AMSE}(\hat{Y}_{HT}) = E_m\left[V_1\left(\sum_{g\in s_\text{I}} \frac{\mu_g}{\pi_g}\right) + \sum_{g\in U_\text{I}} \frac{1}{\pi_g} V_2\left(\hat{\mu}_{HTg}\right)\right] + E_m\left[V_1\left(\sum_{g\in s_\text{I}} \frac{\epsilon_g}{\pi_g}\right) + \sum_{g\in U_\text{I}} \frac{1}{\pi_g} V_2\left(\hat{\epsilon}_{HTg}\right)\right]$$

Using general formula for sampling variance of HT-estimator in the second term

$$\begin{aligned}\text{AMSE}(\hat{Y}_{HT}) =& V_1\left(\sum_{g\in s_{\text{I}}}\frac{\mu_g}{\pi_g}\right)+\sum_{g\in U_{\text{I}}}\frac{1}{\pi_g}V_2\left(\hat{\mu}_{HTg}\right)+\\ & E_m\left[\sum_{g,h=1}^{N_{\text{I}}}\left(\frac{\pi_{gh}}{\pi_g\pi_h}-1\right)\epsilon_g\epsilon_h+\sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\sum_{i,j=1}^{N_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)\epsilon_{gi}\epsilon_{gj}\right]\end{aligned}$$

where first two term of are constant under the model, now we solve model expectation of last term as follow

$$\begin{aligned}E_m&\left[\sum_{g=1}^{N_{\text{I}}}\left(\frac{1}{\pi_g}-1\right)\left(\sum_{i=1}^{N_g}\epsilon_{gi}^2+2\sum_{i<j=1}^{N_g}\epsilon_{gi}\epsilon_{gj}\right)+2\sum_{g<h=1}^{N_{\text{I}}}\left(\frac{\pi_{gh}}{\pi_g\pi_h}-1\right)\sum_{i=1}^{N_g}\sum_{j=1}^{N_h}\epsilon_{gi}\epsilon_{hj}\right.\\ &\left.+\sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\left\{\sum_{i=1}^{N_g}\left(\frac{1}{\pi_{i|g}}-1\right)\epsilon_{gi}^2+2\sum_{i<j=1}^{N_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)\epsilon_{gi}\epsilon_{gj}\right\}\right]\end{aligned}$$

after applying model expectation, we get

$$\begin{aligned}=&\sum_{g=1}^{N_{\text{I}}}\left(\frac{1}{\pi_g}-1\right)\left(\sum_{i=1}^{N_g}\sigma_i^2+2\sum_{i<j=1}^{N_g}\sigma_{ij}\right)+2\sum_{g<h=1}^{N_{\text{I}}}\left(\frac{\pi_{gh}}{\pi_g\pi_h}-1\right)\sum_{i=1}^{N_g}\sum_{j=1}^{N_h}\sigma_{ij}\\ &+\sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\left\{\sum_{i=1}^{N_g}\left(\frac{1}{\pi_{i|g}}-1\right)\sigma_i^2+2\sum_{i<j=1}^{N_g}\left(\frac{\pi_{ij|g}}{\pi_{i|g}\pi_{j|g}}-1\right)\sigma_{ij}\right\}\end{aligned}$$

Under two-stage element sampling by πps-SRS: $\pi_{i|g}=n_g/N_g$ and $\pi_{ij|g}=n_g(n_g-1)/N_g(N_g-1)$. Under the two-level model with constant error variance we have: $\sigma_i^2\equiv\sigma^2$, $\sigma_{ij}=\rho\sigma^2$ for $i,j\in U_g$, $\sigma_{ij}=0$ otherwise, therefore we get

$$\begin{aligned}=&\sum_{g=1}^{N_{\text{I}}}\left(\frac{1}{\pi_g}-1\right)\left(N_g\sigma^2+N_g(N_g-1)\sigma^2\rho\right)+0\\ &+\sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\left\{N_g\left(\frac{N_g}{n_g}-1\right)\sigma^2+N_g(N_g-1)\left(-\frac{N_g-n_g}{n_g(N_g-1)}\right)\sigma^2\rho\right\}\end{aligned}$$

$$=\sum_{g=1}^{N_{\text{I}}}\left(\frac{1}{\pi_g}-1\right)N_g\{1+(N_g-1)\rho\}\sigma^2+\sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\left(\frac{N_g}{n_g}-1\right)N_g(1-\rho)\sigma^2$$

The final expression for AMSE$(\hat{Y}_{HT})$ is given by

$$\begin{aligned}\text{AMSE}(\hat{Y}_{HT}) =& V_1\left(\sum_{g\in s_{\text{I}}}\frac{\mu_g}{\pi_g}\right)+\sum_{g\in U_{\text{I}}}\frac{1}{\pi_g}V_2\left(\frac{N_g}{n_g}\sum_{i\in s_g}\mu_{gi}\right)+\\ &\sum_{g=1}^{N_{\text{I}}}\left(\frac{1}{\pi_g}-1\right)N_g\{1+(N_g-1)\rho\}\sigma^2+\sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\left(\frac{N_g}{n_g}-1\right)N_g(1-\rho)\sigma^2\end{aligned} \tag{A.2}$$

Approximate AMSE of GREG-estimator under two-stage elements sampling by πps-SRS also follows from the above expression, given by

$$\text{AMSE}(\hat{Y}_{GR}) \approx \sum_{g=1}^{N_{\text{I}}}\left(\frac{1}{\pi_g}-1\right)N_g\{1+(N_g-1)\rho\}\sigma^2+\sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\left(\frac{N_g}{n_g}-1\right)N_g(1-\rho)\sigma^2 \tag{A.3}$$

ASME under two-stage cluster sampling Again using general formulation of sampling variance under two-stage cluster sampling design, we get

$$V(\hat{Y}_{HT}^{2Sc}) = V_1\left(\sum_{g\in s_{\text{I}}}\frac{Y_g}{\pi_g}\right)+\sum_{g\in U_{\text{I}}}\frac{1}{\pi_g}V_2\left(\sum_{k=1}^{n_{\text{II}g}}\frac{Y_k}{\pi_{k|g}}\right)$$

