https://doi.org/10.35662/unine-thesis-3198

Université de Neuchâtel

Faculty of Science
Institute of Statistics

New Approaches in Survey Statistics: Balanced Sampling, Calibration and Imputation

Thesis submitted in fulfillment of the requirements
for the degree of
Doctorat ès Sciences

by

Arnaud Tripet

Dissertation Committee

Prof.	Yves Tillé	University of Neuchâtel	Thesis Director
Prof.	Alina Matei	University of Neuchâtel	Jury President
Prof.	David Haziza	University of Ottawa	Rapporteur
Prof.	Camelia Goga	Marie et Louis Pasteur University	Rapporteur
Dr.	Caren Hasler	University of Neuchâtel & University of Zürich	Rapporteur

Thesis Defended on *June, 24th 2025*

unine
Université de Neuchâtel
Faculté des sciences

Rue Emile-Argand 11
CH-2000 Neuchâtel
doctorat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Monsieur Arnaud TRIPET

Titre :

"New Approaches in Survey Statistics: Balanced Sampling, Calibration and Imputation"

sur le rapport des membres du jury composé comme suit :

- **Prof. Yves Tillé,** directeur de thèse, Université de Neuchâtel, Suisse
- **Prof. tit. Alina Matei,** Université de Neuchâtel, Suisse
- **Prof. David Haziza,** University of Ottawa, Canada
- **Prof. Camélia Goga,** Université Bourgogne-Franche Comté, France
- **Dr Caren Hasler,** Université de Neuchâtel, Suisse

Neuchâtel, le 27 août 2025

Le Doyen, Prof. P. Brunner

Remerciements

Tout d'abord, je tiens à exprimer ma profonde gratitude à mon directeur et superviseur de thèse, Yves Tillé, qui m'a fait confiance dès le départ. Il a su m'éclairer, m'apprendre et me soutenir dans ma thèse mais aussi au-delà de mes recherches.

Je tiens également à remercier sincèrement les membres de mon jury de thèse, Prof. Alina Matei, Prof. Camélia Goga, Dr. Caren Hasler et Prof. David Haziza, pour le temps et les efforts qu'ils ont consacrés à l'évaluation de mon travail.

En outre, je suis reconnaissant envers Dr. Caren Hasler et Dr. Esther Eustache, avec qui j'ai eu l'occasion de collaborer sur plusieurs projets qui sont devenus des parties intégrantes de cette thèse.

Enfin, je remercie mes collègues Alina, Andrea, Caren, Corine, Ejub, Esther, Lionel, Michael, Pierre-Yves, Raphaël, Ziqing, ma famille et mes amis qui m'ont écouté et supporté pendant ces presque 4 années de doctorat.

Abstract

This thesis explores new approaches and methods in survey statistics. It focuses on balanced sampling, calibration, and imputation. It begins with an introduction to the key concepts of survey methodology, and more specifically, to the three main themes of the thesis. The first part focuses on extending the cube method by incorporating inequality constraints. This allows the selection of samples that are not only balanced on auxiliary variables, but also meet minimum size requirements in specific groups. Applications include, for example, category bounding, controlled matrix rounding or spatial sampling. The second part focuses on calibration and is divided into two articles. The first article addresses the harmonization of survey weights when different variables each have their own weighting system. Two strategies are compared: one based on calibration and the other using optimal transport. The second article uses bagging and principal component decomposition to solve issues arising from high-dimensional calibration. The third part focuses on imputation and introduces a hybrid method that combines SwissCheese, a donor-based balanced hot-deck approach, with missForest, a predictor based on random forests.

Keywords : balanced sampling, calibration, high-dimension, imputation, inequality constraints, nonresponse.

Résumé

Cette thèse explore de nouvelles approches et méthodes en statistique d'enquête. Elle se concentre sur l'échantillonnage équilibré, le calage et l'imputation. Elle débute par une introduction aux concepts clefs de la méthodologie d'enquête et plus particulièrement aux trois grands axes de la thèse. La première partie porte sur l'extension de la méthode du cube en y intégrant des contraintes d'inégalités. Cela permet de sélectionner des échantillons non seulement équilibrés sur les variables auxiliaires, mais aussi respectant des tailles minimales dans certains groupes. Les applications incluent, par exemple, la délimitation de catégories, l'arrondissement contrôlé de matrices ou encore l'échantillonnage spatial. La deuxième partie traite du calage et se divise en deux articles. Le premier aborde l'harmonisation des poids d'enquête lorsque différentes variables possèdent chacune leur propre système de pondération. Deux stratégies sont comparées : l'une est basée sur le calage et l'autre sur le transport optimal. Le second article utilise le Bagging et la décomposition en composantes principales pour résoudre les problèmes liés au calage en grande dimension. La troisième partie est consacrée à l'imputation et introduit une méthode hybride combinant SwissCheese, une méthode de hot-deck équilibrée basée sur des donneurs, et missForest, un prédicteur basé sur les forêts aléatoires.

Mots-clefs : calage, contraintes d'inégalités, échantillonnage équilibré, grande dimension, imputation, non-réponse.

Contents

Chapter 1

Introduction to Survey Statistics

Survey statistics are an essential element in making analyses and drawing conclusions based on finite populations from samples. This discipline covers a number of topics, including sampling designs, estimation methods, calibration, variance estimation and the treatment of nonresponse. The major steps in the development of survey statistics began with the introduction of the concept of sampling at the end of the 19th century by Kiær (1896), then formalized by Bowley (1926) and later by Neyman (1934). Hansen and Hurwitz (1943), Horvitz and Thompson (1952), and Godambe (1955) developed and theorized key estimators and established the foundations of design-based inference, while Brewer (1963) and Royall (1970) introduced model-based approaches. Basu (1971) later questioned the sufficiency of randomization, and Kish (1965) and Cochran (1953) helped consolidate the theory into practice. Särndal et al. (1992) introduced the model-assisted approach, a hybrid between the model-based and design-based perspectives. In 1992, Deville and Särndal developed calibration methods. Deville and Tillé (2004) proposed the cube method for balanced sampling and Haziza contributed significantly to the treatment of nonresponse (see among others Haziza and Rao, 2003, 2006; Haziza, 2009).

1.1 Finite Population and Sampling

In survey sampling, we consider a finite population $U = \{1, 2, ..., N\}$, where N is the total number of units. A sample $s \subset U$ of size n is drawn using a probabilistic sampling mechanism. Each unit $k \in U$ has an inclusion probability $\pi_k = P(k \in s)$. These probabilities ensure that each unit has a known chance of being selected, which helps to ensure unbiased estimators.

Inclusion probabilities are fundamental to the theory of finite population sampling. The first-order inclusion probability π_k refers to the probability that unit $k \in U$ is selected in the sample s, that is,

$$\pi_k = \mathbb{P}(k \in s).$$

The second-order inclusion probability π_{kl} corresponds to the joint probability that both units k and l are simultaneously selected in a sample:

$$\pi_{kl} = \mathbb{P}(k \in s, l \in s).$$

These two types of probabilities are essential for defining unbiased estimators and computing their variances. One of the main objective in survey sampling is to estimate the population total

$$Y = \sum_{k \in U} y_k,$$

where y_k denotes the value of a study variable for unit k. Under any sampling design with known and strictly positive inclusion probabilities, the Horvitz-Thompson estimator (Horvitz

and Thompson, 1952) provides an unbiased estimator of the total:

$$\hat{Y}_{\mathrm{HT}} = \sum_{k \in s} \frac{y_k}{\pi_k}, \text{ if } \pi_k > 0, k \in U.$$

This estimator accounts for the unequal probabilities of selection that may arise from complex sampling designs.

To improve the efficiency and precision of estimators linked to the variable of interest, or to reduce their variance, it is common practice to incorporate auxiliary information. Let $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})$ be the vector of p auxiliary variables for unit k. These auxiliary variables are known for a large part or all of the population, and are generally correlated or related to the study variable y_k. They can be continuous or categorical and are used to improve estimation through various methods such as stratification, calibration, regression estimation or balanced sampling.

In survey statistics, there are two dominant paradigms : the design-based approach and the model-based approach. In the first one, the randomness comes from the sampling design and every inference is made subject to the finite population. This idea was formalized by Neyman (1934). In contrast, the model-based approach treats population values as the results of a stochastic model with inference based on the assumed distribution. This approach is more efficient, but requires the model to be correctly specified. The distinction between the two lies in the source of the randomness: on the sampling design or on the design and assumed superpopulation model. Särndal et al. (1992) introduce a hybrid approach, called model-assisted, which uses a working model to improve efficiency while preserving design-based inference.

1.2 Balanced Sampling and Cube Method

Sampling designs define how units are selected from the population. The choice of sampling design can influence the accuracy of estimators. Some of the most common designs include the simple random sampling (SRS) where each sample of the same size have the same probability of being selected. The systematic sampling where units are ordered and every k-th unit is selected. The unequal probability sampling where units are selected with different inclusion probabilities often proportional to an auxiliary variable. The stratified sampling where the population is divided into homogeneous subgroups called strata, and samples are drawn independently within each stratum. Or also the cluster sampling where entire clusters are randomly selected instead of individual units. However the sampling design we are interested in here is balanced sampling.

Balanced sampling is a specific sampling design that ensures the Horvitz-Thompson estimators of the total of auxiliary variables of the selected sample is equal or almost equal to their known population totals Royall (1976); Royall and Pfeffermann (1982). This results in a lower variance in the total without requiring larger sample sizes.

To summarize, a sample is said to be balanced if

$$\sum_{k \in s} \frac{\mathbf{x}_k}{\pi_k} \approx \sum_{k \in U} \mathbf{x}_k,$$

with $\mathbf{x}_k$ known for the entire population. The cube method, introduced by Deville and Tillé (2004, 2005), is a widely used technique for drawing a balanced sample. This method selects balanced samples across multiple auxiliary variables. The cube method is flexible because it adapts to different sampling constraints, such as equal or unequal inclusion probabilities and also stratified designs. The method can be divided into two distinct phases: the flight phase and the landing phase. The flight phase is a random walk in which the inclusion probabilities are iteratively adjusted while maintaining the balance equations. During this phase, the algorithm progressively modifies the inclusion probabilities of the units, pulling them towards 1 or 0, to

get closer to a sample that satisfies the balance constraints. Then, the landing phase finalizes the selection by rounding the inclusion probabilities to 0 or 1, guaranteeing that a sample is obtained. The final sample remains as close as possible to the balanced state obtained during the flight phase.

The cube method has also inspired extensions to many areas (Tillé, 2011), for example in environmental and geographical applications through spatially balanced sampling, where the goal is to ensure that selected units are well spread over space Grafström et al. (2012); Robertson et al. (2013).

1.3 Calibration

Another way of satisfying constraints on the known population is through calibration. Introduced by Deville and Särndal (1992), calibration is a technique in survey statistics that adjusts survey weights $d_k = 1/w_k$ so that they better reflect known population totals. Let $\mathbf{x}_k$ be a vector of auxiliary variables associated with unit k, and let X_j denote their known population totals. Calibration adjusts the weights w_k by minimizing a pseudo-distance function:

$$\sum_{k \in s} G(w_k, d_k),$$

such that

$$G(w_k, d_k) \geq 0,$$

$$G(d_k, d_k) = 0$$

and $G(., d_k)$ is strictly convex, and subject to the constraint:

$$\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{X}.$$

The function $G(w_k, d_k)$ defines the pseudo-distance between the calibrated weights w_k and the original design weights d_k. By carefully choosing this function, different calibration methods can be implemented, allowing one to maintain statistical efficiency while preventing extreme weight adjustments (Deville and Särndal, 1992; Särndal et al., 1992).

A fundamental application of calibration is the estimation of a population total. Given a survey variable y_k, the Horvitz-Thompson estimator of the total is:

$$\hat{Y}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}.$$

Using calibrated weights w_k, the estimator becomes:

$$\hat{Y}_{cal} = \sum_{k \in s} w_k y_k,$$

thus incorporating auxiliary information into the estimation process, leading to reduced variance and improved robustness (Deville and Särndal, 1992).

Several calibration techniques have been developed, differing in how weights are adjusted to achieve balance.

One of the simplest is linear calibration, where the adjusted weights are obtained through a linear transformation of the initial weights. This method ensures that weights remain close to the original sampling weights while satisfying the calibration constraints. The adjusted weights take the form

$$w_k = d_k \left(1 + \boldsymbol{\lambda}^T \mathbf{x}_k\right),$$

where $\boldsymbol{\lambda}$ is determined to satisfy the calibration equations (see Deville and Särndal, 1992). Linear calibration minimizes a quadratic distance between original and adjusted weights and is particularly suitable when auxiliary variables are continuous.

A model-assisted extension is the generalized regression estimator (GREG), which incorporates a linear regression model relating the study variable to auxiliary information. Introduced by Cassel et al. (1976) and further developed by Särndal et al. (1992), it is expressed as:

$$\hat{Y}_{\text{GREG}} = \hat{Y}_{\text{HT}} + \sum_j \hat{B}_j (X_j - \hat{X}_j),$$

where $\hat{B}_j$ are estimated regression coefficients. The GREG estimator reduces bias and variance, especially when the auxiliary variables strongly predict the outcome (Kalton and Flores-Cervantes, 2003).

Another widely used technique is exponential calibration, particularly in the form of raking ratio estimation or iterative proportional fitting. Initially proposed by Deming and Stephan (1940) and later formalized within the calibration framework by Deville et al. (1993), raking adjusts weights multiplicatively to match known marginal totals of categorical auxiliary variables. The new weights satisfy:

$$\sum_{k \in s} w_k x_{kj} = X_j, \quad \forall j,$$

with X_j denoting the known total for category j. Exponential calibration minimizes a Kullback-Leibler divergence and is particularly efficient in post-stratification contexts involving categorical data.

Other advanced calibration techniques include: Logistic calibration, which modifies weights using a logistic-type function to avoid extreme values and has been discussed in detail by Deville and Särndal (1992). Ridge calibration, which adds a regularization term to stabilize weight adjustments and is particularly effective under multicollinearity or weak auxiliary information (Chambers, 1996; Beaumont and Bocci, 2008). Empirical likelihood calibration, which applies a nonparametric likelihood framework to optimally constrain the weights while respecting calibration conditions (Chen and Sitter, 1999; Kim, 2009). Calibration is a key tool in survey sampling, offering flexibility in adjusting weights while ensuring consistency with known information (Haziza and Beaumont, 2017).

1.4 Nonresponse and Imputation

Nonresponse occurs when some units in the sample fail to provide data. There are many reasons for this: refusal or inability to contact a selected respondent, data collection problems, errors or technical difficulties. Nonresponse can lead to bias in the estimate, or even complicate it in the case of cross-estimation. Historically, Rubin (1976) introduced the missing at random (MAR). Nonresponse is generally classified into unit nonresponse and item nonresponse (Kalton and Kasprzyk, 1986). Unit nonresponse occurs when an entire sampled unit does not provide any data, rendering it completely unusable. In contrast, item nonresponse occurs when only some variables are missing for a given unit.

To address the issue of nonresponse, different methods are used. One approach is reweighting, where the weights of respondents are adjusted to compensate for nonrespondents (Folsom and Singh, 2000; Särndal and Lundström, 2005; Särndal, 2007; Kott, 2006). Another approach is to use auxiliary information through calibration techniques to align estimates with known population totals, thereby reducing nonresponse bias (Brick, 2013; Haziza and Lesage, 2016). A third widely used technique to handle item nonresponse is imputation. The idea is to replace missing values with plausible estimates in order to reduce the impact of missingness on estimation while preserving the structure of the dataset (Chen and Haziza, 2019; Andridge and Little, 2010).

Several imputation methods have been developed. A commonly used approach is hot-deck imputation, also called donor imputation, where missing values are replaced with observed values from similar units in the dataset. This method ensures that imputed values are realistic, as they come directly from actual respondents (Rancourt and Chen, 1994; Chen and Shao, 1997; Shao and Wang, 2008; Kim and Fuller, 2004; Fuller and Kim, 2005; Chauvet et al., 2011; Eustache et al., 2024). Various other imputation approaches have been developed, such as model-based imputation, which relies on statistical models to predict missing values (David and Sukhatme, 1974; Rao and Sitter, 1995; Shao, 2000; Little, 1988). Multiple imputation is a technique in which several plausible values are generated for each missing data point (Rubin, 1991; Schafer and Graham, 2002; van Buuren and Groothuis-Oudshoorn, 2011). There is also machine learning-based imputation, which leverages predictive models such as random forests or neural networks to estimate missing values (Stekhoven and Bühlmann, 2012; Dagdoug et al., 2023a, 2025).

1.5 Thesis plan

This thesis presents several original contributions to survey statistics, extending existing methods of balanced sampling, calibration and imputation.

In Chapter 2, regarding balanced sampling, an algorithm that extends the classical cube method of Deville and Tillé (2004) by incorporating inequality constraints is proposed. This allows to impose conditions such as minimum or maximum sample sizes in subdomains or to handle non-integer inclusion probability sums in overlapping categories. The method also solves practical problems such as matrix rounding and spatial spreading by translating them into balanced sampling problems subject to inequality constraints.

In terms of calibration, in Chapter 3 harmonized weights strategies for multiple variables of interest are explored. We compare a joint calibration approach and a method based on optimal transport, which both aim to find a common weight system between variable-specific weights. These approaches are particularly useful when cross-tabulations or joint analyses are required. It ensures coherence across the estimated margins and their intersections. Then in Chapter 4, the challenge of calibration with a large number of auxiliary variables is addressed using a method that combines bagging and principal component decomposition. This approach stabilizes the calibration weights by limiting their dispersion and controls the variance of the total estimator. It also allows for flexibility in application to multiple variables of interest.

In Chapter 5, in the area of nonresponse and imputation, a hybrid imputation method that combines random hot-deck method with the predictive power of machine learning models, such as random forests, is developed. This method preserves the benefits of balanced donor imputation while improving its performance in high-dimensional settings, thanks to the refinement of the donor selection via model-based proximity. It offers a flexible tool for dealing with item nonresponse when donor imputation is needed.

Finally, in Chapter 6, a general discussion and conclusion is made. The main contributions of the thesis are summarized, and possible directions for future research in balanced sampling, calibration, and imputation are outlined.

Chapter 2

Balanced Sampling With Inequalities: Application to Category Bounding, Matrix Rounding, and Spread Sampling

Abstract

In this paper, we propose a novel algorithm for balanced sample selection with linear inequality constraints, ensuring that estimators remain within fixed bounds. This algorithm extends the cube method of Deville and Tillé, allowing the selection of a sample from a database where Horvitz-Thompson estimators of totals are equal or nearly equal to the true population totals. The new algorithm has several key applications, including imposing minimum sample sizes for small areas and constraining sample sizes in potentially overlapping categories. It also addresses the controlled rounding matrix problem and links to systematic sampling with unequal probabilities. It can also be used to select doubly stratified samples when the sums of the inclusion probabilities in the strata are not integer. Additionally, the algorithm enables the selection of spatially spread samples. Simulations demonstrate that this new method performs comparably to other spread sampling techniques.[1]

Keywords : cube method, inclusion probabilities, spatial sampling, systematic sampling.

[1]This chapter is based on the article: Tripet, A., & Tillé, Y. (2025). Balanced Sampling With Inequalities: Application to Category Bounding, Matrix Rounding, and Spread Sampling. Journal of the American Statistical Association, 1–21.

2.1 Introduction

When a sample is selected from a register or a sampling frame, a sample is said to be balanced on auxiliary variables $x_1, \ldots, x_p$ if the Horvitz-Thompson estimators of the totals of these auxiliary variables are equal or almost equal to the true population totals. The concept of selecting a balanced sample dates back to the inception of survey sampling theory. Kiær (1896, 1899, 1903, 1905) was the pioneer in proposing the idea of sample selection by quotas, which he termed "representative enumeration". Similarly, Gini and Galvani (1929) applied balanced sample selection in official statistics by choosing 29 Italian districts (*circondari*) out of 214 to best reflect various population averages (Langel and Tillé, 2011; Tillé, 2016; Brewer, 2013).

This method faced significant criticism from Jerzy Neyman because the sample was not chosen randomly (Bellhouse, 1988). Yates (1949) and Thionet (1953) proposed methods where a sample is selected and then improved by successively replacing units to achieve a balanced state. Hájek (1964, 1981) introduced rejective sampling, which involves selecting multiple samples until a sufficiently balanced one is obtained. However, this approach has the drawback of altering the inclusion probabilities of the units, making it impossible to calculate them accurately afterward (Choudhry and Singh, 1979; Dupačová, 1979; Fuller, 2009; Legg and Yu, 2010; Boistard et al., 2012; Fuller et al., 2017). However, they can still be approximate using Monte Carlo approximations (Chauvet et al., 2017).

Deville and Tillé (2004, 2005) proposed the cube method, which allows for the selection of a balanced sample across several auxiliary variables with either equal or unequal inclusion probabilities. Several R packages facilitate the direct selection of a balanced sample (Tillé and Matei, 2021; Grafström and Lisic, 2019; Jauslin et al., 2021). These packages are particularly user-friendly, as their functions rely on just two arguments: the matrix of balancing variables and the vector of inclusion probabilities. Selecting a sample is an integer problem because each unit must be either included or excluded. Therefore, achieving an approximately balanced sample is often necessary.

When the number of balancing variables is very large, it may be necessary to relax the constraints a little. Breidt and Chauvet (2012) have proposed the use of penalized balancing guided by the use of a mixed model. In this paper, we propose an alternative. An algorithm that allows us to select balanced samples that also includes linear inequality constraints. Estimators of totals can then be imposed to remain between fixed bounds. A similar idea was proposed by Oliva-Avilés et al. (2020) who used estimators calibrated on inequality constraints at the estimation stage to improve accuracy and ensure consistency with the expected orderings. In our proposed algorithm, these inequality constraints are directly imposed at the level of the sampling design.

The proposed method is useful when there is a very large number of constraints and it is impossible to meet them all exactly. This algorithm offers great versatility for a variety of applications. For example, it can be used to balance categories when the sum of the inclusion probabilities in these categories is not integer. In this case, we can select a sample whose number of selected units in the category is either the largest integer less than this sum, or the smallest integer greater than this sum.

Another application of this algorithm is to guarantee the minimum sample size in a small group by imposing inequality constraints. This algorithm adapts to various scenarios, including overlapping or unpartitioned groups, and also extends to stratified sampling designs. The algorithm can also be used to randomly round a table whose cell frequencies lie in the interval $[0, 1]$ while respecting the margins, the challenge boils down to the well-known "controlled matrix problem" (see among others Bacharach, 1966; Fellegi, 1975; Cox, 1987; Doerr et al., 2006). Balanced sampling with inequality constraints offers a quick solution to this problem by exactly respecting the probabilities given in the table.

The algorithm can also be used in the context of spatial sampling, whereby, we use inequality

conditions for the neighbourhoods of each unit. This simple approach facilitates the development of an efficient spatial sampling method providing samples that are well spread out in space. To illustrate the performance of this method, we evaluate it, in Section 2.10, against other frequently used spatial sampling methods: The Local Pivotal Method (LPM) proposed by Grafström et al. (2012); Grafström and Lisic (2019), the Weakly Associated Vector Sampling (WAVE) method by Jauslin and Tillé (2020, 2019), and the Generalized Random Tessellation Stratified (GRTS) method by Stevens and Olsen (2004).

The article is structured as follows. In Section 2, we introduce the basics and explain the principles of balanced sampling. In Section 3, we describe the flight phase of the cube method. In Section 4, we present the proposed algorithm, which incorporates inequality constraints, and discuss their consequences. The subsequent sections focus on practical applications of the new algorithm. In Section 5, we show how to bound sample sizes in potentially overlapping categories and in stratified settings. In Section 6, we explain how the algorithm can solve the controlled matrix-rounding problem. In Section 7, we demonstrate that unequal probability systematic sampling satisfies the inequality constraints, thereby motivating the development of a new spread sampling method. In Section 8, we detail this spread sampling approach and provide both a practical example and a spatial-sampling simulation study. In Section 9, we discuss variance estimation under the proposed design. An extensive Monte-Carlo simulation study is reported in Section 10. Finally, in Section 11 we conclude with a discussion of the implications of our results.

2.2 The Problem of Balanced Sampling

The balanced sampling problem consists in selecting a constrained sample from a population $U = \{1, \ldots, k, \ldots, N\}$. For each population unit, the values of p auxiliary variables are assumed to be known. For unit k, the values of these variables are grouped into a vector $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})^\top \in \mathbb{R}^p$. We therefore have a register with known information from which to select a sample. The inclusion probabilities of the units are fixed a priori and denoted by $\pi_k, k \in U$. We also denote $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top \in [0,1]^N$ the vector of inclusion probabilities for all units.

A sample is a subset of the population. A random sample is denoted by a vector of non necessarily independent Bernoulli random variables $\mathbf{s} = (s_1, \ldots, s_k, \ldots, s_N)^\top$. A sampling method with fixed inclusion probabilities is balanced if

$$\sum_{k \in U} \frac{\mathbf{x}_k s_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k \tag{2.1}$$

and $\mathrm{E}(\mathbf{s}) = \boldsymbol{\pi}$. The balancing equations given in (2.1) are difficult to satisfy exactly. In most cases, this equality can only be approximately satisfied.

The balancing equations can be written as

$$\mathbf{A}\mathbf{s} = \mathbf{A}\boldsymbol{\pi}, \tag{2.2}$$

where $\mathbf{A} = (\mathbf{x}_1/\pi_1, \ldots \mathbf{x}_N/\pi_N)$ is a $p \times N$ matrix called the constraint matrix. A sample that approximately satisfies (2.2) can be obtained using the cube method of Deville and Tillé (2004). Our aim is to add inequality constraints to this problem. That is, we also want to ensure that $\mathbf{B}\ \mathbf{s} \leq \mathbf{r}$, where $\mathbf{B}$ is a $q \times N$ matrix and $\mathbf{r}$ is a vector of $\mathbb{R}^q$ such that $\mathbf{B}\boldsymbol{\pi} \leq \mathbf{r}$. Inequality constraints are more flexible than equality constraints as they are easier to satisfy.

2.3 The Flight Phase of the Cube Method

The cube method described in Deville and Tillé (2004) can be used to obtain a balanced random sample by means of a two-phase algorithm: the flight phase and the landing phase. The flight

phase is a random walk through the polytope K_1, where

$$K_1 = \{\mathbf{u} \in [0,1]^N \mid \mathbf{A}\mathbf{u} = \mathbf{A}\boldsymbol{\pi}\}.$$

This random walk ends on one of the vertices of K_1. If this vertex is a sample, the algorithm stops. If this vertex is not a sample, then the landing phase consists in randomly selecting a sample close to the vertex of K_1 given by the flight phase. The cube method respects the inclusion probabilities in expectation.

The flight phase of the cube method described in Algorithm 1 consists of constructing a sequence of vectors $\boldsymbol{\pi}(0)$, $\boldsymbol{\pi}(1)$, $\boldsymbol{\pi}(2)$, ..., such that

1. $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$.
2. $\boldsymbol{\pi}(t) \in K$ for all $t = 0, 1, 2, \ldots$.
3. $\pi_k(t+1) = \pi_k(t)$ if $\pi_k(t) \in \{0, 1\}$.
4. $\mathrm{E}\{\boldsymbol{\pi}(t+1)\} = \boldsymbol{\pi}(t)$ for all $t = 0, 1, 2, \ldots$.

Compared to $\boldsymbol{\pi}(t)$, the vector $\boldsymbol{\pi}(t+1)$ has at least one more component that takes an integer value. Thus in N steps at most, a vertex of K_1 is reached.

Algorithm 1 Basic step of the flight phase of the cube method

$\boldsymbol{\pi}(0) = \boldsymbol{\pi}$
For, $t = 0, 1, 2, 3, \ldots$
In order to go from $\boldsymbol{\pi}(t)$ to $\boldsymbol{\pi}(t+1)$, we proceed as follow:

1. Search a vector $\mathbf{u}(t)$ such that:
   (a) $\mathbf{u}(t)$ is in the kernel of $\mathbf{A}$,
   (b) $u_k(t) = 0$ for all k such that $\pi_k(t) = 0$ or $\pi_k(t) = 1$.
2. Identify the largest values of λ_1 and λ_2 such that

$$\boldsymbol{\pi}(t) + \lambda_1 \mathbf{u}(t) \in [0,1]^N \text{ and } \boldsymbol{\pi}(t) - \lambda_2 \mathbf{u}(t) \in [0,1]^N.$$

3. Define

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1 \mathbf{u}(t) & \text{with probability } \lambda_2/(\lambda_1+\lambda_2) \\ \boldsymbol{\pi}(t) - \lambda_2 \mathbf{u}(t) & \text{with probability } \lambda_1/(\lambda_1+\lambda_2). \end{cases}$$

The algorithm stops at step T when it is not possible to find a vector $\mathbf{u}(T)$ such that $\mathbf{u}(T)$ is in the kernel of $\mathbf{A}$ and $u_k(T) = 0$ for all k such that $\pi_k(T) = 0$ or $\pi_k(T) = 1$. Vector $\boldsymbol{\pi}(T)$ is therefore a vertex of K_1 chosen at random so that $\mathrm{E}\{\boldsymbol{\pi}(T)\} = \boldsymbol{\pi}$.

There are several ways of implementing the flight phase, depending on the choice of vector $\mathbf{u}(t)$. This vector may be either chosen randomly or deterministically. The flight phase can also be applied to part of the population, resulting in a faster algorithm, as proposed in Chauvet and Tillé (2006).

Let $\boldsymbol{\pi}^*$ be the vector obtained at the end of the flight phase. This vector contains at most p components which are different from 0 or 1. The constraints must then be relaxed to obtain a sample. Deville and Tillé (2004) have proposed two methods for the landing phase in order to obtain a sample, i.e. a vector composed solely of 0s and 1s. The first consists of identifying an optimal design on the non-integer units using linear programming. The second consists of relaxing the balancing constraints one by one.

2.4 Cube Method with Inequality Constraints

Methods already exist for imposing inequality constraints on random samples. Fuller (2009) proposed a rejective procedure, whereby a sample selected using a basic sampling design is rejected unless the difference between the sample mean and the population mean of an auxiliary vector lies within a specified distance. The main drawback of Fuller's method is that successive rejections alter the target inclusion probabilities. Therefore, unit-level inclusion probabilities cannot be controlled exactly. The problem becomes particularly acute when the population size is small and the inclusion probabilities are unequal. Rejection methods modify inclusion probabilities. Units with extreme values of the auxiliary variables are less likely to be selected in rejective sampling (see for example Legg and Yu, 2010). Modifying inclusion probabilities biases estimators. Thus, dispersion parameters and quantiles can be severely biased due to these modifications. Unlike the procedure of Fuller (2009), the proposed method respects the inclusion probabilities exactly.

