https://doi.org/10.35662/unine-thesis-3078

Université de Neuchâtel

Nonresponse in sample surveys: new estimation and inference methods

---

Thesis submitted to satisfy the requirements for the degree of Doctorat ès Sciences

by

Esther EUSTACHE

Accepted by the dissertation committee

Dr. Caren Hasler	*Université de Neuchâtel, Switzerland*	Jury president
Prof. Yves Tillé	*Université de Neuchâtel, Switzerland*	Thesis director
Prof. Camélia Goga	*Université de Franche-Comté, France*	Rapporteur
Dr. Alessio Guandalini	*Istituto nazionale di statistica, Italy*	Rapporteur
Prof. Anne Ruiz-Gazen	*Université Toulouse 1 Capitole, France*	Rapporteur

Defended on November 30, 2023

unine
Université de Neuchâtel
Faculté des sciences

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

IMPRIMATUR POUR THÈSE DE DOCTORAT

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

Madame Esther EUSTACHE

Titre :

"Nonresponse in sample surveys : new estimation and inference methods"

sur le rapport des membres du jury composé comme suit :

- Prof. Yves Tillé, directeur de thèse, Université de Neuchâtel,
- Dr Caren Hasler, rapporteure, Université de Neuchâtel
- Prof. Camelia Goga, rapporteure, Université de Franche-Comté, France
- Prof. Anne Ruiz-Gazen, rapporteure, Université Toulouse, France
- Dr Alessio Guandalini, rapporteur, Instituto nazionale di statistica, Rome, Italie

Neuchâtel, le 1er décembre 2023

Le Doyen, Prof. R. Bshary

Je dédie ce manuscrit à mon papi
qui s'est éteint le 7 octobre 2023.

Abstract

In this manuscript, the topic of nonresponse in surveys is studied through several research projects. We address different subjects as treatment of nonresponse in data sets, estimation in presence of nonresponse and analysis of variance estimators. The first chapter of this thesis is devoted to an introduction to survey sampling and to the concepts used in the rest of this manuscript. In the second chapter, a new method for donor imputation of nonresponse is proposed. In the chapter 3, we present a new estimator of population total in the presence of nonresponse, which combines model-assisted estimators and estimators weighted by nonresponse. In a fourth chapter, we raise a problem related to variance estimators of total estimator when the ratio between the number of variables and the number of observations is large. We approximate the bias of the variance estimators in order to adjust the estimators. The chapter 5 of this manuscript deals with a different subject. We propose a new spatiotemporal sampling method that takes into account both sources of spatial and temporal autocorrelations. The method enables us to select a sample that is well spread in time and in space.

Keywords: imputation, model-assisted estimator, jackknife variance estimator, spatiotemporal sampling.

Résumé

Dans ce manuscrit, le sujet de la non-réponse dans les enquêtes est étudié à travers plusieurs projets de recherche. Nous abordons différents sujets tels que le traitement de la non-réponse dans des ensembles de données, l'estimation en présence de non-réponse et l'analyse de la variance d'estimateurs. Le premier chapitre de cette thèse est consacré à une introduction à la statistique d'enquête, et notamment aux différents concepts utilisés dans le reste de ce manuscrit. Dans le deuxième chapitre, une nouvelle méthode d'imputation des non-réponses par donneurs est proposée. Au chapitre 3, nous présentons un nouvel estimateur du total en présence de non-réponses, qui combine l'estimateur assisté par un modèle et l'estimateur ajusté par repondération pour la non-réponse. Dans un quatrième chapitre, nous soulevons un problème lié aux estimateur de variance d'estimateurs de totaux lorsque le rapport entre le nombre de variables et le nombre d'observations est grand. Nous approximons le biais de l'estimateur de variance afin d'ajuster les estimateurs. Le chapitre 5 de ce manuscrit traite d'un tout autre sujet. Nous y proposons une nouvelle méthode d'échantillonnage spatio-temporel qui prend en compte les deux sources d'autocorrélation spatiale et temporelle. La méthode permet de sélectionner un échantillon bien étalé à la fois dans le temps et dans l'espace.

Mots-clés : imputation, estimateur assisté par modèle, estimateur de variance jackknife, échantillonnage spatiotemporel.

Remerciements

En premier lieu, je voudrais exprimer ma profonde reconnaissance à Yves, mon directeur de thèse, pour sa confiance, sa bienveillance et sa bonne humeur quotidienne. J'ai énormément appris à ses côtés durant ces quatres années de thèse, en théorie des sondages mais aussi dans beaucoup d'autres domaines. Les connaissances, l'intuition et la mémoire d'Yves me surprendront toujours ! Grâce à lui, j'ai également pu mener à bien ce double projet qui me tenait tant à cœur, melant thèse et sport de haut niveau. Yves m'a toujours soutenu et je sais qu'être mon directeur de thèse n'a pas été de tout repos. J'espère que ces années à ses côtés m'auront permis d'acquérir la rigueur et la confiance en moi qu'il a si souvent demandées. Merci, Chef.

Je tiens à exprimer ma gratitude à David Haziza, pour m'avoir accueillie quatre mois sur le sol canadien durant ma thèse. Grâce aux nombreux échanges et travaux de recherche réalisés à Ottawa avec David et Mehdi Dagdoug, j'ai pu m'ouvrir à d'autres thématiques. Ce sont des chercheurs passionnés et passionnants, j'ai énormément appris à leur côtés et je les remercie sincèrement.

Je tiens à remercier Camélia Goga, Anne Ruiz-Gazen et Alessio Guandalini d'avoir accepté d'être mes rapporteurs de thèse. Je remercie plus particulièrement Camélia, pour la transmission de sa passion pour les sondages durant mes études de master, ainsi que pour son soutien et son aide afin que je puisse réaliser cette thèse aux côtés d'Yves.

Je souhaite donner un remerciement spécial à ma dernière rapportrice et collègue, Caren Hasler. Caren est une femme inspirante et une chercheuse douée. Je suis chanceuse d'avoir pu évoluer et travailler à ses côtés.

J'adresse également mes remerciements à l'ensemble de mes collègues de l'Institut de Statistique. Merci Raphaël pour la planche à voile, Arnaud pour nos plantations et nos chamailleries, Caren, à nouveau, pour nos débats féministes, Ejub pour tes bons petits gâteaux, Lionel pour ton humour, Michaël pour ta gentillesse. Merci à Alina, Corine et Ziqing. Les vendredis Buvette risquent de beaucoup me manquer.

Mes derniers remerciements vont à ma famille et mes amis, pour leur soutien indéfectible. Merci notamment à ma petite soeur Emeline pour son écoute et sa bienveillance. Finalement, je tiens à remercier Ben, mon binôme au quotidien, pour son écoute et pour me faire évoluer sans cesse. Combiner une thèse et un sport de haut niveau n'aurait jamais été possible sans un entourage aussi présent et conciliant. Je leur suis infiniment reconnaissante.

Contents

Introduction

The objective of survey sampling is to provide information about a population, using only information from a portion of that population. This manuscript focuses solely on sample surveys in *finite populations*. For example, the aim may be to estimate population totals or means. From sample selection and data collection to estimation and final analysis, a large number of different operations are performed as part of a survey, each of which can influence the final estimates. *Nonresponse* is present in almost all surveys. Estimates based on observed data only, omitting nonresponses, tend to be biased, unless the data are missing completely at random (Rubin, 1976). Nonresponse is one of the many sources of error in surveys that may considerably affect the final estimated parameters, adding further bias and causing additional variability. Traditionally, there are two ways of reducing these negative effects: replacing nonresponses with plausible values or reweighting survey respondents to compensate for nonrespondents. Nonresponse can appear in different ways in the collected data, requiring different treatment depending on how it occurs.

The research projects presented below are the result of collaborations between a number of renowned researchers. They enabled me to work on a wide range of subjects in sample surveys. Various stages of a survey where nonresponse complicates conventional methods and requires an adapted process are considered. This thesis is divided into five parts. Chapter 1 is a preamble introducing the notations and basic concepts of sampling necessary for understanding the rest of this manuscript. After this first part, the papers written during this thesis are presented one by one. The last part is a scientific paper that is not related to the subject of nonresponse, but we decided to include it because it represents a significant amount of work produced at the beginning of my thesis.

Chapter 2 of this thesis is devoted to the treatment of nonresponse using a new donor imputation method. This work corresponds to the published papers Eustache, Vallée, and Tillé (2024). It was prompted by a request from the Swiss Federal Office for a single-donor imputation method to avoid problems of inconsistency in multivariate imputations. The proposed imputation method is adapted to the most general case: when nonresponse is multivariate and can appear anywhere in the data set without any pattern. The method provides donor imputation with balancing constraints in the process. It extends the method proposed by Hasler and Tillé (2016) to multivariate nonresponse. The method meets three main requirements: 1. it is a donor imputation method; 2. each unit with nonresponses is imputed by the values of a similar unit; 3. balancing constraints must be satisfied.

Chapters 3 presents the article Eustache and Hasler (2022) which is submitted for publication. It introduces a population total estimator that mixes model-assisted and nonresponse weighted adjustment estimators is proposed. Model-assisted estimators rely on a postulated working model between the survey variable and the auxiliary variables. One of the main advantages of model-assisted estimators is that they retain important properties such as consistency and the absence of asymptotic bias, whether or not the working model is correctly specified. However, if the survey variable contains nonresponses, this class of estimators can no longer be used. A common way of dealing with nonresponse is to reweight survey respondents to compensate for nonrespondents. The proposed estimator is a 'quasi-model-assisted' estimator. We consider nonresponse as a second phase of the survey and reweight the units in model-assisted estimators using the inverse of estimated response probabilities, in order to compensate for the nonrespondents. We provide formulae for asymptotic variance and variance estimators.