Using Eq. (A.1)

$$\begin{aligned}\text{AMSE}(\hat{Y}_{HT}^{2Sc}) =& V_1\left(\sum_{g\in s_{\text{I}}}\frac{\mu_g}{\pi_g}\right)+\sum_{g\in U_{\text{I}}}\frac{1}{\pi_g}V_2\left(\sum_{k=1}^{n_{\text{II}g}}\frac{\mu_{gk}}{\pi_{k|g}}\right)+\\ &E_m\left[\sum_{g,h=1}^{N_{\text{I}}}\left(\frac{\pi_{gh}}{\pi_g\pi_h}-1\right)\epsilon_g\epsilon_h+\sum_{g=1}^{N_{\text{I}}}\frac{1}{\pi_g}\sum_{k,l=1}^{N_{\text{II}g}}\left(\frac{\pi_{kl|g}}{\pi_{k|g}\pi_{l|g}}-1\right)\epsilon_{gk}\epsilon_{gl}\right]\end{aligned}$$

solution to the model expectation of last term is given by

$$E_m\left[\sum_{g=1}^{N_{\mathrm{I}}}\left(\frac{1}{\pi_g}-1\right)\left\{\sum_{k=1}^{N_{\mathrm{II}g}}\sum_{i=1}^{N_{gk}}\epsilon_{gki}^2+2\sum_{k=1}^{N_{\mathrm{II}g}}\sum_{i<j=1}^{N_{gk}}\epsilon_{gki}\epsilon_{gkj}+2\sum_{k<l=1}^{N_{\mathrm{II}g}}\sum_{i=1}^{N_{gk}}\sum_{j=1}^{N_{gl}}\epsilon_{gki}\epsilon_{glj}\right\}\right.$$
$$+2\sum_{g<h=1}^{N_{\mathrm{I}}}\left(\frac{\pi_{gh|g}}{\pi_g\pi_h}-1\right)\sum_{k=1}^{N_{\mathrm{II}g}}\sum_{l=1}^{N_{\mathrm{II}h}}\sum_{i=1}^{N_{gk}}\sum_{j=1}^{N_{hl}}\epsilon_{gki}\epsilon_{hlj}+\sum_{g=1}^{N_{\mathrm{I}}}\frac{1}{\pi_g}\left\{\sum_{k=1}^{N_{\mathrm{II}g}}\left(\frac{1}{\pi_{k|g}}-1\right)\sum_{i=1}^{N_{gk}}\epsilon_{gki}^2\right.$$
$$\left.\left.+2\sum_{k=1}^{N_{\mathrm{II}g}}\left(\frac{1}{\pi_{k|g}}-1\right)\sum_{i<j=1}^{N_{gk}}\epsilon_{gki}\epsilon_{gkj}+2\sum_{k<l=1}^{N_{\mathrm{II}g}}\left(\frac{\pi_{kl|g}}{\pi_{k|g}\pi_{l|g}}-1\right)\sum_{i=1}^{N_{gk}}\sum_{j=1}^{N_{gl}}\epsilon_{gki}\epsilon_{glj}\right\}\right]$$

after applying model expectation, we get

$$=\sum_{g=1}^{N_{\mathrm{I}}}\left(\frac{1}{\pi_g}-1\right)\left\{\sum_{k=1}^{N_{\mathrm{II}g}}\sum_{i=1}^{N_{gk}}\sigma_i^2+2\sum_{k=1}^{N_{\mathrm{II}g}}\sum_{i<j=1}^{N_{gk}}\sigma_{ij}+2\sum_{k<l=1}^{N_{\mathrm{II}g}}\sum_{i=1}^{N_{gk}}\sum_{j=1}^{N_{gl}}\sigma_{ij}\right\}$$
$$+2\sum_{g<h=1}^{N_{\mathrm{I}}}\left(\frac{\pi_{gh|g}}{\pi_g\pi_h}-1\right)\sum_{k=1}^{N_{\mathrm{II}g}}\sum_{l=1}^{N_{\mathrm{II}h}}\sum_{i=1}^{N_{gk}}\sum_{j=1}^{N_{hl}}\sigma_{ij}+\sum_{g=1}^{N_{\mathrm{I}}}\frac{1}{\pi_g}\left\{\sum_{k=1}^{N_{\mathrm{II}g}}\left(\frac{1}{\pi_{k|g}}-1\right)\sum_{i=1}^{N_{gk}}\sigma_i^2\right.$$
$$\left.+2\sum_{k=1}^{N_{\mathrm{II}g}}\left(\frac{1}{\pi_{k|g}}-1\right)\sum_{i<j=1}^{N_{gk}}\sigma_{ij}+2\sum_{k<l=1}^{N_{\mathrm{II}g}}\left(\frac{\pi_{kl|g}}{\pi_{k|g}\pi_{l|g}}-1\right)\sum_{i=1}^{N_{gk}}\sum_{j=1}^{N_{gl}}\sigma_{ij}\right\}$$

Under two-stage cluster sampling by πps-SRS: $\pi_{i|g}=n_{\mathrm{II}g}/N_{\mathrm{II}g}$ and $\pi_{ij|g}=n_{\mathrm{II}g}(n_{\mathrm{II}g}-1)/N_{\mathrm{II}g}(N_{\mathrm{II}g}-1)$. Under two-level with constant error variance: $\sigma_i^2=\sigma^2$, $\sigma_{ij}=\rho\sigma^2$ for $i,j\in U_g$, $\sigma_{ij}=0$ otherwise, therefore we get

$$=\sum_{g=1}^{N_{\mathrm{I}}}\left(\frac{1}{\pi_g}-1\right)\left(1+(N_g-1)\rho\right)N_g\sigma^2+\sum_{g=1}^{N_{\mathrm{I}}}\frac{1}{\pi_g}\left\{\sum_{k=1}^{N_{\mathrm{II}g}}\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}-1\right)N_{gk}\right.$$
$$\left.+\sum_{k=1}^{N_{\mathrm{II}g}}\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}-1\right)N_{gk}(N_{gk}-1)\rho+2\left(-\frac{N_{\mathrm{II}g}-n_{\mathrm{II}g}}{n_{\mathrm{II}g}(N_{\mathrm{II}g}-1)}\right)\sum_{k<l=1}^{N_{\mathrm{II}g}}N_{gk}N_{gl}\rho\right\}\sigma^2$$

Consider the second term of the above expression

$$\begin{aligned}
&\sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g}\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}-1\right)\left\{\sum_{k=1}^{N_{\mathrm{II}g}} N_{gk} + \sum_{k=1}^{N_{\mathrm{II}g}} N_{gk}(N_{gk}-1)\rho - 2\frac{1}{N_{\mathrm{II}g}-1}\sum_{k<l=1}^{N_{\mathrm{II}g}} N_{gk}N_{gl}\rho\right\}\sigma^2 \\
&= \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g}\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}-1\right)\left\{N_g + \sum_{k=1}^{N_{\mathrm{II}g}} N_{gk}^2\rho - \sum_{k=1}^{N_{\mathrm{II}g}} N_{gk}\rho - 2\frac{1}{N_{\mathrm{II}g}-1}\sum_{k<l=1}^{N_{\mathrm{II}g}} N_{gk}N_{gl}\rho\right\}\sigma^2 \\
&= \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g}\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}-1\right)\left\{N_g - N_g\rho + \frac{1}{N_{\mathrm{II}g}-1}\left((N_{\mathrm{II}g}-1)\sum_{k=1}^{N_{\mathrm{II}g}} N_{gk}^2 - 2\sum_{k<l=1}^{N_{\mathrm{II}g}} N_{gk}N_{gl}\right)\rho\right\}\sigma^2 \\
&= \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g}\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}-1\right)\left\{N_g(1-\rho) + \frac{1}{N_{\mathrm{II}g}-1}\left(N_{\mathrm{II}g}\sum_{k=1}^{N_{\mathrm{II}g}} N_{gk}^2 - \left(\sum_{k=1}^{N_{\mathrm{II}g}} N_{gk}\right)^2\right)\rho\right\}\sigma^2 \\
&= \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g}\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}-1\right)\left\{N_g(1-\rho) + \frac{1}{N_{\mathrm{II}g}-1}\left(N_{\mathrm{II}g}\sum_{k=1}^{N_{\mathrm{II}g}} N_{gk}^2 - N_{\mathrm{II}_g}^2\left(\bar{N}_{gk}\right)^2\right)\rho\right\}\sigma^2
\end{aligned}$$