The usual landing phase of the cube method is very efficient if carried out using the linear programming method. However, the linear programming method is limited to handling up to 20 auxiliary variables, which restricts the number of constraints that can be used. The cube method with inequalities allows a large number of variables to be used in order to avoid very unbalanced samples. We can use many more inequality constraints than equality constraints. We can even include more inequality constraints than the number N of observations in the population. The idea of balanced sampling with inequalities is not to propose an even more efficient method, but to be able to add a large number of constraints.

We now want to select a sample, trying to satisfy both equality and inequality constraints simultaneously:

$$\mathbf{A}\,\mathbf{s} = \mathbf{A}\boldsymbol{\pi} \text{ and } \mathbf{B}\,\mathbf{s} \leq \mathbf{r}.$$

Algorithm 2 gives a flight phase which randomly selects a vertex of the polytope K_2, where

$$K_2 = \{\mathbf{u} \in [0,1]^N \mid \mathbf{A}\mathbf{u} = \mathbf{A}\boldsymbol{\pi} \text{ and } \mathbf{B}\mathbf{u} \leq \mathbf{r}\}.$$

Since K_2 is obtained by adding constraints to K_1, $K_2 \subset K_1$. In Algorithm 2, each time the random walk reaches a face of the polytope, the random walk must remain permanently on that face. Therefore, when the vector $\boldsymbol{\pi}(t)$ reaches a face of the polytope, it sticks to that face. The inequality constraints then gradually become equality constraints.

If we define the polytope

$$K_3 = \{\mathbf{u} \in [0,1]^N \mid \mathbf{A}\mathbf{u} = \mathbf{A}\boldsymbol{\pi} \text{ and } \mathbf{B}\mathbf{u} = \mathbf{B}\boldsymbol{\pi}\},$$

we have $K_3 \subset K_2 \subset K_1$. We can therefore first apply a flight phase of 1 of the cube method with the constraints: $\{\mathbf{A}\mathbf{s} = \mathbf{A}\boldsymbol{\pi} \text{ and } \mathbf{B}\mathbf{s} = \mathbf{B}\boldsymbol{\pi}\}$.

The question of whether an exactly balanced sampling design exists is a highly complex one. It boils down to determining which polytope contains only vertices composed solely of 0 and 1. This issue has been addressed in the context of integer linear optimization (see for example Korte and Vygen, 2018). In cases where the existence of an exact solution is not certain, we can first apply Algorithm 1, with the constraints $\{\mathbf{A}\mathbf{s} = \mathbf{A}\boldsymbol{\pi} \text{ and } \mathbf{B}\mathbf{s} = \mathbf{B}\boldsymbol{\pi}\}$. Next, we apply Algorithm 2 to the vector obtained at the end of Algorithm 1, with the constraints $\{\mathbf{A}\mathbf{s} = \mathbf{A}\boldsymbol{\pi} \text{ and } \mathbf{B}\mathbf{s} \leq \mathbf{B}\boldsymbol{\pi}\}$. This allows to explore whether an exact solution can be obtained by Algorithm 1 before adding inequalities. If there is no exact solutions for certain contraints, they are relaxed to allow inequalites; for example, if fixed size in a category is needed and the sum of the inclusion probabilities in this category is not an integer.

At each step, a component of the vector $\boldsymbol{\pi}(t)$ is not necessarily set to 0 or 1. The vector $\boldsymbol{\pi}(t+1)$ may either stop on a face of the hypercube or inside the hypercube on a face of the polytope defined by the inequality constraints. In the latter case, the inequality constraint

Algorithm 2 Basic step of the flight phase of the cube method with inequality constraints

$\boldsymbol{\pi}(0) = \boldsymbol{\pi}, \mathbf{A}(0) = \mathbf{A}, \mathbf{B}(0) = \mathbf{B}$,
For, $t = 0, 1, 2, 3, \ldots$
In order to go from $\boldsymbol{\pi}(t)$ to $\boldsymbol{\pi}(t+1)$, we proceed as follow:

1. Search a vector $\mathbf{u}(t)$ such that
   (a) $\mathbf{u}(t)$ is in the kernel of $\mathbf{A}(t)$
   (b) $u_k(t) = 0$ for all k such that $\pi_k(t) = 0$ or $\pi_k(t) = 1$.
2. Identify the largest values of λ_1 and λ_2 such that

$$\boldsymbol{\pi}(t) + \lambda_1 \mathbf{u}(t) \in [0,1]^N, \quad \boldsymbol{\pi}(t) - \lambda_2 \mathbf{u}(t) \in [0,1]^N,$$

$$\boldsymbol{\pi}(t) + \lambda_1 \mathbf{u}(t) \leq \mathbf{r}, \text{ and } \boldsymbol{\pi}(t) - \lambda_2 \mathbf{u}(t) \leq \mathbf{r}.$$

3. Define

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1 \mathbf{u}(t) & \text{with probability } \lambda_2/(\lambda_1+\lambda_2) \\ \boldsymbol{\pi}(t) - \lambda_2 \mathbf{u}(t) & \text{with probability } \lambda_1/(\lambda_1+\lambda_2). \end{cases}$$

4. All the rows of $\mathbf{B}(t)$ such that $\mathbf{B}(t)\boldsymbol{\pi}(t+1) = \mathbf{r}(t)$ are removed from $\mathbf{B}(t)$ to create $\mathbf{B}(t+1)$ and added to $\mathbf{A}(t)$ to create $\mathbf{A}(t+1)$. The corresponding cells of $\mathbf{r}(t)$ are also removed to define $\mathbf{r}(t+1)$.

The algorithm stops at step T when it is not possible to find a vector $\mathbf{u}(T)$ such that $\mathbf{u}(T)$ is in the kernel of $\mathbf{A}(T)$ and $u_k(T) = 0$ for all k such that $\pi_k(T) = 0$ or $\pi_k(T) = 1$. Vector $\boldsymbol{\pi}(T)$ is therefore a vertex of K_2 chosen at random so that $\mathrm{E}(\boldsymbol{\pi}(T)) = \boldsymbol{\pi}$.

becomes an equality constraint. As the algorithm progresses, these inequality constraints are transformed into equality constraints, gradually converging towards the usual configuration of the flight phase of the standard cube method, where, at each step, a component of the vector is set to 0 or 1. Therefore, the number of steps in the proposed method is strictly limited to $2N$.

One of the tricky questions is the appropriate choice of $\mathbf{r}$. If we impose the inequality constraint $\mathbf{Bs} \leq \mathbf{r}$, we must have $\mathbf{B}\boldsymbol{\pi} \leq \mathbf{r}$. If this is not the case, the vector of inclusion probabilities does not belong to the polytope. Consequently, the algorithm cannot even be started.

If matrix $\mathbf{B}$ indicates categories, then we can take the smallest integers greater than the sum of the inclusion probabilities in these categories, i.e. $\mathbf{Bs} \leq \mathbf{r} = \lceil \mathbf{B}\boldsymbol{\pi} \rceil$, where $\lceil \cdot \rceil$ is the ceiling function. We can also take the largest integers smaller than the sum of the inclusion probabilities in these categories. In this case, the direction of the inequality is reversed. We obtain. $\mathbf{Bs} \geq \mathbf{r} = \lfloor \mathbf{B}\boldsymbol{\pi} \rfloor$, or equivalently $-\mathbf{Bs} \leq -\lfloor \mathbf{B}\boldsymbol{\pi} \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function. We therefore recommend choosing $\mathbf{r}$ in such a way that the constraints can be satisfied.

2.5 Minimum Group Sizes, Stratification and Rounding Problem

In many survey problems, we are interested in estimates in small groups or in small geographical areas. The whole methodology known as "small area estimation" (Rao and Molina, 2015) has been developed to construct accurate estimates in small entities. If the small areas of interest are identified in advance, potential issues can be avoided at the design stage. With the technique

of balanced sampling on inequalities, we can ensure a minimum sample size in small entities by imposing the constraint

$$\sum_{k \in G_h} s_k \geq \left\lfloor \sum_{k \in G_h} \pi_k \right\rfloor .$$

where $G_h \subset U, h = 1, \ldots, H$ is a group or a small area. The G_h groups can form a partition of the population U, but not necessarily. Groups can also overlap, without this posing any implementation problems. Perfectly feasible inequality constraints can thus be applied to any categorical variable.

A stratification is a partition of the population $U_h \subset U$, with $U = \bigcup_{h=1}^{H} U_h$ and $U_h \cap U_\ell = \emptyset$ if $k \neq \ell$. In stratification, inclusion probabilities depend on the allocation in each stratum U_h of size $N_h, h = 1, \ldots, H$. Whether for proportional or optimal allocation, there is no reason why the sum of the inclusion probabilities in each stratum should be an integer. With balanced sampling on inequality constraints, we can simply impose

$$\left\lfloor \sum_{k \in U_h} \pi_k \right\rfloor \leq \sum_{k \in U_h} s_k \leq \left\lceil \sum_{k \in U_h} \pi_k \right\rceil , \text{ for } h = 1, \ldots, H$$

and that

$$\sum_{k \in U} s_k = n.$$

In general, balancing constraints on categorical variables can overlap. Balanced sampling can therefore be applied to several stratifications whose strata categories overlap. In this way, inequalities can be imposed on the margins of a contingency table obtained by sampling. In case of double stratification $U_1, \ldots, U_h, \ldots U_H$ and $V_1, \ldots, V_i, \ldots, V_I$, we can impose

$$\left\lfloor \sum_{k \in U_h} \pi_k \right\rfloor \leq \sum_{k \in U_h} s_k \leq \left\lceil \sum_{k \in U_h} \pi_k \right\rceil , \text{ for } h = 1, \ldots, H,$$

$$\left\lfloor \sum_{k \in V_i} \pi_k \right\rfloor \leq \sum_{k \in V_i} s_k \leq \left\lceil \sum_{k \in V_i} \pi_k \right\rceil , \text{ for } i = 1, \ldots, I,$$

and that

$$\sum_{k \in U} s_k = n. \tag{2.3}$$

We thus have the equality constraint given in (2.3) and $2 \times (H + I)$ inequality constraints. The inclusion probabilities are then exactly satisfied and the inequality constraints can be exactly satisfied.

Example 1. *In the MU284 population of 284 Swedish areas (Särndal et al., 1992), we constructed a categorical variable by splitting the P75 variable "Population in 1975" into 4 classes. This variable is crossed with the REG variable consisting of the 8 Swedish regions. We select $n = 50$ areas with equal probability. The inclusion probabilities are thus $\pi_k = 50/284$. In the table crossing the two variables, the expectation of the number of units to be selected is given in Table 2.1.*

Thanks to the balanced sampling method with inequality constraints, we can impose that all the margins of the sample table are equal to the rounding down or up. We therefore have an equality constraint to obtain the fixed sample size and $(8 + 4)/2 = 24$ inequality constraints to frame all marginal totals by integers. For example, we obtained Table 2.2. The inequality constraints can be satisfied exactly. There is therefore no need for a landing phase.

Table 2.1: Expected number of units to select in each category

REG/P75	(0,10]	(10,15]	(15,29]	(29,700]	Total
1	0.18	0.88	1.23	2.11	4.40
2	2.64	1.58	2.29	1.94	8.45
3	0.88	1.76	1.76	1.23	5.63
4	0.53	2.64	1.76	1.76	6.69
5	3.52	1.94	1.94	2.46	9.86
6	1.94	1.94	1.94	1.41	7.22
7	0.70	0.53	0.88	0.53	2.64
8	3.52	0.18	0.53	0.88	5.11
Total	13.91	11.44	12.32	12.32	50.00

Table 2.2: Example of sample selected with balanced sampling with inequality constraints

REG/P75	(0,10]	(10,15]	(15,29]	(29,700]	Total
1	0	1	1	3	5
2	3	1	3	1	8
3	1	2	1	1	5
4	0	3	2	2	7
5	4	1	2	3	10
6	1	3	2	1	7
7	1	1	0	0	2
8	4	0	1	1	6
	14	12	12	12	50

2.6 The Controlled Matrix Rounding Problem

The controlled rounding matrix problem involves randomly rounding a probability matrix, whose values lie between 0 and 1, to 0 or 1. The process must remain unbiased, meaning the expectation of the resulting matrix equals that of the original matrix. In addition, the margins of the table must be preserved by rounding them to the nearest integer either below or above their original values (see among others Bacharach, 1966; Fellegi, 1975; Cox, 1987; Doerr et al., 2006). Balanced sampling with inequality constraints makes it possible to quickly solve this problem in a random way by exactly respecting the probabilities given in the table.

For instance, if we take, as suggested in Cox (1987), the example of Cochran (1977, p. 124), which represents a population of 165 schools stratified by city size into five classes and by average expenditure per pupil into four classes, and divide it by 16.5 and then remove the integer parts, we obtain a controlled matrix problem as shown in Table 2.3.

Table 2.3: Example of controlled matrix problem

	A	B	C	D	Total
I	0.91	0.27	0.03	0.55	1.76
II	0.61	0.48	0.79	0.42	2.30
III	0.36	0.55	0.30	0.48	1.70
IV	0.24	0.18	0.36	0.36	1.15
V	0.18	0.12	0.30	0.48	1.09
Total	2.30	1.61	1.79	2.30	8.00

To apply our method, the table values are aligned in the vector of inclusion probabilities

of $\mathbb{R}^{20}$. An equality constraint is used to obtain the fixed sample size 8, and $2 \times (4+5) = 18$ constraints are used to frame the margins with integers. Our algorithm solves this problem efficiently, one of the solutions is shown in Table 2.4. Indeed, each margin is respected with either upper or lower rounding.

Table 2.4: Example of a solution to the controlled matrix problem using balanced sampling with inequality constraints

	A	B	C	D	Total
I	1	0	0	1	2
II	1	1	0	1	3
III	0	0	0	1	1
IV	0	1	0	0	1
V	0	0	1	0	1
Total	2	2	1	3	8

2.7 Unequal Probability Systematic Sampling

Consider a sequence of N inclusion probabilities $\pi_1, \ldots, \pi_N$ which sum to n. Consider also the cumulative inclusion probabilities:

$$V_k = \sum_{j=1}^{k} \pi_j,$$

for $j = 1, \ldots, N$ with $V_0 = 0, V_1 = \pi_1$, and $V_N = n$. In order to select a sample using a systematic design with unequal probabilities, we generate a continuous uniform variable u in the interval [0,1]. Next, we select the units $k_1, \ldots, k_j, \ldots, k_n$ such that

$$V_{k_j - 1} < u + (j-1) \leq V_{k_j}, j = 1, \ldots, n.$$

Consider now $v_k = (V_k \bmod 1)$ and $v_{(j)}, j = 1, \ldots, N$ the ordered v_k. Pea et al. (2007) showed that unequal systematic sampling is a minimal support design, i.e. at most N samples have a non-zero probability of being selected. The probabilities of the jth sample are

$$p(\mathbf{s}_j) = v_{(j)} - v_{(j-1)}, j \in \{1, \ldots, N \mid v_{(j)} - v_{(j-1)} \neq 0\},$$

where sample $\mathbf{s}_j$ contains the units k such that intervals $\Big(v_{(j-1)} + i - 1, \;\; v_{(j)} + i - 1\Big]$, $i = 1, \ldots, n$ are included in intervals $(V_{k-1}, V_k]$.

Result 1. *A necessary condition for a sampling design to be systematic is that $E(s_k) = \pi_k$, for all $k \in U$,*

$$\left\lfloor \sum_{t=1}^{T} \pi_{\{(k+t-1) \bmod N\}+1} \right\rfloor \leq \sum_{t=1}^{T} s_{\{(k+t-1) \bmod N\}+1} \leq \left\lceil \sum_{t=1}^{T} \pi_{\{(k+t-1) \bmod N\}+1} \right\rceil, \tag{2.4}$$

for all $k \in U, T = 1, \ldots, N-1$, and

$$\sum_{k \in U} s_k = n.$$

Proof. In an interval of length L included in $[0, n]$, we will select $\lfloor L \rfloor$ or $\lceil L \rceil$ units. □

The condition given in Result 1 shows that systematic sampling with unequal inclusion probabilities is related to balanced sampling with inequalities. There are $2N(N-1)$ inequalities in (2.4), but they are largely redundant. We ran a set of simulations. After removing redundancies, we selected a large number of balanced samples on these inequalities. We only selected systematic samples. However, we have not been able to establish a sufficient condition to prove that these inequality constraints lead to systematic sampling.

2.8 Spread Sampling

In one dimension, systematic samples are well spread out in the sense that if a unit k is selected, the following units will not be selected until the sum of their inclusion probabilities (with the unit k) is larger than or equal to 1. We can therefore draw inspiration from systematic sampling to propose a well spread out method in a space of two or more dimensions. Well-spread sampling has recently been the subject of numerous publications, since if the observations are spatially autocorrelated, the spread considerably increases the precision of the total and mean estimators (Stevens and Olsen, 1999, 2003, 2004; Grafström, 2011; Grafström et al., 2012; Grafström and Lundström, 2013; Grafström and Tillé, 2013; Grafström et al., 2014; Dickson and Tillé, 2016; Jauslin and Tillé, 2020; Eustache et al., 2022; Jauslin et al., 2022). Figure 2.1 graphically illustrates the importance of choosing an appropriate sampling method when spatially sampling. The graph on the left shows a spatially well-spread sample, selected using the balanced sampling with inequalities algorithm. We can see that the sampled points, represented in black, are well spread over the area, which means that population coverage is complete, and avoids excessive clustering. On the other hand, the sample in the graph on the right has been selected by cluster sampling and shows us a sample that is poorly spread in space. The points are not scattered but grouped together in certain zones, which, as indicated above, can lead to problems in the accuracy of the estimates.

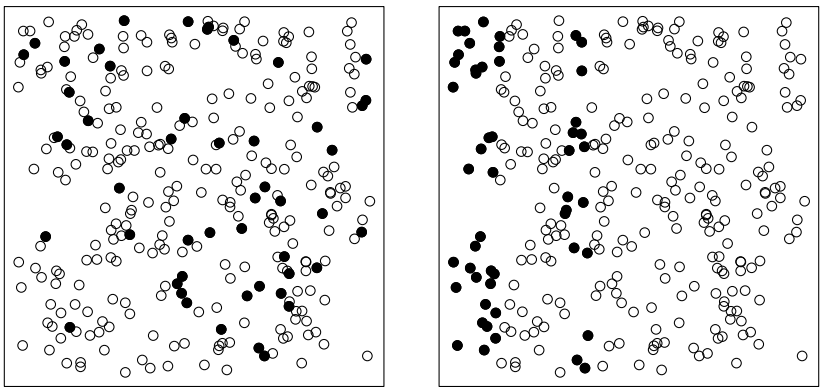

Figure 2.1: Illustration of spatial sampling. The left plot shows a well-spread sample obtained using the balanced sampling algorithm with inequalities. The right plot shows a poorly spread sample.

Suppose one has defined a distance between the units $d(k,\ell)$, for all $k,\ell \in U$. We define:

- C_k^{b-} the largest set of units closest to k (including unit k) such that $\sum_{\ell \in C_k} \pi_k \leq b$. More formally, consider a permutation σ^k of $\{1,\dots,N\}$ such that if $i < j$ then $d(\sigma^k(i),k) \leq d(\sigma^k(j),k)$. Define also

$$L_k^- = \max_\ell \left\{ \ell \mid \sum_{i=1}^{\ell} \pi_{\sigma^k(i)} \leq b \right\} \text{ and } C_k^{b-} = \{\sigma^k(i) : i \in 1,\cdots,L_k^-\}.$$

- C_k^{b+} the smallest set of units closest to k (including unit k) such that $\sum_{\ell \in C_k} \pi_k \geq b$.

  More formally, by using the same permutation σ^k , define

$$L_k^+ = \min_\ell \left\{ \ell \mid \sum_{i=1}^{\ell} \pi_{\sigma^k(i)} \geq b \right\} \text{ and } C_k^{b+} = \{\sigma^k(i) : i \in 1, \cdots, L_k^+\}.$$

In order to obtain a sample well spread out in space, we can apply balanced sampling on inequalities on the following constraints

$$\sum_{\ell \in C_k^{1-}} s_\ell \leq 1, k = 1, \ldots, N, \text{ and } \sum_{k \in U} s_k = n$$

or on

$$\sum_{\ell \in C_k^{1+}} s_\ell \geq 1, k = 1, \ldots, N, \text{ and } \sum_{k \in U} s_k = n.$$

By imposing these inequality constraints, we can ensure that in the neighbourhood of each unit, the number of units selected does not exceed a fixed value or is at least equal to a fixed value. The sample will therefore be well spread out in space. The proposed method can thus be applied to spatial sampling.

In the case of spatial neighborhood constraints, the cube method with equalities is not applicable, as the number of constraints exceeds the number of observations. Therefore, the direct application of the landing phase is also inapplicable, as it involves managing too many variables. This scenario clearly illustrates that the cube method with inequalities paves the way to applications that were previously impossible.

2.9 Point and Variance Estimation

To make inference about the total of a variable of interest y, defined as $Y = \sum_{k \in U} y_k$, we can use the Horvitz-Thompson estimator given by

$$\widehat{Y} = \sum_{k \in U} \frac{y_k s_k}{\pi_k}$$

(Horvitz and Thompson, 1952). In the case of the cube method, second-order inclusion probabilities cannot be calculated, even in simple scenarios. For balancing with inequalities, which is an even more complex problem, computing even an approximate expression for the second-order inclusion probabilities is extremely challenging if not impossible.

A first method, which we recommend for small sample sizes, involves approximating the first- and second-order inclusion probabilities through Monte Carlo simulations, as proposed by Breidt and Chauvet (2011). We select M samples $\mathbf{s}_1, \ldots, \mathbf{s}_M$ and compute

$$\widetilde{\boldsymbol{\pi}} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{s}_i, \quad \widetilde{\boldsymbol{\Pi}} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{s}_i \mathbf{s}_i^\top, \quad \widetilde{\boldsymbol{\Delta}} = \widetilde{\boldsymbol{\Pi}} - \widetilde{\boldsymbol{\pi}} \widetilde{\boldsymbol{\pi}}^\top.$$

We can then estimate the variance with the Sen-Yates-Grundy estimator (Sen, 1953; Yates and Grundy, 1953)

$$\widehat{\text{var}}(\widehat{Y}) = -\frac{1}{2} \sum_{k \in U} \sum_{\ell \in U} s_k s_\ell \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\widetilde{\Delta}_{k\ell}}{\widetilde{\pi}_{k\ell}},$$

where $\widetilde{\pi}_{k\ell}$ and $\widetilde{\Delta}_{k\ell}$ denote respectively the element (k, ℓ) of $\widetilde{\pi}$ and $\widetilde{\Delta}$.

We recommend carrying out at least $M = 10,000$ simulations. Breidt and Chauvet (2011) also proposed a method that uses the martingale structure of the sampling algorithm to approximate second-order inclusion probabilities. This approach can also be applied when balancing with inequalities.

In the case of large sample sizes, we can rely on the results of Deville and Tillé (2005), who proposed a heuristic variance estimation method. This method is based on the hypothesis that balanced sampling can be viewed as a Poisson sampling design conditional on the balancing constraints. When the sample is balanced on the variables $\mathbf{As} = \mathbf{A}\boldsymbol{\pi}$, with $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_N)$, an estimator of the variance of $\widehat{Y}$ is given by

$$\widehat{\text{var}}(\widehat{Y}) = \sum_{k \in U} c_k e_{k,y|A}^2,$$

where $e_{k,y|A}$ is the residual of a weighted regression with y_k/π_k as the response variable and $\mathbf{a}_k = \mathbf{x}_k/\pi_k$, as the predictors;

$$e_{k,y|A} = \frac{y_k}{\pi_k} - \mathbf{a}_k^\top \widehat{\boldsymbol{\Gamma}}_{y|A},$$

where

$$\widehat{\boldsymbol{\Gamma}}_{y|A} = \left(\sum_{k \in U} s_k c_k \mathbf{a}_k \mathbf{a}_k^\top \right)^{-1} \sum_{k \in U} \frac{s_k c_k \mathbf{a}_k y_k}{\pi_k}$$

and

$$c_k = \frac{n}{n-p}(1 - \pi_k).$$

In the case where the sample is balanced with in addition constraints of inequalities on the variables $\mathbf{As} = \mathbf{A}\boldsymbol{\pi}$ and $\mathbf{Bs} \leq \mathbf{r}$, with $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_N)$ and $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_N)$, the estimator of the variance of the estimator of the total is:

$$\widehat{\text{var}}(\widehat{Y}) = \sum_{k \in U} c_k e_{k,y|AB}^2 + \left\{ \sum_{k \in U} \begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{pmatrix}^\top (s_k - \pi_k) \widehat{\boldsymbol{\Gamma}}_{y|AB} \right\}^2, \tag{2.5}$$

where $e_{k,y|AB}$ are the residuals of a weighted regression of the variable of interest y_k/π_k by the auxiliary variables $\mathbf{a}_k = \mathbf{x}_k/\pi_k$ and $\mathbf{b}_k$, i.e.

$$e_k = \frac{y_k}{\pi_k} - \begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{pmatrix}^\top \widehat{\boldsymbol{\Gamma}}_{y|AB},$$

where

$$\widehat{\boldsymbol{\Gamma}}_{y|AB} = \left\{ \sum_{k \in U} s_k c_k \begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{pmatrix} \begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{pmatrix}^\top \right\}^{-1} \sum_{k \in U} s_k c_k \begin{pmatrix} \mathbf{a}_k \\ \mathbf{b}_k \end{pmatrix} \frac{y_k}{\pi_k},$$

$$c_k = \frac{n}{n-p-q}(1 - \pi_k),$$

and q is the number of rows of $\mathbf{B}$.

The second term of (2.5) accounts for the fact that the sample is only approximately balanced, meaning a rounding error may persist in both $\mathbf{A}$ and $\mathbf{B}$. This term takes the impact of the rounding problem on variance into account. If we decide to calibrate the weights $1/\pi_k$ on the population totals on all equality and inequality balancing variables, following the methodology of Deville and Särndal (1992), the variance can then be estimated by neglecting the second term of the estimator given in (2.5).

Example 2.9.1. *To evaluate the accuracy of these variance estimators, we conduct 200,000 simulations using the Swiss municipalities database available in the R sampling package (Tillé and Matei, 2021). Samples of* $n = 400$ *communes are selected with inclusion probabilities proportional to population size. The sample is balanced on several variables: total population, number of men, number of 1-person households, number of 2-person households, number of 3-person households, and the surface area of the commune. The aim is to estimate the total wooded area in Switzerland. Additionally, a minimum number of selected communes is imposed in each canton. For canton* U_i*, at least* $\left\lfloor \sum_{k \in U_i} \pi_k \right\rfloor$ *communes had to be included in the sample. The results are summarized in Table 2.5. The proposed estimator performs well, with a slight underestimation of the Sen-Yates-Grundy variance estimator, due to the fact that 1,353 joint inclusion probabilities (out of 4,191,960) are equal to zero.*

Table 2.5: Simulation results based on the Swiss municipalities database. For the variance and the Sen-Yates-Grundy variance estimator, the first- and second-order inclusion probabilities estimated via simulations.

Variance of the Horvitz-Thompson estimator	11232
Variance computed with the Sen-Yates-Grundy formula	11232
Mean under simulations of the Sen-Yates-Grundy variance estimator	10704
Mean under simulations of the estimator proposed in (2.5)	11202

In some specific cases, alternative variance estimators may be required. It has been observed that the number of constraints can sometimes exceed the sample size, in this case, formula (2.5) is inapplicable. Furthermore, in some cases, such as spread sampling, a large proportion of second-order inclusion probabilities are zero, making the Sen-Yates-Grundy variance estimator based on their simulated approximation not applicable either. In such situations, we can use variance estimation methods based on heuristic reasoning, as developed in Stevens and Olsen (2003), Grafström et al. (2012), or Grafström and Tillé (2013).

2.10 Simulation Study

A simulation study comprising $M = 10,000$ iterations is carried out to compare our spread sampling method with three widely used high-performance spatial sampling methods: WAVE, LPM and GRTS. The GRTS method proposed by Stevens and Olsen (2004) is a sampling method that uses a quadrant-recursive function to map the two-dimensional population into one dimension. The sample is then selected systematically. The LPM was proposed by Grafström et al. (2012). This method applies the pivotal method of Deville and Tillé (1998) and Srinivasan (2001) on neighbouring units. A repulsion in the selection of neighbouring units makes it possible to obtain a well-spread sample. The WAVE method is based on the search for directions that are weakly or not at all correlated with the variables indicating the contiguity of each unit. This method spreads out very well but requires considerable computing time once the population is large.

Our study uses the "Meuse" river dataset available in the R package "sp" (Pebesma and Bivand, 2005; Bivand et al., 2013; Burrough et al., 2015), a well-established reference dataset, of size $N = 155$, in the environmental sciences. The main objective of the study is to estimate the concentration of cadmium, a heavy metal pollutant, under two distinct sampling scenarios. The sample size in both scenarios is set to 20% of the dataset, resulting in $n = 0.2 \times 155 = 31$. In the first scenario, the sample is selected with equal inclusion probabilities, i.e. $\pi_k = n/N$. In the second scenario, the inclusion probabilities are proportional to the copper concentration variable.

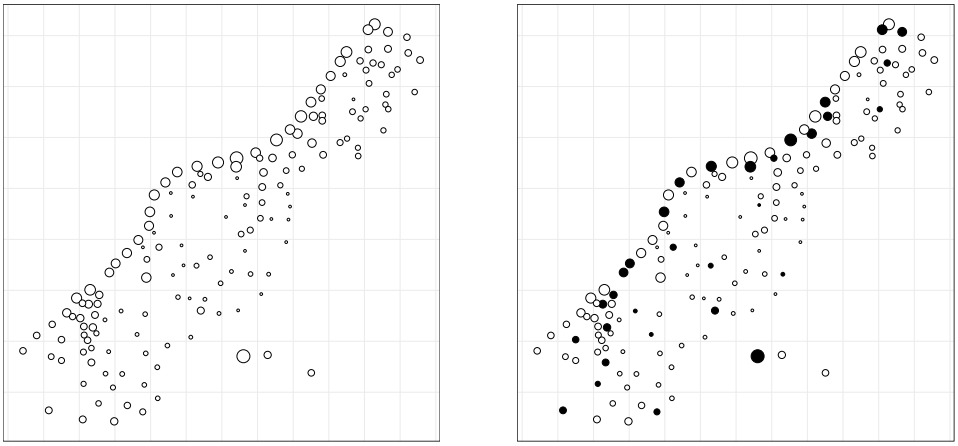

Figure 2.2: Spatial distribution of cadmium concentrations in the Meuse river dataset, black points represent an example of a selected sample.