Chapter 4 presents the work in progress Eustache, Dagdoug, and Haziza (2023) which will be soon submitted. It is devoted to the study of variance estimators of total and mean estimators. We focus on jackknife variance estimator proposed in Berger and Skinner (2005) and Berger and Rao (2006) and on estimator based on the first-order Taylor expansion. We consider the model-assisted total estimator and the imputed mean estimator, with deterministic linear regression as working model. We highlight a dimensional problem of the resampling jackknife method commonly used to estimate the variance of the imputed mean estimator. A high-dimensional setting refers to a situation where the number of predictors used in the working model is of the same order of magnitude as the sample size. A new expression for the generalized jackknife estimator is proposed to avoid this problem.

Finally, Chapter 5 differs slightly from the other parts of this manuscript as it does not deal with nonresponse. Chapter 5 corresponds to the published paper Eustache, Jauslin, and Tillé (2022) in which we present a new spatiotemporal sampling design for equal and unequal inclusion probabilities. In spatiotemporal data, there are two sources of autocorrelation – spatial and temporal. The proposed sampling design manages both autocorrelations by selecting units that are spread in time and in space, using systematic sampling and the pivotal method. We provide simulations on a real *odonata* dataset to demonstrate the effectiveness of the method.

Chapter 1

Introduction to Sample Surveys

1.1 Introduction

This chapter is devoted to an introduction to statistical methods and basic concepts that are used in the following chapters of this manuscript. To better understand the issues involved in a survey, the sampling strategy is presented in Section 1.2. In Section 1.3, we discuss the statistical properties that enable us to obtain an accurate and efficient estimator. Auxiliary information is very useful in a survey and can be used in different stages. We present how auxiliary information can be used at the estimation stage in Section 1.4. Nonresponse is a major problem that is difficult to avoid during a survey. Section 1.5 is devoted to nonresponse. Section 1.6 provides an introduction to variance estimation in the context of complete response, while Section 1.7 discusses this topic in the case of imputation. For more information on sampling theory, the reader is advised to read Tillé (2020) and Särndal, Swensson, and Wretman (1992).

1.2 Sampling strategy

Let us consider a set U of $N \in \mathbb{N}^*$ discrete elements. In survey sampling, this set is called the *finite population of interest*. The elements of U are labeled according to their position in the population, i.e. the k-th element of U is labeled by k and $U = \{1, \ldots, k, \ldots, N\}$.

Let y_k be the value of a survey variable Y for element $k \in U$. The aim of a survey is to estimate an unknown parameter, denoted by θ_y, of the population of interest, related to survey variable Y. This unknown parameter is a function $\theta(.)$ of values $\{y_k\}_{k\in U}$, such that $\theta_y = \theta(\{y_k\}_{k\in U})$. Common parameters of interest are the population total

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

and the population mean $\mu_y = t_y/N$. Almost always, the Y-value is not known for all elements of the population, due to cost, time or simply the impossibility of collecting data, making this parameter impossible to compute. A sampling procedure is therefore essential. In particular, it provides a point estimate of the parameter of interest, using the Y-values of only a subset s, called a *sample*, of the population.

The first major choice in a survey is the selection of the subset $s \subset U$. Throughout this work, each sample s refers to a *probability sample*, meaning that it has been selected using a *probability sampling scheme* without replacement. We define $\mathscr{S}$ as the set containing all non-empty subsets of U. A probability sampling scheme without replacement assigns to each sample s a probability $p(s)$ of being selected. The function $p(.)$ is called the *sampling design*. It is a probability distribution function on $\mathscr{S}$, with $p(s) \geq 0, s \in \mathscr{S}$ and $\sum_{s\in\mathscr{S}} p(s) = 1$. Note that s can be viewed as a realization of a random variable S whose definition set is $\mathscr{S}$, and

$$\mathbb{P}(S = s) = p(s), \quad s \subset \mathscr{S}.$$

The number n_s of elements selected in sample s can be fixed or random, depending on $p(.)$. If n_s is fixed, the sampling design is said to be of *fixed size*, and we have $n_s = n$, regardless of

the selected sample s. In this case, only samples $s \subset \mathscr{S}$ with a number of elements n have a probability larger than 0 of being selected.

One way of representing the selected elements other than the subset s is to use a *sample membership indicator variable*. The sample membership indicator variable of an element k is $I_k(S) := \mathbb{1}_{k \in S}$. For simplicity, we use I_k instead of $I_k(S)$. In the case of a sampling design without replacement, the variable I_k follows a Bernoulli distribution of parameter π_k, such that

$$\mathbb{P}(I_k = 1) = \pi_k \quad \text{and} \quad \mathbb{P}(I_k = 0) = 1 - \pi_k,$$

and thus

$$\mathbb{E}(I_k) = \pi_k \quad \text{and} \quad \mathbb{V}(I_k) = \pi_k(1 - \pi_k).$$

The *first-order inclusion probability* of element k in the sample is π_k. The *second-order inclusion probability* $\pi_{k\ell}$ is the probability that both elements k and $\ell \in U$ are selected in the sample. Throughout this manuscript, we assume

$$\pi_k > 0, \quad \text{for all } k \in U,$$

guaranteeing that each population unit has a chance of being selected in the sample, and also $\pi_{k\ell} > 0$, for all pairs $(k, \ell) \in U \times U$. Depending on the sampling design $p(.)$, the corresponding inclusion probabilities may be equal or unequal and are generally known or chosen prior to sampling. The following examples introduce some well-used sampling designs. More complex sampling designs are presented in Tillé (2020).

Example 1.2.1 (Simple random sampling without replacement)**.** *Simple random sampling without replacement gives every possible sample of size n in U the same probability of being selected. The size n is fixed and chosen beforehand. We have* $p(s) = \binom{N}{n}^{-1}$ *if s is of size n, 0 otherwise. All elements of the population have the same probability* $\pi = n/N$ *of being selected, so* $\pi_k = \pi > 0$ *for all* $k \in U$*, and*

$$\pi_{k\ell} = \frac{n(n-1)}{N(N-1)},$$

for all $k, \ell \in U$ *and* $k \neq \ell$*.*

Example 1.2.2 (Bernoulli sampling)**.** *Bernoulli sampling gives every element of the population the same probability* π *of being selected, so* $\pi_k = \pi > 0$ *for all* $k \in U$*. The sample membership indicators* $\{I_k\}_{k \in U}$ *are independent and identically distributed random variables. It follows that* $\pi_{k\ell} = \pi^2$ *for all* $k, \ell \in U$ *and* $k \neq \ell$*. The Bernoulli design is expressed by*

$$p(s) = \pi^{n_s}(1 - \pi)^{N - n_s}.$$

The sample size n_s *is a random variable, that follows a binomial distribution with mean* $N\pi$ *and variance* $N\pi(1 - \pi)$*. Not controlling the sample size during a survey can be a disadvantage. For example, the cost of the survey or the number of surveyors required cannot be known in advance.*

Example 1.2.3 (Poisson sampling)**.** *Unlike simple random sampling and Bernoulli sampling, Poisson sampling allows for unequal inclusion probabilities. Poisson sampling assigns to each sample* $s \in \mathscr{S}$ *the probability*

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \in U \setminus s} (1 - \pi_k)$$

of being the selected sample. The sample size n_s *is a random variable. Because of the independence of variables* $\{I_k\}_{k \in U}$*,* $\pi_{k\ell} = \pi_k \pi_\ell$*. As the inclusion probabilities are not limited to a specific value, It is worth noting that if the inclusion probabilities are equal, the Poisson sampling is equivalent to Bernoulli sampling.*

The second major choice in a survey is the statistic used to compute an estimate of the unknown parameter of interest θ_y, based on values collected on the sample. An *estimator* $\widehat{\theta}_y$ of θ_y is a *statistic*, a function of the Y-values on the random sample S, $\widehat{\theta}_y = \widehat{\theta}_y(\{y_k\}_{k\in S})$. Let $\widehat{\theta}_y(s)$ denote the value taken by the estimator on the selected sample s. It provides a *point estimate* of θ_y. Well-known estimators of the population total t_y and the population mean μ_y are respectively the *Horvitz-Thompson estimator* (Horvitz and Thompson, 1952)

$$\widehat{t}_{y,\pi} := \sum_{k\in S} \frac{y_k}{\pi_k} \tag{1.1}$$

and the *Hájek estimator* (Hájek, 1971)

$$\widehat{\mu}_{y,H} := \frac{1}{\widehat{N}} \sum_{k\in S} \frac{y_k}{\pi_k}$$

where $\widehat{N} = \sum_{k\in S} \pi_k^{-1}$.

The choices of the sampling design $p(.)$ and the estimator $\widehat{\theta}_y$ are not dissociated from each other. In sample surveys, the aim is to find the *sampling strategy* $(p, \widehat{\theta}_y)$ that will give the most accurate point estimate of the parameter possible.