$$\begin{aligned}
&= \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g}\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}-1\right)\left\{N_g(1-\rho) + \frac{N_{\mathrm{II}g}}{N_{\mathrm{II}g}-1}\left(\sum_{k=1}^{N_{\mathrm{II}g}} N_{gk}^2 - N_{\mathrm{II}_g}\left(\bar{N}_{gk}\right)^2\right)\rho\right\}\sigma^2 \\
&= \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g}\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}-1\right)\left\{N_g(1-\rho) + N_{\mathrm{II}g}S_{N_{gk}}^2\rho\right\}\sigma^2
\end{aligned}$$

where $S_{N_{gk}}^2$ is variance of second-stage cluster sizes within gth PSU. Final expression for AMSE of HT-estimator under two-stage cluster sampling by πps-SRS is given by

$$\begin{aligned}
\mathrm{AMSE}(\hat{Y}_{HT}) =& V_1\left(\sum_{g\in s_{\mathrm{I}}} \frac{\mu_g}{\pi_g}\right) + \sum_{g\in U_{\mathrm{I}}} \frac{1}{\pi_g} V_2\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}\sum_{k=1}^{n_{\mathrm{II}g}} \mu_{gk}\right) + \\
& \sum_{g=1}^{N_{\mathrm{I}}} \frac{1}{\pi_g}\left(\frac{N_{\mathrm{II}g}}{n_{\mathrm{II}g}}-1\right)\left\{N_g(1-\rho) + N_{\mathrm{II}g}S_{N_{gk}}^2\rho\right\}\sigma^2 \qquad \text{(A.4)}
\end{aligned}$$

Similarly, approximate AMSE of GREG-estimator under two-stage cluster sampling by

πps-SRS can be written

$$\begin{aligned}\text{AMSE}(\hat{Y}_{GR}) \approx & \sum_{g=1}^{N_{\text{I}}} \left(\frac{1}{\pi_g} - 1 \right) \left(1 + (N_g - 1)\rho\right) N_g \sigma^2 \\ & + \sum_{g=1}^{N_{\text{I}}} \frac{1}{\pi_g} \left(\frac{N_{\text{II}g}}{n_{\text{II}g}} - 1 \right) \left\{ N_g(1-\rho) + N_{\text{II}g} S^2_{N_{gk}} \rho \right\} \sigma^2 \end{aligned} \tag{A.5}$$

A.2 Descriptives for simulated population of households in Southampton

Table A.1: Composition of households and respective proportions in Southampton based on UK census 2011..

	Composition	Proportion
1	One person Aged 65 and over	0.1397
2	One person Aged under 65	0.1835
3	Couple without children	0.2119
4	Couple with dependent children	0.2153
5	Couple with all non-dependent children	0.0642
6	Lone parent with dependent children	0.0647
7	Lone parent with all non-dependent children	0.0317
8	Multiperson with dependent children	0.0230
9	Multiperson others	0.0660

Table A.2: Household sizes and their proportions in Southampton based on UK census 2011.

Size	1	2	3	4	5	6	7	8 or more
Proportion	0.2936	0.3578	0.1505	0.1335	0.0448	0.0144	0.0035	0.0019

Table A.3: Household tenure types and respective proportions in Southampton based on UK census 2011.

	Household tenure type	
1	Owned: Owned outright	0.3186
2	Owned: Owned with a mortgage or loan or shared ownership	0.3397
3	Rented: Social rented	0.1617
4	Rented: Private rented or living rent free	0.1801

Table A.4: Gender of household reference person (HRP) and their proportion in Southampton based on UK census 2011.

Gener of HRP	Male	Female
Proportion	0.6169	0.3831

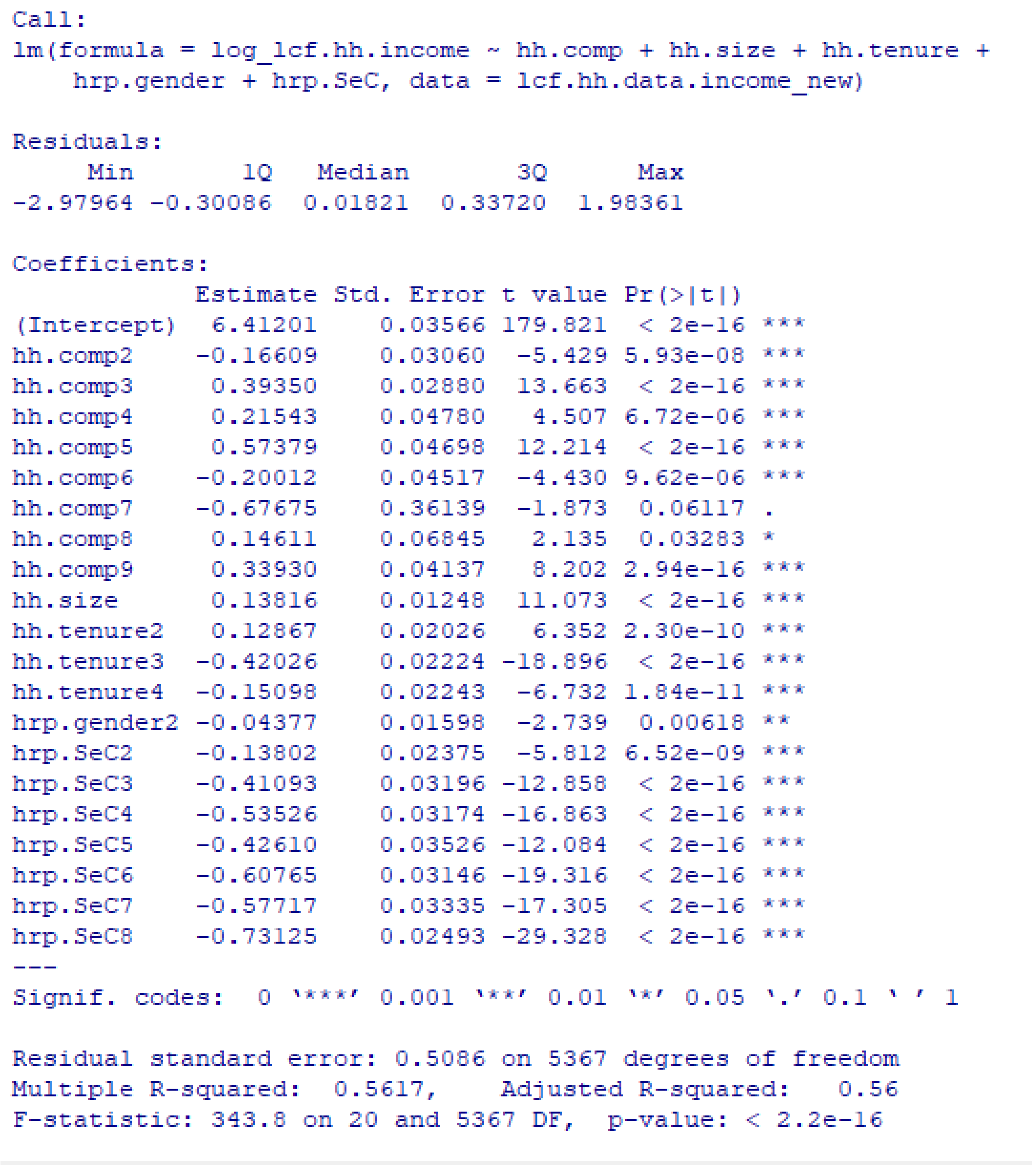
Table A.5: National Statistics Socio-economic Status (NS-SeC) of household reference person (HRP) and proportion of households for Southampton based on UK census 2011.