Figure 2.2 is a spatial representation of the "Meuse" river dataset. Each point represents a location where a cadmium concentration was recorded. The size of each point is proportional to the cadmium concentration recorded, with larger points indicating higher concentrations. Points coloured black in the graph on the right represent a sample selected using our algorithm.

To assess the quality of the spatial samples generated by the different methods in the simulation study, we use several measures: the Moran index, the measure of spatial balance using Voronoï polygons and the total estimate of the variable of interest, cadmium concentration.

First suggested by Stevens and Olsen (2004), then used by Grafström et al. (2012), the measure of spatial balance can be assessed using Voronoï polygons and can be used to determine whether a sample is well spread or not (see Grafström and Lundström, 2013). Voronoï polygons provide a partition of a space into regions based on the distance to a specific set of points. The Voronoï polygon of each sample unit includes those population units that are closer to it than to any other sample unit. If the sample is perfectly spread, the sum of the inclusion probabilities $v_k, k \in S$ of these units is equal or close to 1. Then, to assess the spatial balance of a sample, we calculate the variance of the expectation of the sum of these inclusion probabilities. The Spatial Balance measure (SB) is defined as:

$$\mathrm{SB} = \frac{1}{n} \sum_{k \in S} (v_k - 1)^2.$$

So the closer SB is to 0, the more spread the sample. Tillé et al. (2018) modified the Moran index (Moran, 1950) to obtain a measure of dispersion between 0 and 1. The closer the modified Moran index is to 1, the greater the spatial clustering. The closer the Moran index is to -1, the more dispersed the sample.

In the simulation study, we evaluate the performance of the Horvitz-Thompson estimator for our variable of interest, cadmium level, by Monte Carlo simulations. We compute the MC bias

$$\widehat{\mathrm{Bias}}_{MC}\left(\widehat{Y}\right) = \frac{1}{M} \sum_{m=1}^{M} \left(\widehat{Y}_m - Y\right),$$

the MC variance

$$\widehat{\mathrm{Var}}_{MC}\left(\widehat{Y}\right) = \frac{1}{M} \sum_{m=1}^{M} \left(\widehat{Y}_m - \frac{1}{M} \sum_{m'=1}^{M} \widehat{Y}_m\right)^2$$

and the MC Mean Squared Error (MSE)

$$\widehat{\mathrm{MSE}}_{MC}\left(\widehat{\theta}\right) = \frac{1}{M} \sum_{m=1}^{M} \left(\widehat{Y}_m - Y\right)^2,$$

The true total is denoted by θ, and $\widehat{\theta}_m$ represents the estimate of the total obtained in the mth simulation iteration.

In order to see whether one approach distinguishes itself from the others, we compare three versions of our balanced sampling algorithm with inequality constraints, with the previously mentioned methods: LPM, GRTS and WAVE. These different variants are all based on the process described in Section 2.8 and differ in the constraints on which the inequalities are applied. In the first version, called CubeI$_{G1}$, we use the upper sum constraints:

$$\sum_{\ell \in C_k^{1-}} s_\ell \leq 1, \quad k = 1, \ldots, N, \quad \text{and} \quad \sum_{k \in U} s_k = n.$$

In the second version, called CubeI$_{L1}$, we use the lower sum constraints.

$$\sum_{\ell \in C_k^{1+}} s_\ell \geq 1, \quad k = 1, \ldots, N, \quad \text{and} \quad \sum_{k \in U} s_k = n.$$

In the third version, called CubeI$_{LG1}$, we use a combination of both the upper and lower sum constraints. In addition, to study the results of a scenario where spatial sampling constraints are not considered, we also compare our methods to Simple Random Sampling Without Replacement (SRSWOR) when dealing with equal inclusion probabilities and to the maximum entropy sampling (also called conditional Poisson sampling) method (MaxEnt) for unequal inclusion probabilities (see, for instance, Tillé, 2006, chapter 6).

Table 2.6 shows the SB measure and Moran index for different sampling methods in equal and unequal probability scenarios. In the scenario with equal inclusion probabilities, the balanced sampling with inequality constraints methods show SB values between 0.17 and 0.18, indicating a spatial balance that the three methods achieve similar results. In addition, these results are comparable to widely used methods, and even better than the GRTS method. The Moran index for the CubeI methods lies between -0.402 and -0.403, which also shows spatial dispersion. Again, this value is very similar to that of the LPM, and this time much better than GRTS. Overall, WAVE achieves the best results every time. As expected, SRSWOR in the equal probability scenario shows a significantly higher SB value and a Moran index very close to 0. In the scenario with unequal inclusion probabilities, our methods maintain SB values between 0.16 and 0.17, as well as a Moran index between -0.23 and -0.26, demonstrating consistent spatial balance and reasonable spatial dispersion with results again similar to the LPM and superior to GRTS. As for SRSWOR in the case of equal inclusion probability, the MaxEnt method has an SB value much higher than the others and a Moran index very close to 0, i.e. much less favourable results than the other methods.

Table 2.6: Spatial Balance (SB) and Moran's index for different sampling methods with equal and unequal probabilities

	SB	Moran's I		SB	Moran's I
Equal probabilities			Unequal probabilities		
CubeI$_{L1}$	0.18	-0.40	CubeI$_{L1}$	0.16	-0.26
CubeI$_{G1}$	0.17	-0.40	CubeI$_{G1}$	0.16	-0.23
CubeI$_{LG1}$	0.17	-0.40	CubeI$_{LG1}$	0.17	-0.23
GRTS	0.18	-0.22	GRTS	0.19	-0.09
WAVE	0.11	-0.66	WAVE	0.11	-0.43
LPM	0.12	-0.43	LPM	0.12	-0.29
SRSWOR	0.43	-0.02	MaxEnt	0.34	0.06

Table 2.7: Estimation of the total with different sampling methods with equal and unequal probabilities

	CubeI$_{L1}$	CubeI$_{G1}$	CubeI$_{LG1}$	GRTS	WAVE	LPM	SRS/MaxEnt
Population Total	503	503	503	503	503	503	503
Equal							
Bias	0.51	-0.12	-0.40	1.41	-0.35	0.18	-0.30
Variance	5366	5389	5484	6031	4645	4915	7479
MSE	5367	5389	5484	6033	4645	4915	7479
Unequal							
Bias	-0.08	-0.11	0.06	0.62	0.05	0.14	0.60
Variance	1125	1027	1112	1216	985	1042	1439
MSE	1125	1027	1112	1216	985	1042	1440

Table 2.7 provides estimates of the total with different sampling methods with equal or unequal inclusion probability scenarios. As the Horvitz-Thompson estimators are unbiased, almost all of the MSE measured in simulation is due to variance. Once again, in both scenarios, the results of our methods are comparable to those of other methods. It consistently outperforms GRTS and occasionally surpasses LPM. The WAVE method, on the other hand, seems to consistently deliver the best results. However, the WAVE method can be very time-consuming to calculate. Overall, methods based on inequality constraints show competitive spatial balance, dispersion, accuracy and precision, comparable to methods widely used in spatial sampling. In all cases, they outperform the GRTS method, once again underlining their value in this field. Analysis of the CubeI methods reveals that the results of the three methods are very similar in all scenarios. Each method gives comparable results in terms of spatial balance and Moran index. When estimating the total, the three methods have similar MSEs, with each method sometimes outperforming the others depending on the specific scenario.

2.11 Discussion

In this paper, we introduced a new balanced sample selection algorithm that integrates both equality and inequality constraints, enhancing versatility and efficiency. The algorithm is easy to apply and can efficiently handle a range of scenarios, from category balancing with non-integer inclusion probabilities, to random matrix rounding and double stratification with non-integer probability sums.

The proposed method can also be used for spread sampling. Our simulation study, comparing this algorithm with commonly used spatial sampling methods such as LPM, WAVE and GRTS, highlights its potential as an interesting alternative. In several cases, it outperformed existing methods in terms of spatial balance and estimation accuracy. These simulations confirm the effectiveness of the algorithm in providing well-spread samples and accurate estimates, validating its broad applicability in a variety of research areas.

In conclusion, this new balanced sampling algorithm with inequality constraints offers researchers and practitioners a simple and robust tool for diverse sampling needs. Its demonstrated versatility and efficiency make it a valuable addition to research methodologies, with potential for numerous additional applications. R code is given in the supplementary material.

Acknowledgements

We are grateful to the reviewers and the associate editor for their insightful comments and constructive suggestions, which substantially improved the clarity and quality of the manuscript. The authors would like to thank the Swiss Federal Statistical Office for partially funding this

research. The conclusions and observations in this article are not necessarily those of the Swiss Federal Statistical Office.

Funding

This research was partially supported by the Swiss Federal Statistical Office. No other specific grant was received from public, commercial or not-for-profit funding agencies.

Disclosure Statement

The authors report that there are no competing financial or non-financial interests to declare

Chapter 3

Calibration and Optimal Transport Approaches for Harmonizing Survey Weights

Abstract
Originally, the construction of weighting systems to estimate parameters of interest in survey data does not depend on the variables of interest. However, recent research raises questions about the practice of using specific weighting systems for each variable of interest, thus deviating from the universal nature of the weights proposed by survey data processing methods. This article examines the challenge of harmonizing weights for variables with different weighting systems in survey data processing, by presenting and comparing two methods: one that creates a common weight for both variables using the calibration method, and one that uses optimal transport to match the variables. A simulation study is carried out to evaluate the performance of these methods in different sampling scenarios and with continuous and categorical variables. The results of the simulation study show that the optimal transport method for weight harmonization can provide accurate parameter estimates in different scenarios, particularly in situations where disparities between sample and population distributions are large. It therefore appears to be a more versatile solution.[1]

Keywords : cross-analysis, sampling, surveys, unification, weighting system.

3.1 Introduction

Deville and Särndal (1992) had a significant impact on survey data processing. They introduced a weighting system to construct an estimator that can be used for multiple variables of interest. The generalized regression estimator given in Särndal et al. (1992) is a special case of the calibration estimator. It involves a regression between a variable of interest and auxiliary variables and can be written using a weighting system that does not depend on the variable of interest. To apply calibrated estimators, the weights are calculated and added to the sample database. Then, the parameters of interest are estimated by weighting the variables.

Some generalizations of the calibration method retain this property. For example, the ridge calibration proposed by Chambers (1996), Beaumont and Bocci (2008) and Rao and Singh (2009) produce weights that do not depend on the variable of interest. However, recent work

[1]This chapter is based on: Tripet, A., & Tillé, Y., (2025). Calibration and optimal transport approaches for harmonizing survey weights, Statistical Methods & Applications, 34(2), 195–210.

has explored weighting systems that depend directly on the variable of interest. For example, Breidt and Opsomer (2000) proposed estimators assisted by a model predicted using the local polynomial method.

Other papers, such as Guggemos and Tillé (2010), Goga and Shehzad (2014) and Mayor-Gallego et al. (2019), defined penalized and partially penalized calibration estimators. McConville et al. (2017) and Chen et al. (2019) introduced regression and calibrated estimators based on the Lasso penalization. Breidt and Opsomer (2017) proposed a general approach where model prediction techniques are applied in a model-assisted approach. Dagdoug et al. (2023a,b) developed weights based on random forest predictions. Beaumont (2008) suggested modelling the weights to smooth them by using their predicted values. This trend leads to a situation where each variable of interest requires a specific weighting system.

These methods often lead to different systems of weights for different variables, which can improve efficiency but lack the universal character of the Deville and Särndal (1992) calibration method. A single set of auxiliary variables is often used to ensure a unified weighting system, seeking a compromise. The idea of using different systems of weights for different variables is usually motivated by efficiency considerations. However, a compromise or harmonized weight is necessary when estimating population parameters that involve the cross-product of many variables of interest. Indeed, it is beneficial when several variables need to be studied simultaneously and also ensures the consistency of the weighting system between variables, which promotes a comparable analytical framework. This consistency is crucial in cross-tabulations, where the aim is to examine the relationships between several variables.

Unification methods using calibration have already been proposed in this field. Wu and Sitter (2001) examine the effective use of auxiliary information in finite population estimation, proposing a unified model calibration framework. Montanari et al. (2009) perform calibration both on the values of auxiliary variables and the fitted values of variables of interest obtained from parametric or non-parametric models. Montanari and Ranalli (2005) extend model calibration using more general superpopulation models and non-parametric methods using neural network learning and local polynomial smoothing. An expansion of the calibration method is also extended to two-frame surveys, making it possible to include auxiliary information from two sources as described in Ranalli et al. (2016). Santacatterina and Bottai (2018) and Burgard et al. (2019) present methods to generalize the calibration method for obtaining consistent and efficient estimates from a variety of data sources and obtaining optimal weights, by minimizing the Euclidean distance from the target weights.

In this article, we explore scenarios in which two variables have different weighting systems and a harmonized weighting system is created so that it can be used for both variables. We compare two methods: one constructs a common weighting system for the two variables using calibration. The other uses the optimal transportation problem to match the variables. Optimal transport, derived from Wasserstein distance, quantifies the optimal way to transport one probability distribution into another while minimizing a cost function. It has applications in a variety of fields (see Villani et al., 2009). Garès et al. (2020) and Garès and Omer (2022) explored how optimal transport can be used to merge databases using covariates in different application domains, and showed the versatility of optimal transport. Jauslin and Tillé (2023) focused on statistical matching, a method for combining two statistical sources, such as samples. They presented an efficient method for matching two samples, even if they have different weighting systems.

These two methods are studied and compared in a simulation study. To evaluate their performance, samples are drawn from a population and the common weights created by each method are used to estimate the correlation coefficient, the χ^2 statistic and Cramér's V. The methods are compared to a situation where no weights are used, or to a situation where the weights of one of the two variables are used. This shows the usefulness of creating common weights.

3.2 Problem and Notation

Suppose that a random sample S is selected in a population $U = \{1, \ldots, N\}$. The first order inclusion probabilities are $\pi_k = Pr(k \in S)$. Suppose that two variables of interest, y_{1k} and y_{2k}, were each given, respectively, a particular weight, $w_{1k} \geq 0$ and $w_{2k} \geq 0$, by one of the methods described in the introduction. In order to estimate

$$Y_1 = \sum_{k \in U} y_{1k} \text{ and } Y_2 = \sum_{k \in U} y_{2k},$$

the following estimators are used

$$\widehat{Y}_1 = \sum_{k \in S} w_{1k} y_{1k} \text{ and } \widehat{Y}_2 = \sum_{k \in S} w_{2k} y_{2k}.$$

In addition, we assume that:

$$\sum_{k \in S} w_{1k} = \sum_{k \in S} w_{2k}.$$

However, some population parameters involve the product of both variables, for example the covariance

$$S_{12}^2 = \frac{1}{N-1} \sum_{k \in U} \left(y_{1k} - \overline{Y}_1 \right) \left(y_{2k} - \overline{Y}_2 \right),$$

where $\overline{Y}_1 = Y_1/N$ and $\overline{Y}_2 = Y_2/N$ or various other parameters such as correlation coefficients or regression coefficients. We can also consider the cases where $\mathbf{y}_{1k}$ and $\mathbf{y}_{2k}$ are multivariate. If variables are qualitative, all components of $\mathbf{y}_{1k}$ and $\mathbf{y}_{2k}$ are zero except for the ones indicating the categories to which the unit k belongs. The following contingency table can be then defined :

$$\mathbf{T} = \sum_{k \in U} \mathbf{y}_{1k} \mathbf{y}_{2k}^{\top}.$$

In our case, the challenge is to estimate $\mathbf{T}$ while respecting the estimators of the table margins which can be estimated separately :

$$\widehat{\mathbf{Y}}_1 = \sum_{k \in S} w_{1k} \mathbf{y}_{1k} \text{ and } \widehat{\mathbf{Y}}_2 = \sum_{k \in S} w_{2k} \mathbf{y}_{2k}.$$

Derived from the estimation of $\mathbf{T}$, we can estimate the χ^2 statistic and Cramér's V, that provide insights of the strength and significance of associations between the categorical variables.

The χ^2 chi-square statistic is used to test the independence of two categorical variables. It measures the difference between the observed numbers and the numbers expected if the variables were independent, and is defined by:

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} is the observed frequency in each category, and E_{ij} is the expected frequency in each category under the assumption of independence. The expected frequency E_{ij} is defined by:

$$E_{ij} = \frac{R_i \times C_j}{N}$$

where R_i is the sum in rows and C_j is the sum in columns. Cramér's V is a measure of association between two categorical variables and is based on the χ^2 statistic. The formula for Cramér's V is:

$$V = \sqrt{\frac{\chi^2}{N \ \min(I-1, J-1)}}$$

where χ^2 is the chi-squared statistic, N is the total number of observations, I is the number of rows and J is the number of columns. Cramér's V ranges from 0 to 1. A value close to 0 indicates a lack of association between the variables, while a value close to 1 indicates a strong association.

3.3 Calibration approach

A first solution consists of creating a common weighting system for both variables. Indeed, if we compute the geometric mean of the weights $\sqrt{w_{1k}w_{2k}}$, we can search weights $\widetilde{w}_k \geq 0$ close to this geometric mean satisfying the constraints given by the two variables:

$$\sum_{k\in S}\widetilde{w}_k y_{1k} = \sum_{k\in S} w_{1k}y_{1k}, \sum_{k\in S}\widetilde{w}_k y_{2k} = \sum_{k\in S} w_{2k}y_{2k} \text{ and } \sum_{k\in S}\widetilde{w}_k = \sum_{k\in S} w_{1k} = \sum_{k\in S} w_{2k}. \tag{3.1}$$

We can proceed by minimizing

$$\sum_{k\in S} G(\widetilde{w}_k, \sqrt{w_{1k}w_{2k}}),$$

where $G(.,.)$ is the distance proposed by Deville and Särndal (1992) that minimize $\widetilde{w}_k$ subject to the constraints given in (3.1), for example, the Kullback-Leibler divergence:

$$G(\widetilde{w}_k, \sqrt{w_{2k}w_{1k}}) = \sum_{k\in S}\widetilde{w}_k \log\frac{\widetilde{w}_k}{\sqrt{w_{1k}w_{2k}}}.$$

Choosing the geometric mean as the reference weight offers a natural compromise between the two initial weight systems. The advantage of the geometric mean is that if one of the two weights is zero, then the geometric mean of these two weights is also zero. If an individual did not respond to one of the two surveys, they can simply be given a zero weight. In this case, the average weight will also be zero.

However, other choices are possible. For example, the design weights could be used as a reference, but this could result in a less balanced system initially if the two weight systems, w_{1k} and w_{2k}, differ strongly. The arithmetic mean, $(w_{1k} + w_{2k})/2$, could also be considered as an alternative, although it lacks the multiplicative properties of the geometric mean. For these reasons, the geometric mean was chosen as the best compromise. Nevertheless, a more detailed exploration of other approaches could be considered in future work.

If $\mathbf{y}_{2k}$ and $\mathbf{y}_{1k}$ are vectors, then we calibrate respectively on

$$\sum_{k\in S}\widetilde{w}_k \mathbf{y}_{2k} = \sum_{k\in S} w_{1k}\mathbf{y}_{2k}, \sum_{k\in S}\widetilde{w}_k \mathbf{y}_{2k} = \sum_{k\in S} w_{2k}\mathbf{y}_{2k} \text{ and } \sum_{k\in S}\widetilde{w}_k = \sum_{k\in S} w_{2k} = \sum_{k\in S} w_{1k}.$$

If these variables are categorical, then the estimated table

$$\widehat{\mathbf{T}} = \sum_{k\in S}\widetilde{w}_k \mathbf{y}_{1k}\mathbf{y}_{2k}^\top$$

has the marginals totals equal to

$$\widehat{\mathbf{Y}}_1 = \sum_{k\in S} w_{1k}\mathbf{y}_{1k} \text{ and } \widehat{\mathbf{Y}}_2 = \sum_{k\in S} w_{2k}\mathbf{y}_{2k}.$$

The estimation of the cells in the table is therefore consistent with the marginal distributions which can each be estimated separately. Regarding the correlation coefficient ρ, a problem of consistency in the estimation of the variance is observed depending on whether they are estimated using the original weights or the calibrated weights. Indeed if

$$\rho = \frac{S_{12}}{S_1 S_2},$$

where

$$S_{12} = \frac{1}{N-1}\sum_{k\in U}(y_{2k} - \overline{Y}_2)(y_{2k} - \overline{Y}_1),$$

$$S_2^2 = \frac{1}{N-1}\sum_{k\in U}(y_{2k} - \overline{Y}_2)^2, \qquad S_1^2 = \frac{1}{N-1}\sum_{k\in U}(y_{1k} - \overline{Y}_1)^2,$$

$$\overline{Y}_1 = \frac{1}{N}\sum_{k\in U} y_{1k} \qquad \text{and } v\overline{Y}_2 = \frac{1}{N}\sum_{k\in U} y_{2k}.$$

We can estimate

$$\widehat{\widehat{N}} = \sum_{k\in S} w_{1k} = \sum_{k\in S} \widetilde{w}_k = \sum_{k\in S} w_{2k},$$

$$\widehat{\overline{Y}}_1 = \frac{\sum_{k\in S} w_{1k}y_{1k}}{\sum_{k\in S} w_k} = \frac{\sum_{k\in S} \widetilde{w}_k y_{1k}}{\sum_{k\in S} \widetilde{w}_k}, \widehat{\overline{Y}}_2 = \frac{\sum_{k\in S} w_{2k}y_{2k}}{\sum_{k\in S} w_{2k}} = \frac{\sum_{k\in S} \widetilde{w}_k y_{2k}}{\sum_{k\in S} \widetilde{w}_k},$$

$$\widehat{S}_{y_1y_2} = \frac{1}{\widehat{N}-1}\sum_{k\in S} \widetilde{w}_k(y_{2k} - \widehat{\overline{Y}}_2)(y_{1k} - \widehat{\overline{Y}_1}),$$

$$\widehat{S}_1^2 = \frac{1}{\widehat{N}-1}\sum_{k\in S} \widetilde{w}_k(y_{1k} - \widehat{\overline{Y}}_1)^2, \widehat{S}_2^2 = \frac{1}{\widehat{N}-1}\sum_{k\in S} \widetilde{w}_k(y_{2k} - \widehat{\overline{Y}}_2)^2.$$

Nevertheless,

$$\frac{1}{\widehat{N}-1}\sum_{k\in S} \widetilde{w}_k(y_{1k} - \widehat{\overline{Y}}_1)^2 \neq \frac{1}{\widehat{N}-1}\sum_{k\in S} w_{1k}(y_{1k} - \widehat{\overline{Y}}_1)^2,$$

and

$$\frac{1}{\widehat{N}-1}\sum_{k\in S} \widetilde{w}_k(y_{2k} - \widehat{\overline{Y}}_2)^2 \neq \frac{1}{\widehat{N}-1}\sum_{k\in S} w_{2k}(y_{2k} - \widehat{\overline{Y}}_2)^2.$$

However, the problem can be solved by also calibrating the $\widetilde{w}_k$ weights on the constraints

$$\sum_{k\in S} \widetilde{w}_k y_{1k}^2 = \sum_{k\in S} w_k y_{1k}^2, \text{ and } \sum_{k\in S} \widetilde{w}_k y_{2k}^2 = \sum_{k\in S} w_{2k} y_{2k}^2.$$

It is possible to generalize this method to more than two variables. However, the method has limitations, as each new variable implies the addition of a calibration constraint.

3.4 Optimal Transport approach

A completely different method is to use the transportation problem to match variables. A distance or cost function is defined to measure the cost of transporting a unit from one point to another. Usually the Euclidean distance or its variations are used. In our case, we are looking for units that are close enough to be matched, the distance between all pairs of units is defined, for example, the square of the following distance:

$$d^2(k,\ell) = \frac{(y_{1k} - y_{1,\ell})^2}{\widehat{S}_1^2} + \frac{(y_{2k} - y_{2,\ell})^2}{\widehat{S}_2^2}, k, \ell \in S,$$

where $\widehat{S}_1^2$ and $\widehat{S}_2^2$ are the variance estimation of the different variables of interest defined below in Expressions (3.2) and (3.3).

We then look for the weights $\widetilde{w}_{k\ell} \geq 0$ that minimize

$$\sum_{k\in S}\sum_{\ell\in S} \widetilde{w}_{k\ell} d(k,\ell)$$

subject to

$$\sum_{\ell\in S} \widetilde{w}_{k\ell} = w_{1k} \text{ and } \sum_{k\in S} \widetilde{w}_{k\ell} = w_{2,\ell}.$$

The aim is to minimize the total cost of transport, while ensuring that the total weight assigned to each unit corresponds to the original weights.

The weights $\widetilde{w}_{k\ell}$ enable us to estimate the parameters

$$\widehat{Y}_1 = \sum_{k\in S}\sum_{\ell\in S}\widetilde{w}_{k\ell}y_{1k} = \sum_{k\in S} w_{1k}y_{1k}, \widehat{Y}_2 = \sum_{k\in S}\sum_{\ell\in S}\widetilde{w}_{k\ell}y_{2k} = \sum_{\ell\in S} w_{2,\ell}y_{2,\ell},$$

$$\widehat{N} = \sum_{k\in S}\sum_{\ell\in S}\widetilde{w}_{k\ell} = \sum_{k\in S} w_{1k} = \sum_{\ell\in S} w_{2,\ell},$$

$$\widehat{\overline{Y}}_1 = \frac{\widehat{Y}_1}{\widehat{N}} \text{ and } \widehat{\overline{Y}}_2 = \frac{\widehat{Y}_2}{\widehat{N}}.$$

Covariance and variances can be estimated by

$$\widehat{S}_{y_2y_1} = \frac{1}{\widehat{N}-1}\sum_{k\in S}\sum_{\ell\in S}\widetilde{w}_{k\ell}(y_{1k}-\widehat{\overline{Y}}_1)(y_{2,\ell}-\widehat{\overline{Y}}_2),$$

$$\widehat{S}_1^2 = \frac{1}{\widehat{N}-1}\sum_{k\in S}\sum_{\ell\in S}\widetilde{w}_{k\ell}(y_{1k}-\widehat{\overline{Y}}_1)^2 = \frac{1}{\widehat{N}-1}\sum_{k\in S} w_{1k}(y_{1k}-\widehat{\overline{Y}}_1)^2, \tag{3.2}$$

$$\widehat{S}_2^2 = \frac{1}{\widehat{N}-1}\sum_{k\in S}\sum_{\ell\in S}\widetilde{w}_{k\ell}(y_{2k}-\widehat{\overline{Y}}_2)^2 = \frac{1}{\widehat{N}-1}\sum_{k\in S} w_{2k}(y_{2k}-\widehat{\overline{Y}}_2)^2. \tag{3.3}$$

Therefore, by using the general weights $\widetilde{w}_{k\ell}$ we obtain the marginal variances. The same applies to estimating a contingency table. By using

$$\sum_{k\in S}\sum_{\ell\in S}\widetilde{w}_{k\ell}\mathbf{y}_{1k}\mathbf{y}_{\ell}^{(2)\top}$$

we obtain a table with the same marginal distributions as those computed separately for each variable. Indeed, we have

$$\sum_{k\in S}\sum_{\ell\in S}\widetilde{w}_{k\ell}\mathbf{y}_{1k} = \sum_{k\in S} w_{1k}\mathbf{y}_{1k} \text{ and } \sum_{k\in S}\sum_{\ell\in S}\widetilde{w}_{k\ell}\mathbf{y}_{2k} = \sum_{\ell\in S} w_{2,\ell}\mathbf{y}_{2,\ell}.$$

Regarding the estimated correlation coefficient

$$\widehat{\rho} = \frac{\widehat{S}_{y_2y_1}}{\widehat{S}_2\widehat{S}_1},$$

it may pose a problem in the case where $y_{1k} = y_{2k}$ as its value will not necessarily be 1. However, if the variables of interest are the same, the weights should presumably be identical and so the problem would no longer exist. Another difficulty is that the method is difficult to generalise to a large number of variables. Generalization is theoretically possible, but it faces a combinatorial explosion problem.

3.5 Simulation Study

A simulation study is conducted to evaluates the performance of the two weight harmonization methods presented in Sections 3.3 and 3.4 under different scenarios. The study involves generating a synthetic population, constructing variables, designing two sampling scenarios (non-informative and informative), and applying Poisson sampling to compute weights.

We generated a population of size $N = 1,000$ consisting of two cases:

- Case 1: Two categorical variables, y_1 and y_2.
- Case 2: Two continuous variables, y_3 and y_4.

Two auxilliary variables, x_1 and x_2, are generated following a normal distribution $\mathcal{N}(0,3)$. A hidden variable $z \sim \mathcal{N}(0,4)$ is also generated in order to create a link between the variables of interest unrelated to x_1 and x_2. Noise variables e_1 and e_2 are generated, each following $\mathcal{N}(0,2)$.

Categorical variables are constructed by dividing continuous variables into terciles, resulting in categories of equivalent size:

$$y_1 = \begin{cases} 1 & \text{if } \frac{2x_1}{5} + z + e_1 \leq \mathcal{Q}(p = \frac{1}{3}), \\ 3 & \text{if } \frac{2x_1}{5} + z + e_1 \geq \mathcal{Q}(p = \frac{2}{3}), \\ 2 & \text{otherwise} \end{cases}$$

and

$$y_2 = \begin{cases} 1 & \text{if } \frac{2x_2}{3} + \left(\frac{z}{2}\right)^2 + e_2 \leq \mathcal{Q}(p = \frac{1}{3}), \\ 3 & \text{if } \frac{2x_2}{3} + \left(\frac{z}{2}\right)^2 + e_2 \geq \mathcal{Q}(p = \frac{2}{3}), \\ 2 & \text{otherwise.} \end{cases}$$

Continuous variables are linear or non-linear transformations of Z:

$$y_3 = \frac{2x_1}{5} + z + e_1,$$

$$y_4 = \frac{2x_2}{3} + \left(\frac{z}{2}\right)^2 + e_2.$$

To understand the rationale behind the construction of the variables, the study experimented with different generations of artificial datasets. We observed that highly correlated auxiliary variables made the weights w_1 and w_2 too similar, making the methods less meaningful. Conversely, weakly correlated variables of interest also reduced the significance of the methods. To solve this problem, correlated variables of interest were created while keeping auxiliary variables x uncorrelated by introducing the common hidden variable z. Squaring z in the construction of y_2 and y_4 introduced non-linear relationships within the variable pairs of interest (y_1, y_2) and (y_3, y_4), balancing a significant correlation of the variables of interest while maintaining the variety of the auxiliary variables. This setup ensures significant correlation between the variables of interest, while maintaining uncorrelated auxiliary variables x_1 and x_2.