1.3 Statistical properties of an estimator

At the end of a survey, the final point estimate $\widehat{\theta}_y(s)$ should be as close as possible to the real value. However, it varies from one sample to another. A good estimator not only gives estimates close to the true values, but also varies little. The *expectation* and the *variance* with respect to the sampling design $p(.)$ of $\widehat{\theta}_y$ are respectively

$$\mathbb{E}_p(\widehat{\theta}_y) = \sum_{s\in\mathscr{S}} p(s)\widehat{\theta}_y(s) \quad \text{and} \quad \mathbb{V}_p(\widehat{\theta}_y) = \sum_{s\in\mathscr{S}} p(s)\left\{\widehat{\theta}_y(s) - \mathbb{E}_p(\widehat{\theta}_y)\right\}^2.$$

Two important measures for the quality of an estimator are the *bias* and the *mean squared error*. The bias of $\widehat{\theta}_y$ under the sampling design is defined as $B_p(\widehat{\theta}_y) = \mathbb{E}_p(\widehat{\theta}_y) - \theta_y$. When it comes to sampling, it is important for an estimator to be *design-unbiased*, i.e. $B_p(\widehat{\theta}_y) = 0$, for all variables of interest Y, or approximately unbiased, i.e. to have a bias close to 0 for large sample sizes. The mean squared error (MSE) of $\widehat{\theta}_y$ under the sampling design is defined by

$$MSE_p(\widehat{\theta}_y) = \sum_{s\in\mathscr{S}} p(s)\left\{\widehat{\theta}_y(s) - \theta_y\right\}^2.$$

It can be demonstrated that $MSE_p(\widehat{\theta}_y) = \mathbb{V}_p(\widehat{\theta}_y) + B_p(\widehat{\theta}_y)^2$. A low MSE implies both low variance and low bias, indicating good estimator behavior.

The presented measure of efficiency consider that the source of randomness comes from the sampling design, i.e, the sample S is a random variable. This approach is called *design-based* inference. Another approach, formalized in Brewer (1963), Royall (1970) and others, is called *model based* and considers $\{y_k\}_{k\in N}$ as the realized outcome of random variables. The main difference between design-based and model-based philosophies is the source of randomness in the estimator (Särndal, 1978).

Let us focus on properties of the population total estimator $\widehat{t}_{y,\pi}$ of (1.1). First, $\widehat{t}_{y,\pi}$ is a design-unbiased estimator of t_y (Horvitz and Thompson, 1952). Its variance is given by

$$\mathbb{V}_p(\widehat{t}_{y,\pi}) = \sum_{k\in U}\sum_{\ell\in U} \Delta_{k\ell}\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell},$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell$. This variance can not be computed since Y-values are not known for the whole population elements. So it can be estimated by the design-unbiased estimator

$$\widehat{V}_\pi(\widehat{t}_{y,\pi}) := \sum_{k\in S}\sum_{\ell\in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}}\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell}.$$

Depending on the chosen sampling design, the expression of the variance of $\widehat{t}_{y,\pi}$ and that of the variance estimator can be simplified. Under Bernoulli sampling design, the variance of estimator $\widehat{t}_{y,\pi}$ takes the form

$$\mathbb{V}_{\pi,BE}(\widehat{t}_{y,\pi}) = \frac{1-\pi}{\pi}\sum_{k\in U} y_k^2$$

and a design-unbiased variance estimator is

$$\widehat{V}_{\pi,BE}(\widehat{t}_{y,\pi}) := \frac{1-\pi}{\pi^2}\sum_{k\in s} y_k^2.$$

Under simple random sampling without replacement design, variance of the $\widehat{t}_{y,\pi}$ estimator takes the form

$$\mathbb{V}_{p,SRSWOR}(\widehat{t}_{y,\pi}) = \frac{N(N-n)}{n}S_{yU}^2,$$

and a design-unbiased variance estimator is

$$\widehat{V}_{\pi,SRSWOR}(\widehat{t}_{y,\pi}) := \frac{N(N-n)}{n}S_{ys}^2,$$

where S_{yU}^2 and S_{ys}^2 are the population and the sample adjusted variances, such that

$$S_{yU}^2 = \frac{1}{N-1}\sum_{k\in U}\left\{y_k - \frac{1}{N}\sum_{k\in U} y_k\right\}^2 \text{ and } S_{ys}^2 = \frac{1}{n-1}\sum_{k\in s}\left\{y_k - \frac{1}{n}\sum_{k\in s} y_k\right\}^2.$$

In the case where the Y-values are not centered on S, the values of the variance and the variance estimator, for the Bernoulli sampling and for the simple random sampling without replacement, will be very different. It shows the importance of the choice of the sampling design, as it influences the precision of the estimator. Similarly, the choice of the estimator is important depending on the sampling design used. For instance, under Bernoulli sampling, the variance of $\widehat{t}_{y,\pi}$ can become large because the sample size n_s is random. In this case, a better estimator is given by the *population total Hájek estimator* (Hájek, 1971) $\widehat{t}_{y,H} := n\widehat{t}_{y,\pi}/n_s$. This supports the fact that the selection of the sample and the estimator must depend on each other, as discussed in Section 1.2.

1.4 Auxiliary information at the estimation stage

The term *auxiliary information* refers to any information available prior to sampling. These may be totals, proportions or means of certain variables over the population - or sub-populations - or the value of a variable known for all population units. The main objective is to make the best possible use of this information to improve the accuracy of the final estimate. Auxiliary

information can be used at the sampling stage or at the estimation stage. At the moment, we only consider the use of auxiliary information at the estimation stage.

At the estimation stage, the auxiliary information is used in the estimator formula. We consider p auxiliary variables $X_1, \ldots, X_p$. Let $\mathbf{x}_k \in \mathbb{R}^p$ denote the vector of values taken by the auxiliary variables for element $k \in U$. We consider two levels of available auxiliary information: 1) U-level, when the auxiliary variables are known for all population units; 2) S-level, when the auxiliary variables are known only for the sampled units. We assume that values of these p variables are known at U-level or S-level. It should be noted that the auxiliary variables are at least known in the sample.

This part focuses on the estimation stage, so we assume that the sampling stage has already been completed. A sample $s \in U$ was selected using a random sampling design $p(.)$ and data $\{y_k\}_{k\in s}$ are then available. The auxiliary variables are used in the final estimator formula in an attempt to significantly reduce the variance compared with the $\widehat{t}_{y,\pi}$ estimator, which does not use auxiliary information.

1.4.1 Model-assisted estimation

One way of using auxiliary information at the estimation stage is *model-assisted estimation*. This approach makes use of a set of predicted values to improve the efficiency estimators (see e.g. Särndal, Swensson, and Wretman, 1992 and Breidt and Opsomer, 2017).

Let us consider the situation where auxiliary information is known at U-level (see Section 1.4). Model-assisted estimation starts by postulating that the finite population of y_k's conditioned on $\mathbf{x}_k$'s is a realization of an infinite superpopulation $\mathscr{E}$, in which

$$\mathscr{E} : y_k = m(\mathbf{x}_k) + \varepsilon_k, \quad k \in U \tag{1.2}$$

where $m(.)$ is an unknown function, ε_k denotes a sequence of independent and identically distributed random variables such that $\mathbb{E}_m(\varepsilon_k) = 0$ and $\mathbb{V}_m(\varepsilon_k) = \sigma^2$, and subscript m indicates that the expectation and variance are computed under model $\mathscr{E}$.

To introduce model-assisted estimators, we start by looking at a closely related estimator: the *difference estimator*. We assume that auxiliary information is known at level-U, so vector $\mathbf{x}_k$ is known for the whole population. So we can form a set of 'predicted' Y-values, using function $m(.)$ and $\mathbf{x}_k$; y_k can be predicted by $m(\mathbf{x}_k)$. The unknown population total t_y can be written as

$$t_y = \sum_{k\in U} m(\mathbf{x}_k) - \sum_{k\in U} \{y_k - m(\mathbf{x}_k)\}$$

(Särndal, Swensson, and Wretman, 1992). Since $\sum_{k\in U} \{y_k - m(\mathbf{x}_k)\}$ depends on unknown values $\{y_k\}_{k\in U\setminus S}$, one idea is to use the Horvitz-Thomson estimator. The results is the *difference estimator* (Cassel, Särndal, and Wretman, 1976)

$$\widehat{t}_{y,\text{diff}} := \sum_{k\in U} m(\mathbf{x}_k) - \sum_{k\in s} \frac{y_k - m(\mathbf{x}_k)}{\pi_k}.$$

In reality, the function $m(.)$ is also unknown, which makes the difference estimator impossible to compute. The idea of model-assisted estimators is to replace the unknown function $m(.)$ in the difference estimator by an estimator $\widehat{m}(.)$. The estimator $\widehat{m}(.)$ is fitted on the available data $\{(y_k, \mathbf{x}_k)\}_{k\in s}$, in order to be able to predict y_k using $\mathbf{x}_k$ for all population element $k \in U$. The *model-assisted estimator* of t_y (Särndal, Swensson, and Wretman, 1992) is

$$\widehat{t}_{y,\text{ma}} := \sum_{k\in U} \widehat{m}(\mathbf{x}_k) - \sum_{k\in s} \frac{y_k - \widehat{m}(\mathbf{x}_k)}{\pi_k}. \tag{1.3}$$

Model-assisted estimation starts in Särndal (1980) with the *generalized regression estimator*, which considers a linear regression as function $m(.)$.