	National Statistics Socio-economic Status	Proportion
1	Higher managerial, administrative and professional occupations	0.1549
2	Lower managerial, administrative and professional occupations	0.2414
3	Intermediate occupations	0.1155
4	Small employers and own account workers	0.1119
5	Lower supervisory and technical occupations	0.0854
6	Semi-routine occupations	0.1246
7	Routine occupations	0.1109
8	Never worked and long-term unemployed	0.0301
9	L15 Full-time students	0.0254

A.3 Summary of fitted regression models using LCF survey data 2017-18

We fitted linear regression model for “log of household income” following Skentelbery (2010) and logistic model for “internet connection” using five covariates for both of them. The data set consists of 5407 cases. After removing cases with zero household income 5397 cases left. For the regression models we used *Bonferroni test*, see Kutner et al. (2005) based on studentized residuals for the assessment of potential outlying values. After removing the 11 outliers, models were fitted again and summary of those given bellow.

Figure A.1: Linear regression for natural log of household income based on LCF survey 2017-18 data set



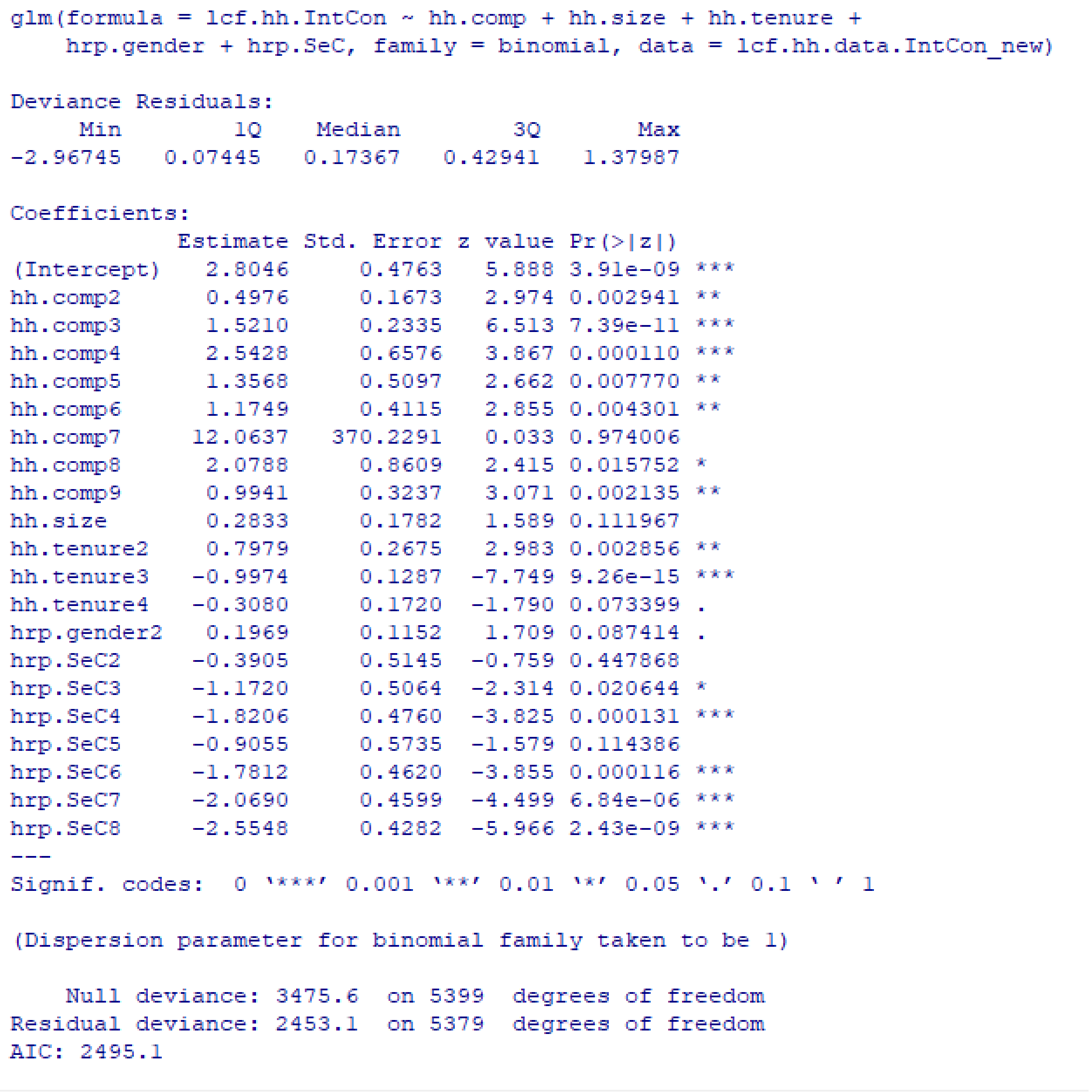
```
Call:
lm(formula = log_lcf.hh.income ~ hh.comp + hh.size + hh.tenure +
    hrp.gender + hrp.SeC, data = lcf.hh.data.income_new)

Residuals:
     Min       1Q   Median       3Q      Max
-2.97964 -0.30086  0.01821  0.33720  1.98361

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.41201    0.03566 179.821  < 2e-16 ***
hh.comp2    -0.16609    0.03060  -5.429 5.93e-08 ***
hh.comp3     0.39350    0.02880  13.663  < 2e-16 ***
hh.comp4     0.21543    0.04780   4.507 6.72e-06 ***
hh.comp5     0.57379    0.04698  12.214  < 2e-16 ***
hh.comp6    -0.20012    0.04517  -4.430 9.62e-06 ***
hh.comp7    -0.67675    0.36139  -1.873  0.06117 .
hh.comp8     0.14611    0.06845   2.135  0.03283 *
hh.comp9     0.33930    0.04137   8.202 2.94e-16 ***
hh.size      0.13816    0.01248  11.073  < 2e-16 ***
hh.tenure2   0.12867    0.02026   6.352 2.30e-10 ***
hh.tenure3  -0.42026    0.02224 -18.896  < 2e-16 ***
hh.tenure4  -0.15098    0.02243  -6.732 1.84e-11 ***
hrp.gender2 -0.04377    0.01598  -2.739  0.00618 **
hrp.SeC2    -0.13802    0.02375  -5.812 6.52e-09 ***
hrp.SeC3    -0.41093    0.03196 -12.858  < 2e-16 ***
hrp.SeC4    -0.53526    0.03174 -16.863  < 2e-16 ***
hrp.SeC5    -0.42610    0.03526 -12.084  < 2e-16 ***
hrp.SeC6    -0.60765    0.03146 -19.316  < 2e-16 ***
hrp.SeC7    -0.57717    0.03335 -17.305  < 2e-16 ***
hrp.SeC8    -0.73125    0.02493 -29.328  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5086 on 5367 degrees of freedom
Multiple R-squared:  0.5617,    Adjusted R-squared:   0.56
F-statistic: 343.8 on 20 and 5367 DF,  p-value: < 2.2e-16
```