Our simulation study focuses on two distinct scenarios in order to assess the accuracy of the methods:

- Scenario 1 (or the non-informative sample scenario): In this scenario, we draw a sample whose distribution closely matches that of the original population data. In essence, this sample is relatively similar to the population.
- Scenario 2 (or the informative sample scenario): Conversely, the second scenario involves drawing a sample that is deliberately different from the original population data. This sample is intentionally distinct, making it dissimilar to the population.

The aim of these two scenarios is to provide a solid evaluation of the methods under a variety of conditions. By applying the methods separately to groups of categorical and continuous variables in each case, we can evaluate their performance on different data types and scenarios. Samples of size $n \simeq 100$ are drawn using Poisson sampling under each scenarios. In Scenario 1, the first-order inclusion probabilities are computed using a sampling design proportional to:

$$\sqrt{(x_1 - \min(x_1) + 1)(x_2 - \min(x_2) + 1)}$$

for continuous variables and to

$$\sqrt{|\ x_1 x_2\ |}$$

for categorical variables. The sampling design depends on x_1 and x_2, which allows us to generate a sample that is relatively similar to the population. In Scenario 2, the sampling design depends on z, a variable common to y_1 and y_1 and independent of x_1 and x_2, making it possible to generate a sample that is distinctly different from the population. The first-order inclusion probabilities are computed proportional to

$$| 1/z |$$

for categorical variables and to

$$(z - \min(z) + 1)^3$$

for continuous variables. Figure 3.1 and Figure 3.2 compare the population and sample distributions for both scenarios, highlighting the differences introduced by the sampling designs. To align with the approach used for creating variables of interest, unequal inclusion probabilities were introduced, and samples were drawn using Poisson sampling. This decision aims to generate even more varied weights, accentuated by slight variations in sample size. Weights of w_1 and w_2 are generated by calibrating the basic sampling design weights $d_k = 1/\pi_k$ on the population totals of x_1 and x_2 and on N.

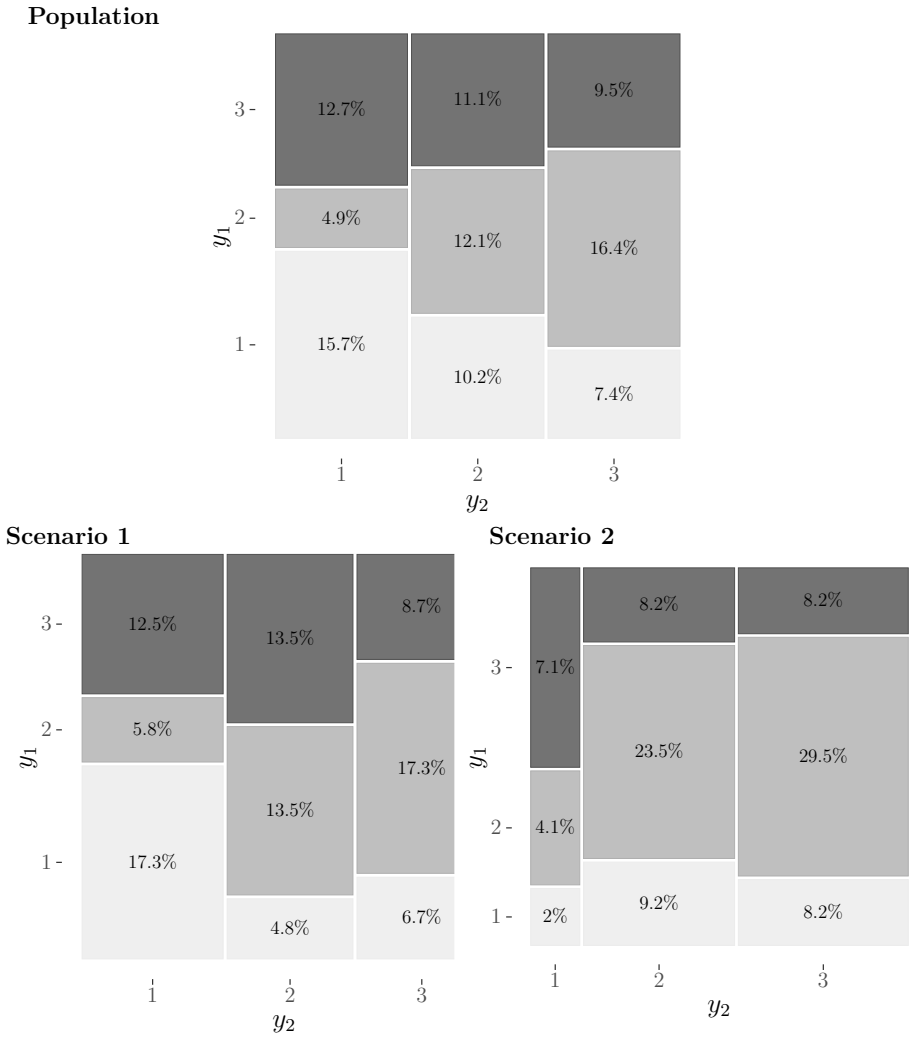

Figure 3.1: Mosaic plot of y_1 and y_2 for the population and for two samples drawn according to the criteria of scenario 1 and scenario 2, corresponding to Case 1 with categorical variables.

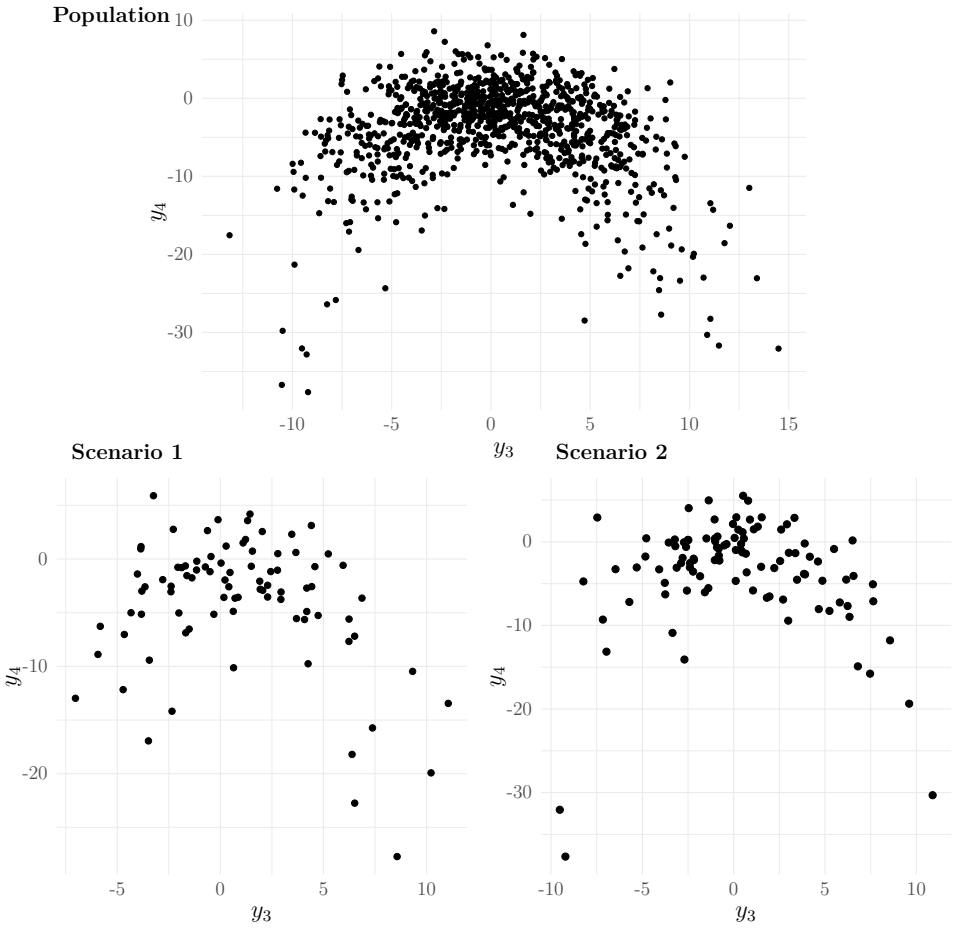

Figure 3.2: Scatter plot of y_3 and y_4 for the population and for 2 samples drawn according to the criteria of scenario 1 and scenario 2, corresponding to Case 2 with continuous variables.

To evaluate the accuracy of the methods for the categorical variables (y_1 and y_2), we estimate the χ^2 statistic, which measures the association between the categorical variables. In addition, we compute Cramér's V. For the continuous variables (y_3 and y_4), we estimate the correlation coefficient between the two continuous variables.

The two harmonized weights are compared to the weights of each variable and to situations where either w_1, w_2 or no weights (NW) are taken into account. We repeat the estimation on $M = 10,000$ samples. We then compute the bias

$$\widehat{\text{Bias}}_{MC}\left(\widehat{\theta}\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_m - \theta\right),$$

the variance

$$\widehat{\text{Var}}_{MC}\left(\widehat{\theta}\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_m - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\theta}_m\right)^2$$

and the mean squared error (MSE)

$$\widehat{\text{MSE}}_{MC}\left(\widehat{\theta}\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_m - \theta\right)^2,$$

between these estimators and the true value calculated on the population, where θ is the parameter to be estimated and $\widehat{\theta}_m$ is the m-th estimate of this parameter.

Table 3.1: Bias, variance and MSE of the estimated χ^2 statistic between y_1 and y_2 using harmonized weights informative scenario

Scenario	Non-Informative		Informative	
	χ^2	Cramér's V	χ^2	Cramér's V
True Value	97.482	0.221	97.482	0.221
Bias				
NW	-83.435	0.039	-90.486	-0.042
w_1	38.503	0.033	94.352	0.077
w_2	38.528	0.033	93.942	0.077
Calibration	38.518	0.033	93.502	0.076
Optimal Matching	34.607	0.030	82.016	0.067
Var				
NW	34.626	0.003	18.028	0.003
w_1	3763.264	0.004	10630.136	0.007
w_2	3761.851	0.004	10782.033	0.007
Calibration	3771.985	0.004	10635.197	0.007
Optimal Matching	3462.794	0.003	9422.303	0.007
MSE				
NW	6995.971	0.005	8205.658	0.005
w_1	5245.746	0.005	19532.497	0.013
w_2	5246.262	0.005	19607.179	0.013
Calibration	5255.643	0.005	19377.749	0.013
Optimal Matching	4660.423	0.004	16148.877	0.011

3.6 Results

The results of simulations evaluating categorical variables for the different scenarios can be found in Table 3.1. Simulation results for continuous variables can be found in Table 3.2. In Table 3.1 evaluating the chi-square statistic and Cramér's V, as expected the "informative scenario" poses greater problems for all estimation methods than the "non-informative scenario", with a much larger variance and MSE. Optimal matching consistently outperforms the other methods in both scenarios, with equivalent or lower bias, variance and MSE. In the "informative scenario", the case without weighting struggles with a high bias and MSE, but retains a lower variance. Calibration, while performing better in terms of bias in the "non-informative scenario", fails to manage variance and MSE, particularly in the "informative scenario".

The tables evaluating correlation estimation reveal a similar trend. Optimal matching maintains its dominance in both scenarios, displaying lower bias, variance and MSE than the other methods, except for non-informative scenario bias. In the "informative scenario", estimation without weighting weights faces a significant bias, while Calibration has a higher bias and MSE. Estimates using only w_1 and w_2 as weights are also in trouble in both scenarios, with mixed results.

When comparing estimation methods in each scenario, optimal matching consistently outperforms calibration as well as alternative methods in all tables. It offers a solution that appears more robust for more accurate estimation, showing a balanced compromise between bias, variance and MSE.

If we compare the "informative scenario" and the "non-informative scenario", it is clear that the "informative case" scenario poses greater problems for all estimation methods. However, despite these challenges, optimal matching maintains its superior performance in both scenarios.

Table 3.2: Bias, variance and MSE of the estimated correlation between y_3 and y_4 using harmonized weights non-informative scenario and informative scenario

Scenario	Non-Informative	Informative
	ρ	ρ
True Value	-0.105	-0.105
Bias		
NW	0.018	-0.510
w_1	-0.003	-0.178
w_2	-0.002	-0.181
Calibration	-0.003	-0.190
Optimal Matching	-0.001	-0.180
Var		
NW	0.022	0.004
w_1	0.032	0.048
w_2	0.032	0.049
Calibration	0.032	0.046
Optimal Matching	0.031	0.045
MSE		
NW	0.022	0.264
w_1	0.032	0.081
w_2	0.032	0.081
Calibration	0.032	0.082
Optimal Matching	0.031	0.077

To conclude this simulation and case study, optimal matching emerges as the most reliable and versatile estimation method in the various cases and scenarios proposed, offering consistent accuracy and demonstrating its ability to deal effectively with informative scenarios.

3.7 Discussion

In the context of cross-analysis, adopting a harmonized weighting system is advantageous. Instead of assigning separate weights to each variable of interest, a unified system simplifies the analysis process and ensures consistent, fair comparisons.

The simulation study shows that the optimal transport method for weight harmonization appears to perform best when dealing with significant differences between sample and population distributions. It also remains effective in standard scenarios for continuous and categorical variables. This method is particularly advantageous in real cases where actual correlations are unknown, outperforming or at least equalling alternative methods. The calibration method also produces encouraging results, but the method is highly dependent on the weights of the variables.

The promising potential lies in extending this method to those involving missing values, which could be valuable avenues for further exploration. For example, the nonresponse issue could be handled by assigning a zero weight for nonresponding units in the calibration step approach. Another issue is the generalization of the method to more than two variables. While the calibration method does not seem to pose any problems in this context, the optimal transport method seems more complex and would require some thought. In particular, the calculation of the distance to be used. Additionally, the method is flexible enough to be applied to specific

subsets, such as sub-domains or strata.

In the future, it is possible to apply the optimal transport method to different correlation estimators or datasets to discover its effectiveness in different contexts. These results collectively underline the versatility of the method, making it a compelling option for various scenarios and justifying further research.

One reviewer pointed out that the method could be of significant interest to the field of official statistics. By examining different scenarios and focusing on the integrated use of two or more surveys, we can provide estimates for one or more target variables. Combining records from multiple surveys with the definition of a single weight could improve the efficiency of estimates and enhance the coherence of the overall system of official statistics.

Data Availability Statement

In order to evaluate the method for various configurations, the data was generated according to the models described in the article. Programs in R language are available on request.

Chapter 4

Calibration with Bagging of the Principal Components on a Large Number of Auxiliary Variables

Abstract
Calibration is a widely used method in survey sampling to adjust weights so that estimated totals of some chosen calibration variables match known population totals or totals obtained from other sources. When a large number of auxiliary variables are included as calibration variables, the variance of the total estimator can increase, and the calibration weights can become highly dispersed. To address these issues, we propose a solution inspired by bagging and principal component decomposition. With our approach, the principal components of the auxiliary variables are constructed. Several samples of calibration variables are selected without replacement and with unequal probabilities from among the principal components. For each sample, a system of weights is obtained. The final weights are the average weights of these different weighting systems. With our proposed method, it is possible to calibrate exactly for some of the main auxiliary variables. For the other auxiliary variables, the weights cannot be calibrated exactly. The proposed method allows us to obtain a total estimator whose variance does not explode when new auxiliary variables are added and to obtain very low scatter weights. Finally, our proposed method allows us to obtain a single weighting system that can be applied to several variables of interest of a survey. We evaluate the proposed total using a simulation study on real survey data from the Swiss Survey on Income and Living Conditions. The results show that the proposed solution significantly reduces the weight variability and the variance of the total estimator compared to competing total estimators for some variables of interest. [1]

Keywords : calibration variables selection, estimation, high dimension, weight system, weighting.

4.1 Introduction

Calibration, as introduced by Deville and Särndal (1992), is a widely used method in survey sampling that adjusts the sampling weights so that the estimated totals of auxiliary variables match known population totals. The estimator produced by these adjusted weights is referred

[1] This chapter is based on the article: Hasler, C., Tripet, A., & Tillé, Y. (2025). Calibration with Bagging of the Principal Components on a Large Number of Auxiliary Variables. Submitted for publication

to as the calibrated estimator. The main objectives of calibration are twofold: (1) to ensure consistency between survey estimated totals and known totals or totals derived from external sources, and (2) to reduce the variance of the total estimator compared to the Horvitz-Thompson estimator (Horvitz and Thompson, 1952).

With the increasing availability of large-scale datasets, calibration is often performed on a high number of auxiliary variables, an approach referred to as high-dimensional calibration. In this context, however, calibration faces two major issues. First, the reduction of the variance of the total estimator can no longer be achieved. Indeed, as noted by Silva and Skinner (1997), the Mean Squared Error (MSE) of the calibrated estimator initially decreases as more auxiliary variables are added. However, beyond a certain point, it begins to increase. Second, as highlighted by Haziza and Beaumont (2017), a large number of auxiliary variables often leads to highly dispersed calibration weights. This can result in unstable estimates for variables that are weakly correlated with the calibration variables.

Several strategies have been proposed to address these two issues. One common approach consists of selecting a subset of auxiliary variables that minimizes the estimated MSE. This can be done through best possible subset selection or forward selection procedures (Silva and Skinner, 1997; Chauvet and Goga, 2022), by discarding some calibration constraints (Bankier et al., 1992), or by assessing the contribution of each auxiliary variable via the Shapley decomposition and retaining only the most impactful variables (Guandalini and Ceccarelli, 2022).

An alternative to variable selection is to relax some calibration constraints. This leads to so-called soft calibration. In this idea, Guggemos and Tillé (2010) use soft calibration with mixed models. Other methods incorporate penalization, such as ridge regression approaches proposed by Beaumont and Bocci (2008), Rao and Singh (1997), and extended by Montanari et al. (2009). Burgard et al. (2019) propose a generalized calibration method that remains applicable even in high-dimensional settings using soft calibration and box-constraints. Williams and Savitsky (2024) propose range-restricted soft calibration independent of any particular variable of interest. Methods based on empirical likelihood have also been suggested to obtain range-restricted weights while relaxing benchmark constraints (Chen et al., 2002), or by explicitly incorporating constraint deviations into the objective function (Fetter et al., 2005).

Another approach, suggested by Wu and Sitter (2001), uses model calibration, where the calibration variables are the predicted values of the variables of interest obtained from a working model. Finally, some authors propose reducing the dimension of the calibration problem through principal component analysis (Cardot et al., 2017). In this case, calibration is performed not on the raw auxiliary variables but on a reduced number of principal components.

In this article, we propose a new calibration approach specifically designed for settings involving a large number of auxiliary variables. The method combines the bagging technique, introduced by Breiman (1994), with a preliminary dimension reduction step using principal component analysis (Pearson, 1901; Hotelling, 1936). We select a large number of samples of calibration variables without replacement and with unequal probabilities from among the principal components. We perform calibration on each sample of principal component. We obtain as many system of weights as the number of selected samples. The final weights are the average weights of these different systems of weights. Our approach offers several advantages. First, it reduces the instability of calibration weights generally observed in high-dimensional settings. Second, it reduces the risk of increased variance often associated with the inclusion of many auxiliary variables. Finally, as calibration is performed independently of any variable of interest, the resulting weights can be used for different estimation purposes or different variables of interest.

The article is organized as follows: in Section 4.2 we introduce the notation and review the different methods on which our approach is based. In Section 4.3 we present the proposed calibration method in detail. In Section 4.4 we illustrate the empirical performance of our methods through a simulation study based on real data. We assess the accuracy of the total

estimators and the dispersion of the weights. In Section 4.5 we conclude with a discussion of the main findings.

4.2 Framework

4.2.1 Notation

We consider a finite population U of size N, and denote by k a generic unit with $k \in \{1, \ldots, N\}$. Let S be a random sample of size n selected from U with a sampling design that assigns to unit k an inclusion probability of appearing in the sample $\pi_k = \Pr(k \in S)$. We assume that $\pi_k > 0$ for all $k \in U$ and denote by $d_k = 1/\pi_k$ the design weight of unit k. Let y_k be the value of a variable of interest y for unit $k \in U$. Value y_k is only observed for units contained in the sample. The aim is to estimate the population total

$$t_y = \sum_{k \in U} y_k$$

of variable of interest y. With no additional information, the total t_y can be estimated by the *expansion estimator* or Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952)

$$\widehat{t}_{yd} = \sum_{k \in S} d_k y_k.$$

The HT estimator is design-unbiased, meaning unbiased under the sampling design, provided that $\pi_k > 0$ for all $k \in U$. We assume the variable of interest is univariate for simplicity. However, most surveys aim to collect information on several variables of interest. The methodology presented in this article extends naturally to that case.

4.2.2 Calibration

Consider a vector $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kq})^\top$ of q auxiliary variables for each unit $k \in U$. Let $\mathbf{X} \in \mathbb{R}^{N \times q}$ be the data matrix obtained by stacking the $\mathbf{x}_k$, $k \in U$. That is, each row corresponds to a unit and each column to a variable. Without loss of generality, we will suppose in what follows that the columns of $\mathbf{X}$ are centered and scaled (they have mean 0 and variance 1). Denote by $\mathbf{t}_\mathbf{x} = \sum_{k \in U} \mathbf{x}_k$ the population total of these auxiliary variables. Suppose that $\mathbf{x}_k$ is available for all population units $k \in U$. We will relax this assumption in Section 4.3, and suppose instead that $\mathbf{x}_k$ is available only for sampled units $k \in S$, while the population total $\mathbf{t}_\mathbf{x}$ remains known.

Calibration, first introduced by Deville and Särndal (1992), allows the auxiliary information to be efficiently used to improve the HT estimator. It consists in modifying the initial design weights $d_k = 1/\pi_k$ into new weights w_k that are as close as possible, in an average sense for a given metric, to the initial design weights, while satisfying the calibration equation

$$\mathbf{t}_\mathbf{x} = \sum_{k \in S} w_k \mathbf{x}_k.$$

This constraint ensures that the weighted sum of the auxiliary variables in the sample matches their known population totals, thereby increasing the coherence and efficiency of the estimator.

The weights can be written as $w_k = d_k g_k = g_k/\pi_k$, where g_k solves

$$\sum_{k \in U} \mathbf{x}_k = \sum_{k \in S} \frac{g_k}{\pi_k} \mathbf{x}_k.$$

A common and convenient choice for the calibration distance is the chi-squared distance, defined as

$$\Phi_S(\mathbf{w}) = \sum_{k \in S} \frac{(w_k - d_k)^2}{q_k d_k},$$

where $1/q_k$ is an individual weight associated with each unit $k \in S$, allowing more or less importance to be assigned to certain units. When there is no reason to privilege any unit, the weights are set to $q_k = 1$ for all k (Deville and Särndal, 1992). Given this criterion, the calibration weights $w_k = g_k/\pi_k = d_k g_k$ are defined as the solution to the minimization problem

$$\mathbf{w} = \arg\min_{\mathbf{w}} \Phi_S(\mathbf{w}),$$

subject to the calibration equation above.

While calibration may introduce a small bias compared to the Horvitz-Thompson estimator, it often leads to a significant reduction in variance, particularly when the auxiliary variables $\mathbf{x}$ are strongly correlated with the variable of interest y. This tradeoff typically results in improved mean squared error (MSE). However, when the number of auxiliary variables becomes large, calibration may lead to instability. Indeed, weights can become extremely variable and the variance of the estimator may increase instead of decreasing.

4.2.3 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction method introduced by Pearson (1901) and later extended by Hotelling (1936). It is mainly used to reduce the dimension of a dataset while preserving as much of its variability as possible. It transforms a dataset with a large set of correlated variables into a new set of uncorrelated variables called principal components, so that a smaller number of variables captures the greater part possible of the variability in the initial variables.

Let $\mathbf{X} \in \mathbb{R}^{N \times q}$ be the data matrix, where each row corresponds to a unit and each column to a variable. The goal of PCA is to find an orthonormal basis $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_q)$ such that the first few of the transformed variables, called principal components, in the columns of $\mathbf{Z} = \mathbf{XV}$ capture the largest portion possible of the variance in the initial variables $\mathbf{X}$. The vectors $\mathbf{v}_j$ are solutions to the eigenvalue problem

$$\mathbf{\Sigma}\mathbf{v}_j = \lambda_j \mathbf{v}_j, \quad j = 1, \ldots, q,$$

where $\mathbf{\Sigma}$ is the population covariance matrix of $\mathbf{X}$

$$\mathbf{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

The eigenvalues λ_j represent the portion of the variance in the initial data matrix explained by principal component j defined as

$$\mathbf{Z}_j = \mathbf{X}\mathbf{v}_j.$$

The first c principal components represents a projection of the initial data onto a lower-dimensional space. This transformation achieves a dimensionality reduction in the sense that it retains the maximum possible variance in a small number of components.

4.2.4 Bagging

Bootstrap Aggregating, or Bagging, was introduced by Breiman (1996) as a variance reduction technique to improve stability of estimators, particularly in high-dimensional settings. The method is based on bootstrapping, a resampling method. It generates multiple training samples by resampling with replacement from the initial sample. Separate models are trained on these bootstrap samples and their predictions are aggregated to create a final estimator.

Let $\widehat{\theta}$ be a base estimator trained on a dataset. Bagging generates B bootstrap samples of observations drawn with replacement from the dataset. An estimate $\widehat{\theta}^{(b)}$ is obtained for each sample. The final bagging estimator is

$$\widehat{\theta}_{\text{bag}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{(b)}.$$

By averaging multiple estimators, bagging reduces variance and overfitting.

4.3 Decomposition Into Main Components and Bagging-Inspired Calibration

When the number of auxiliary variables is large, traditional calibration methods can lead to highly scattered weights and increased variance of the total estimator. To address this, we propose a bagging-inspired approach combined with principal component decomposition to improve stability and efficiency in high-dimensional contexts. We propose to select a large number of samples of calibration variables without replacement and with unequal probabilities among the principal components of the auxiliary variables. We perform calibration on each sample of principal components. In the remainder of this section, we present our proposed method, discuss the choice of the parameters involved in the method, list some advantages and limitations of the method, and show that the resulting calibration estimator can be written as a model-assisted.

In the current context, using terms bootstrap and bagging is somewhat inaccurate. Indeed, bootstrapping traditionally refers to a procedure by which subsamples of units are selected with replacement from an initial sample. Bagging refers to aggregating estimates coming from several bootstrap samples. In the current context, we select samples of principal components, i.e. of variables, without replacement and we aggregate the weights coming from the different samples. We will still refer to these procedures to as bootstrapping and bagging for simplicity.

4.3.1 Calibration via Bagging on Principal Components

In order to obtain final weights, we start by computing the principal components of the centered and scaled matrix of auxiliary variables $\mathbf{X}$. We obtain the principal components $\mathbf{Z}_j = \mathbf{X}\mathbf{v}_j$ and eigenvalues λ_j, $j = 1, \ldots, q$, as described in Section 4.2.3. Denote by $\mathbf{z}_k$ the values taken by the q principal components for unit k. Then, we apply the bagging (our own version of it) a large number of times B as follows. At each iteration $b = 1, \ldots, B$, we select a subset of size c without replacement among these components with unequal probabilities proportional to $(\lambda_j / \sum \lambda_j)^{\alpha}$. Parameter c sets the number of components selected at each iteration, α controls the contrast in the inclusion probabilities of the components. In Section 4.3.2, we discuss these quantities and suggest guideline for their choice. Let $\mathbf{z}_k^{(b)}$ denote the values taken by the selected components for unit k. Calibration weights $w_k^{(b)}$ are computed by solving the calibration equation

$$\sum_{k\in S} w_k^{(b)}\mathbf{z}_k^{(b)} = \sum_{k\in U}\mathbf{z}_k^{(b)},$$

using a distance function such as the chi-square distance, see Section 4.2.2. At the end of the B iterations, we have B systems of weights $w_k^{(b)}$, $b = 1, \ldots, B$. The final weights are obtained by averaging these B systems of weights. That is,

$$\widehat{w}_k = \frac{1}{B}\sum_{b=1}^{B} w_k^{(b)}.$$

The full procedure is detailed in Algorithm 3. For any variable of interest y, the final estimator is defined as

$$\widehat{t}_{bp} = \sum_{k \in S} \widehat{w}_k y_k.$$

Subscripts b and p refer to bagging and PCA, respectively. Estimator $\widehat{t}_{bp}$ can alternatively be written

$$\widehat{t}_{bp} = \frac{1}{B} \sum_{b=1}^{B} \widehat{t}_{\text{cal}}^{(b)},$$

where

$$\widehat{t}_{\text{cal}}^{(b)} = \sum_{k \in S} w_k^{(b)} y_k.$$

Applying bagging to principal components instead of the original variables ensures better control of weight dispersion and yields a more stable estimator.

Algorithm 3 PCA–Based Bagging Calibration Algorithm

1: Input: Auxiliary variable matrix $\mathbf{X}$, inclusion probabilities π_k, sample size n, number of bootstrap iterations B, number of principal components selected at each iteration c, adjustement parameter α
2: Compute principal components $\mathbf{Z}_j = \mathbf{X}\mathbf{v}_j$ and eigenvalues λ_j, $j = 1, \ldots, q$.
3: **for** $b = 1$ to B **do**
4: Select a subset of c principal components without replacement and unequal probabilities proportional to $(\lambda_j / \sum \lambda_j)^\alpha$
5: Compute calibration weights $w_k^{(b)}$ by solving the calibration equation

$$\sum_{k \in S} w_k^{(b)} \mathbf{z}_k^{(b)} = \sum_{k \in U} \mathbf{z}_k^{(b)}.$$

6: **end for**
7: Compute the final weight for each unit:

$$\widehat{w}_k = \frac{1}{B} \sum_{b=1}^{B} w_k^{(b)}.$$

8: Output: final weights $\widehat{w}_k$.

Our proposed estimator is hence the average of B calibrated estimators. However, our proposed estimator is usually not exactly a calibrated estimator. Their weights are exactly calibrated only on those principal components that are included in each and every sample of principal components. See Section 4.3.3 for a possible approach to obtain exact calibration on some important variables. For more details on the derivation and properties, see Section 4.3.4.