1.4.2 Calibrated estimation

Another class of estimators that uses auxiliary information are *calibrated estimators*. Calibrated estimators do not necessarily require auxiliary information to be known at U-level, but do require known population totals. An estimator of the form

$$\widehat{t}_{y,cal} = \sum_{k \in S} w_k y_k,$$

where weights $\{w_k\}_{k \in S}$ satisfy

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k,$$

is a calibrated estimator. The weight w_k is then a function of the $\{\mathbf{x}_k\}_{k \in S}$. Many different weights can satisfy this equation. Deville and Särndal (1992) propose to introduce a notion of pseudo-distance so that the solution is unique. The weights w_ks solve the minimization problem

$$\begin{cases} \underset{w_k \in \mathbb{R}}{\text{minimize}} & \sum_{k \in S} G(w_k, \pi_k^{-1}), \\ \text{subject to} & \sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k, \end{cases}$$

where $G(.,.)$ is a pseudo-distance function. Weights w_ks are as close as possible to the inverse of the inclusion probabilities π_k^{-1}s in an attempt to control the bias. A wide range of functions can be used (see Deville and Särndal, 1992). Kim and Park (2006) propose a review of calibration estimators, study asymptotic properties and discuss variance estimation.

1.5 Nonresponse

1.5.1 Nonresponse modelling

Nonresponse is present in almost all surveys. The presence of nonresponse means that the value of the variable of interest Y is unavailable for certain units in the sample. These units are called *nonrespondents*. Otherwise, units for which the variable of interest is known are called the set of *respondents*. Let $s_r \subset s$ be the set containing the respondents to variable Y, and $s_m = s \backslash s_r$ the set containing the nonrespondents. Let us define R_k, the response membership indicator variable, such that $R_k(S_r) := \mathbb{1}_{k \in s_r}$. For simplicity, we use R_k instead of $R_k(S_r)$. We suppose that each variable R_k follows a Bernoulli distribution of parameter $p_k \in]0,1]$, such that $\mathbb{P}(R_k = 1) = p_k$ and $\mathbb{P}(R_k = 0) = 1 - p_k$. This random process is called the *nonresponse mechanism*. We also define $\mathscr{S}_r$ as the set of non-empty subsets of s. Nonresponse can be modelled in two different ways: using the *two-phase approach* or the *reverse approach*, that are described below and illustrated in Figure 1.1.

Nonresponse can be seen as a second phase of the survey. In the *two-phase approach*, we assume nonresponse as a second sampling phase. Two sampling operations are carried out one after the other to obtain the final sample of respondents: 1) a sample s is randomly selected in U; 2) a sample s_r is randomly selected in s. The nonresponse mechanism can be seen as a sampling process, assigning each possible sample $s_r \in \mathscr{S}_r$ a probability $q(s_r \mid s)$ of being the set of responding units. So s_r is a realization of a random variable S_r such that

$$\mathbb{P}(S_r = s_r \mid S = s) = q(s_r \mid s), \quad s_r \in \mathscr{S}_r.$$

The response probability to variable Y of a unit k is the probability that k is selected in sample s_r given that k is selected in s, that is

$$p_k = \mathbb{P}(k \in s_r \mid k \in s).$$

A survey can then be divided into two distinct samplings: the selection of n_s units to make up the random sample S using the sampling design $p(s)$, and the selection of respondents to make up the random sample S_r using the nonresponse mechanism $q(s_r \mid s)$.

In the *reverse approach* that was first proposed in Fay (1991), we assume that the sampling and the nonresponse designs are two independent processes. The term 'reversed' is used because this approach can reversed the order of sampling and nonresponse compared to the classic two-phase approach, due to independence. The procedure is: 1) the population U is randomly divided in two sub-populations, the population of respondents U_r and the population of nonrespondents U_m; 2) a sample s is selected in U. Finally, we obtain $s_r = U_r \cap s$. In this case, whether or not a unit would respond to Y does not depend on its selection in the sample. This implies

$$p_k = \mathbb{P}(k \in s_r \mid k \in s) = \mathbb{P}(k \in U_r) \cdot \mathbb{P}(k \in s).$$

In this work, we assume the independence of the R_ks from one unit to another. In other words, each unit may or may not respond to variable Y independently of the others. In this case and if the sampling design and nonresponse mechanism are independent, the nonresponse mechanism is simply a Poisson sampling design with inclusion probabilities equal to the unknown response probabilities p_k.

The nonresponses can be classified into three different types (Rubin, 1976). They can be *completely missing at random*, if their absence is linked neither to the variable of interest nor to auxiliary variables. If the missing values are not linked to the variable of interest after removing the contribution of auxiliary variables, nonresponses are said to be *missing at random*. Finally, if nonresponses are *not missing at random*, their absence is directly linked to the variable Y, whatever the auxiliary variables. This manuscript is devoted to missing at random nonresponse.

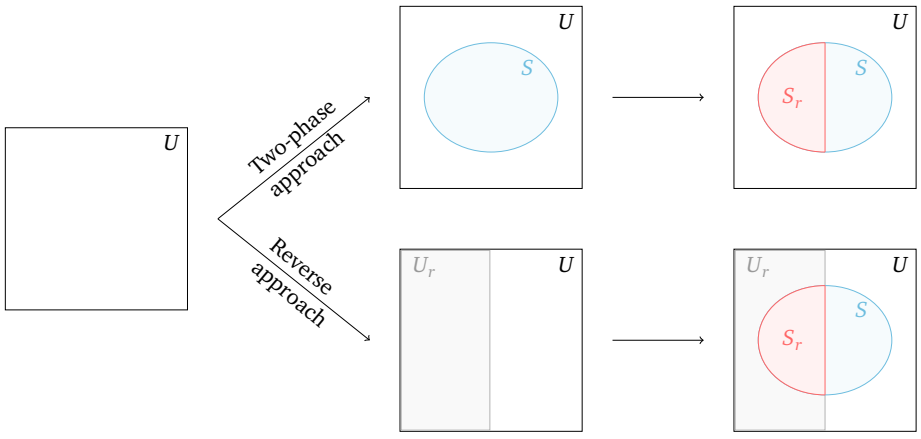

FIGURE 1.1: Representation of the two approaches to modelling nonresponse.

1.5.2 Multivariate nonresponse

Suppose we are interested in estimating parameters of more than one survey variable. In this case, the Y-value does not contain a single value, but is a collection of $Q \in \mathbb{N}^*$ population variables $Y_1, Y_2, \ldots, Y_Q$.

When nonresponses appear in more than one survey variable, we speak of *multivariate nonresponse*. Multivariate nonresponse should be treated with greater caution than univariate nonresponse. Indeed, the relationship between the survey variables must also be taken into account when dealing with multivariate nonresponses.

Multivariate nonresponse is categorized into: *unit nonresponse* and *item nonresponse*. Unit nonresponse is the total absence of response to the survey variables from a unit in the survey sample. Item nonresponse is the absence of part of the information from a sampled unit for one or more survey variables. In this work we focus on item nonresponse.

Let us define R_k^q, the response membership indicator to variable $q \in \{1,\dots,Q\}$, such that $R_k^q(S_r) = 1$ if y_{kq}, the Y_q-value of unit k, is available and $R_k^q(S_r) = 0$ otherwise. In the case of multivariate nonresponse, a unit $k \in s$ is a respondent, i.e. $k \in s_r$, if $R_k^q = 1$ for all survey variable q.

1.5.3 Estimation in the presence of unit nonresponse

Dealing with nonresponse is no easy task. In presence of nonresponses, all the aforementioned total estimators are unavailable. Two main approaches to handle nonresponse can be considered: 1) *reweighting for unit nonresponse*: only the set of responses is used, and the expression of the estimator is modified to compensate for the nonresponses; 2) *imputation*: the nonresponses are replaced with imputed values in the data set, and the estimator is calculated with the observed and estimated data.

If nonresponse is considered as a second phase of the survey, as explained in Section 1.5.1, and p_k was known, the *two-phase estimator*

$$\widehat{t}_{y,\pi p} := \sum_{k\in S_r} \frac{y_k}{\pi_k p_k}.$$

could be used. The respondent weights are increased by considering the inclusion probabilities multiplied by the response probabilities instead of the inclusion probabilities alone. However, during a survey, the response probabilities $\{p_k\}_{k\in U}$ are unknown. A first idea is therefore to replace the probabilities $\{p_k\}_{k\in S_r}$ by point estimates $\{\widehat{p}_k\}_{k\in S_r}$, leading to the *nonresponse weighting adjusted* (NWA) estimator or propensity score adjusted estimator

$$\widehat{t}_{y,NWA} := \sum_{k\in S_r} \frac{y_k}{\pi_k \widehat{p}_k}. \tag{1.4}$$

Estimation of the response probabilities is generally carried out by modelling nonresponse. A *response model* is the set of assumptions about the true response mechanism, subject to point estimators of response probabilities which are obtained. We assume that the inclusion probabilities and auxiliary variables can be related using the following general model

$$\mathbb{P}(k \in s_r \mid k \in s) = h(\mathbf{x}_k^\top \mathbf{b}),$$

where $h : \mathbb{R} \to]0;1]$ is a derivable and increasing function and $\mathbf{b} \in \mathbb{R}^J$. A commonly used function to model nonresponse is the logistic function. The parameter $\mathbf{b}$ can be estimated using least squared method, maximum likelihood or calibration. Särndal and Lundström (2005) and Haziza and Beaumont (2017) provide overviews of nonresponse weighted adjustment.

The second approach, imputation, consists of completing the nonresponses of a data set by predicting them. A missing value y_k, $k \in S_m$, is replaced in the collected data by an estimate $\widehat{y}_k$, called *imputed value*. There are parametric and non-parametric imputation methods. Chen and Haziza (2019) review imputation methods for item nonresponse.