Figure A.2: Logistic regression for binary variable “internet connection” based on LCF survey 2017-18 data set



```
glm(formula = lcf.hh.IntCon ~ hh.comp + hh.size + hh.tenure +
    hrp.gender + hrp.SeC, family = binomial, data = lcf.hh.data.IntCon_new)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.96745   0.07445   0.17367   0.42941   1.37987

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.8046     0.4763   5.888 3.91e-09 ***
hh.comp2      0.4976     0.1673   2.974 0.002941 **
hh.comp3      1.5210     0.2335   6.513 7.39e-11 ***
hh.comp4      2.5428     0.6576   3.867 0.000110 ***
hh.comp5      1.3568     0.5097   2.662 0.007770 **
hh.comp6      1.1749     0.4115   2.855 0.004301 **
hh.comp7     12.0637   370.2291   0.033 0.974006
hh.comp8      2.0788     0.8609   2.415 0.015752 *
hh.comp9      0.9941     0.3237   3.071 0.002135 **
hh.size       0.2833     0.1782   1.589 0.111967
hh.tenure2    0.7979     0.2675   2.983 0.002856 **
hh.tenure3   -0.9974     0.1287  -7.749 9.26e-15 ***
hh.tenure4   -0.3080     0.1720  -1.790 0.073399 .
hrp.gender2   0.1969     0.1152   1.709 0.087414 .
hrp.SeC2     -0.3905     0.5145  -0.759 0.447868
hrp.SeC3     -1.1720     0.5064  -2.314 0.020644 *
hrp.SeC4     -1.8206     0.4760  -3.825 0.000131 ***
hrp.SeC5     -0.9055     0.5735  -1.579 0.114386
hrp.SeC6     -1.7812     0.4620  -3.855 0.000116 ***
hrp.SeC7     -2.0690     0.4599  -4.499 6.84e-06 ***
hrp.SeC8     -2.5548     0.4282  -5.966 2.43e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3475.6  on 5399  degrees of freedom
Residual deviance: 2453.1  on 5379  degrees of freedom
AIC: 2495.1
```

A.4 Simulated Annealing

The simulated annealing (SA) is a stochastic optimization algorithm, for more details see Aarts and Korst (1988). It aims at the global optimum solution of the problem. It starts with an *initial value* of the solution from the *solution space*, and a large initial value of

Algorithm 1 Simulated annealing algorithm to minimize cost function $C(\boldsymbol{\lambda}_t)$.

```
1: Initialize:
     • Initial solution vector $\boldsymbol{\lambda}(0) = \boldsymbol{\lambda}$,
     • Initial value of temperature parameter is $temp_0$, which is obtained form Algorithm 2,
2: Generate: Generate a new solution vector $\boldsymbol{\lambda}_n \in \boldsymbol{S}$ using the random generation mechanism,
3: Cost: Calculate cost for the new solutions, $C(\boldsymbol{\lambda}_n)$,
4: Acceptance: Let $\boldsymbol{\lambda}_c$ is current solution, that is, $\boldsymbol{\lambda}(t) = \boldsymbol{\lambda}_c$. The new solution $\boldsymbol{\lambda}_n$ is selected according following criterion:
     $\boldsymbol{\lambda}(t+1) = \begin{cases} \boldsymbol{\lambda}_n & \text{if } C(\boldsymbol{\lambda}_n) \leq f(\boldsymbol{\lambda}_c), \\ \boldsymbol{\lambda}_n & \text{if } C(\boldsymbol{\lambda}_n) > f(\boldsymbol{\lambda}_c) \text{ and } U(0,1) < \exp\left(-\frac{f(\boldsymbol{\lambda}_n)-f(\boldsymbol{\lambda}_c)}{temp_r}\right), \\ \boldsymbol{\lambda}_c & \text{otherwise.} \end{cases}$
   where $U(0,1)$ is a uniform random variable,
5: Repeat: Step 2-4 are repeated $L_r$ times, where $L_r = 50$ and $r = 0, 1, 2, ...$
6: Update:
     • $temp_{r+1} = 0.95(temp_r)$,
     • Update minimum solution vector, that is, if $C(\boldsymbol{\lambda}_n) < C(\boldsymbol{\lambda}_{min})$ then $\boldsymbol{\lambda}_{min} = \boldsymbol{\lambda}_n$
7: Terminate: algorithm stops if either of the following condition reached:
     • If $t \geq t_{max} = 50000$,
     • If there is no improvement in terms minimum cost for consecutive 1000 iterations.
```

another parameter called *temperature*. A new solution is obtained by a *random generation mechanism* in the neighbourhood of the current solution. If the new solution has cost value smaller than the cost of the current solution then new solution is accepted as current solution. Otherwise new solution is accepted as current solution according to certain *probability distribution*. A fixed number of iterations are run for each of the temperature. The value of the temperature decreases to zero according to a specific *cooling schedule*. As the value of temperature decreases the probability of accepting solutions with large cost values decreases. In the early stages of the algorithm, large values of temperature allow accepting solutions with high cost which aims to avoid the local minima.

In Step 2 of the two-step cube method, SA-algorithm is used to minimise realised total imbalance for a given realised cube sample space of size K. Features of the SA-algorithm for this minimisation problem are described bellow:

- **Solution space $\boldsymbol{S}$:** It consists of all possible sampling distributions denoted by $\boldsymbol{\lambda}_t$ vector of size K such that empirical inclusion probabilities implied by the cube method $\boldsymbol{\pi}(\boldsymbol{\lambda})$ are achieved. That is,

$$\boldsymbol{S} = \{\boldsymbol{\lambda}_t | \mathbf{A}\boldsymbol{\lambda}_t = \boldsymbol{\pi}(\boldsymbol{\lambda}), \mathbf{1}_K^T\boldsymbol{\lambda}_t = 1\}$$

  where $\mathbf{A}$ is $N \times q$ matrix with elements $a_{ik} = 1$ if $i \in s_k$, otherwise $a_{ik} = 0$, $i \in U$ and $k = 1, ..., K$.