4.3.2 Choice of Parameters

The choice of the number of principal components c remains a critical parameter in this method. A too small value of c results in weights far from being calibrated to the total of the principal components and a total estimator that does not benefit from a reduced variance in case the principal components are highly correlated to the variable on interest. As the number of selected

principal component c increases, the weights become closer and closer to being calibrated to the total of the principal components (when $c = q$, the weights are exactly calibrated). At some point, c is too large and too many principal components are selected. The problems associated with high-dimensional settings resurface, such as instable calibration weights and highly variable total estimators. Striking the right balance between dimensionality reduction and information retention helps maximizing the efficiency of the method and depends on the data at hand.

As a rule of thumb, we suggest $c = \sqrt{n}$. Alternatively, we can fix a proportion of variance of the initial auxiliary variables to be explained by the first c components. For instance, we may want to select the smallest c so that the first c components explain 60% of the variance in the auxiliary variables. This selection is universal with respect to the variables of interest. Or we can also choose the smallest c that explains a fixed fraction of the variation in a specific variable variable of interest. The resulting weights depend on one specific variable on interest and my not be suitable for another variable of interest.

Another key parameter to determine is the exponent applied to the eigenvalues α. This parameter controls how contrasted the probabilities of inclusion of the principal components are. A high value of α yields highly contrasted inclusion probabilities. The first components are much more likely to be selected than the last components. As α lowers, the contrast decreases. The inclusion probability of the components get closer and closer. The choice $\alpha = 0$ assigns the same inclusion probability to all components. When α is too low, too much emphasis is put on lower-ranked components, that do not explain much of the variability in the auxiliary variables, and may add noise to the calibration process. As a general guideline, we recommend $\alpha = 1/2$. With this choice, the probability that a component is selected is proportional to the portion of standard deviation explained by this component. Other choices may yield better results depending on the data at hand.

4.3.3 Advantages and Limitations

The method we propose has several advantages. First, it is independent of any variable of interest. The resulting weights can therefore be applied simultaneously to several variables of interest. It is also the case of the method of Cardot et al. (2017). Second, for variables of interest strongly linearly related to the principal components that are frequently selected in the bootstrap samples, the variance of the total estimator that we propose is lower than that of the HT estimator. Third, the obtained weights are very stable and close to the initial design weights.

However, there are also certain limitations. The calibration weights obtained with our method are not exactly calibrated, unless we decide to impose exact calibration on a set of important auxiliary variables (as presented below). In addition, if the linear relationship between a variable of interest and the first few principal components (those the most often selected in the bootstrap samples) is weak, then our total estimator may be worse than the HT estimator for this variable of interest. Our estimator is biased and may not be more efficient than the HT for that variable of interest.

The closest competitor to our method is the method of Cardot et al. (2017). With their method, the weights are exactly calibrated on the first c principal components. With our method, the weights are nor exactly calibrated on any principal components. Therefore, the total estimator of Cardot et al. (2017) is more efficient than ours for variables of interest for which a large proportion of variance is explained by the first c components. Our total estimator is more efficient when the $q - c$ other principal components still explain a substantial portion of the variance in the variable of interest, in addition to the variance explained by the first c principal components. An advantage of our method over the method of Cardot et al. (2017) is that the weights obtained with our method are more stable than those obtained with the method of Cardot et al. (2017) for a same number of principal components c considered.

As mentioned above, the calibration weights obtained with our method are not exactly calibrated. When there is a need to obtain weights exactly calibrated on some important auxiliary variables, our method can be adapted using a simple procedure described in Cardot et al. (2017) and detailed in this paragraph. Suppose that we want to calibrate the weights exactly on c_1 auxiliary variables, where $c_1 < c < q$. We compute the residuals of the regression of the remaining $q - c_1$ auxiliary variables on these important auxiliary variables. We hence create $q - c_1$ variables of residuals. We apply PCA to these $q - c_1$ variables of residuals and obtain $q - c_1$ principal components. These principal components are orthogonal to one another and to the subspace generated by the c_1 important auxiliary variables. Then, in each bootstrap sample, we select all c_1 important auxiliary variables and a sample of $c - c_1$ principal components.

Finally, we have assumed that the auxiliary variables are known for all population units. We now briefly explain how this assumption can be relaxed. Our method can indeed be applied when the user knows only the values of the auxiliary variables for sample units and the population total of these variables. To do so, we can apply the procedure of Cardot et al. (2017). Their procedure consists of estimating the matrix $\mathbf{\Sigma}$ of population covariance matrix of $\mathbf{X}$ using the design weights and the sample values. Then the eigenvectors and eigenvalues of this matrix are used to obtain estimated principal components. The method is then applied using these estimated principal components instead of the true principal components.

4.3.4 Our Proposed Estimator as a Model-Assisted Estimator

In this section, we show that our proposed estimator can be written as a model-assisted estimator. Our estimator is

$$\widehat{t}_{bp} = \sum_{k \in S} \frac{g_k y_k}{\pi_k} = \frac{1}{B} \sum_{b=1}^{B} \widehat{t}_{\text{cal}}^{(b)}, \tag{4.1}$$

where

$$\widehat{t}_{\text{cal}}^{(b)} = \sum_{k \in S} \frac{g_k^{(b)} y_k}{\pi_k},$$

with $g_k^{(b)}$ the solution to the calibration equation

$$\sum_{k \in U} \mathbf{z}_k^{(b)} = \sum_{k \in S} \frac{g_k^{(b)} \mathbf{z}_k^{(b)}}{\pi_k}.$$

With the chi-squared distance

$$\Phi_S(\mathbf{w}) = \sum_{k \in S} \frac{(w_k - d_k)^2}{q_k d_k}$$

and the choice $q_k = 1$, the calibration weights $w_k^{(b)} = g_k^{(b)}/\pi_k = d_k g_k^{(b)}$ are

$$\mathbf{w}^{(b)} = \arg\min_{\mathbf{w}} \Phi_S(\mathbf{w})$$

subject to the calibration equation above. The solution is

$$g_k^{(b)} = 1 + \mathbf{z}_k^{(b)\top} \left(\sum_{k \in S} \frac{1}{\pi_k} \mathbf{z}_k^{(b)} \mathbf{z}_k^{(b)\top} \right)^{-1} \left(\sum_{k \in U} \mathbf{z}_k^{(b)} - \sum_{k \in S} \frac{1}{\pi_k} \mathbf{z}_k^{(b)} \right),$$

see Deville and Särndal (1992), Equation (1.3). The calibrated estimator $\widehat{t}_{\text{cal}}^{(b)}$ can be written as

$$\widehat{t}_{\text{cal}}^{(b)} = \sum_{k \in S} \frac{y_k}{\pi_k} + \sum_{k \in S} \frac{\mathbf{z}_k^{(b)\top}}{\pi_k} \widehat{\mathbf{T}}_{\mathbf{z}\pi}^{(b)-1} \mathbf{t}_{\mathbf{z}}^{(b)} y_k - \sum_{k \in S} \frac{\mathbf{z}_k^{(b)\top}}{\pi_k} \widehat{\mathbf{T}}_{\mathbf{z}\pi}^{(b)-1} \widehat{\mathbf{t}}_{\mathbf{z}\pi}^{(b)} y_k,$$

where

$$\begin{aligned}
\widehat{\mathbf{T}}_{\mathbf{z}\pi}^{(b)} &= \sum_{k \in S} \frac{1}{\pi_k} \mathbf{z}_k^{(b)} \mathbf{z}_k^{(b)\top}, \\
\mathbf{t}_{\mathbf{z}}^{(b)} &= \sum_{k \in U} \mathbf{z}_k^{(b)}, \\
\widehat{\mathbf{t}}_{\mathbf{z}\pi}^{(b)} &= \sum_{k \in S} \frac{1}{\pi_k} \mathbf{z}_k^{(b)}.
\end{aligned}$$

Rearranging, we obtain

$$\widehat{t}_{\text{cal}}^{(b)} = \sum_{k \in U} \mathbf{z}_k^{(b)\top} \widehat{\mathbf{B}}_S^{(b)} + \sum_{k \in S} \frac{y_k - \mathbf{z}_k^{(b)\top} \widehat{\mathbf{B}}_S^{(b)}}{\pi_k}. \tag{4.2}$$

where

$$\widehat{\mathbf{B}}_S^{(b)} = \left(\sum_{k \in S} \frac{1}{\pi_k} \mathbf{z}_k^{(b)} \mathbf{z}_k^{(b)\top} \right)^{-1} \sum_{k \in S} \frac{1}{\pi_k} \mathbf{z}_k^{(b)} y_k$$

Finally, using Equations (4.1) and (4.2), our proposed estimator $\widehat{t}_{bp}$ can be rewritten as a model-assisted estimator as follows

$$\widehat{t}_{bp} = \sum_{k \in U} \widehat{m}(\mathbf{z}_k) + \sum_{k \in S} \frac{y_k - \widehat{m}(\mathbf{z}_k)}{\pi_k},$$

where

$$\widehat{m}(\mathbf{z}_k) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{z}_k^{(b)\top} \widehat{\mathbf{B}}_S^{(b)}.$$

The predicted value $\widehat{m}(\mathbf{z}_k)$ is the average over all bootstrap samples $b = 1, \ldots, B$ of the predictions $\mathbf{z}_k^{(b)\top} \widehat{\mathbf{B}}_S^{(b)}$ of the linear regressions of the variable of interest on the components included in the bootstrap samples.

4.4 Simulation Study

4.4.1 Data Presentation and Preparation

We conduct a simulation study on a real dataset to assess the accuracy of our approach in a practical context. We analyze multiple scenarios to ensure robustness of the presented results. We consider the Swiss Survey on Income and Living Conditions (SILC) data (Swiss Federal Statistical Office, 2015). The considered dataset consists of 89 variables recorded for $N = 425$ households. Among the 89 variables, $p = 87$ serve as auxiliary variables, with 64 being binary (dummy) and 23 being continuous. The remaining 2 are the variables of interest. Among the auxiliary variables we find for instance a variable indicating whether the household is located in Zurich or not, a variable that records the number of persons in the household, and the mortgage amount of the household. The variables of interest are:

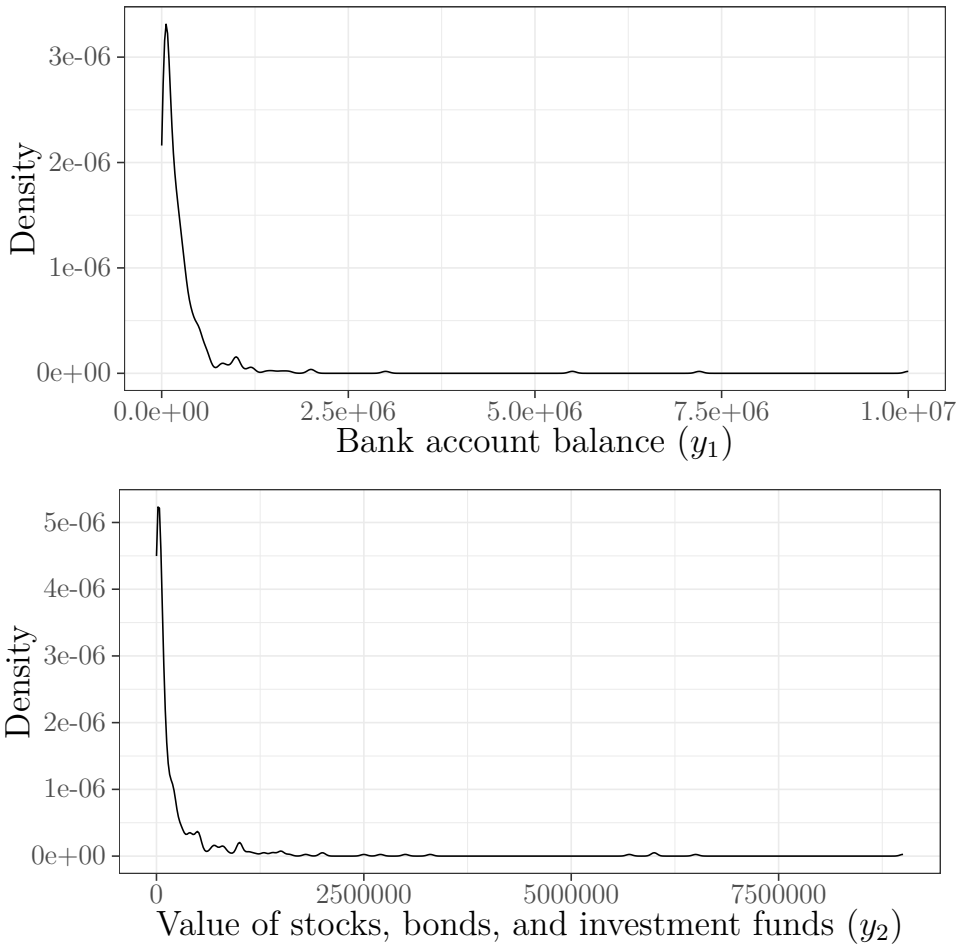

Figure 4.1: Density estimates of Bank account balance, y_1 (top panel) and Value of stocks, bonds, and investment funds, y_2 (bottom panel) for 425 households.

- y_1: bank account balance and
- y_2: value of stocks, bonds, and investment funds.

Figure 4.1 shows density estimates of the variables of interest. We see that these variables of interest are highly skewed to the right with large outliers.

In order to test the influence of extreme values on our proposed approach, we create two additional variables of interest as follows. We create a variable of interest y_3 that is a copy of y_1, except that all values above 500,000 are replaced with random draws from the remaining distribution. We repeat the same procedure to create a copy y_4 of y_2. Figure 4.2 shows density estimates of the created variables of interest y_3 and y_4. We see that y_3 and y_4 are still right-skewed, but much less so than y_1 and y_2. Moreover, y_3 and y_4 do not show any clear outliers.

Each auxiliary variable is then scaled by subtracting its mean and dividing by its standard deviations. We use the scaled auxiliary variables in what follows. We apply the PCA decomposition to the matrix $\mathbf{X}$ of scaled auxiliary variables as described in Section 4.2.3. The obtained eigenvalues λ_i, $i \in 1, \ldots, p$ are ranked in decreasing order $\lambda_1 = 7.14 > \lambda_2 = 5.84 > \ldots > \lambda_{87} = 6.48 \cdot 10^{-7}$. Their values range from approximately 0 to 7.14, with a mean value of 1 and a median value of 0.65. We choose the probabilities of inclusion of the principal components in the set of calibration variables to be proportional to the square root of the corresponding eigenvalues. That is, if we select c components that act as calibration variables for each bagging step, then the probability that principal component Z_i is a calibration variable is proportional to $c \cdot \left[\lambda_i \left(\sum_{j=1}^{p} \lambda_j\right)^{-1}\right]^{\alpha}$, with $\alpha = 1/2$. With this choice, the probability of inclusion of a

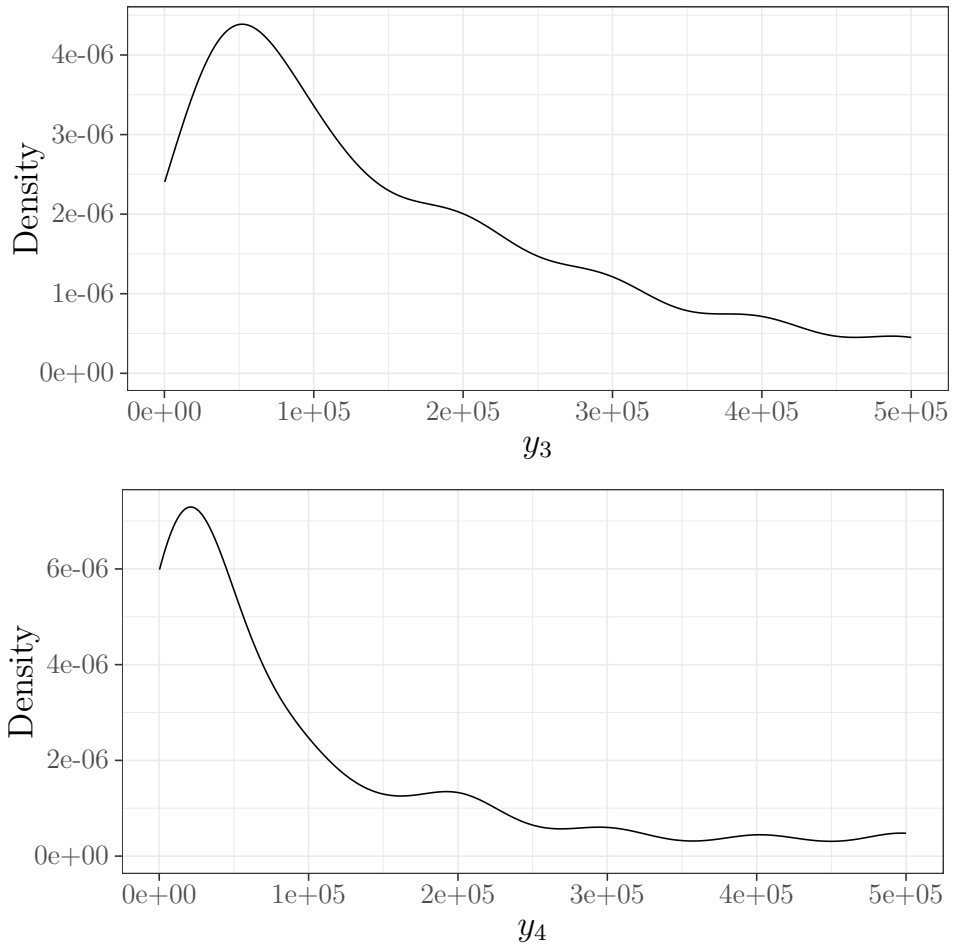

Figure 4.2: Density estimates of bank account balance without values above 500,000, y_3 (top panel) and Value of stocks, bonds, and investment funds without values above 500,000, y_4 (bottom panel) for 425 households.

component is proportional to the part of the standard deviation in the auxiliary variables that is explained by that component. For this study, we select $c = 10$ components that act as calibration variables for each bagging step. The first 10 components explain 43% of the variability in the auxiliary variables.

Table 4.1 contains the R-squared values for linear regressions of y_1 to y_4 on all the auxiliary variables and the first 10 principal components. Comparing the values for y_1 (respectively y_2) and those for y_3 (respectively y_4), we can see that cutting the tail of the distribution decreases by more than a half the proportion of the variance in the variables of interest explained by the linear models. Moreover, we also see a drop in the proportion of the variance in the variable of interest explained by the model when the first 10 principal components are considered as compared to when all the auxiliary variables are considered.

4.4.2 Simulation Design

We carry out 10,000 simulation runs as follows. For each simulation run, a sample of 20% of the households is selected with simple random sample without replacement. This corresponds to a sample size of $n = 85$. The total of the four variables of interest is computed using five different estimators. All five estimators can be written as calibration estimators on the form $\sum_{k \in S} \frac{g_k}{\pi_k} y_k$ for different choices of coefficients g_k. These estimators are described below.

1. (CAL) The usual calibration estimator. This estimator is obtained with all auxiliary variables as calibration variables.

Table 4.1: R-squared values for linear regressions of y_1 to y_4 on all auxiliary variables and the first 10 principal components.

	All Auxiliary Variables	**First 10 Principal Components**
y_1	0.6699	0.2301
y_2	0.6371	0.3050
y_3	0.2889	0.0780
y_4	0.2713	0.0894

2. (PCA) The calibration estimator with the first $c = 10$ principal components used as calibration variables. This is the method proposed in Cardot et al. (2017).

3. (BAG) The calibration estimator with $c = 10$ auxiliary variables selected at random with equal probabilities for each bootstrap step. We use $B = 500$ bootstrap steps.

4. (BAG+PCA) The calibration estimator with $c = 10$ principal components selected at random with probabilities proportional to the square root of the eigenvalues of the sample covariance matrix as explained in the previous section. We use $B = 500$ bootstrap steps. We use functions "UPMEpiktildefrompik", "UPMEqfromw", and "UPMEsfromq" of R package "sampling" (Tillé and Matei, 2023) in order to select the samples of principal components with fixed sample size $c = 10$ and unequal probabilities.

5. (HT) The Horvitz-Thompson estimator, which is obtained with $g_k = 1$ for all $k \in S$.

For the first four estimators, we consider linear calibration, that is the chi-squared distance, and use function "calib" of R package "sampling" (Tillé and Matei, 2023) to obtain the coefficients g_k. We include a constant calibration variables for all four estimators.

4.4.3 Measures of Comparison and Results for the Point Estimator

For a generic mean estimator $\widehat{t}$, we compute the following measures of comparison.

- The Monte Carlo relative Bias (RB)

$$\text{RB} = \frac{I^{-1}\sum_{i=1}^{I}\widehat{t}^{(i)} - t}{t},$$

  where $\widehat{t}^{(i)}$ is the value of $\widehat{t}$ obtained at simulation run i, $i = 1, \ldots, I$, $I = 10,000$ is the number of simulations, and t the population total that estimator $\widehat{t}$ intends to estimate.

- The Monte Carlo Relative Standard Deviation (RSD)

$$\text{RSD} = \frac{\left[(I-1)^{-1}\sum_{i=1}^{I}\left(\widehat{t}^{(i)} - \bar{\widehat{t}}^{(\cdot)}\right)^2\right]^{1/2}}{t},$$

  where $\bar{\widehat{t}}^{(\cdot)} = I^{-1}\sum_{i=1}^{I}\widehat{t}^{(i)}$ is the mean value of $\widehat{t}$ over all simulation runs.

- Monte Carlo Relative Root Mean Square Error (RRMSE)

$$\text{RRMSE} = \frac{\left[(I-1)^{-1}\sum_{i=1}^{I}\left(\widehat{t}^{(i)} - t\right)^2\right]^{1/2}}{t}.$$

- The Monte Carlo Variance relative to the Monte Variance of the HT estimator

$$\text{VARrHT} = \frac{(I-1)^{-1}\sum_{i=1}^{I}\left(\widehat{t}^{(i)} - \bar{\widehat{t}}^{(\cdot)}\right)^2}{(I-1)^{-1}\sum_{i=1}^{I}\left(\widehat{t}_{HT}^{(i)} - \bar{\widehat{t}}_{HT}^{(\cdot)}\right)^2}.$$

The results are presented in Table 4.2. In this context of high dimension, the usual calibration estimator (CAL) performs poorly. It is highly biased, with a bias (in absolute value) between approximately 3 and 50 times the value of the true totals. It is also highly unstable, with a standard deviation higher than 3,000 times the value of the true totals for each of the four variables of interest. The estimator with bagging of the auxiliary variables with equal probabilities performs better than the usual calibration estimator but still performs badly. In the worth case, its bias (in absolute value) amounts to 1.3 times the value of the true total. It is also unstable with a standard deviation between approximately five times to 15 times the value of the true totals.

The method of Cardot et al. (2017) and our method both provide excellent results in terms of bias with relative biases (in absolute value) smaller than 4% of the value of the true totals. Our proposed estimator provides the most stable estimators with a variance smaller than that of the HT estimator for all four variables of interest. For variables of interest y_1 and y_2 the variance of our proposed total estimator is approximately 74% of that of the HT estimator. That is, there is a reduction of 26% of the variance as compared to the HT estimator. For variables of interest y_3 and y_4, the variance of our proposed total estimator is almost the same as that of the HT estimator. The reason is that a linear model predicting the variable of interest based on the auxiliary variables and on the principal components explains a greater part of the variance in y_1 and y_2 than in y_3 and y_4. The R-squared of some linear models are in Table 4.1. We can also note that our proposed estimator (BAG+PCA) provides better results than the estimator of Cardot et al. (2017) (PCA). The reason is that the linear model of the variables of interest on the first ten principal components (the components on which their estimator is calibrated) explain much less of the variability in the variables of interest than the linear models on all the auxiliary variables. That is, some of the variability in the variables of interest is explained by the remaining components. Our proposed method uses these components, the method of Cardot et al. (2017) does not. In a context where the first principal components explain a larger portion of the variability in the variables of interest, the method of Cardot et al. (2017) performs better than it does in the current context.

We conducted a robustness check to examine how our method behaves under different values of α and c, see Section 4.4.5.

4.4.4 Measure of Comparison and Results for the Calibration Coefficients g_k

For each simulation run, we compute the Coefficient of Variation (CV) of the calibration coefficients g_k for each of the four calibration estimators (CAL, PCA, BAG, BAG+PCA) defined as the ratio of the standard deviation of the g_k's to their mean. We obtain 10,000 CVs for each estimator. Table 4.3 contains summary statistics of the CVs of the coefficients g_k over 10,000 simulation runs for the four calibration methods. We can see that the CVs obtained with estimators CAL and BAG are very high compared to the CVs of PCA and BAG+PCA. Highly dispersed coefficients may be problematic, especially if the coefficients are used to compute other parameters of interest such as totals in small domains. This illustrates how CAL and BAG are inappropriate in the context of high dimension. PCA and BAG+PCA provide the smallest CV with values between 0.11 and 1.23 for PCA and between 0.08 and 0.24 for BAG+PCA. Our proposed estimator, BAG+PCA, provides the best results with CVs much smaller. This illustrates

Table 4.2: Monte Carlo relative Bias (RB), Relative Standard Deviation (RSD), Relative Root Mean Square Error (RRMSE), and Variance relative to the Monte Variance of the HT estimator (VARrHT) for five estimators and four variables of interest.

	RB	**RSD**	**RRMSE**	**VARrHT**
		y_1		
CAL	−50.74	4279.47	4279.77	3.18e8
PCA	0.04	0.30	0.30	1.59
BAG	−1.01	15.22	15.25	4.02e3
BAG+PCA	−0.03	0.20	0.21	0.728
HT	0.00	0.24	0.24	1.00
		y_2		
CAL	2.93	5039.30	5039.30	3.50e8
PCA	−0.01	0.26	0.26	0.961
BAG	−1.30	15.11	15.17	3.15e3
BAG+PCA	−0.04	0.23	0.24	0.740
HT	−0.00	0.27	0.27	1.00
		y_3		
CAL	−11.11	3349.43	3349.45	1.69e9
PCA	0.00	0.09	0.09	1.19
BAG	0.96	5.82	5.90	5.11e3
BAG+PCA	0.00	0.08	0.08	0.983
HT	0.00	0.08	0.08	1.00
		y_4		
CAL	33.71	5172.17	5172.28	1.85e9
PCA	0.02	0.13	0.13	1.18
BAG	0.07	8.85	8.85	5.42e3
BAG+PCA	0.00	0.12	0.12	0.976
HT	0.00	0.12	0.12	1.00

how our proposed method allow to stabilize the calibration coefficients g_k, and hence the final weights, in the context of high dimension.

4.4.5 Choice of the Number of Principal Components c and Exponent α

We conducted a robustness check to examine how our method behaves under different values of c and α. We run 10,000 simulations as explained earlier in the current section for different values of c and α. Below are the results for y_1. The results for the other three variables of interest are similar.

We first investigate the effect of varying the number of components selected in the bootstrap

Table 4.3: Summary Statistics of the CVs of the coefficients g_k over 10,000 simulation runs for four methods.

Method	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
CAL	1.30	6.20	271.10	12591.30	9544.20	2120381.80
PCA	0.11	0.30	0.37	0.39	0.47	1.23
BAG	0.11	0.23	15.09	36.42	40.05	4750.14
BAG+PCA	0.08	0.12	0.13	0.13	0.14	0.24

samples c. Figure 4.3 shows the Monte Carlo Relative Root Mean Square Error (RRMSE) of the total estimator of y_1 for different values of c. We can see that, as c grows, the RRMSE first decreases until a point where it starts to increase. For this variable of interest, the minimum is attained at around $c = 20$. The Monte Carlo Relative Bias of the total estimator of y_1 is smaller than or equal to 4% (in absolute value) for all the tested values of c, with no clear pattern.

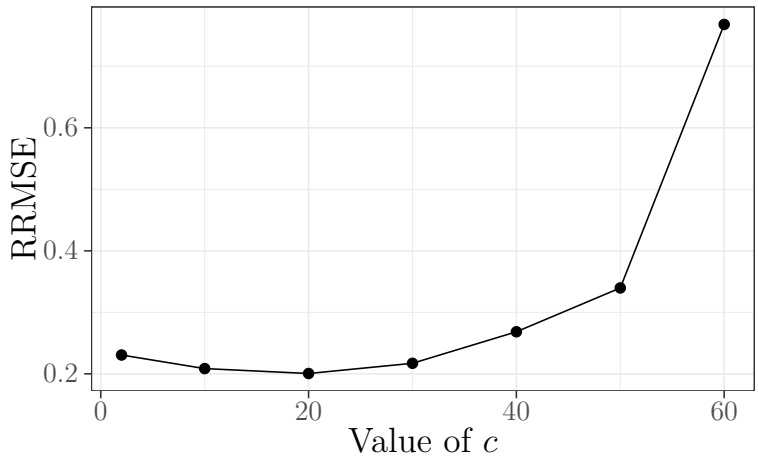

Figure 4.3: Monte Carlo Relative Root Mean Square Error (RRMSE) of the total estimator for different values of c.

For every simulation run, we compute the minimum and the maximum of the coefficients g_k. Then we compute the mean over the 10,000 simulations of the minimum and maximum weights. We also compute the minimum and maximum weight over all 10,000 simulations. We repeat the procedure for different values of c. The results are in Figure 4.4. We see that, as more components are added to the bootstrap samples (and therefore to the calibration), the coefficients g_k become more and more dispersed.

We repeat the same procedure to investigate the effect of varying α. The results are in Figures 4.5 and 4.6. The Monte-Carlo Relative Bias of the total estimator of y_1 is smaller than or equal to 3% (in absolute value) for all tested values of α, with no particular pattern. The total estimator of y_1 has the smallest RRMSE around $\alpha = 1/2$. Moreover, its RRMSE is larger than selecting the components with equal probabilities ($\alpha = 0$) when α is larger than 1. This phenomenon depends on the variable of interest, on the prediction power of the first principal components, and on the choice of c. It shows however that giving higher probabilities of being selected to the first principal components in not necessarily a good option. We also see that the coefficients g_k become more and more dispersed as α increases. This phenomenon also depends on the variable of interest, on the prediction power of the first principal components, and on the choice of c. The general conclusion is that our proposed method can perform well when the parameters c and α are appropriately selected.