In *parametric imputation*, a working model must be postulated that links auxiliary variables to the survey variable Y in order to predict missing values $\{y_k\}_{k\in s_m}$. Given that the nonresponse is assumed to be missing at random, the distribution of variable Y is the same for the sets of respondents S_r and of nonrespondents S_m. In this case, the model (1.2) linking values $\{y_k\}_{k\in U}$ and $\{\mathbf{x}_k\}_{k\in U}$ can be considered. Because of nonresponses, the function $m(.)$ can not be estimated using values of the sampled units as in Section 1.4. An estimator $\widehat{m}_r(.)$ is then fitted on the available data $\{(y_k,\mathbf{x}_k)\}_{k\in s_r}$, in order to be able to predict value y_k using $\mathbf{x}_k$ for all $k\in s_m$. The predicted Y-value of unit k is $\widehat{y}_k := \widehat{m}_r(\mathbf{x}_k)$ and the imputed Y-values for each unit $k\in S$ is

$$\widetilde{y}_k := \begin{cases} y_k & \text{if } k\in S_r, \\ \widehat{y}_k & \text{if } k\in S_m. \end{cases}$$

To estimate the total t_y and the mean μ_y, simple estimators are respectively the *imputed estimator*

$$\widehat{t}_{y,imp} := \sum_{k\in S}\frac{\widetilde{y}_k}{\pi_k} = \sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{y}_k}{\pi_k}$$

and

$$\widehat{\mu}_{y,imp} := \frac{1}{\widehat{N}}\sum_{k\in S}\frac{\widetilde{y}_k}{\pi_k} = \frac{1}{\widehat{N}}\sum_{k\in S_r}\frac{y_k}{\pi_k} + \frac{1}{\widehat{N}}\sum_{k\in S_m}\frac{\widehat{y}_k}{\pi_k}.$$

Many $m(.)$ functions can be considered to impute Y values, as for example a linear regression.

Other *non-parametric* methods are *hot deck imputations* (see Andridge and Little, 2010). They do not require the specification of a working model, and use the respondents' observed values as imputed values. The idea is to impute the nonresponse of a unit $k\in S_m$, called the *recipient*, using values of certain respondents that are similar units, called *donors*. There are many hot deck imputation procedures, such as K-nearest neighbours, sequential hot deck imputation or random hot deck imputation. The advantage of hot deck procedures is that the imputed values are observed values, whereas predicting nonresponse can lead to unrealistic imputed values. However, in order to find donors for a nonrespondent, a notion of similarity must be introduced via a distance function, such as the Euclidean function on the available values. In some cases, it may be difficult to compute distances, as for example in high-dimensional data sets.

1.6 Introduction to variance estimation with complete respon-se

The measure most commonly used to assess the precision of a survey estimator is its variance. It usually depends on unknown parameters of the population. Variance estimators are generally adapted to the sampling design and the survey estimator. The estimator variance depends on both the formulae of the estimator and the sampling design. In Section 1.3, we present a variance estimator when θ is a total of a variable over the population. However, the parameter of interest θ is usually a more complex function, such as a ratio of totals, a population variance or a median. This makes it even more difficult to know and estimate the variance of $\widehat{\theta}$. In this case, there are several solutions to deal with the problem. A first class of methods involves approximating the variance by replacing the complex parameter of interest by the total of another variable, known as the linearized variable, using the *Taylor linearization technique*. Another class of methods use resampling to estimate variance through simulations. The *jackknife resampling method* is introduced in Section 1.7.

1.6.1 Taylor linearization technique for a function of totals

Let us consider the case where the parameter of interest θ_y is a function of $Q \in \mathbb{N}^*$ totals, such that

$$\theta_y = f(t_1, \ldots, t_Q),$$

where t_q is the population total of variable Y_q, $q \in \{1, \ldots, Q\}$. We assume that values $\{y_{qk}\}_{k \in S}$ of each variable Y_q are known for the sampled units. If $f(.)$ is linear, the parameter of interest can be expressed as

$$\theta_y = a_0 + \sum_{q=1}^{Q} a_q t_q,$$

where $a_0 \in \mathbb{R}$ and $a_q \in \mathbb{R}$, $q \in \{1, \ldots, Q\}$. An unbiased estimator of θ_y is then

$$\widehat{\theta}_y = a_0 + \sum_{q=1}^{Q} a_q \widehat{t}_{q,\pi} = a_0 + \sum_{k \in S} \frac{v_k}{\pi_k},$$

where $\widehat{t}_{q,\pi} = \sum_{k \in S} \pi_k^{-1} y_{kq}$ and $v_k = \sum_{q=1}^{Q} a_q y_{kq}$. So the variance of $\widehat{\theta}_y$ is equal to the variance of the Horvitz-Thompson estimator of the total $\sum_{k \in U} v_k / \pi_k$. So the variance of $\widehat{\theta}_y$ is

$$\mathbb{V}(\widehat{\theta}_y) = \sum_{k \in U} \sum_{\ell \in U} \Delta_{k\ell} \frac{v_k}{\pi_k} \frac{v_\ell}{\pi_\ell}, \tag{1.5}$$

and can be estimated by

$$\widehat{V}(\widehat{\theta}_y) = \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{v_k}{\pi_k} \frac{v_\ell}{\pi_\ell}. \tag{1.6}$$

The variable v_k is the linearized variable of y_k.

We now turn to the usual case where $f(.)$ is non-linear. The idea of the *Taylor linearization technique* is to estimate the non-linear estimator $\widehat{\theta}_y$ by a linear expression $\widehat{\theta}_T$, using Taylor series, in order to reduce to the case of a linear function of totals and thus be able to approximate the variance. The first-order Taylor approximation of $f(.)$, expanding around the point $(t_1, \ldots, t_Q)$, is

$$\widehat{\theta}_y \approx \widehat{\theta}_T = \theta_y + \sum_{q=1}^{Q} \left(\widehat{t}_{q,\pi} - t_q\right) \left.\frac{\partial f}{\partial \widehat{t}_q}\right|_{(\widehat{t}_{1,\pi}, \ldots, \widehat{t}_{Q,\pi}) = (t_1, \ldots, t_Q)}.$$

The variance of $\widehat{\theta}_T$ under the sampling design is equal to (1.5), with

$$a_q = \left.\frac{\partial f}{\partial \widehat{t}_q}\right|_{(\widehat{t}_{1,\pi}, \ldots, \widehat{t}_{Q,\pi}) = (t_1, \ldots, t_Q)},$$

and can then be estimated by (1.6). However, value v_k contains unknown quantities $\{a_q\}$ depending on unknown population totals. So, the last step to obtain a variance estimator of $\widehat{\theta}_T$ is to replace unknown totals by some estimators, leading to a variance estimator for $\widehat{\theta}_y$, that is

$$\widehat{V}(\widehat{\theta}_y) = \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{v}_k}{\pi_k} \frac{\widehat{v}_\ell}{\pi_\ell},$$

where $\widehat{v}_k = \sum_{q=1}^{Q} \widehat{a}_{q,\pi} y_{kq}$. The $\{\widehat{a}_q\}$ estimators are obtained by replacing the unknown totals by the Horvitz-Thompson.

1.6.2 Linearization by sample membership indicators

Taylor linearization technique can lead to multiple variance estimators that are asymptotically design unbiased under repeated sampling (Wolter, 2007). The choice among variance estimators must then be based on other considerations such as conditional properties of the variance estimators. Estimator $\widehat{\theta}_y$ can always be expressed as a function of the $\{I_k\}_{k\in U}$, such that $\widehat{\theta}_y = g(I_1,\ldots,I_N)$, $g : [0,1]^N \mapsto \mathbb{R}$. The sample membership indicators are the only source of randomness in the sampling design. Graf (2011) has developed a Taylor linearization method that consists in deriving $\widehat{\theta}_y$ at point $(I_1,\ldots,I_N)$, assuming that it is derivable at this point. Note that $\mathbb{E}[I_k] = \pi_k$, for all $k \in U$. The first-order Taylor approximation of $\widehat{\theta}_y$, expanding around the point $(\pi_1,\ldots,\pi_n)$, is

$$\widehat{\theta}_y \approx g(\pi_1,\ldots,\pi_N) + \sum_{k\in U}(I_k - \pi_k)\,\widetilde{a}_k,$$

where

$$\widetilde{a}_k = \left.\frac{\partial \widehat{\theta}_y}{\partial I_k}\right|_{(I_1,\ldots,I_N)=(\pi_1,\ldots,\pi_N)}.$$

The linearized variable $\widetilde{a}_k$ is not random under the sampling design, so it can be used to approximate the variance of $\widehat{\theta}$, such that

$$\mathbb{V}_p(\widehat{\theta}) \approx \sum_{k\in U}\sum_{\ell\in U}\Delta_{k\ell}\widetilde{a}_k\widetilde{a}_\ell.$$

As many $\widetilde{a}_k$s are unknown because they do not belong to the sample, $\widetilde{a}_k$ must be replaced by an estimator $\widehat{\widetilde{a}}_k$ in the expression to finally obtain a variance estimator of $\widehat{\theta}$.

1.7 Variance estimation in the context of nonresponse imputation

Variance estimation in the presence of nonresponse is one of the most difficult issues in survey research. In the case of imputation, it is common practice to compute variance estimates using standard formulae, treating the imputed values as if they were observed values, without taking into account the variance due to nonresponse. This procedure can lead to serious underestimation of the true variance of the estimates. The variance of the estimator can depend on four different sources of randomness: 1) the sample membership indicator variables $\{I_k\}_{k\in U}$; 2) the superpopulation model $\mathscr{E}$ as presented in Section 1.4.1; 3) the response membership indicator variables $\{R_k\}_{k\in U}$; 4) the imputation, in the case where the process is random, as presented in Section 1.5.3.