- **Random generation mechanism:** Let $\boldsymbol{\lambda}_c$ denotes the current solution. A new solution vector $\boldsymbol{\lambda}_n$ using current solution vector $\boldsymbol{\lambda}_c$ is generated as follows:
  - Generate a random vector $\boldsymbol{\delta}_c$ of K elements from uniform distribution such that $\boldsymbol{\delta}_c \sim U(-\mathbf{a}, \mathbf{a})$, where $\mathbf{a}$ is a vector of K elements defined as $a_k = min(\lambda_{ck}, 1 - \lambda_{ck}, \lambda_k)$, $k = 1, ..., K$ and $\lambda_k \equiv 1/K$.
  - Calculate the vector $\boldsymbol{\lambda}_c' = \{\lambda_{ck}\}_{k=1}^K = \begin{cases} \lambda_{ck} + \delta_{ck} & \text{if } \lambda_{ck} + \delta_{ck} > 0, \\ \lambda_{ck} & \text{otherwise,} \end{cases}$
  - Compute $\mathbf{g}$ weights based on calibration equations $\mathbf{A}\boldsymbol{\lambda}_c' = \boldsymbol{\pi}(\boldsymbol{\lambda})$. To compute the calibration weights, `R`-function `gencalib(method="logit")` is used from R-package `sampling` (Deville, 2000; Estevao and Särndal, 2000; Tillé and Matei, 2021).
  - Compute the new solution vector $\boldsymbol{\lambda}_n = \mathbf{g} \times \boldsymbol{\lambda}_c'$

- **Cost function:** Cost value for the current solution vector $\boldsymbol{\lambda}_c$ is calculated as

$$C(\boldsymbol{\lambda}_c) = tr(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\lambda}_c})$$

- **Acceptance criteria:** Let $\boldsymbol{\lambda}_c$ is current solution vector at stage t, that is $\boldsymbol{\lambda}(t) = \boldsymbol{\lambda}_c$. The new vector $\boldsymbol{\lambda}_n$ at stage $t+1$ is accepted according to following probability criteria:

$$P[\boldsymbol{\lambda}(t+1) = \boldsymbol{\lambda}_n] = \begin{cases} 1 & \text{if} \quad C(\boldsymbol{\lambda}_n) \leq C(\boldsymbol{\lambda}_c) \\ \exp\left(\frac{C(\boldsymbol{\lambda}_n) - C(\boldsymbol{\lambda}_c)}{temp}\right) & \text{if} \quad C(\boldsymbol{\lambda}_n) > C(\boldsymbol{\lambda}_c) \end{cases} \tag{A.6}$$

  where *temp* is temperature parameter defined in cooling schedule bellow.

- **Cooling schedule:** It defines the initial value, decrement function of temperature parameter and number of iterations for each value of temperature. It also defines the termination criteria of the algorithm.

i. To find the *initial value* of temperature parameter, the SA-algorithm is run with increasing temperature parameter until the all proposed solutions are accepted. The resulting value is considered as initial value of temperature parameter.

ii. There are many *temperature decreasing functions* range from simple to complex forms, we shall consider as simple geometric function $temp_{r+1} = d \times temp_r$, where d is a positive value which is smaller than 1.

iii. *Number of iterations* $L_r = L_0$ for the given temperature value $temp_r$ is fixed.

iv. There are many criteria for termination of an algorithm depending time, number of iterations and quality of solutions. We fix maximum number of iterations for this algorithm, denoted by t_{max}. Optimum solution is recorded at each iteration and the algorithm stops when there is no improvement in terms of optimum solution for a fixed number of iterations.

Algorithm 2 computes the initial value of temperature parameter for Algorithm 1.

```
1: **Initialize:**
     • Initial solution vector $\boldsymbol{\lambda}(0) = \boldsymbol{\lambda}$,
     • Initial value of temperature parameter is $temp_0 = 1$,
2: **Generate:** Generate a new solution vector $\boldsymbol{\lambda}_n \in \boldsymbol{S}$ using the random generation mechanism,
3: **Cost:** Calculate cost for the new solutions, $C(\boldsymbol{\lambda}_n)$,
4: **Acceptance:** Let $\boldsymbol{\lambda}_c$ is current solution, then the new solution $\boldsymbol{\lambda}_n$ is selected according to following criterion:
     $$\boldsymbol{\lambda}(t+1) = \begin{cases} \boldsymbol{\lambda}_n & \text{if } C(\boldsymbol{\lambda}_n) \leq f(\boldsymbol{\lambda}_c), \\ \boldsymbol{\lambda}_n & \text{if } C(\boldsymbol{\lambda}_n) > f(\boldsymbol{\lambda}_c) \text{ and } U(0,1) < \exp\left(-\frac{f(\boldsymbol{\lambda}_n) - f(\boldsymbol{\lambda}_c)}{temp_r}\right), \\ \boldsymbol{\lambda}_c & \text{otherwise.} \end{cases}$$
     where $U(0,1)$ is a uniform random variable,
5: **Repeat:** Step 2-4 are repeated $L_r$ times, where $L_r = 50$ and $r = 0, 1, 2, ...$,
6: **Update:**
     • $temp_{r+1} = 1.5(temp_r)$,
     • Calculate acceptance ratio $AR = \frac{\text{No. of accepted solutions}}{L_r}$,
7: **Terminate:** Steps 2-6 are repeated until $AR = 1$ (rounded up to three decimal points).
```

Algorithm 1 describes process of minimisation using simulation annealing. Algorithm 2 describes process of computing initial value for temperature parameter in the simulated annealing algorithm.

A.5 MU284 data description

Table A.6: Description of variables in the Swedish municipalities data MU284.

2	P85	1985 population (in thousands).
3	P75	1975 population (in thousands).
4	RMT85	revenues from 1985 municipal taxation (in millions of kronor).
5	CS82	number of Conservative seats in municipal council.
6	SS82	number of Social-Democratic seats in municipal council.
7	S82	total number of seats in municipal council.
8	ME84	number of municipal employees in 1984.
9	REV84	real estate values according to 1984 assessment (in millions of kronor).
10	REG	geographic region indicator.
11	CL	cluster indicator (a cluster consists of a set of neighboring).

Table A.7: Correlation matrix for the modified Clustered MU284 data set used in the simulation study for two-step cube method.

	P75	RMT85	ME84	P85	CS82	S82	REV84	SIZE	S82-CS82-SS82	CS82-SS82
P75	1.00	0.98	0.98	1.00	0.68	0.82	0.93	0.63	0.57	-0.46
RMT85	0.98	1.00	1.00	0.99	0.68	0.76	0.93	0.56	0.47	-0.41
ME84	0.98	1.00	1.00	0.99	0.68	0.76	0.93	0.56	0.47	-0.41
P85	1.00	0.99	0.99	1.00	0.72	0.80	0.95	0.62	0.55	-0.40
CS82	0.68	0.68	0.68	0.72	1.00	0.64	0.65	0.51	0.44	0.14
S82	0.82	0.76	0.76	0.80	0.64	1.00	0.79	0.93	0.88	-0.64
REV84	0.93	0.93	0.93	0.95	0.65	0.79	1.00	0.61	0.56	-0.45
SIZE	0.63	0.56	0.56	0.62	0.51	0.93	0.61	1.00	0.88	-0.65
S82-CS82-SS82	0.57	0.47	0.47	0.55	0.44	0.88	0.56	0.88	1.00	-0.53
CS82-SS82	-0.46	-0.41	-0.41	-0.40	0.14	-0.64	-0.45	-0.65	-0.53	1.00

Bibliography

Aarts, E. and Korst, J. (1988). *Simulated annealing and Boltzmann machines.* New York, John Wiley and Sons Inc.