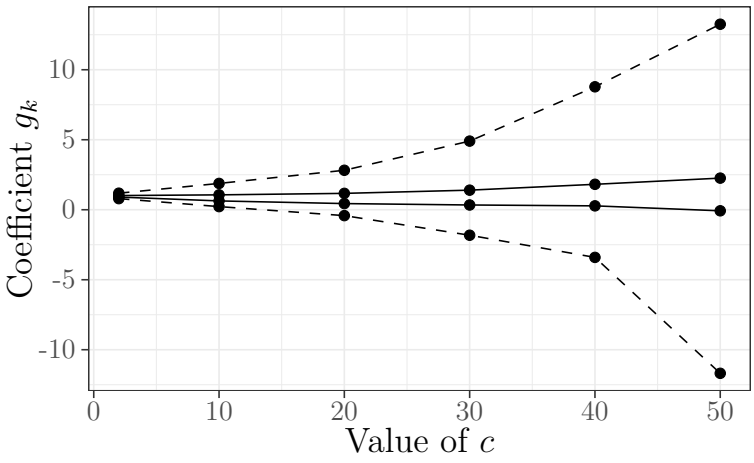

Figure 4.4: Mean over 10,000 simulations of the minimum and maximum coefficients g_k (solid lines); minimum and maximum coefficients g_k over 10,000 simulations (dashed). Different values of c are considered.

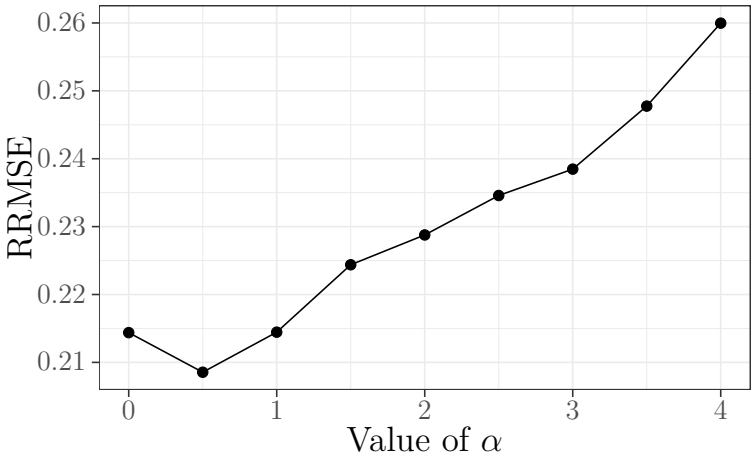

Figure 4.5: Monte Carlo Relative Root Mean Square Error (RRMSE) of the total estimator for different values of α.

4.5 Discussion

In this paper, we introduce a new calibration approach for high-dimensional settings, based on bagging and principal component decomposition. Conventional calibration often fails when the number of auxiliary variables is high, leading to unstable weights and total estimators. To overcome this problem, we perform multiple calibrations using random subsets of auxiliary variables and aggregate the resulting weights. This aggregation strategy improves stability and delivers good performance for different variables of interest.

Our simulation results show several advantages of the proposed method. First, the final weights obtained with bagging are considerably less dispersed than with standard calibration or calibration based on PCA alone. Second, the variance of the total estimator remains small, even in the presence of many irrelevant or redundant auxiliary variables. Moreover, the final weights are not tailored to a specific variable of interest y, allowing them to be reused for multiple variables, which is particularly useful surveys with many outputs.

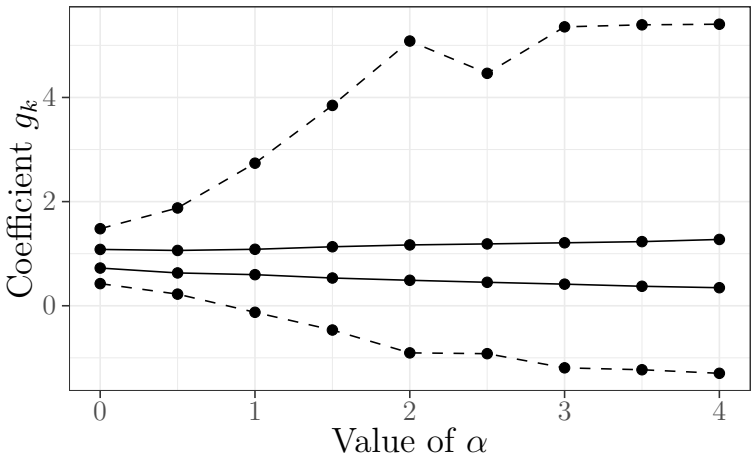

Figure 4.6: Mean over 10,000 simulations of the minimum and maximum coefficients g_k (solid lines); minimum and maximum coefficients g_k over 10,000 simulations. Different values of α are considered.

Acknowledgements

This research was financially supported by the Swiss Federal Statistical Office. The views expressed in this article are those of the author solely and do not necessarily reflect those of the aforementioned organization.

Chapter 5

Improving Donor Imputation Using the Prediction Power of Random Forests : a Combination of SwissCheese and MissForest

Abstract
Imputation procedures are frequently used to treat nonresponse. With random hot deck imputation, missing values are replaced by valid observed values from other units in the same dataset. The recently-developed balanced nearest neighbor imputation method, implemented in the SwissCheese R package, generates random hot deck imputation under certain balancing constraints to decrease the variance of the total estimator, in the presence of multivariate nonresponse. The method relies on a notion of neighbourhood between units, utilizing a distance measure that becomes difficult to define in high dimensions. In contrast to hot deck imputation methods, many imputation procedures obtain replacement values from prediction models fit from observed data. The missForest method, which uses random forests as prediction models, is an example of this approach. In this article, we propose a new approach that uses the two methods in a complementary manner. We refine the distance measure in the SwissCheese method using missForest predictions. Through a simulation study on empirical data from the Swiss Survey on Income and Living Conditions, we demonstrate reductions in Monte Carlo variance, bias and mean squared error of the totals obtained by our proposed imputed estimator compared to those obtained using SwissCheese alone.[1]

Keywords : balancing constraint, item nonresponse, machine learning, non-monotone nonresponse, random imputation.

Statement of significance

In this paper, we propose a novel imputation procedure that combines a random hot deck procedure (SwissCheese) and a prediction procedure based on machine learning (missForest).

[1]This chapter is based on: Tripet, A., Eustache, E., & Tillé, Y., (2024). Improving Donor Imputation Using the Prediction Power of Random Forests: a Combination of SwissCheese and missForest, Journal of Survey Statistics and Methodology, 12(5), 1389–1404.

Our approach retains the benefits of hot deck imputation (plausible imputed values, realistic multivariate relationships) while improving the utilized measures of similarity with accurate predictions. We conduct a simulation study on a real dataset of the Swiss Survey on Income and Living Conditions. We show that the combinations of the two methods outperform the basic SwissCheese method and are almost as good as the predictive missForest method.

5.1 Introduction

Survey data is often subject to nonresponse. Unit nonresponse occurs when all the information on a sampled unit is missing while item nonresponse occurs when only some of the data collected from the sampled units are missing. Nonresponse can have a monotonic pattern in the dataset as in longitudinal studies (Hasler et al., 2018) and can be multivariate by appearing in more than one survey variable. In this article, we focus on multivariate item nonresponse without any monotonic pattern, i.e. non-monotone nonresponse or Swiss cheese nonresponse (Marker et al., 2002; Peytchev, 2012), for surveys that collect more than one variable. We define a respondent as a unit providing valid values for all items, and a nonrespondent as a unit with missing data for at least one item.

Nonresponse implies risks of bias, loss of information, and errors in inference as noted in Chen and Haziza (2019). Imputation, which involves replacing a nonresponse or invalid value with a plausible value, is often used in practice. A common method of imputation is to use predictive models, such as the fully conditional specification procedure, which proceeds iteratively by imputing the variables affected by nonresponse one by one. Many imputation methods are based on fully-conditional specifications; well-known examples include sequential regression (Raghunathan et al., 2001) or multivariate imputation by chained equation (van Buuren and Groothuis-Oudshoorn, 2011). Another well-known predictive method is missForest (Stekhoven and Bühlmann, 2012), which uses random forest models (Breiman, 2001a).

In contrast, random hot deck imputation methods use observed values to fill in nonresponses, see Judkins (1997) and Andridge and Little (2010) for an overview. Hot deck methods match a respondent (donor) to a nonrespondent (recipient), replacing the missing respondent data with donor values. The matching procedures often rely on a measure of similarity between respondents and nonrespondents. In this paper, we examine the SwissCheese method (Eustache et al., 2024), a hot deck imputation method, which is a multivariate extension first proposed in Hasler and Tillé (2016). Its main feature is the selection of a unique donor for each nonrespondent, while satisfying balancing constraints to reduce the variance of the total estimator. The selection of a single donor for each nonrespondent: 1. Generates imputed values that are observed and therefore realistic; 2. Preserves the relationship between the variables and reduces the impossible combination of imputed values in the case of multivariate nonresponses. In the SwissCheese method, the notion of similarity for donor selection is based on the Euclidean distance calculated on the available data. However, this distance suffers from dimensionality (Zimek et al., 2012), since increasing the number of variables adds complexity as discussed in Breiman (2001b). Furthermore, distance can be difficult to compute if the amount of information for certain units is limited, compromising the choice of donor.

In this paper, we combine the SwissCheese and missForest imputation methods, using the missForest imputations in the SwissCheese distance formula. This combined method remains a random hot deck imputation method. We conduct a simulation study on a real dataset of the Swiss survey on income and living conditions to compare imputed estimates of totals, first and ninth deciles, and correlations between survey variables.

The paper is organized as follows. Notations and brief explanations of missForest and SwissCheese imputations are introduced in Section 5.2. The proposed method combining SwissCheese and missForest is presented in Section 5.3. Section 5.4 contains the simulation study on real data of the Swiss survey on income and living conditions. We finish with a discussion in Section 5.5.

5.2 SwissCheese and MissForest Methods

Consider a finite population $U = \{1, \ldots, N\}$. Let $S \subset U$ be a sample of size n selected in U according to a sampling design. The inclusion probability of unit k in S is denoted by π_k, $k \in U$. For each sampled unit k, $J \in \mathbb{N}^*$ survey variables are collected in a vector $\mathbf{x}_k = (x_{k1}, \ldots, x_{kJ})^\top$. We consider multivariate nonresponse, i.e. some values of $\mathbf{x}_k$ may be missing. Let $\mathbf{r}_k = (r_{k1}, \ldots, r_{kJ})^\top$ be the J-vector of variable response indicators, such that $r_{kj} = 1$ if the variable j is observed for unit k and 0 otherwise, $j \in \{1, \cdots, J\}$. The set of respondents, denoted by $S_r \subset S$, contains units without any nonresponse in their vector $\mathbf{x}_k$ and the set of nonrespondents $S_m = S \backslash S_r$ contains units with at least one nonresponse in their vector $\mathbf{x}_k$. Values n_r and n_m denote the size of S_r and S_m respectively. Note that the variables in $\mathbf{x}_k$ can be categorical, but must be converted to dummy variables.

For each variable j, the aim is to estimate an unknown parameter θ_j of the population U, by a point estimate $\widehat{\theta}_j(S)$ based on values $\{x_{kj}\}_{k \in S}$. For the sake of simplicity, we write $\widehat{\theta}_j$ for $\widehat{\theta}_j(S)$. If x_{kj} is unavailable for some units $k \in S$, a solution is to impute nonresponses by valid values and compute an estimator $\widehat{\theta}_j$ using the completed dataset. For example, the total

$$t_j = \sum_{k \in U} x_{kj},$$

can be estimated by the imputed estimator

$$\widehat{t}_j := \sum_{k \in S} \pi_k^{-1} \{ r_{kj} x_{kj} + (1 - r_{kj}) x_{kj}^* \},$$

where x_{kj}^* denote the imputed value of x_{kj}.

Many imputation methods use predictive models. When nonresponse is multivariate, the process is iterative: models are fit by changing the set of covariates (independent variables) and the variable requiring imputation (the dependent variable). Then, the nonresponses are imputed using these models. Stekhoven and Bühlmann (2012) propose missForest that uses random forests, a collection (ensemble) of decision trees, as machine learning method (Breiman, 2001a). In general, random forests perform well for prediction, even if the number of observations n is almost equal or less than the number of variables J, or when variables are unrelated (Waljee et al., 2013). Indeed, it can manage the sparsity and the noise that some variables may contain (Biau, 2012). MissForest imputation starts with an initial imputation by the mean, then iteratively imputes nonresponses variable by variable, in ascending order of the number of nonresponses, using random forest models. At each step, the variable $j \in \{1, \cdots, J\}$ subject to nonresponse is selected, and a random forest is fit with variables $j' \in \{1, \cdots, J\} \backslash j$ as predictors and j as dependant variable, yielding new prediction for values $\{x_{kj} \mid r_{kj} = 0\}$. The procedure cycles through each variable computing revised imputed values conditional on the other (fully-imputed) variables, until a predetermined stopping criteria or a maximum number of iterations is reached.

Hot deck imputations replace nonresponses with a function of observed values. Eustache et al. (2024) propose the SwissCheese imputation method. This method assigns a single donor to each nonrespondent, and the donor's values impute its nonresponses. As in many hot deck procedures, this similarity between the selected donor and recipient is determined by a distance function, calculated from the values observed in the two units. Let $\psi_{uv} \in [0, 1]$ denote the probability that respondent $u \in S_r$ is the donor of nonrespondent $v \in S_m$. SwissCheese imputation starts by computing these imputation probabilities by solving the linear program with constraints

$$\begin{cases} \underset{\psi_{uv} \in [0,1]}{\text{minimize}} & \displaystyle\sum_{v \in S_m} \sum_{u \in S_r} d(u, v) \psi_{uv}, & \\ \text{subject to} & \displaystyle\sum_{u \in S_r} \psi_{uv} = 1, & v \in S_m, \\ & \displaystyle\sum_{v \in S_m} r_{vj} \sum_{u \in S_r} \psi_{uv} x_{uj} = \sum_{v \in S_m} r_{vj} x_{vj}, & j \in \{1, \ldots, J\}, \end{cases} \tag{5.1}$$

where

$$d(u,v) \coloneqq \left\{ \sum_{j=1}^{J} r_{vj}(x_{uj} - x_{vj})^2 \right\}^{1/2}.$$

Quantity $d(u,v)$ represents the distance between the respondent u and the nonrespondent v. Note that the continuous variables must be normalized and scaled before computing the distance. Smaller distances represent greater similarity, and in this case the estimated imputation probability (ψ_{uv}) also increases. The probabilities $\{\psi_{uv}\}$ solve a minimisation problem, and are therefore optimal solutions because they are subject to constraints. To find a solution to the linear program (5.1), note that the ratio J/n_r must not be too large.

The first constraint of (5.1) enforces selection of only one donor per nonrespondent. Furthermore, the method includes balancing constraints for the imputation. The second constraint holds the total estimate constant for each survey variable. After the computation of the imputation probabilities, a stratified sampling, satisfying the imputation probabilities and the two constraints of (5.1), is applied in order to select only one donor per nonrespondent. Consider the bipartite set of S_r and S_m, $U^* = S_r \times S_m$, of size $n_r \cdot n_m$. A stratum $U_v^* = \{(u,v)|u \in S_r\}$ of U^* containing the set of the n_r possible donors is assigned to each nonrespondent v. Then, we select a sample of cells in U_v^* using a sampling design of fixed size with $\{\psi_{uv}\}$ as inclusion probabilities. Stratified sampling generates the imputation indicators ϕ_{uv}, such that $\phi_{uv} = 1$ if respondent u is the donor selected for nonrespondent v, and $\phi_{uv} = 0$ otherwise. The sample is of size one since constraint $\sum_{u \in S_r} \phi_{uv} = 1$ must be satisfied. Final selected cell represents the final donor of nonrespondent v. Steps of the SwissCheese imputation are summarized in Algorithm (4).

Algorithm 4 SwissCheese Imputation Method

1. Compute distance $d(u,v)$ between units $u \in S_r$ and $v \in S_m$.
2. Compute the imputation probabilities $\psi_{uv} \in [0,1]$ by solving the linear program (5.1).
3. Compute the imputation indicators $\phi_{uv} \in \{0,1\}$ using a stratified sampling design satisfying imputation probabilities ψ_{uv} and constraints of linear program (5.1).
4. Impute each unavailable value x_{vj} by value $x_{vj}^* := \sum_{u \in S_r} \phi_{uv} x_{uj}$ for all $v \in S_m$.

This balanced donor-based method has several advantages. Firstly, the imputed values are observed values and are therefore realistic. Secondly, the principle of consistency between values from the same observation is preserved since the imputed values come from the same donor. A third advantage lies in the calibration and balancing constraints, which reduce the variability of the various estimated parameters.

5.3 Method Combination

In SwissCheese imputation, the donor selection is strongly related to the distance function $d(.,.)$ (see Section 5.2). The distance value $d(u,v)$ between a respondent $u \in S_r$ and a nonrespondent $v \in S_m$ is computed on the available values. This distance may suffer from 1) too many nonresponses in the vectors $\mathbf{x}_v$, $v \in S_m$; 2) the number of survey variables J is too large. In the first case, the recipient's lack of information makes matching with a similar donor difficult. In the latter case, the distance function can be misleading, as few selected variables can dominate the matching process. Both situations can lead to the selection of donors dissimilar to the recipients, which can compromise the rest of the imputation process, since this is the first step

in the method, see Algorithme (4). As the SwissCheese and missForest imputation methods have complementary characteristics, we propose to combine them to refine the notion of similarity between respondents and nonrespondents, in the first stage of Algorithm (4), by exploiting the predictive power of missForest, even in the two cases mentioned above (Tang and Ishwaran, 2017). In other words, we need to modify the distance function $d(.,.)$ to reduce its drawbacks by using missForest.

We propose to use the MissForest procedure to carry out an initial imputation of nonresponses, and then compute the Euclidean distance between units using both the observed and the imputed values, such that

$$d_1(u,v) = \left[\sum_{j=1}^{J} r_{vj}(x_{uj} - x_{vj})^2 + \sum_{j=1}^{J}(1 - r_{vj}) \cdot (x_{uj} - x_{vj}^{\circ})^2\right]^{1/2},$$

where x_{vj}° is the imputed value of x_{vj} using the missForest imputation. Unlike $d(.,.)$, the distance $d_1(.,.)$ does not suffer from the first case, as missForest generates an accurate prediction even in these complicated cases.

However, the inconvenience encountered when the number of variables J is large, indicated in Section refsec:notation, is still present with the distance $d_1(.,.)$. Then, we propose a second variant to distance $d(.,.)$ which reduces the disadvantages of the two cases mentioned above. Let $\mathcal{G} \subset \{1,\ldots,J\}$ denote the set of $G \leq J$ variables subject to nonresponse. This variant $d_2(.,.)$ differs from $d_1(.,.)$ by computing the Euclidean distance only on variables containing nonresponses, which are pre-imputed by missForest, such that

$$d_2(u,v) = \left[\sum_{j\in\mathcal{G}} r_{vj}(x_{uj} - x_{vj})^2 + \sum_{j\in\mathcal{G}}(1 - r_{vj}) \cdot (x_{uj} - x_{vj}^{\circ})^2\right]^{1/2}.$$

Although there is a loss of information because some variables are not included in the calculation of $d_2(.,.)$, some information is still summarised in the pre-imputation carried out by missForest. If there is no fully-observed variables, note that $J = G$ and $d_2(.,.)$ is equivalent to $d_1(.,.)$.

These two distance functions lead to new versions of the SwissCheese methods in which the first step is modified. The first and the second new versions use respectively functions $d_1(.,.)$ and $d_2(.,.)$ instead of $d(.,.)$ in Step 1 of Algorithm (4). They combine the missForest and the SwissCheese imputations and are still random hot deck imputation methods.

As pointed out by a referee, there are several similarities between our proposed methods and predictive mean matching (PMM). PMM is a hot deck imputation method first proposed in Rubin (1986) and Little (1988). It uses a predicted value for a certain variable j to match donors and recipients, with predictions made for all sampled units, generally using a model with fully-observed auxiliary variables. Then, it randomly selects a donor to impute the nonresponses by its observed value. For more details, see van Buuren (2018) and Vink et al. (2014). In the same way as the methods we propose, PMM pre-imputes the nonresponses and uses a notion of distance between units to impute the dataset. One difference is that the SwissCheese method contains balancing constraints. Moreover, unlike our proposed approach, PMM imputes missing values on a variable-by-variable basis, not ensuring that the same respondent will whereas we propose simultaneous multivariate imputation from the donor.

5.4 Simulation Study

In this section, we present a simulation study comparing SwissCheese, missForest, and our two new versions of SwissCheese, which combine missForest imputation in distance calculations (see Section 5.3), under different nonresponse scenarios. The R packages SwissCheese (Eustache

et al., 2021) and missForest (Stekhoven, 2022) are used to perform the simulations. We consider a real dataset from a survey of income and living conditions provided by the Swiss Federal Statistical Office, containing 600 units and 85 variables. We create a population U of $N = 300$ units selected using simple random sampling without replacement, treated as a census so that $U = S$ and $N = n$. Our census contains five variables, each subjected to nonresponse; four continuous variables, namely bank account balances (X_1), value of stocks, bonds, investment funds (X_2), valuables (X_3), and total mortgages (X_4) and one categorical variable, regular inter-household transfers (X_5). Table 5.1 presents the totals on population U for these variables (i.e. the true values).

Table 5.1: Population totals of study variables

	X_1	X_2	X_3	X_4	X_5
Population totals	3524.63	3311.37	3040.93	3876.74	177.00

We generate nonresponses in these five variables, and the goal is to estimate unknown parameters θ_j of variable X_j on population U, $j \in \{1, \ldots, 5\}$. We focus on four parameters: the total, the first and ninth deciles, and the Pearson correlations between variables. The nonresponse is induced in two phases. First, we randomly select 80% of the units in the population, randomly inducing nonresponse in these selected units in order to obtain an expected nonresponse rate of 25% in each variable. As nonresponse is induced separately for each variable, this approach guarantees that there are enough respondents (n_r) for hot deck imputation, reserving 20% of the population as fully-observed units. The probability of being selected in these 80% units is given by

$$p_k^0 = \left\{1 + \exp\left[-\boldsymbol{\lambda}_0^\top (1, x_{k1}, x_{k2}, x_{k3}, x_{k4}, x_{k5})^\top\right]\right\}^{-1}$$

where x_{kj} is the value of variable X_j for unit k and $\boldsymbol{\lambda}_0^\top = (2.017, -0.007, -0.025,$
$-0.036, 0.016, -0.191)$. Vector $\boldsymbol{\lambda}_0$ is set so that the expected number rate of units selected, i.e. the mean of the $\{p_k^0\}$ on the population, is equal to 0.8. A unit k is randomly selected in the 80% of the first phase using a random draw of a Bernoulli variable, with probability p_k^0.

The probability that unit k responds to item j, i.e. the probability that $r_{kj} = 1$, is

$$p_{kj} = 1 - p_k^0 \cdot [1 + -\lambda_j(x_{kj})]^{-1},$$

where λ_j is the j^th element of vector $\boldsymbol{\lambda}^\top = (-0.067, -0.072, -0.078, -0.061, -1.533)$. Value λ_j is set so that the mean of the $\{p_{kj}\}$ on the population is equal to 0.75. Probabilities $\{p_{kj}\}$ and $\{p_k^0\}$ are strictly between 0 and 1. In the second phase, nonresponses are generated in each variable j using a random draw of a Bernoulli variable with probability p_{kj}. According to this procedure, it is likely that more than one variable is subject to nonresponse in the same unit designated k. The expected percentage of respondents (i.e. units with no missing value) is 32% and the expected value of n_r is 96. As nonresponse is directly correlated with the variables of interest, the studied response mechanism is not missing at random (Rubin, 1976).

We estimate the totals of variables X_1, X_2, X_3, X_4 and X_5 in different scenarios, each varying the number of fully-observed auxiliary variables (J') denoted by $\widetilde{X}_{j'}$, $j' \in \{1, \ldots, J'\}$, where $J' = 0, 2, 4, 6, 8, 10, 12, 14, 16, 18$ or 20. This yields $(5 + J')$ variables, $k \in U$, which includes the five study variables described above, each subject to nonresponse.

The J' auxiliary variables are determined after computing the Pearson correlations between each survey variable $X_j, j \in \{1, \ldots, 5\}$ and the remaining 81 survey variables. We restrict the selection set of auxiliary variables to the 20 variables that have the largest correlation with at least one of the five study variables; correlation coefficients range from 0.396 to 0.123 in the selection set. Finally, we designated half of the J' variables from the upper half of this selection

set and the remainder from the lower half. For example, with $J' = 2$, the first auxiliary variable $\widetilde{X}_1$ is the one in the 20 that has the largest correlation with one of the survey variables (0.396), and the second auxiliary variable $\widetilde{X}_2$ is the one in the 20 that has the lowest correlation with one of the survey variables (0.123). This ensures variability in the prediction strength of the covariate set. To assess this collective prediction power, we fit linear regressions predicting each study variable using the covariate set assigned to each scenario; adjusted R^2 statistics for each model are presented in Table 5.2.

Table 5.2: Adjusted R-Squared of the linear regression between variables of interest X_1, X_2, X_3 and X_4 and each scenario considering between 2 and 20 auxiliary variables.

Scenario	2	4	6	8	10	12	14	16	18	20
X_1	0.21	0.28	0.30	0.31	0.32	0.34	0.33	0.34	0.33	0.34
X_2	0.40	0.41	0.42	0.43	0.45	0.45	0.46	0.47	0.46	0.47
X_3	0.05	0.09	0.09	0.09	0.11	0.18	0.17	0.17	0.15	0.17
X_4	0.00	0.07	0.08	0.08	0.10	0.20	0.21	0.21	0.21	0.23

We compare five imputation methods:

- Balanced nearest neighbor imputation or SwissCheese;
- Sequential Random Forest or missForest;
- The new version of SwissCheese based on distance d_1 (see Section 5.3);
- The new version of SwissCheese based on distance d_2 (see Section 5.3);
- PMM (see Section 5.3).

For PMM imputation, we use the function MICE of the van Buuren et al. (2015) package, with default parameters.

We obtain $M = 10,000$ imputations per method and thus M imputed point estimators $\widehat{\theta}_j$ of the unknown parameter θ_j, for each variable j. Performance of each imputation method is measured by Monte Carlo bias, empirical variance, and mean squared error (MSE), computed as

$$\widehat{\text{Bias}}_{MC}\left(\widehat{\theta}_j\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_j^m - \theta_j\right),$$

$$\widehat{\text{Var}}_{MC}\left(\widehat{\theta}_j\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_j^m - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\theta}_j^{m'}\right)^2$$

and

$$\widehat{\text{MSE}}_{MC}\left(\widehat{\theta}_j\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_j^m - \theta_j\right)^2,$$

where $\widehat{\theta}_j^m$ is the value of the imputed point estimator of the parameter θ_j at simulation $m \in \{1, \ldots, M\}$.

Figures 5.1, 5.2 and 5.3 present the Monte Carlo bias, empirical variance, and MSE of the totals for the five study variables for each considered scenario. Tables 1-3, in Appendix B, provide these evaluation statistics for $J' = 0, 10$, and 20 auxiliary variables. See Tables 4-6 in Appendix B statistics for corresponding evaluation for deciles, and Tables 7-9 for corresponding evaluation statistics for correlation coefficients.

First, for continuous variables X_1, X_2, X_3 and X_4, the MSE of the total estimator is mainly explained by the variance as the bias is negligible. Compared with the scenario without auxiliary

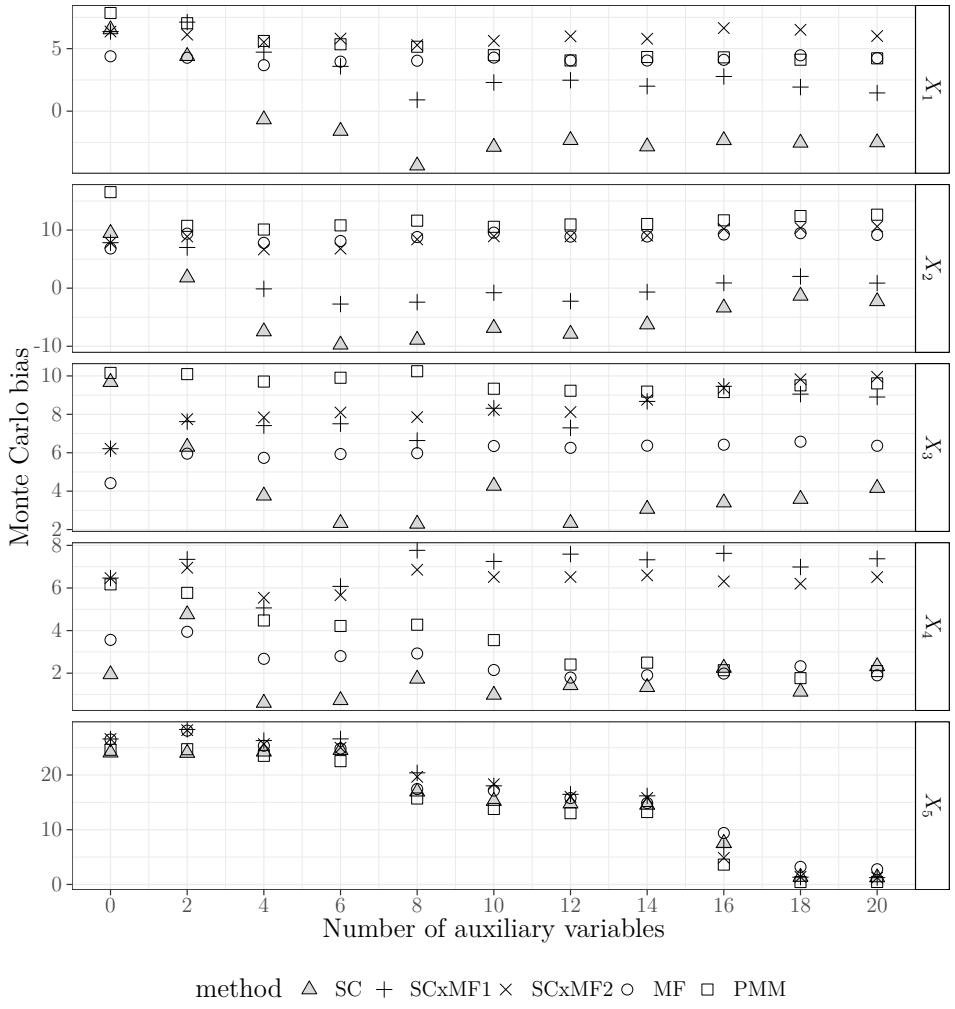

Figure 5.1: Monte Carlo bias of the total estimators of variables X_1, X_2, X_3, X_4 and X_5 for five different imputation methods: SwissCheese (SC), missForest (MF) combinations of the two methods (SCxMF_1 and SCxMF_2) and predictive mean matching (PMM). Different scenarios considering between 0 and 20 auxiliary variables are compared.

variables, the addition of a maximum of 4 auxiliary variables reduces the MSE of each method, particularly for variables X_1 and X_2. The addition of highly correlated and fully-observed variables provides additional information and thus improves imputation. In addition, we add the most correlated variables in order, so that the more variables we add, the more uninteresting variables we add.