So far, we have only considered estimators of the variance due to the sampling design. The estimators presented in Section 1.4.1 also contain variance due to the $\mathscr{E}$ superpopulation model. We have not discussed the variance of these estimators, but both sources of randomness must be taken into account when estimating their variance. Next sections consider the first three sources of randomness.

1.7.1 The methods of Särndal and Shao Steel

We consider the case where values $\{y_k\}_{k\in U}$ are random. We assume missing at random nonresponse, that is the first moment of the imputation model, $\mathbb{E}(\mathbf{y}_U \mid \mathbf{X}_U)$, is correctly specified. Let $\widehat{\theta}_{y,imp} = \widehat{\theta}(\{\widetilde{y}_k\}_{k\in S})$ the imputed estimator corresponding to estimator $\widehat{\theta}_y$. The imputed estimator has the same form as the corresponding estimator, but replacing nonresponses $\{y_k\}_{S_m}$ by imputed values. For example, in the case of a population total, the imputed estimator of $\widehat{t}_{y,\pi}$ is $\widehat{t}_{y,imp}$ (see Section 1.5.3).

We present two methods for estimating the variance of $\widehat{\theta}_{y,imp}$: the method of Särndal (1992) and the method of Shao and Steel (1999). The first is based on the two-phase approach to modeling nonresponse, while the second is based on the reverse approach (see 1.5.1). These methods decompose the variance expression so that it can be estimated and then calculated. In almost all cases, estimating the terms requires Taylor linearizations. Since the estimator can depend on three sources of randomness, a Taylor linearization can be done on I_k, r_k, and ε_k. The method of Graf (2011), described in Section 1.6.2, may be used.

For the sake of simplicity, we replace operator $\mathbb{E}_m\mathbb{E}_p\mathbb{E}_q(.)$ by $\mathbb{E}_{mpq}$.In the method of Särndal, the total variance of an imputed estimator $\widehat{\theta}_y$ is expressed as

$$\begin{aligned}\mathbb{V}(\widehat{\theta}_y) &= \mathbb{E}_{mpq}\left[(\widehat{\theta}_{y,imp}-\theta_y)^2\right]\\ &= \mathbb{E}_{mpq}\left[(\widehat{\theta}_y-\theta_y+\widehat{\theta}_{y,imp}-\widehat{\theta}_y)^2\right]\\ &= \mathbb{E}_m\mathbb{V}_p(\widehat{\theta}_y)+\mathbb{E}_q\mathbb{E}_p\mathbb{V}_m(\widehat{\theta}_{y,imp}-\widehat{\theta}_y)+2\mathbb{E}_p\mathbb{E}_q\mathrm{Cov_m}\{(\widehat{\theta}_{\mathrm{y}}-\theta_{\mathrm{y}})(\widehat{\theta}_{\mathrm{y,imp}}-\widehat{\theta}_{\mathrm{y}})\}.\end{aligned}$$

In the last equality, the first term is the variance due to the sampling design, the second term is the variance due to the nonresponse and the third term is the covariance between the sampling and nonresponse errors. This variance decomposition enables us to estimate each term separately, to finally obtain an estimator of the total variance.

Unlike the method of Särndal, the method of Shao Steel requires another assumption that the sampling and the nonresponse designs are two independent processes. The total variance of an imputed estimator $\widehat{\theta}_y$ can be expressed as

$$\begin{aligned}\mathbb{V}(\widehat{\theta}_y) &= \mathbb{E}_{mpq}\left[(\widehat{\theta}_{y,imp}-\theta_y)^2\right]\\ &= \mathbb{E}_{mpq}\left[\left\{\widehat{\theta}_{y,imp}-\mathbb{E}_p(\widehat{\theta}_{y,imp})+\mathbb{E}_p(\widehat{\theta}_{y,imp})-\widehat{\theta}_y\right\}^2\right]\\ &= \mathbb{E}_q\mathbb{E}_m\mathbb{V}_p(\widehat{\theta}_{y,imp})+\mathbb{E}_q\mathbb{V}_m\mathbb{E}_p(\widehat{\theta}_{y,imp}-\theta_y).\end{aligned}$$

The last equality allows us to determine a variance estimator by estimating each term separately.

1.7.2 Jackknife procedure

The *jackknife technique* comes from another domain than survey sampling and is mainly used for two different purposes: bias reduction and variance estimation. Quenouille (1949a) proposed the method as a means to reduce the bias of an estimator, in an infinite population context. Durbin (1959) was the first to propose the jackknife technique for finite populations. Next, Tukey (1958) suggested the use of jackknife to produce variance estimates. In broad terms, the idea of the jackknife variance estimation is to compute estimates of the parameter of interest from each of several sub-samples of the original sample, and then estimate the variance of the original sample estimator from the variability between the sub-samples estimates.

The jackknife procedure is a resampling method that starts by creating n sub-samples of size $n-1$, denoted $S^{(\ell)}$, $\ell \in S$, such that $S^{(\ell)} = S\backslash\{\ell\}$. Next, point estimate of $\widehat{\theta}_y$ is computed as if the original sample were $S^{(\ell)}$ instead of S, leading to point estimate $\widehat{\theta}_y^{(\ell)}$. In this section, we consider the case where the parameter of interest is μ_y. We first focus on jackknife variance estimation in the case of complete response. We want to estimate the variance of the Hájek estimator, described in Section 1.2. For example, the Hájek estimator $\widehat{\mu}_{y,H}$ omitting unit $\ell \in S$ is equal to

$$\widehat{\mu}_{y,H}^{(\ell)} := \frac{1}{\widehat{N}^{(\ell)}}\sum_{k\in S^{(\ell)}}\frac{y_k}{\pi_k^{(\ell)}},$$

where $\pi_k^{(\ell)}$ denotes the inclusion probabilities adapted so that the selected sample is $S^{(\ell)}$ and $\widehat{N}^{(\ell)} = \sum_{k\in S^{(\ell)}} 1/\pi_k^{(\ell)}$. If the inclusion probabilities π_k are all equal to π, as in Bernoulli sampling and simple random sampling without replacement (see Section 1.2), we have $\pi_k^{(\ell)} = n(n-1)^{-1}\pi$, for all $k \in U$, and $\widehat{N}^{(\ell)} = (n-1)^2(n\pi)^{-1}$. Then, the n point estimates $\widehat{\mu}_y^{(\ell)}$ must be computed. If collected data $\{y_k\}_{k\in S}$ and $\{\mathbf{x}_k\}_{k\in S}$ are independent and identically distributed, a variance estimator of $\widehat{\mu}_{y,H}$ is the jackknife variance estimate proposed in Tukey (1958)

$$\widehat{V}_{Tuk}(\widehat{\mu}_{y,H}) := \frac{n-1}{n}\sum_{k\in S}\left(\widehat{\mu}_{y,H}^{(k)} - \frac{1}{n}\sum_{\ell\in S}\widehat{\mu}_{y,H}^{(\ell)}\right)^2.$$

The 'independent and identically distributed'-condition is satisfied in the case of simple random sampling with replacement, which is rarely the case in sample surveys. Lee (1973) develops the method to handle stratified simple random sampling without replacement. The restriction of the jackknife method to stratified multi-stage designs limits its applicability, unlike estimators based on linearization, for example. In a general paper, Campbell (1980) proposes a jackknife variance estimator equivalent to a standard linearization variance for designs with unequal probabilities. A condition for using the estimator of Campbell (1980) is that the parameter θ_y can be expressed as a function $g_1 : \mathbb{R}^Q \mapsto \mathbb{R}$ of $Q > 0$ means, that is $\theta_y = g_1(\mu_1, \ldots, \mu_Q)$. This estimator is highlighted and studied in depth in Berger and Skinner (2005) and is given by

$$\widehat{V}_{camp}(\widehat{\mu}_{y,H}) := \sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}(1-w_k)\left(\widehat{\mu}_{y,H} - \widehat{\mu}_{y,H}^{(k)}\right)(1-w_\ell)\left(\widehat{\mu}_{y,H} - \widehat{\mu}_{y,H}^{(\ell)}\right). \tag{1.7}$$

where $w_k = (\widehat{N}\pi_k)^{-1}$.

1.7.3 Jackknife in presence of nonresponse

We now turn to the case of nonresponse. We consider the imputation framework of Section 1.5.3. We want to estimate the variance estimator of the imputed estimator $\widehat{\mu}_{y,imp}$. Following the same procedure, each unit ℓ excluded from sample S may belong to the set of respondents or nonrespondents. Depending on this, the expression of the estimator $\widehat{\mu}_{y,imp}^{(\ell)}$ will not be the same. Recall that $\widehat{m}_r(.)$ denotes the estimator of the $m(.)$ function fitted on the data from the responding units in the S sample, i.e. S_r. Let $\widehat{m}_r^{(\ell)}(.)$ denote the estimator of function $m(.)$ fitted on the data from the responding units in the sample $S^{(\ell)}$, i.e. of units in $S^{(\ell)} \cap S_r$.