Abi, N. (2019). *Spatially balanced sampling methods in household surveys.* PhD thesis, University of Canterbury, School of Mathematics and Statistics.

Altieri, L. and Cocchi, D. (2021). Spatial sampling for non-compact patterns. *International Statistical Review*, 89(3):532–549. https://doi.org/10.1111/insr.12445.

Ardilly, P. (1991). Échantillonnage représentatif optimum à probabilités inégales. *Annales d'Économie et de Statistique*, pages 91–113.

Arnab, R. (2017). Chapter 3 - simple random sampling. In Arnab, R., editor, *Survey Sampling Theory and Applications*, pages 51–88. Academic Press.

Bellhouse, D. R. (1977). Some optimal designs for sampling in two dimensions. *Biometrika*, 64(3):605–611.

Benedetti, R., Dickson, M. M., Espa, G., Pantalone, F., and Piersimoni, F. (2022). A simulated annealing-based algorithm for selecting balanced samples. *Computational Statistics*, 37(1):491–505.

Benedetti, R., Espa, G., and Taufer, E. (2017a). Model-based variance estimation in non-measurable spatial designs. *Journal of Statistical Planning and Inference*, 181:52–61.

Benedetti, R. and Piersimoni, F. (2017). A spatially balanced design with probability function proportional to the within sample distance. *Biometrical Journal*, 59(5):1067–1084.

Benedetti, R., Piersimoni, F., and Postiglione, P. (2017b). Alternative and complementary approaches to spatially balanced samples. *Metron*, 75(3):249–264.

Benedetti, R., Piersimoni, F., and Postiglione, P. (2017c). Spatially balanced sampling: a review and a reappraisal. *International Statistical Review*, 85(3):439–454.

Benedetti, R., Piersimoni, F., Postiglione, P., et al. (2015). *Sampling spatial units for agricultural surveys*. Springer.

Berger, Y. G. (1998a). Rate of convergence for asymptotic variance of the horvitz–thompson estimator. *Journal of Statistical Planning and Inference*, 74(1):149–168.

Berger, Y. G. (1998b). Variance estimation using list sequential scheme for unequal probability sampling. *Journal of Official Statistics*, 14(3):315.

Berger, Y. G. (2005). A variance estimator for systematic sampling from a deliberately ordered population. *Communications in Statistics-Theory and Methods*, 34(7):1533–1541.

Berger, Y. G., Muñoz, J. F., and Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totalsan application to the extended regression estimator and the regression composite estimator. *Computational statistics & data analysis*, 53(7):2596–2604.

Bivand, R., Nowosad, J., and Lovelace, R. (2021). *spData: Datasets for Spatial Analysis*. R package version 2.0.1.

Bondesson, L. and Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics*, 35(3):466–483.

Bowley, A. L. (1926). Measurement of the precision attained in sampling. *Bulletin de L'Institute International de Statistique*, XXII.

Breidt, F. J. and Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141(1):479–487.

Breidt, F. J. and Chauvet, G. (2012). Penalized balanced sampling. *Biometrika*, 99(4):945–958.

Brewer, K. R. W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5(1):5–13.

Brewer, K. R. W. and Gregoire, T. G. (2009). Introduction to survey sampling. In *Handbook of Statistics*, volume 29, pages 9–37. Elsevier.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with unequal probabilities*, volume 15. Springer Science & Business Media.

Cassel, C. M., Särndal, C.-E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

Chao, M. T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69(3):653–656.

Chauvet, G., Haziza, D., and Lesage, É. (2017). Examining some aspects of balanced sampling in surveys. *Statistica Sinica*, pages 313–334.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm for balanced sampling. *Computational Statistics*, 21(1):53–62.

Chen, B., Wang, J., Zhao, H., and Principe, J. (2016). Insights into entropy as a measure of multivariate variability. *Entropy*, 18(5):196.

Chun, Y. and Griffith, D. A. (2018). Impacts of negative spatial autocorrelation on frequency distributions. *Chilean Journal of Statistics*, 9(1):3–17.

Cliff, A. D. and Ord, J. K. (1973). *Spatial autocorrelation.* London : Pion. Includes index.

Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models & Applications.* Pion.

Cochran, W. G. (1939). The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34(207):492–510.

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics*, 17(2):164–177.

Cochran, W. G. (1977). *Sampling techniques.* John Wiley & Sons, 3rd edition.

Das, A. C. (1950). Two dimensional systematic sampling and the associated stratified and random sampling. *Sankhyā: The Indian Journal of Statistics*, pages 95–108.

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. In *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, pages 5–7.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey methodology*, 25(2):193–204.

Deville, J.-C. (2000). Generalized calibration and application to weighting for nonresponse. In *COMPSTAT*, pages 65–76. Springer.

Deville, J.-C., Grosbras, J., and Roth, N. (1988). Efficient sampling algorithms and balanced samples. In *Compstat*, pages 255–266. Springer.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85(1):89–101.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4):893–912.

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128(2):569–591.

Dickson, M. M., Benedetti, R., Giuliani, D., and Espa, G. (2014). The use of spatial sampling designs in business surveys. *Open Journal of Statistics*, 2014.

Dickson, M. M., Grafström, A., Giuliani, D., and Espa, G. (2019). Efficiency and feasibility of sampling schemes in establishment surveys. *Mathematical Population Studies*, 26(2):114–122.

Dumelle, M., Kincaid, T., and Olsen, T. (2021). *spsurvey: Spatial Sampling Design and Analysi.* R package version 5.1.0.

Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *Journal of the Royal Statistical Society: Series B (Methodological)*, 15(2):262–269.

Estevao, V. M. and Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16(4):379.

Filipponi, D., Piersimoni, F., Benedetti, R., Dickson, M. M., Espa, G., and Giuliani, D. (2019). Sampling design and analysis using geo-referenced data. In Zhang, L.-C. and Chambers, R. L., editors, *Analysis of Integrated Data*, chapter 10, pages 219–245. CRC Press, Boca Raton.

Fuller, W. A. (2009a). *Sampling Statistics.* Wiley Series in Survey Methodology. Wiley.

Fuller, W. A. (2009b). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4):933–944.

Gabler, S. and Schweigkoffer, R. (1990). The existence of sampling designs with preassigned inclusion probabilities. *Metrika*, 37(1):87–96.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):269–278.

Godambe, V. P. and Joshi, V. M. (1965). Admissibility and bayes estimation in sampling finite populations. I. *The Annals of Mathematical Statistics*, 36(6):1707–1722.

Goldstein, H. (1991). Multilevel modelling of survey data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 40(2):235–244.

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

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

Grafström, A. and Lundström, N. L. (2013). Why well spread probability samples are balanced. *Open Journal of Statistics*, 3(1):36–41.

Grafström, A., Lundström, N. L., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–520.

Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24(2):120–131.

Gräler, B., Pebesma, E., and Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *The R Journal*, 8:204–218.

Griffith, D. A. (2003). *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization*. Advances in Spatial Science. Springer.

Griffith, D. A. (2011). Positive spatial autocorrelation impacts on attribute variable frequency distribution. *Chilean Journal of Statistics*, 2(2):3–28.