First, for continuous variables X_1, X_2, X_3 and X_4, the MSE of the total estimator is mainly explained by the variance, as the bias is negligible. The addition of highly correlated, fully observed variables provides additional information and improves imputation. This can be seen as early as the addition of the first auxiliary variables, which reduces the MSE of each method compared to the scenario without auxiliary variables. This is particularly visible for the variables X_1 and X_2.

For each considered scenario and continuous study variable, missForest method has the smallest MSE, demonstrating the efficiency of this imputation method, regardless of the number of auxiliary variables. With the donor imputation methods, the variance increases as the number of auxiliary variables increases. The two proposed new methods yield lower MSEs than the (basic) SwissCheese method, providing evidence that contribution of the missForest pre-imputation reduces the Monte Carlo variance of the total estimates. When the number of auxiliary variables is low, the Monte Carlo variance and MSE of the two new proposed methods are close to the

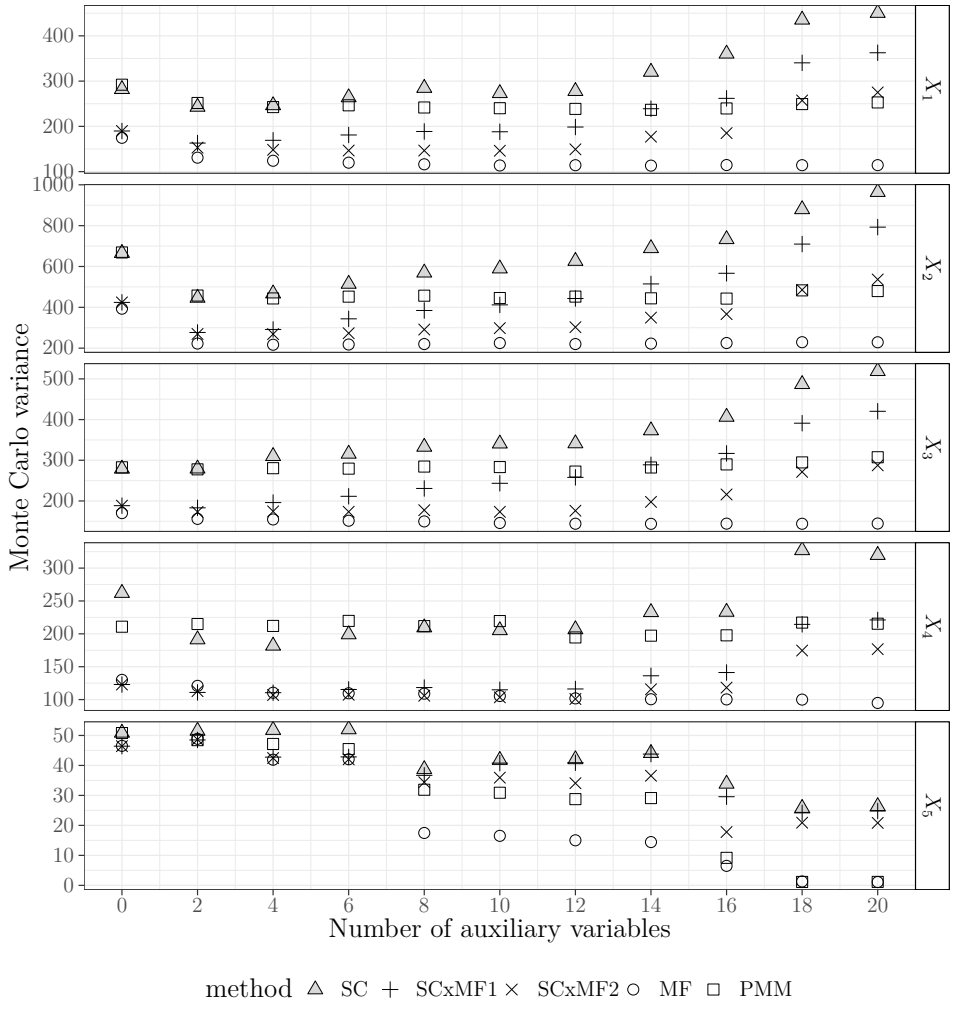

Figure 5.2: Monte Carlo variance of the total estimators of variables X_1, X_2, X_3, X_4 and X_5 for five different imputation methods: SwissCheese (SC), missForest (MF) combinations of the two methods (SCxMF_1 and SCxMF_2) and predictive mean matching (PMM). Different scenarios considering between 0 and 20 auxiliary variables are compared.

ones of missForest. The second version of the method combining SwissCheese and missForest outperforms the first version for each continuous variable of interest. Its variances and MSE are lower and increase as the number of auxiliary variables increases. The PMM method obtains a higher MSE than all the other methods except SwissCheese, when the number of auxiliary variables is less than 12. Nevertheless, the method seems to be less impacted by the addition of auxiliary variables. It even obtains a lower MSE than the new versions of SwissCheese we propose, for the variable of interest X_1, when $J' = 18$ and 20.

Then, on the side of the categorical variable X_5, the MSE is mainly the bias when the number of auxiliary variables is less than 14. As the number of auxiliary variables increases, the variance likewise increases. Indeed, at $J' = 20$, the variance is the dominant term in the MSE. As more highly predictive auxiliary variables are added, the bias, variance, and MSE of the categorical variable totals decreases. When the number of auxiliary variables increases, the variance, the bias and thus the MSE decrease for all methods. It can be explained by the additions of some variables that are highly related with the categorical variable. That said, the MSEs of the total of X_5 are quite close for all methods in corresponding scenarios, with quicker reductions with missForest and PMM than with SwissChesse as the number of auxiliary variables increases. Indeed, the methods using SwissCheese retain a larger variance in the total estimate.

In conclusion, the simulations show that missForest produces accurate imputed totals, even

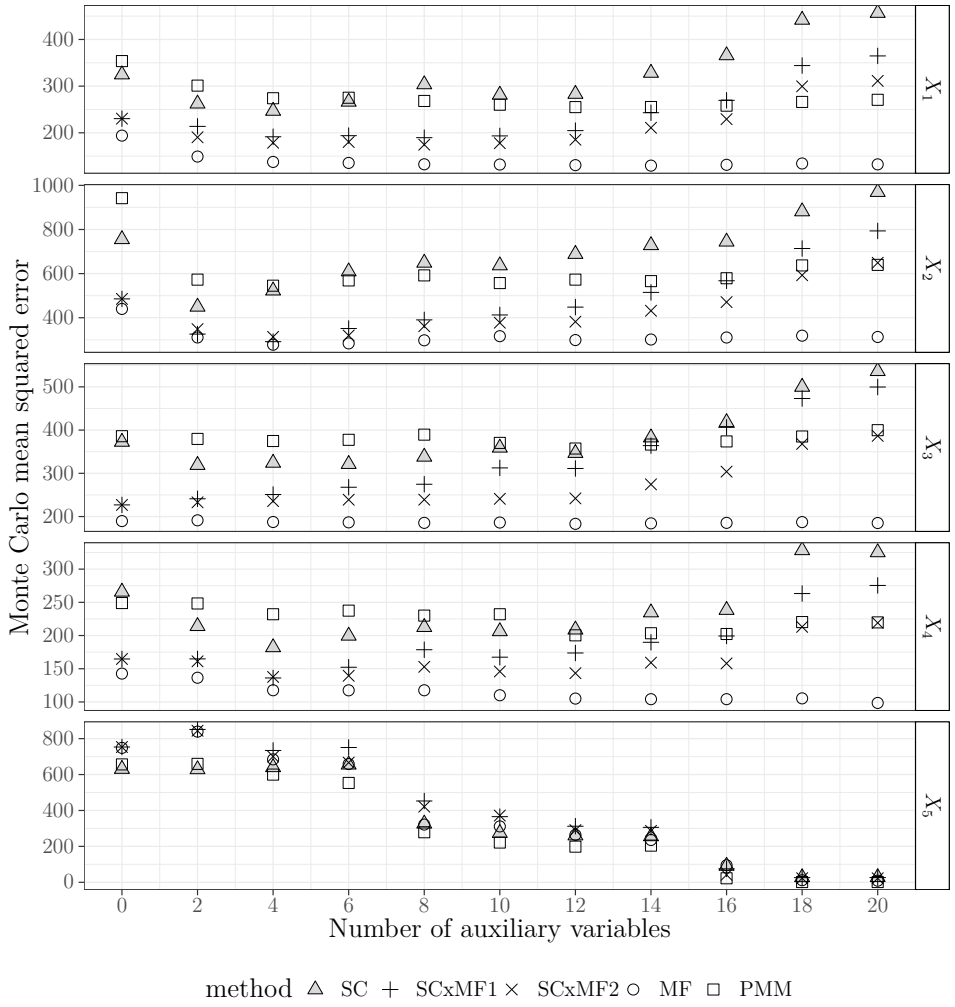

Figure 5.3: Monte Carlo mean squared error of the total estimators of variables X_1, X_2, X_3, X_4 and X_5 for five different imputation methods: SwissCheese (SC), missForest (MF) combinations of the two methods ($SCxMF_1$ and $SCxMF_2$) and predictive mean matching (PMM). Different scenarios considering between 0 and 20 auxiliary variables are compared.

when the number of auxiliary variables is high. Often, imputation by observed values is required, to avoid imputing by impossible values, for example in data sets on living conditions. In this case, our proposed second version of missForest and SwissCheese combination provides a good compromise, guaranteeing plausible imputed values as imputations use observed values while reducing the variability over the 'basic' SwissCheese method, and avoiding the pitfalls of a misspecified distance function as the number of auxiliary matching variables increases.

5.5 Discussion

Single donor imputation may be difficult to manage depending on the structure of the dataset, the number of respondents, and the degree of nonresponse in the data. Prediction can be further complicated in a multivariate setting, when selected combinations of variables are unrealistic or impossible. In this case, donor allocation methods such as SwissCheese are preferable. That said, databases are becoming increasingly complex. Surveys can collect a large number of variables, and additional auxiliary data sources are multiplying rapidly. New predictive methods that utilize machine learning, such as missForest, can easily handle this complexity while achieving

generally accurate predictions. Nevertheless, these methods are sometimes referred to as 'black boxes', because they are often difficult or even impossible to interpret.

In this paper, we combine both approaches, producing a method that takes advantage of each method's relative strengths and largely overcoming each method's deficiencies. In addition, to take advantage of a large set of auxiliary variables, the proposed combination method should be more robust to model misspecification than the PMM method, given that missForest predictions are based on (non-parametric) random forests (Morris et al., 2014). For continuous variables, our simulation results showed less sensitivity in matching donors with a large number of auxiliary variables with our proposed combined method compared to a simple application of SwissCheese. However, for imputation of categorical variables, the proposed combined methods seem to increase variance of the total estimators. Moreover, if the number of fully observed units n_r is high or if the ratio J/n_r is high, the choice of donor becomes difficult or even impossible, resulting in poor imputed values.

The results obtained are promising. Revisiting distance computation by predicting and then using the predicted values seems to be an attractive approach, whether using machine learning or other predictive methods.

Finally, the inclusion of the variance of the modeled parameters into the overall variance estimate should be investigated. The determination of an explicit variance estimator needs to be studied further, perhaps considering pseudo-population bootstrap variance estimator (Chen and Haziza, 2017).

Acknowledgements

The authors would like to thank the three reviewers for their constructive comments, which enabled them to improve this article considerably. The authors are particularly grateful to the Editor-in-Chief, Katherine Thompson, for her careful proofreading and insightful comments, which greatly improved understanding of the article. The authors would also like to thank the Swiss Federal Statistical Office for providing the data. The conclusions and observations contained in this article are not necessarily shared by the Swiss Federal Statistical Office. The authors would also like to thank Dr. Caren P. Hasler for her assiduous proofreading of the article and her corrections.

Chapter 6

Conclusion

This thesis proposed new approaches in survey statistics, structured around three themes: balanced sampling, calibration, and imputation. Each part contributed methodological innovations and practical solutions to current challenges in survey design and estimation.

In the first part, we introduced a new extension of the cube method that incorporates linear inequality constraints. This allows for the selection of balanced samples while ensuring the respect of lower bounds on domain-specific sample sizes. The resulting algorithm not only retains the efficiency and flexibility of the original cube method but also proves suitable for applications involving category bounding, matrix rounding, or spatial spread sampling. Simulation studies demonstrated that our method maintains desirable statistical properties and enables more controlled and representative designs in practice. The use of this new method seems to be versatile and can certainly be applied to many fields.

The second part addressed the challenge of harmonizing multiple weighting systems in surveys. We proposed two strategies for constructing a unified set of weights: one based on calibration, the other on optimal transport. The comparison of these approaches highlighted the versatility and robustness of the optimal transport method, especially in high-discrepancy scenarios between sample and population distributions. This method is particularly promising if it is generalized. We further developed a solution to the high-dimensional calibration problem using bagging and principal component analysis, reducing computational burden and improving estimator stability in complex settings. In high-dimensional settings, we proposed a new method based on bagging and principal component decomposition. The idea is to draw several samples of principal components and to compute calibration weights on each of these samples. The final weights are obtained by averaging the resulting weights. The proposed method reduces weight variability, limits the risk of increased variance, and yields a single set of weights that can be used to estimate several totals. It is especially well-suited for high-dimensional contexts where traditional calibration methods may fail.

The third part focused on item nonresponse. It presented a hybrid imputation method combining SwissCheese with missForest. The integration of model-based predictions into donor selection refines the notion of similarity and enhances the quality of the imputations. Simulations confirmed reductions in bias, variance, and mean squared error compared to traditional donor-based methods, demonstrating the practical value of this synergy if we wish to use a donor imputation method.

Altogether, the contributions of this thesis provide new approaches and tools directly applicable to current survey practices. The proposed methods are also often versatile and can be applied to a variety of domains.

Overall, this thesis aims to provide an overview on survey sampling by examining and addressing some of its key components: sampling, calibration and imputation. And not just in isolation but also as interconnected tools. Indeed, these different approaches are not mutually exclusive. On the contrary, they can be combined to develop robust and coherent survey designs

adapted to the complex challenges of the real world, where there is a growing desire to work with high-dimensional dataset, while at the same time having to cope with a growth in missing values. One of the main aims of this thesis has been to maintain a practical orientation. The aim is for the proposed methods to be not only theoretical but also directly applicable to real datasets and survey contexts. R implementations and simulation studies underline this applied perspective.

The methods presented show promising results and are versatile, however, limitations may remain. For example, calculation costs for large-scale applications. Some of these methods, such as the imputation method in Chapter 5, or sampling with inequality constraints in Chapter 2, can be computationally time-consuming. Also, some methods suggest rather particular characteristics, such as Chapter 3 with scenarios where the distribution of samples is very different from those of their populations. In terms of continuity and future prospects, it would be interesting to push these methods in more generalized contexts, as in Chapter 3. A more advanced calculation with a variance estimator for each of the methods should also be considered.

Appendices

Appendix A

Balanced Sampling With Inequalities - R Code

```
#####################################################
# Balanced Sampling With Inequalities
# R Code and Examples
#####################################################

# Load necessary libraries
require(sampling)
# Computes inclusion probabilities

require(MASS)
# Provides the Null function

require(hitandrun)
# Handles inequality constraints

require(StratifiedSampling)
# Implements the Cube method for balanced sampling

#####################################################
# Function Definitions
#####################################################

#########################################################
# Computes the inequality constraints for balanced sampling
# Arguments:
# - X: The dataset (coordinates, auxiliary variables, etc.)
# - pik: Vector of inclusion probabilities
# - cstr: inequality constraint type ("less", "greater",
# or "two.sided")
# - bound: The boundary for constraints (default is 1)
# Output:
# - A matrix defining the inequality constraints
#########################################################

ineq.cstr <- function(X, pik, cstr = "less", bound = 1) {
```

```
if (!cstr %in% c("greater", "less", "two.sided")) {
stop("'cstr' should be one of 'two.sided', 'less',
or 'greater'")
}

N <- length(pik)
DD <- as.matrix(dist(X)) # Compute distance matrix

size <- ifelse(cstr == "two.sided", 2, 1)
IND <- array(0, c(N, size * N * length(bound)))
M <- 1

# Apply inequality constraints
for (bound_i in bound) {
for (i in ((1:N) + (N * (1:(size * length(bound)) - 1)[M]))) {
o <- order(DD[, i == ((1:N) +
(N * (1:(size * length(bound)) - 1)[M]))])
cond <- cumsum(pik[o]) <= bound_i
IND[o, i][cond] <- ifelse(cstr == "greater", -1, 1)
}
M <- M + 1
}

return(IND)
}

#####################################################
# Fast Cube Sampling with Inequality Constraints
# Implements the flight phase of the Cube method
# Arguments:
# - X: Dataset
# - pik: Inclusion probabilities
# - B: Inequality constraint matrix
# - r: Vector of bounds for inequality constraints
# Output:
# - Adjusted inclusion probabilities after the flight phase
#####################################################

fast.flight.cube.ineq<-function(X,pik,B,r,deepness=1,
EPS=0.00000001){

Kernel <- function(A) residuals(lm(rnorm(nrow(A))~A-1))

A=as.matrix(X/pik)
if(nrow(A)==0) A=matrix(0,c(length(pik),1))

######

B=as.matrix(B/rep(1,length(pik)))
TT=!(nrow(B)==0)
if(TT) c=c(t(B)%*%pik)
```

```
######

if(TT){
TE=abs(c(t(B)%*%pik)-r)<=EPS
A=cbind(A,B[,TE])
r=r[!TE]
B=B[,!TE]
c=c[!TE]
}

B=as.matrix(B/rep(1,length(pik)))
TT=!(nrow(B)==0)

if(is.null(A)) prof=deepness else{if(is.matrix(A))prof=ncol(A)+
        deepness else
        prof=1+deepness}
TEST=(EPS<pik) & (pik<1-EPS)
prof2=min(sum(TEST),prof)
if(TT & prof2!=0) {BR=matrix(B[TEST,],c(sum(TEST),
length(B[TEST,])/sum(TEST)))[1:prof2,];
if(ncol(B)!=0) BR=matrix(BR,c(length(BR)/ncol(B),ncol(B)));}
if(prof2==0) a=0 else {pikR=pik[TEST][1:prof2];
AR=matrix(A[TEST,],c(sum(TEST),
        length(A[TEST,])/sum(TEST)))[1:prof2,];
AR=matrix(AR,c(length(AR)/ncol(A),ncol(A)));
if(nrow(AR)==1 | sum(abs(AR))==0) NN=matrix(1,c(1,1)) else
        NN=Null(matrix(AR,c(length(AR)/ncol(A),ncol(A))))}
if(ncol(NN)==0){a <- 0}else{
a <- 1
u <- NN[,1]
}

while(a>0.000001){

#####
#####


l1=min(pmax((1-pikR)/u,-pikR/u))
l2=min(pmax((pikR-1)/u,pikR/u))
if(TT ){
c=c(t(B)%*%pik)
cR=c(t(BR)%*%pikR)
pet=(r-c)/c(t(BR)%*%u)
if(sum(pet>0)>0) nu1=min(pet[pet>0])  else nu1=100000000000
if(sum(pet<0)>0) nu2=min(-pet[pet<0]) else nu2=100000000000
l1=min(l1,nu1)
l2=min(l2,nu2)
}
pik[TEST][1:prof2]=if (runif(1)<l2/(l1+l2)) pikR+l1*u else
```

```
        pikR-l2*u
TEST=(EPS<pik) & (pik<1-EPS)

#####

if(TT){
TE=abs(c(t(B)%*%pik)-r)<=EPS
A=cbind(A,B[,TE])
r=r[!TE]
B=B[,!TE]
c=c[!TE]
}
B=as.matrix(B/rep(1,length(pik)))
TT=!(nrow(B)==0)

###

prof=ncol(A)+deepness
prof2=min(sum(TEST),prof)
if(TT & prof2!=0) {BR=matrix(B[TEST,],c(sum(TEST),
length(B[TEST,])/sum(TEST)))[1:prof2,];
if(ncol(B)!=0) BR=matrix(BR,c(length(BR)/ncol(B),ncol(B)));}
if(prof2==0) a=0 else {pikR=pik[TEST][1:prof2];
AR=matrix(A[TEST,],c(sum(TEST),
        sum(A[TEST,])/sum(TEST)))[1:prof2,];
AR=matrix(AR,c(length(AR)/ncol(A),ncol(A)));
if(sum(abs(AR))==0) {NN=matrix(1,c(nrow(AR),1))} else
        {u <- Kernel(matrix(AR,c(length(AR)/ncol(A),ncol(A))))}
a=sum(abs(u))}

####
}

pik

}

#####################################################
# Sampling with Cube Method and Inequality Constraints
#####################################################

cube.ineq <- function(X, pik, B, r, index = 0.01,
deepness = 1, EPS = 1e-8) {

s <- ffphase(cbind(pik, X, B * pik), pik)
s <- fast.flight.cube.ineq(X / pik * s, s, B, r, deepness, EPS)

add.i <- 1
while (any(round(s, 5) != 0 & round(s, 5) != 1)) {
s <- fast.flight.cube.ineq(s, s, B, r * add.i^sign(r),
        deepness, EPS)
```

```
add.i <- add.i + index
}

return(round(s, 5))
}

#####################################################
# Example 1: Matrix Rounding using Cube Method
# with Inequality Constraints
#####################################################

# Define an example matrix
values <- c(15, 21, 17, 9, 10, 8, 13, 7, 6, 9, 5, 8, 4, 3,
 6, 6, 3, 2, 5, 8)
M1 <- matrix(values, 5, 4, byrow = TRUE)
M <- M1 / 16.5
M[M1 / 16.5 > 1] <- (M1 / 16.5)[M1 / 16.5 > 1] - 1

# Create design matrix
X <- cbind(disjunctive(rep(1:5, 4)),
        disjunctive(sort(rep(1:4, 5))))
pik <- c(M) # Inclusion probabilities
B <- cbind(X, -X) # Define constraints
r <- ceiling(t(B) %*% pik) # Define bounds

# Apply the sampling algorithm
s <- fast.flight.cube.ineq(pik, pik, B, r)

# Display results
Z1 <- addmargins(round(matrix(s, dim(M)), 4)) # Rounded matrix
Z2 <- addmargins(M) # Original matrix
Z1 - Z2 # Should contain values between -1 and 1

#####################################################
# Example 2: Spatial Sampling with Inequality Constraints
#####################################################

# Define population parameters
N <- 100  # Population size
n <- 20   # Sample size
X <- cbind(runif(N), runif(N)) # Generate dataset
pik <- inclusionprobabilities(runif(N), n)
# Compute inclusion probabilities

# Ensure sum of inclusion probabilities is equal to n
stopifnot(abs(sum(pik) - n) < 1e-6)

# Generate inequality constraints
LG1 <- ineq.cstr(X, pik, "two.sided", 1)
r <- ceiling(t(LG1) %*% pik)
```

```
# Draw sample
s <- cube.ineq(pik, pik, LG1, r)

# Visualizing the sample
plot(X, main = "Spatial␣Sampling␣with␣Inequality␣Constraints")
# Plot population
points(X[s == 1, ], pch = 19, col = "red")
# Highlight selected sample
```

Appendix B

Combination of SwissCheese and MissForest - Tables

Table B.1: Comparison of the monte carlo bias ($\widehat{\text{Bias}}_{MC}$) , variance ($\widehat{\text{Var}}_{MC}$) and mean squared error ($\widehat{\text{MSE}}_{MC}$) between total estimators of SwissCheese (SC), missForest (MF) combinations of the two methods ($SCxMF_1$ and $SCxMF_2$) and predictive mean matching (PMM) without auxiliary variable.

	X_1	X_2	X_3	X_4	X_5
$\widehat{\text{Bias}}_{MC}$					
SC	6.534	9.479	9.677	1.944	24.083
$SCxMF_1$	6.366	7.828	6.210	6.465	26.588
$SCxMF_2$	6.366	7.828	6.210	6.465	26.588
MF	4.389	6.827	4.415	3.563	26.494
PMM	7.848	16.532	10.157	6.170	24.612
$\widehat{\text{Var}}_{MC}$					
SC	282.120	665.703	278.972	262.035	50.867
$SCxMF_1$	189.812	424.055	188.382	122.907	46.425
$SCxMF_2$	189.812	424.055	188.382	122.907	46.425
MF	174.741	393.256	170.056	129.980	46.523
PMM	292.073	668.701	282.675	210.851	50.807
$\widehat{\text{MSE}}_{MC}$					
SC	324.807	755.546	372.625	265.816	630.838
$SCxMF_1$	230.336	485.338	226.945	164.700	753.342
$SCxMF_2$	230.336	485.338	226.945	164.700	753.342
MF	194.001	439.859	189.545	142.676	748.444
PMM	353.669	942.007	385.847	248.919	656.577

Table B.2: Comparison of the monte carlo bias ($\widehat{\text{Bias}}_{MC}$) , variance ($\widehat{\text{Var}}_{MC}$) and mean squared error ($\widehat{\text{MSE}}_{MC}$) between total estimators of SwissCheese (SC), missForest (MF) combinations of the two methods ($SCxMF_1$ and $SCxMF_2$) and predictive mean matching (PMM) with 10 auxiliary variables.

	X_1	X_2	X_3	X_4	X_5
$\widehat{\text{Bias}}_{MC}$					
SC	-2.862	-6.866	4.273	0.983	15.243
$SCxMF_1$	5.626	8.915	8.230	6.516	18.322
$SCxMF_2$	2.294	-0.788	8.315	7.246	18.022
MF	4.283	9.560	6.350	2.150	17.166
PMM	4.476	10.563	9.333	3.554	13.799
$\widehat{\text{Var}}_{MC}$					
SC	273.116	589.618	340.818	205.318	41.782
$SCxMF_1$	146.115	298.028	173.327	103.412	35.923
$SCxMF_2$	188.074	411.935	243.395	114.874	40.772
MF	113.603	224.964	145.839	105.442	16.514
PMM	240.143	445.519	283.266	219.434	30.894
$\widehat{\text{MSE}}_{MC}$					
SC	281.310	636.762	359.076	206.283	274.118
$SCxMF_1$	177.764	377.505	241.060	145.872	371.606
$SCxMF_2$	193.336	412.555	312.537	167.384	365.581
MF	131.950	316.359	186.166	110.065	311.176
PMM	260.177	557.095	370.375	232.062	221.307

Table B.3: Comparison of the monte carlo bias ($\widehat{\text{Bias}}_{MC}$) , variance ($\widehat{\text{Var}}_{MC}$) and mean squared error ($\widehat{\text{MSE}}_{MC}$) between total estimators of SwissCheese (SC), missForest (MF) combinations of the two methods ($SCxMF_1$ and $SCxMF_2$) and predictive mean matching (PMM) with 20 auxiliary variables.

	X_1	X_2	X_3	X_4	X_5
$\widehat{\text{Bias}}_{MC}$					
SC	-2.500	-2.267	4.159	2.321	1.272
$SCxMF_1$	6.009	10.618	9.952	6.510	1.233
$SCxMF_2$	1.457	0.871	8.901	7.371	1.239
MF	4.242	9.160	6.364	1.901	2.774
PMM	4.224	12.647	9.612	2.101	0.457
$\widehat{\text{Var}}_{MC}$					
SC	450.456	964.261	518.859	319.778	26.372
$SCxMF_1$	275.080	536.432	287.526	176.686	20.828
$SCxMF_2$	362.639	792.693	420.380	221.113	24.802
MF	114.352	228.820	144.530	94.866	1.099
PMM	252.881	479.663	307.414	215.390	1.154
$\widehat{\text{MSE}}_{MC}$					
SC	456.706	969.401	536.160	325.164	27.991
$SCxMF_1$	311.189	649.174	386.577	219.071	22.348
$SCxMF_2$	364.763	793.452	499.603	275.444	26.338
MF	132.350	312.719	185.025	98.479	8.796
PMM	270.726	639.617	399.810	219.803	1.363

Table B.4: Comparison of the Monte Carlo bias ($\widehat{\text{Bias}}_{MC}$), variance ($\widehat{\text{Var}}_{MC}$) and MSE ($\widehat{\text{MSE}}_{MC}$) between the estimator of the first decile in the left-hand table and the ninth decile in the right-hand table SwissCheese (SC) missForest (MF) combinations of the two methods ($SCxMF_1$ and $SCxMF_2$) and predictive mean matching (PMM) without auxiliary variable.

	X_1	X_2	X_3	X_4	X_5
$\widehat{\text{Bias}}_{MC}$					
SC	-0.019	0.062	-0.002	0.029	0
$SCxMF_1$	0.132	0.316	0.050	0.205	0
$SCxMF_2$	0.132	0.316	0.050	0.205	0
MF	0.155	0.355	0.080	0.191	0
PMM	-0.026	0.061	-0.008	0.048	0
$\widehat{\text{Var}}_{MC}$					
SC	0.020	0.049	0.003	0.037	0
$SCxMF_1$	0.018	0.029	0.010	0.011	0
$SCxMF_2$	0.018	0.029	0.010	0.011	0
MF	0.016	0.024	0.012	0.009	0
PMM	0.017	0.042	0.003	0.029	0
$\widehat{\text{MSE}}_{MC}$					
SC	0.021	0.053	0.003	0.038	0
$SCxMF_1$	0.035	0.129	0.012	0.053	0
$SCxMF_2$	0.035	0.129	0.012	0.053	0
MF	0.040	0.149	0.018	0.046	0
PMM	0.018	0.046	0.003	0.031	0

	X_1	X_2	X_3	X_4	X_5
$\widehat{\text{Bias}}_{MC}$					
SC	0.063	-0.107	0.069	0.003	0
$SCxMF_1$	-0.017	-0.221	-0.002	-0.081	0
$SCxMF_2$	-0.017	-0.221	-0.002	-0.081	0
MF	-0.031	-0.287	-0.004	-0.091	0
PMM	0.088	-0.069	0.112	0.021	0
$\widehat{\text{Var}}_{MC}$					
SC	0.010	0.014	0.018	0.008	0
$SCxMF_1$	0.004	0.013	0.001	0.001	0
$SCxMF_2$	0.004	0.013	0.001	0.001	0
MF	0.004	0.010	0.001	0.001	0
PMM	0.012	0.012	0.025	0.009	0
$\widehat{\text{MSE}}_{MC}$					
SC	0.014	0.026	0.022	0.008	0
$SCxMF_1$	0.004	0.062	0.001	0.008	0
$SCxMF_2$	0.004	0.062	0.001	0.008	0
MF	0.005	0.092	0.001	0.010	0
PMM	0.020	0.017	0.038	0.010	0

Table B.5: Comparison of the monte carlo bias ($\widehat{\text{Bias}}_{MC}$), variance ($\widehat{\text{Var}}_{MC}$) and MSE ($\widehat{\text{MSE}}_{MC}$) between the estimator of the first decile in the left-hand table and the ninth decile in the right-hand table SwissCheese (SC) missForest (MF) combinations of the two methods ($SCxMF_1$ and $SCxMF_2$) and predictive mean matching (PMM) with 10 auxiliary variable.