If unit ℓ is a nonrespondent, i.e. $\ell \in S_m$, we have $S^{(\ell)} \cap S_r = S_r$. In this case, the prediction function $\widehat{m}_r^{(\ell)}(.)$ is equivalent to $\widehat{m}_r(.)$ since the unit ℓ does not contribute to the adjustment of the function. In the other case, if $\ell \in S_r$, this has an impact on the prediction function. Indeed, the function is fitted on $S_r \backslash \{\ell\}$ instead of S_r. The expression of the estimator $\widehat{\mu}_{y,imp}^{(\ell)}$ will therefore not be the same depending on whether ℓ is a respondent or a nonrespondent: if $\ell \in S_r$,

$$\widehat{\mu}_{y,imp}^{(\ell)} = \frac{1}{\widehat{N}^{(\ell)}}\sum_{k\in S_r\backslash\{\ell\}}\frac{y_k}{\pi_k^{(\ell)}} + \frac{1}{\widehat{N}^{(\ell)}}\sum_{k\in S_m}\frac{\widehat{m}_r^{(\ell)}(\mathbf{x}_k)}{\pi_k^{(\ell)}},$$

and otherwise, if $\ell \in S_m$,

$$\widehat{\mu}_{y,imp}^{(\ell)} = \widehat{N}^{(\ell)}\sum_{k\in S_r}\frac{y_k}{\pi_k^{(\ell)}} + \widehat{N}^{(\ell)}\sum_{k\in S_m\backslash\{\ell\}}\frac{\widehat{m}_r(\mathbf{x}_k)}{\pi_k^{(\ell)}}.$$

Rao and Shao (1992) propose an adjusted jackknife method for estimating the variance in the case of imputation. This method assumes that units in S are selected with replacement, which

is often not the case. The *generalized jackknife variance estimator* of $\widehat{\mu}_{y,imp}$ is proposed in Berger and Rao (2006) and is adapted to sampling without replacement and a non-negligible sampling fraction. This estimator is given by

$$\widehat{V}_{gen}(\widehat{\mu}_{y,imp}) := \sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}(1-w_k)\left(\widehat{\mu}_{y,imp}-\widehat{\mu}^{(k)}_{y,imp}\right)(1-w_\ell)\left(\widehat{\mu}_{y,imp}-\widehat{\mu}^{(\ell)}_{y,imp}\right). \quad (1.8)$$

It estimates the first term that appears in the variance decomposition of the Shao Steel method (see Section 1.7.1).

1.8 Conclusion

The theory of sample surveys covers a wide range of subjects and methods. This chapter introduced the main concepts used in the other chapters of this manuscript. In Section 1.5.2, we presented the multivariate nonresponse and Section 1.5.3 introduced the concept of imputation. In Chapter 2, we will present a new method for imputing multivariate nonresponse. Chapter 3 will define a quasi-model-estimator that is a blend between the model-assisted estimator (1.3) and a nonresponse weighting adjusted estimator (1.4). In Chapter 4 we discuss variance estimation for total and mean estimators. We focus in particular on the behaviour of variance estimators (1.7) and (1.8) based on the jackknife procedure (see Section 1.7.2) and of the variance estimator of Taylor or Shao Stell (see Section 1.7.1).

Chapter 2

Balanced Donor Imputation Handling Swisscheese Nonresponse

Abstract: The estimator of a parameter of interest can be affected significantly by missing values, which introduce bias and cause additional variability. Swiss cheese nonresponse, also known as nonmonotone nonresponse, is difficult to deal with, because it occurs when each variable of a survey may contain missing values, but without any particular pattern. To reduce the effects of nonresponses, missing values are usually imputed. However, when several variables of a data set need to be imputed, it can be difficult to preserve the distributions of the variables and the relationships between them. In this paper, we propose a new donor imputation method that generalizes the balanced k-nearest neighbor imputation, and is applicable to any configuration of item nonresponses. This new method uses random imputations by donors and is constructed to meet the following requirements. First, all missing values of a unit should be imputed by the same donor. Next, a unit with missing values should be imputed by a neighboring donor. Last, the donors are selected to satisfy some balancing constraints that allows us to decrease the variance of the estimator. The method is divided into two phases. First, we create a stratification by computing a matrix of imputation probabilities using linear programming. Then, we select donors using these imputation probabilities and balanced stratified sampling.

Keywords: donor imputation, linear programming, nonmonotone nonresponse, random imputation.

This chapter is a reprint of article Eustache, Vallée, and Tillé (2024).

2.1 Introduction

In large-scale surveys, nonresponses are often inevitable. There are two types of nonresponse: unit nonresponse, which occurs when all information is missing for a sampled unit, and item nonresponse, which occurs when some, but not all information is missing for a sampled unit. Missing values can affect the estimators of the parameters of interest significantly by introducing bias and causing additional variability. There are two approaches to reducing such effects: the imputation model, in which the missing values are imputed, and the nonresponse model, in which the responding units are reweighted to compensate for the nonresponding units. Although we focus on donor imputation methods, we also show (Proposition 2.6.2) that the estimators can be presented as a reweighting method.

Nonresponses can be univariate or multivariate. In the first case, nonresponses occurs in only one variable, and we can perform imputation using the other fully observed variables. Although several methods exist for univariate imputation, fractional hot deck imputations (FDHIs) are popular in practice (Kim and Fuller, 2004; Fuller and Kim, 2005). Recently, Chen and Haziza (2019) reviewed methods (deterministic and random) for univariate imputation, including multiple and fractional imputations.

In the multivariate case, nonresponses occur in more than one variable of the survey. Here, we need to determine whether nonresponses can appear in all variables, or only in some, and whether or not the nonresponses are monotonic. Monotone nonresponse occurs when the missing values follows a specific pattern in the data set, as in longitudinal studies, where there is attrition.

In the first case, the missing values do not appear in all variables of the data set, and are not monotonic. Several methods have been proposed to deal with this missing pattern (Murray and Reiter, 2014; Sang, Kim, and Lee, 2022). The most general nonresponse pattern is when nonresponses can occur in all survey variables. Here, the difficulty lies in preserving the distributions of the variables and the relationships between them when replacing the missing values. Hasler, Craiu, and Rivest (2018) use grapevine copulas to impute monotone nonresponses, and present an overview of other imputation methods for this pattern.

This work is devoted to methods that can be applied to the most general situation, that is, nonmonotone nonresponse, also known as Swiss cheese nonresponse, which occurs when the survey variables all have missing values, but without a particular pattern. Most existing imputation methods are iterative, because of the presence of nonresponses in all variables. van Buuren (2018) reviewed joint modeling and fully conditional specification (FCS) procedures. An example of these iterative algorithms is a sequence of regression models between the variables developed by Raghunathan et al. (2001). However, Chen (2010) argues that FCS methods may encounter difficulties due to model incompatibilities. Stekhoven and Bühlmann (2011) developed a widely used and efficient iterative imputation method based on random forest models.

Donor imputation methods impute the missing values of a unit using values from other responding units, called donors. The advantage of this method is that the imputed values are plausible, because they are observed for the donor units. Moreover, these methods do not require an iterative system. Yang and Kim (2016) introduced an FHDI for a multivariate nonresponse pattern that is a donor imputation method implemented in the R package FHDI (Im, Cho, and Kim, 2018). Judkins (1997) and Andridge and Little (2010) present overviews of donor imputation methods in both univariate and multivariate cases.

Here, we propose a donor imputation method that includes balancing constraints for Swiss cheese nonresponses. This idea of using balancing constraints for imputing missing values has been considered before. Chauvet, Deville, and Haziza (2011) reduced the imputation variance using balanced sampling. Hasler and Tillé (2016) developed a balanced k-nearest neighbor imputation to deal with an univariate nonresponse. This imputation method has the advantage of satisfying balancing equations on the survey variables. Our method extends the balanced k-nearest neighbor imputation to include Swiss cheese nonresponses. This extension is not trivial, because we need to manage missing values for several variables simultaneously and the model cannot be constructed based on completely observed variables.

The proposed imputation method meets three essential requirements. First, in order to preserve the distributions of the variables, it must be a donor imputation method, which allows us to impute continuous and categorical variables using realistic values. Furthermore, all the missing values of a unit should be imputed by the same donor, in order to preserve the relationships between the variables. Second, a unit with missing values must be imputed by a similar donor to ensure consistency between imputed and observed values. Third, we use balancing constraints to reduce the additional variability of the estimated parameters. Note that the proposed method can also be applied to simpler nonresponse patterns, such as monotone or univariate nonresponses.

We present the context and the requirements of the method in Section 2.2, and the construction of the matrix of imputation probabilities in Section 2.3. We discuss selecting the donors and the imputation process in Section 2.4, and the FHDI method in Section 2.5. In Section 2.6, we examine several properties of the estimator of the total after imputation using our proposed method. A simulation study using the R package SwissCheese (Eustache, Vallée, and Tillé, 2021) is presented in Section 2.7. Section 2.8 concludes the paper.

2.2 Motivations

Consider a finite population U of size N with J variables of interest. A random sample S of size n is selected in U. The first-order inclusion probability of unit i is π_i, the second-order inclusion probability of units i and ℓ is $\pi_{i\ell}$, and $\pi_{ii} = \pi_i$, for any $i, \ell \in U$. The vector of J variables of interest, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ})^\top$, is not necessarily fully observed for all $i \in S$. The vector of response indicators of a unit i is $\mathbf{r}_i = (r_{i1}, \ldots, r_{ij}, \ldots, r_{iJ})^\top$, where r_{ij} is one if the variable j of unit i is observed, and zero otherwise. Consider $S_r \subset S$, the set of $n_r > 0$ units for which the J variables are completely observed. That is, $r_{ij} = 1$, for all $j = 1, \ldots J$ and any $i \in S_r$. Consider $S_m = S \backslash S_r$, a set of $n_m = (n - n_r)$ units, such that some values, but not all, are missing. Throughout this paper, units in S_r are referred to as respondents, and units in S_m are referred to as nonrespondents. The nonresponse is nonmonotone, and thus has no particular pattern. Figure 1 illustrates a data set with Swiss cheese nonresponses. Note that the proposed method can also be applied to simpler configurations, for example, when some variables are not affected by a nonresponse. For example, when a variable j is fully observed, then $r_{ij} = 1$, for all $i \in U$, and the following discussion therefore remains valid.