Griffith, D. A. (2019). Negative spatial autocorrelation: One of the most neglected concepts in spatial statistics. *Stats*, 2(3):388–415.

Griffith, D. A. and Arbia, G. (2010). Detecting negative spatial autocorrelation in georeferenced random variables. *International Journal of Geographical Information Science*, 24(3):417–437.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, pages 1491–1523.

Hájek, J. (1981). *Sampling from a finite population*. M. Dekker.

Hanif, M. and Brewer, K. R. W. (1980). Sampling with unequal probabilities without replacement: a review. *International Statistical Review/Revue Internationale de Statistique*, pages 317–335.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.*, 14(4):333–362.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953a). *Sample survey methods and theory: Volume 1*. Wiley publications in statistics. Wiley.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953b). *Sample survey methods and theory: Volume 2*. Wiley publications in statistics. Wiley.

Hartley, H. O. and Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *The Annals of Mathematical Statistics*, pages 350–374.

Haziza, D., Mecatti, F., and Rao, J. N. K. (2004). Comparison of variance estimators under rao-sampford method: a simulation study. In *Joint Statistical Meeting*, pages 3638–3643. US.

Haziza, D., Mecatti, F., and Rao, J. N. K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron*, 66(1):91–108.

Hedayat, A. S. and Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference*, 44(2):237–247.

Hedayat, A. S., Rao, C. R., and Stufken, J. (1988). Sampling plans excluding contiguous units. *Journal of Statistical Planning and Inference*, 19(2):159–170.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Jauslin, R. and Tillé, Y. (2020). Spatial spread sampling using weakly associated vectors. *Journal of Agricultural, Biological and Environmental Statistics*, 25(3):431–451.

Jones, K. (1993). Using multilevel models for survey analysis. *JOURNAL-MARKET RESEARCH SOCIETY*, 35:249–249.

Kiær, A. N. (1896). Observations and experiments on representative enumeration. *Bulletin de L'Institute International de Statistique*, 9(livre 2):176–83.

Kish, L. (1965). *Survey Sampling*. A Wiley Interscience Publication. Wiley.

Kosiorowski, D. and Zawadzki, Z. (2022). *DepthProc An R Package for Robust Exploration of Multidimensional Economic Phenomena*.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., et al. (2005). *Applied linear statistical models*, volume 5. McGraw-Hill Irwin New York.

Lahiri, D. B. (1951). A method for sample selection providing unbiased ratio estimates. *Bull. Int. Stat. Inst.*, 33(2):133–140.

Leuenberger, M., Eustache, E., Jauslin, R., and Tillé, Y. (2022). Balancing a sample almost perfectly. *Statistics & Probability Letters*, 180:109229.

Lister, A. J. and Scott, C. T. (2009). Use of space-filling curves to select sample locations in natural resource monitoring studies. *Environmental Monitoring and Assessment*, 149(1):71–80.

Madow, W. G. (1949). On the theory of systematic sampling, ii. *The Annals of Mathematical Statistics*, 20(3):333–354.

Mahalanobis, P. C. and Fisher, R. A. (1944). On large-scale sample surveys. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 231(584):329–451.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Mullen, K. M. (2014). Continuous global optimization in r. *Journal of Statistical Software*, 60(i06).

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Office for National Statistics (2011). UK census 2011: data by postcode sectors for Southampton.

Office for National Statistics (2019). Living costs and food survey, 2017-2018.

Pantalone, F., Benedetti, R., and Federica, P. (2019). *Spbsampling: spatially balanced sampling*. R package version 1.3.4.

Pfeffermann, D., Da Silva Moura, F. A., and Do Nascimento Silva, P. L. (2006). Multilevel modelling under informative sampling. *Biometrika*, 93(4):943–959.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 60(1):23–40.

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, pages 355–375.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):805–827.

Rao, J. N. K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3(4):173–80.

Rao, J. N. K., Hartley, H. O., and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2):482–491.

Rao, J. N. K., Verret, F., and Hidiroglou, M. A. (2013). A weighted composite likelihood approach to inference for two-level models from survey data. *Survey Methodology*, 39(2):263–282.

Ribeiro Jr, P. J., Diggle, P. J., Christensen, O., Schlather, M., Bivand, R., and Ripley, B. (2020). *geoR: Analysis of Geostatistical Data*. R package version 1.8-1.

Robertson, B., Brown, J., McDonald, T., and Jaksons, P. (2013). Bas: Balanced acceptance sampling of natural resources. *Biometrics*, 69(3):776–784.

Robertson, B., McDonald, T., Price, C., and Brown, J. (2018). Halton iterative partitioning: spatially balanced sampling via partitioning. *Environmental and Ecological Statistics*, 25(3):305–323.

Rosén, B. (1991). Variance estimation for systematic pps-sampling. *Report: In Statistics Sweden.*

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54(3/4):499–513.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling.* Springer-Verlag New York, 1st edition.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5(1194):127.

Skentelbery, R. (2010). Regression analysis of household expenditure and income. In *Family Spending*, pages 71–78. Springer.

Skinner, C. J. (1989). Domain means, regression and multi-variate analysis. In Skinner, C. J., Holt, D., and Smith, T. M. F., editors, *Analysis of Complex Surveys*, pages 59–88. Wiley.

Skinner, C. J. and de Toledo Vieira, M. (2007). Variance estimation in the analysis of clustered longitudinal survey data. *Survey Methodology*, 33(1):3–12.

Stevens Jr, D. L. (1997). Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics: The official journal of the International Environmetrics Society*, 8(3):167–195.

Stevens Jr, D. L. and Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14(6):593–610.

Stevens Jr, D. L. and Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of The American Statistical Association*, 99(465):262–278.

Thionet, P. (1953). *La Theórie des Sondages.* Paris: INSEE. Imprimerie Nationale.

Tillé, Y. (2006). *Sampling algorithms.* Springer-Verlag New York.

Tillé, Y. (2011). Ten years of balanced sampling with the cube method: an appraisal. *Survey methodology*, 37(2):215–226.

Tillé, Y., Dickson, M. M., Espa, G., and Giuliani, D. (2018). Measuring the spatial balance of a sample: A new measure based on morans i index. *Spatial Statistics*, 23:182–192.

Tillé, Y. and Matei, A. (2021). *sampling: Survey Sampling.* R package version 2.9.

Valliant, R., Dever, J. A., and Kreuter, F. (2015). Effects of Cluster Sizes on Variance Components in Two-Stage Sampling. *Journal of Official Statistics*, 31(4):763–782.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley Series in Probability and Statistics. Wiley.

Wolter, K. (2007). *Introduction to variance estimation*. Springer Science & Business Media, 2nd edition.

Wright, K. (2021). *agridat: Agricultural Datasets*. R package version 1.18.

Xiang, Y., Gubian, S., Suomela, B., and Hoeng, J. (2013). Generalized simulated annealing for global optimization: the GenSA package. *The R Journal*, 5(1):13.

Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, 109(1):12–43.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society: Series B (Methodological)*, 15(2):253–261.

Zhang, L.-C. (2008). On some common practices of systematic sampling. *Journal of Official Statistics*, 24(4):557–569.