	X_1	X_2	X_3	X_4	X_5
$\widehat{\text{Bias}}_{MC}$					
SC	-0.044	0.031	-0.013	0.047	0
$SCxMF_1$	0.137	0.340	0.071	0.233	0
$SCxMF_2$	0.049	0.200	0.025	0.194	0
MF	0.171	0.398	0.122	0.232	0
PMM	-0.037	0.048	-0.010	0.017	0
$\widehat{\text{Var}}_{MC}$					
SC	0.022	0.056	0.006	0.033	0
$SCxMF_1$	0.018	0.024	0.013	0.006	0
$SCxMF_2$	0.014	0.043	0.005	0.012	0
MF	0.016	0.020	0.019	0.005	0
PMM	0.017	0.038	0.003	0.032	0
$\widehat{\text{MSE}}_{MC}$					
SC	0.024	0.057	0.006	0.036	0
$SCxMF_1$	0.037	0.139	0.018	0.060	0
$SCxMF_2$	0.016	0.083	0.006	0.049	0
MF	0.046	0.179	0.034	0.059	0
PMM	0.018	0.041	0.003	0.032	0

	X_1	X_2	X_3	X_4	X_5
$\widehat{\text{Bias}}_{MC}$					
SC	0.028	-0.166	0.085	-0.022	0
$SCxMF_1$	-0.026	-0.235	-0.003	-0.085	0
$SCxMF_2$	-0.002	-0.202	0.012	-0.060	0
MF	-0.043	-0.283	-0.010	-0.096	0
PMM	0.081	-0.079	0.120	0.019	0
$\widehat{\text{Var}}_{MC}$					
SC	0.006	0.015	0.022	0.007	0
$SCxMF_1$	0.003	0.009	0.001	0.001	0
$SCxMF_2$	0.004	0.012	0.004	0.003	0
MF	0.003	0.007	0.002	0.001	0
PMM	0.011	0.012	0.026	0.008	0
$\widehat{\text{MSE}}_{MC}$					
SC	0.007	0.042	0.029	0.007	0
$SCxMF_1$	0.004	0.064	0.001	0.009	0
$SCxMF_2$	0.004	0.053	0.004	0.007	0
MF	0.005	0.087	0.002	0.010	0
PMM	0.017	0.018	0.041	0.009	0

Table B.6: Comparison of the monte carlo bias ($\widehat{\text{Bias}}_{MC}$), variance ($\widehat{\text{Var}}_{MC}$) and MSE ($\widehat{\text{MSE}}_{MC}$) between the estimator of the first decile in the left-hand table and the ninth decile in the right-hand table SwissCheese (SC) missForest (MF) combinations of the two methods ($SCxMF_1$ and $SCxMF_2$) and predictive mean matching (PMM) with 20 auxiliary variable.

	X_1	X_2	X_3	X_4	X_5
$\widehat{\text{Bias}}_{MC}$					
SC	-0.040	0.003	-0.036	0.043	0
$SCxMF_1$	0.126	0.315	0.060	0.201	0
$SCxMF_2$	0.030	0.113	0.008	0.134	0
MF	0.188	0.410	0.120	0.211	0
PMM	-0.051	0.035	-0.015	-0.010	0
$\widehat{\text{Var}}_{MC}$					
SC	0.029	0.070	0.015	0.045	0
$SCxMF_1$	0.024	0.039	0.014	0.020	0
$SCxMF_2$	0.022	0.063	0.007	0.030	0
MF	0.017	0.021	0.019	0.006	0
PMM	0.018	0.040	0.004	0.035	0
$\widehat{\text{MSE}}_{MC}$					
SC	0.030	0.070	0.017	0.047	0
$SCxMF_1$	0.040	0.138	0.018	0.060	0
$SCxMF_2$	0.023	0.076	0.007	0.048	0
MF	0.052	0.189	0.034	0.050	0
PMM	0.021	0.041	0.005	0.035	0

	X_1	X_2	X_3	X_4	X_5
$\widehat{\text{Bias}}_{MC}$					
SC	0.031	-0.130	0.093	-0.009	0
$SCxMF_1$	-0.016	-0.207	0.008	-0.069	0
$SCxMF_2$	0.008	-0.160	0.043	-0.033	0
MF	-0.050	-0.287	-0.008	-0.089	0
PMM	0.097	-0.060	0.162	0.028	0
$\widehat{\text{Var}}_{MC}$					
SC	0.010	0.017	0.024	0.009	0
$SCxMF_1$	0.007	0.013	0.005	0.003	0
$SCxMF_2$	0.008	0.016	0.013	0.007	0
MF	0.004	0.008	0.001	0.001	0
PMM	0.012	0.010	0.030	0.008	0
$\widehat{\text{MSE}}_{MC}$					
SC	0.011	0.034	0.033	0.009	0
$SCxMF_1$	0.007	0.056	0.005	0.008	0
$SCxMF_2$	0.008	0.042	0.015	0.008	0
MF	0.006	0.090	0.001	0.009	0
PMM	0.022	0.014	0.057	0.009	0

Table B.7: Comparison of Monte Carlo mean square error of correlation estimation between variables of interest for SwissCheese (SC), missForest (MF) combinations of the two methods ($SCxMF_1$ and $SCxMF_2$) and predictive mean matching (PMM) without auxiliary variable.

	X_1	X_2	X_3	X_4	X_5
SC					
X_1	1.000	0.003	0.004	0.005	0.004
X_2	0.003	1.000	0.006	0.003	0.007
X_3	0.004	0.003	1.000	0.003	0.006
X_4	0.006	0.003	0.007	1.000	0.006
X_5	0.005	0.004	0.006	0.006	1.000
$SCxMF_1$					
X_1	1.000	0.006	0.005	0.006	0.005
X_2	0.006	1.000	0.005	0.004	0.005
X_3	0.005	0.004	1.000	0.003	0.006
X_4	0.005	0.003	0.005	1.000	0.007
X_5	0.006	0.005	0.006	0.007	1.000
$SCxMF_2$					
X_1	1.000	0.006	0.005	0.006	0.005
X_2	0.006	1.000	0.005	0.004	0.005
X_3	0.005	0.004	1.000	0.003	0.006
X_4	0.005	0.003	0.005	1.000	0.007
X_5	0.006	0.005	0.006	0.007	1.000
MF					
X_1	1.000	0.009	0.006	0.006	0.006
X_2	0.009	1.000	0.005	0.006	0.005
X_3	0.006	0.006	1.000	0.004	0.006
X_4	0.005	0.004	0.005	1.000	0.008
X_5	0.006	0.006	0.006	0.008	1.000
PMM					
X_1	1.000	0.003	0.004	0.005	0.004
X_2	0.003	1.000	0.006	0.003	0.006
X_3	0.004	0.003	1.000	0.004	0.005
X_4	0.006	0.004	0.006	1.000	0.005
X_5	0.005	0.004	0.005	0.005	1.000

Table B.8: Comparison of Monte Carlo mean square error of correlation estimation between variables of interest for SwissCheese (SC), missForest (MF) combinations of the two methods ($SCxMF_1$ and $SCxMF_2$) and predictive mean matching (PMM) with 10 auxiliary variables.

	X_1	X_2	X_3	X_4	X_5
SC					
X_1	1.000	0.003	0.004	0.004	0.005
X_2	0.003	1.000	0.005	0.005	0.008
X_3	0.004	0.005	1.000	0.003	0.005
X_4	0.005	0.003	0.008	1.000	0.004
X_5	0.004	0.005	0.005	0.004	1.000
$SCxMF_1$					
X_1	1.000	0.003	0.003	0.003	0.004
X_2	0.003	1.000	0.004	0.002	0.004
X_3	0.003	0.002	1.000	0.002	0.004
X_4	0.004	0.002	0.004	1.000	0.003
X_5	0.003	0.004	0.004	0.003	1.000
$SCxMF_2$					
X_1	1.000	0.002	0.003	0.003	0.005
X_2	0.002	1.000	0.004	0.002	0.005
X_3	0.003	0.002	1.000	0.002	0.004
X_4	0.004	0.002	0.005	1.000	0.003
X_5	0.003	0.005	0.004	0.003	1.000
MF					
X_1	1.000	0.004	0.003	0.002	0.002
X_2	0.004	1.000	0.003	0.003	0.004
X_3	0.003	0.003	1.000	0.002	0.002
X_4	0.003	0.002	0.004	1.000	0.002
X_5	0.002	0.002	0.002	0.002	1.000
PMM					
X_1	1.000	0.003	0.004	0.004	0.003
X_2	0.003	1.000	0.005	0.003	0.007
X_3	0.004	0.003	1.000	0.003	0.004
X_4	0.005	0.003	0.007	1.000	0.004
X_5	0.004	0.003	0.004	0.004	1.000

Table B.9: Comparison of Monte Carlo mean square error of correlation estimation between variables of interest for SwissCheese (SC), missForest (MF) combinations of the two methods ($SCxMF_1$ and $SCxMF_2$) and predictive mean matching (PMM) with 20 auxiliary variables.

	X_1	X_2	X_3	X_4	X_5
SC					
X_1	1.000	0.006	0.006	0.003	0.003
X_2	0.006	1.000	0.004	0.007	0.008
X_3	0.006	0.007	1.000	0.003	0.004
X_4	0.004	0.003	0.008	1.000	0.003
X_5	0.003	0.003	0.004	0.003	1.000
$SCxMF_1$					
X_1	1.000	0.003	0.003	0.002	0.002
X_2	0.003	1.000	0.004	0.003	0.005
X_3	0.003	0.003	1.000	0.002	0.003
X_4	0.004	0.002	0.005	1.000	0.002
X_5	0.002	0.002	0.003	0.002	1.000
$SCxMF_2$					
X_1	1.000	0.003	0.003	0.002	0.002
X_2	0.003	1.000	0.003	0.004	0.006
X_3	0.003	0.004	1.000	0.002	0.003
X_4	0.003	0.002	0.006	1.000	0.002
X_5	0.002	0.002	0.003	0.002	1.000
MF					
X_1	1.000	0.003	0.003	0.001	0.001
X_2	0.003	1.000	0.003	0.002	0.003
X_3	0.003	0.002	1.000	0.002	0.001
X_4	0.003	0.002	0.003	1.000	0.001
X_5	0.001	0.001	0.001	0.001	1.000
PMM					
X_1	1.000	0.003	0.004	0.002	0.001
X_2	0.003	1.000	0.005	0.003	0.008
X_3	0.004	0.003	1.000	0.003	0.002
X_4	0.005	0.003	0.008	1.000	0.002
X_5	0.002	0.001	0.002	0.002	1.000

List of Figures

List of Tables

List of Algorithms

Bibliography

Andridge, R. R. and R. J. A. Little (2010). A review of dot deck imputation for survey non-response. *International Statistical Review 78*, 40–64.

Bacharach, M. (1966). Matrix rounding problems. *Management Science 12*(9), 732–742.

Bankier, M. D., S. Rathwell, and M. Majkowski (1992). Two step generalized least squares estimation in the 1991 canadian census. Technical report, Working paper, Census Operations Section, Social Surveys Methods Division, Statistics Canada.

Basu, D. (1971). An essay on the logical foundations of survey sampling. In V. P. Godambe and D. A. Sprott (Eds.), *Foundations of Statistical Inference*, Toronto, pp. 203–233. Holt, Rinehart and Winston.

Beaumont, J.-F. (2008, 09). A new approach to weighting and inference in sample surveys. *Biometrika 95*(3), 539–553.

Beaumont, J.-F. and C. Bocci (2008). Another look at ridge calibration. *Metron 66*(1), 5–20.

Bellhouse, D. R. (1988). A brief history of random sampling methods. In P. R. Krishnaiah and C. R. Rao (Eds.), *Handbook of Statistics Volume 6: Sampling*, New York, Amsterdam, pp. 1–14. Elsevier/North-Holland.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research 13*(1), 1063–1095.

Bivand, R. S., E. Pebesma, and V. Gomez-Rubio (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY.

Boistard, H., H. P. Lopuhaä, and A. Ruiz-Gazen (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electronic Journal of Statistics 6*, 1967–1983.

Bowley, A. L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute 22*, 6–62.

Breidt, F. J. and G. Chauvet (2011). Improved variance estimation for balanced samples drawn via the Cube method. *Journal of Statistical Planning and Inference 141*, 479–487.

Breidt, F. J. and G. Chauvet (2012). Penalized balanced sampling. *Biometrika 99*(4), 945–958.

Breidt, F. J. and J. D. Opsomer (2000). Local polynomial regresssion estimators in survey sampling. *The Annals of Statistics 28*(4), 1026–1053.

Breidt, F. J. and J. D. Opsomer (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science 32*(2), 190–205.

Breiman, L. (1994). Bagging predictors. Technical Report 421, University of California, Berkeley.

Breiman, L. (1996). Bagging predictors. *Machine learning 24*, 123–140.

Breiman, L. (2001a). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science 16*(3), 199–231.

Brewer, K. R. W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics 5*, 5–13.

Brewer, K. R. W. (2013). Three controversies in the history of survey sampling. *Survey Methodology 39*(2), 249–262.

Brick, M. J. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics 29*(3), 329–353. cited By 32.

Burgard, J. P., R. Münnich, and M. Rupp (2019). A Generalized Calibration Approach Ensuring Coherent Estimates with Small Area Constraints. Research Papers in Economics 2019-10.

Burrough, P. A., R. A. McDonnell, and C. D. Lloyd (2015). *Principles of Geographical information systems*. Oxford University Press.

Cardot, H., C. Goga, and M.-A. Shehzad (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica 27*(1), 243–260.

Cassel, C. M., C.-E. Särndal, and J. H. Wretman (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika 63*, 615–620.

Chambers, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics 12*, 3–32.

Chauvet, G., J.-C. Deville, and D. Haziza (2011). On balanced random imputation in surveys. *Biometrika 98*, 459–471.

Chauvet, G. and C. Goga (2022). Asymptotic efficiency of the calibration estimator in a high-dimensional data setting. *Journal of Statistical Planning and Inference 217*, 177–187.

Chauvet, G., D. Haziza, and É. Lesage (2017). Examining some aspects of balanced sampling in surveys. *Statistica Sinica*, 313–334.

Chauvet, G. and Y. Tillé (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics 21*, 9–31.

Chen, J. and J. Shao (1997). Nearest neighbor imputation for survey data. *Journal of Official Statistics 13*(2), 123–132.

Chen, J. and R. R. Sitter (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica 9*(2), 385–406.

Chen, J., R. R. Sitter, and C. Wu (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika 89*(1), 230–237.

Chen, J. K. T., R. Valliant, and M. R. Elliott (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 68*(3), 657–681.

Chen, S. and D. Haziza (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika 104*(2), 439–453.

Chen, S. and D. Haziza (2019). Recent developments in dealing with item non-response in surveys: A critical review. *International Statistical Review 87*, S192–S218.

Choudhry, G. H. and M. P. Singh (1979). Sampling with unequal probabilities and without replacement – a rejective method. *Survey Methodology 5*, 162–177.

Cochran, W. G. (1953). *Sampling Techniques*. New York: Wiley.

Cochran, W. G. (1977). *Sampling Techniques*. New York: Wiley.

Cox, L. H. (1987, June). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association 82*, 520–524.

Dagdoug, M., C. Goga, and D. Haziza (2023a). Imputation Procedures in Surveys Using Nonparametric and Machine Learning Methods: An Empirical Comparison. *Journal of Survey Statistics and Methodology 11*(1), 141–188.

Dagdoug, M., C. Goga, and D. Haziza (2023b). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association 118*(542), 1234–1251.

Dagdoug, M., C. Goga, and D. Haziza (2025). Statistical inference in the presence of imputed survey data through regression trees and random forests. *Scandinavian Journal of Statistics*.

David, I. P. and B. Sukhatme (1974). On the bias and mean square error of the ratio estimator. *Journal of the American Statistical Association 69*(346), 464–466.

Deming, W. E. and F. F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics 11*(4), 427–444.

Deville, J.-C., C.-E. Särndal, and O. Sautory (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association 88*, 1013–1020.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*(418), 376–382.

Deville, J.-C. and Y. Tillé (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika 85*, 89–101.

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

Deville, J.-C. and Y. Tillé (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference 128*, 569–591.

Dickson, M. M. and Y. Tillé (2016). Ordered spatial sampling by means of the traveling salesman problem. *Computational Statistics 31*(4), 1359–1372.

Doerr, B., T. Friedrich, C. Klein, and R. Osbild (2006). Unbiased matrix rounding. In *Algorithm Theory–SWAT 2006: 10th Scandinavian Workshop on Algorithm Theory, Riga, Latvia, July 6-8, 2006. Proceedings 10*, pp. 102–112. Springer.

Dupačová, J. (1979). A note on rejective sampling. In *Contribution to Statistics (Jaroslav Hájek memorial volume)*, pp. 71–78. Academia Prague.

Eustache, E., R. Jauslin, and Y. Tillé (2022). Spatiotemporal sampling with spatial spreading and rotation of units in time. *Spatial Statistics Accepted*, 1–15.

Eustache, E., A.-A. Vallée, and Y. Tillé (2021). *The SwissCheese R Package*. R package version beta.

Eustache, E., A.-A. Vallée, and Y. Tillé (2024). Balanced donor imputation handling swiss cheese nonresponse. *Statistica Sinica 34*, 637–655.

Fellegi, I. P. (1975). Controlled random rounding. *Survey Methodology*, 123–133.

Fetter, M., J. Gentle, and C. Perry (2005). Calibration adjustments when not all targets can be met. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 3031–3035.

Folsom, R. E. and A. C. Singh (2000). The generalized exponential model for design weight calibration for extreme values, nonresponse and poststratification. In *Section on Survey Research Methods, American Statistical Association*, pp. 598–603.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika 96*, 933–944.

Fuller, W. A. and J. K. Kim (2005). Hot-deck imputation for the response model. *Survey Methodology 31*, 139–149.

Fuller, W. A., J. C. Legg, and Y. Li (2017). Bootstrap variance estimation for rejective sampling. *Journal of the American Statistical Association 112*(520), 1562–1570.

Garès, V., C. Dimeglio, G. Guernec, R. Fantin, B. Lepage, M. R. Kosorok, and N. Savy (2020). On the use of optimal transportation theory to recode variables and application to database merging. *The International Journal of Biostatistics 16*(1), 20180106.

Garès, V. and J. Omer (2022). Regularized optimal transport of covariates and outcomes in data recoding. *Journal of the American Statistical Association 117*(537), 320–333.

Gini, C. and L. Galvani (1929). Di una applicazione del metodo rappresentativo al censimento italiano della popolazione (1. dicembre 1921). *Annali di Statistica Series 6, 4*, 1–107.

Godambe, V. P. (1955). A unified theory of sampling from finite population. *Journal of the Royal Statistical Society B17*, 269–278.

Goga, C. and M. A. Shehzad (2014). A note on partially penalized calibration. *Pakistan Journal of Statistics 30*(4), 429–438.

Grafström, A. (2011). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference 142*, 139–147.

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

Grafström, A. and N. L. P. Lundström (2013). Why well spread probability samples are balanced? *Open Journal of Statistics 3*(1), 36–41.

Grafström, A., N. L. P. Lundström, and L. Schelin (2012). Spatially balanced sampling through the pivotal method. *Biometrics 68*(2), 514–520.

Grafström, A., S. Saarela, and L. T. Ene (2014). Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Canadian Journal of Forest Research 44*(10), 1156–1164.

Grafström, A. and Y. Tillé (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics 14*(2), 120–131.

Guandalini, A. and C. Ceccarelli (2022). Impact measurement and dimension reduction of auxiliary variables in calibration estimator using the shapley decomposition. *Statistical Methods & Applications 31*, 759–784.

Guggemos, F. and Y. Tillé (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference 140*(11), 3199–3212.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics 35*, 1491–1523.

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.

Hansen, M. H. and W. N. Hurwitz (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics 14*, 333–362.

Hasler, C., R. V. Craiu, and L.-P. Rivest (2018). Vine copulas for imputation of monotone non-response. *International Statistical Review 86*(3), 488–511.

Hasler, C. and Y. Tillé (2016). Balanced k-nearest neighbor imputation. *Statistics 105*, 11–23.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In D. Pfeffermann and C. R. Rao (Eds.), *Sample surveys: Design, methods and applications*, New York, Amsterdam, pp. 215–246. Elsevier/North-Holland.

Haziza, D. and J.-F. Beaumont (2017). Construction of weights in surveys: A review. *Statistical Science 32*(2).

Haziza, D. and É. Lesage (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics 32*(1), 129–145.

Haziza, D. and J. N. K. Rao (2003). Inference for totals in cluster sampling under mean imputation for missing data. In *Proceedings of the Statistics Canada Symposium*.

Haziza, D. and J. N. K. Rao (2006). A nonresponse model approach to inference under imputation for missing survey data. *Techniques d'Enquête 32*(1), 59–71.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–685.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika 28*(3/4), 321.

Jauslin, R., E. Eustache, B. Panahbehagh, and Y. Tillé (2021). *StratifiedSampling: Different Methods for Stratified Sampling*. Vienna, Austria: R Foundation for Statistical Computing. R package version 0.3.0.

Jauslin, R., B. Panahbehagh, and Y. Tillé (2022). Sequential spatially balanced sampling. *Environmetrics 33*, 1–17.

Jauslin, R. and Y. Tillé (2023). An efficient approach for statistical matching of survey data through calibration, optimal transport and balanced sampling. *Journal of Statistical Planning and Inference 225*, 121–131.

Jauslin, R. and Y. Tillé (2019). *WaveSampling: Weakly Associated Vectors (WAVE) Sampling*. R package version 0.1.0.

Jauslin, R. and Y. Tillé (2020). Spatial spread sampling using weakly associated vectors. *Journal of Agricultural, Biological and Environmental Statistics 25*(3), 431–451.

Judkins, D. R. (1997). Imputing for Swiss cheese patterns of missing data. In *Proceedings of Statistics Canada Symposium*, Volume 97, Statistics Canada, pp. 143–148.

Kalton, G. and I. Flores-Cervantes (2003). Weighting methods. *Journal of Official Statistics 19*(2), 81–97.

Kalton, G. and D. Kasprzyk (1986). The treatment of survey missing data. *Survey Methodology 12*, 1–16.

Kiær, A. N. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique 9*, 176–183.

Kiær, A. N. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique 11*, 180–185.

Kiær, A. N. (1903). Sur les méthodes représentatives ou typologiques. *Bulletin de l'Institut International de Statistique 13*, 66–78.

Kiær, A. N. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique 14*, 119–134.

Kim, J.-K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica 19*(1), 145–157.

Kim, J. K. and W. A. Fuller (2004). Fractional hot-deck imputation. *Biometrika 91*, 559–578.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Korte, B. and J. Vygen (2018). *Combinatorial Optimization: Theory and Algorithms*. Algorithms and Combinatorics. Springer Berlin Heidelberg.

Kott, P. S. (2006). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association 101*(475), 1268–1276.

Langel, M. and Y. Tillé (2011). Corrado Gini, a pioneer in balanced sampling and inequality theory. *Metron 69*, 45–65.

Legg, J. C. and C. L. Yu (2010). Comparison of sample set restriction procedures. *Survey Methodology 36*, 69–79.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics 6*, 287–296.

Marker, D. A., D. R. Judkins, and M. Winglee (2002). Large-scale imputation for complex surveys. *Survey nonresponse 329341*, 329—341.

Mayor-Gallego, J., J. Moreno-Rebollo, and M. Jiménez-Gamero (2019). Estimation of the finite population distribution function using a global penalized calibration method. *AStA Advances in Statistical Analysis 103*(1), 1–35.

McConville, K. S., F. J. Breidt, T. C. M. Lee, and G. G. Moisen (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology 5*(2), 131–158.

Montanari, G. E. and M. G. Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association 100*(472), 1429–1442.

Montanari, G. E., M. G. Ranalli, et al. (2009). Multiple and ridge model calibration for sample surveys. In *Workshop on Calibration and Estimation in Surveys: Proceedings*, pp. 1–13. Statistics Canada.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika 37*(1/2), 17–23.

Morris, T. P., I. R. White, and P. Royston (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC medical research methodology 14*, 1–13.

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*, 558–606.

Oliva-Avilés, C., M. C. Meyer, and J. D. Opsomer (2020). Estimation and inference of domain means subject to qualitative constraints. *Survey Methodology 46*(2), 145–180.

Pea, J., L. Qualité, and Y. Tillé (2007). Systematic sampling is a minimal support design. *Computational Statistics & Data Analysis 51*, 5591–5602.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine 2*, 559–572.

Pebesma, E. J. and R. S. Bivand (2005, November). Classes and methods for spatial data in R. *R News 5*(2), 9–13.

Peytchev, A. (2012, 03). Multiple Imputation for Unit Nonresponse and Measurement Error. *Public Opinion Quarterly 76*(2), 214–237.

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, and P. W. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology 27*(1), 85–95.

Ranalli, M. G., A. Arcos, M. d. M. Rueda, and A. Teodoro (2016). Calibration estimation in dual-frame surveys. *Statistical Methods & Applications 25*, 321–349.

Rancourt, E. and J. Chen (1994). Nearest neighbor imputation for categorical variables. *Proceedings of the Survey Research Methods Section, ASA*, 300–305.

Rao, J. N. K. and I. Molina (2015). *Small Area Estimation*. New York: Wiley.

Rao, J. N. K. and A. C. Singh (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. In *Section on Survey Research Methods, American Statistical Association, Washington DC, 57-65.*

Rao, J. N. K. and A. C. Singh (2009). Range restricted weight calibration for survey data using ridge regression. *Pakistan Journal of Statistics 25*(4).

Rao, J. N. K. and R. R. Sitter (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika 82*, 453–460.

Robertson, B. L., J. A. Brown, T. L. McDonald, and P. Jaksons (2013). Bas: Balanced acceptance sampling of natural resources. *Biometrics 69*(3), 776–784.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika 57*, 377–387.

Royall, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association 71*, 657–664.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics 4*(1), 87–94.

Rubin, D. B. (1991). Multiple imputation after 18+ years. *Journal of the American Statistical Association 91*, 473–489.

Santacatterina, M. and M. Bottai (2018). Optimal probability weights for inference with constrained precision. *Journal of the American Statistical Association 113*(523), 983–991.

Särndal, C.-E. (2007). The calibration approach un survey theory and practice. *Survey Methodology 33*, 99–119.

Särndal, C.-E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse.* New York: Wiley.

Särndal, C.-E., B. Swensson, and J. H. Wretman (1992). *Model Assisted Survey Sampling.* New York: Springer.

Schafer, J. L. and J. W. Graham (2002). Missing data: Our view of the state of the art. *Psychological Methods 7*(2), 147–177.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics 5*, 119–127.

Shao, J. (2000). Cold deck and ratio imputation. *Survey Methodology 26*, 79–85.

Shao, J. and H. Wang (2008). Confidence intervals based of survey data with nearest neighbor imputation. *Statistica Sinica 18*, 281–298.

Silva, P. L. N. and C. J. Skinner (1997). Variable selection for regression estimation in finite populations. *Survey Methodology 23*(1), 23–32.

Srinivasan, A. (2001). Distributions on level-sets with applications to approximation algorithms. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pp. 588–597. IEEE.

Stekhoven, D. J. (2022). *missForest: Nonparametric Missing Value Imputation using Random Forest*. R package version 1.5.

Stekhoven, D. J. and P. Bühlmann (2012). Missforest – non-parametric missing value imputation for mixed-type data. *Bioinformatics 28*(1), 112–118.

Stevens, Jr., D. L. and A. R. Olsen (1999). Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics 4*, 415–428.

Stevens, Jr., D. L. and A. R. Olsen (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics 14*(6), 593–610.

Stevens, Jr., D. L. and A. R. Olsen (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association 99*(465), 262–278.

Swiss Federal Statistical Office (2015). Survey on Income and Living Conditions in Switzerland. Data set of the 2015 survey.

Tang, F. and H. Ishwaran (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal 10*(6), 363–377.

Thionet, P. (1953). *La théorie des sondages*. Paris: Institut National de la Statistique et des Études Économiques, Études théoriques vol. 5, Imprimerie nationale.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.

Tillé, Y. (2011). Ten years of balanced sampling with the cube method: an appraisal. *Survey Methodology 37*, 215–226.

Tillé, Y. (2016). The legacy of Corrado Gini in survey sampling and inequality theory. *Metron 74*(2), 167–174.

Tillé, Y., M. M. Dickson, G. Espa, and D. Giuliani (2018). Measuring the spatial balance of a sample: A new measure based on the Moran's I index. *Spatial Statistics 23*, 182–192.

Tillé, Y. and A. Matei (2021). *sampling: Survey Sampling*. R package version 2.9.

Tillé, Y. and A. Matei (2023). *sampling: Survey Sampling*. R package version 2.10.

van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition*. Chapman and Hall/CRC.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software 45*(3), 1–67.

van Buuren, S., K. Groothuis-Oudshoorn, A. Robitzsch, G. Vink, L. Doove, S. Jolani, et al. (2015). Package 'mice'. *Computer software*.

Villani, C. et al. (2009). *Optimal Transport: Old and New*, Volume 338. New-York: Springer.

Vink, G., L. E. Frank, J. Pannekoek, and S. Van Buuren (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica 68*(1), 61–90.

Waljee, A. K., A. Mukherjee, A. G. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu, and P. D. Higgins (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open 3*(8), 1–7.

Williams, M. R. and T. D. Savitsky (2024). Optimization for calibration of survey weights under a large number of conflicting constraints. *Journal of Computational and Graphical Statistics 33*(3), 1047–1060.

Wu, C. and R. R. Sitter (2001). A model calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*, 185–193.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. London: Charles Griffin.

Yates, F. and P. M. Grundy (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B15*, 235–261.

Zimek, A., E. Schubert, and H.-P. Kriegel (2012, August). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining 5*(5), 363–387.