When no vector $\mathbf{x}_i$ suffers from nonresponse, an unbiased estimator of the population total of the variable j, $T_j = \sum_{i \in U} x_{ij}$, is given by the Horvitz-Thompson estimator

$$\widehat{T}_j^{HT} = \sum_{i \in S} d_i x_{ij},$$

where $d_i = \pi_i^{-1}$ is the sampling weight of unit i. If values are missing in the data set, then they can be imputed, where the imputed value of unit i for a variable j is denoted by x_{ij}^*. Then, T_j is estimated by

$$\widehat{T}_j = \sum_{i \in S} \{ r_{ij} d_i x_{ij} + (1 - r_{ij}) d_i x_{ij}^* \}.$$

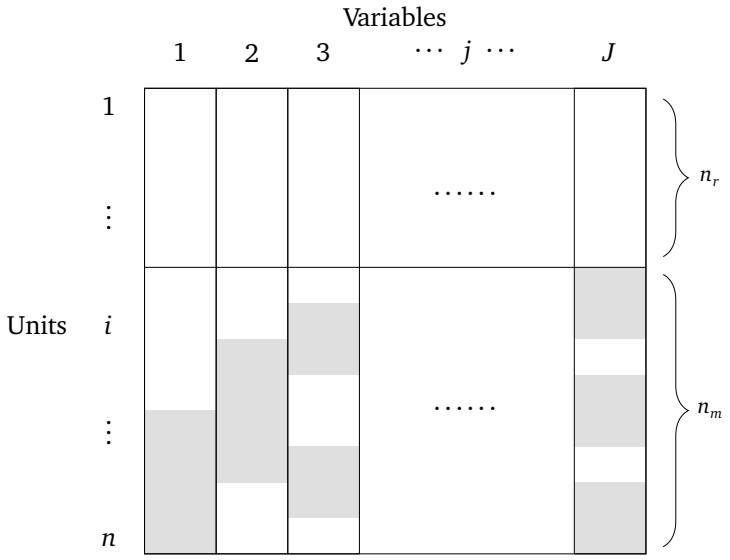

FIGURE 1: Representation of Swiss cheese nonresponse in a data set of n units and J variables. The first n_r rows correspond to the respondents, and the subsequent n_m rows correspond to nonrespondents. The gray rectangles cover the missing values.

The proposed method ensures coherence and accuracy in the imputed data set, and is based on the following three requirements:

(i) The imputed values should be selected from the values of the n_r fully observed units: a donor imputation method should be used. Furthermore, all missing values of a nonrespondent should be imputed using the same donor.

(ii) The donors should be as close as possible to the nonrespondents, in terms of the distance between survey variables.

(iii) If the observed values of the nonrespondents are imputed, the estimator of the total of each survey variable should remain unchanged.

Requirement (i) ensures that the imputed values are observed, and therefore realistic, for both categorical and continuous variables. Furthermore, a random imputation method tends to preserve the distributions of the variables. To illustrate Requirement (i), consider $J = 3$ variables and a nonrespondent $v \in S_m$, such that $\mathbf{r}_v = (1,0,0)^\top$. The missing values of unit v, x_{v2} and x_{v3}, are imputed using observed values selected from the same donor. This means that x_{v2} and x_{v3} are imputed by x_{u2} and x_{u3}, respectively, of a selected donor $u \in S_r$. Requirement (i) aims to preserve the relationships between variables.

Requirement (ii) allows the imputation of a nonrespondent using a similar unit. This ensures coherence between the imputed and the observed values of a nonrespondent. For instance, if we are recording the sex and height of people, the missing height of a man should be imputed using the height of another man. Requirement (ii) also aims to preserve the relationships between variables.

The idea behind Requirement (iii) is that the observed information would remain unchanged if the units with missing values were completely imputed. The estimators based on known values would not be affected. This requirement reduces the variance of the estimators.

To implement a donor imputation method, each fully observed unit receives a probability of donating its values to each nonrespondent. Next, we select one donor per nonrespondent, based on these imputation probabilities. The imputation probabilities satisfying Requirements (i)–(iii) are discussed further in Section 2.3. The selection of donors is discussed in Section 2.4.

2.3 Imputation probabilities

2.3.1 Matrix of imputation probabilities

The first step of a donor imputation method is to assign imputation probabilities to the units in the set of respondents. Consider $\boldsymbol{\psi} = (\psi_{uv})$, where $(u,v) \in S_r \times S_m$, the matrix of imputation probabilities. The element ψ_{uv} is the probability that respondent u is the donor selected to impute the missing items of nonrespondent v, with $\psi_{uv} \in [0,1]$. We need to impose some additional constraints on the imputation probabilities in order to meet Requirements (i)–(iii).

First, only one donor should be randomly selected for a unit $v \in S_m$; see Requirement (i). To this end, if the donors are chosen using balanced sampling, as suggested in Section 2.4, it is sufficient to ensure that the imputation probabilities associated with a nonrespondent sum to one; that is

$$\sum_{u \in S_r} \psi_{uv} = 1, \quad v \in S_m. \tag{2.3.1}$$

Requirement (iii) suggests that if the observed values of any $v \in S_m$ are imputed, the estimator of the total of each variable remains equal to the total of the observed values in S_m. Therefore, the imputation probabilities are chosen so that if the known values of the units in S_m were imputed by the expectation of their imputed values, the estimator of the total would correspond to that

based on the observed values. This means that the imputation probabilities must satisfy

$$\sum_{v \in S_m} d_v r_{vj} \sum_{u \in S_r} \psi_{uv} x_{uj} = \sum_{v \in S_m} d_v r_{vj} x_{vj}, \quad j \in \{1, \ldots, J\}, \tag{2.3.2}$$

see also Figure 2. The right-hand side of Equation (2.3.2) is the estimated total of the jth variable based on the observed values in S_m; see Figure 2b. The left-hand side of Equation (2.3.2) is the same estimated total, but calculated using imputed values in S_m. Each observed value x_{vj}, such that $v \in S_m$ and $r_{vj} = 1$, is imputed by

$$x^*_{vj} = \sum_{u \in S_r} \psi_{uv} x_{uj}.$$

The hatched region in Figure 2c represents these values. Then, the total of these imputed values corresponds to the left-hand side of Equation (2.3.2); see Figure 2c.

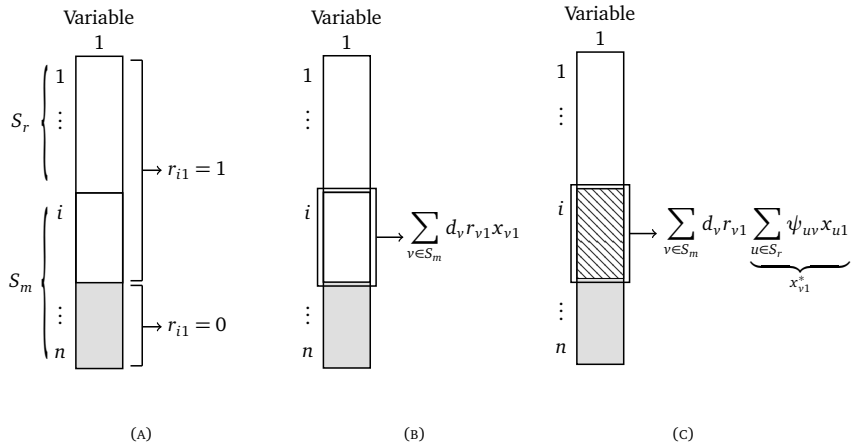

FIGURE 2: The balancing constraints of the balanced SwissCheese imputation for variable $j = 1$. The gray rectangles cover the missing values. Figure 2a represents the variable and the sets of the respondents S_r and the nonrespondents S_m. For unit i with a missing value at variable one, the corresponding response indicator r_{i1} is zero. Figure 2b represents the right-hand side of Equation (2.3.2) for the variable $j = 1$. In Figure 2c, the observed values in S_m are imputed and represented in the hatched region. The left-hand side of Equation (2.3.2) is represented, and x^*_{v1} is the imputed value for nonrespondent v.

Requirement (ii) implies that the donor must be similar to the nonrespondent, where similarity is defined in terms of the distance between units. Let $d(\cdot, \cdot)$ denote a distance function. The closer the distance $d(u, v)$ is to zero, the more similar the units u and v are. After computing the distance between a nonrespondent v and all responding units in S_r, those with the smallest distances to v should have the highest probabilities of being a donor for v.

In other words, for each nonrespondent $v \in S_m$, we want to select the donor $u \in S_r$ that minimizes the product $d(u, v)\psi_{uv}$. For instance, the distance between a respondent u and a nonrespondent v could be the Euclidean distance where the variables with missing values do not contribute to the distance, such that

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

The variable must be standardized before calculating the distance because of possible differences in the magnitudes.

The matrix $\boldsymbol{\psi}$ satisfying equations (2.3.1) and (2.3.2) can be found by solving the linear program

$$\begin{cases} \underset{\psi_{uv}\in[0,1]}{\text{minimize}} & \displaystyle\sum_{v\in S_m}\sum_{u\in S_r} d(u,v)\psi_{uv}, & \\ \text{subject to} & \displaystyle\sum_{u\in S_r}\psi_{uv}=1, & v\in S_m, \\ & \displaystyle\sum_{v\in S_m} d_v r_{vj}\sum_{u\in S_r}\psi_{uv}x_{uj}=\sum_{v\in S_m} d_v r_{vj}x_{vj}, & j=1,\ldots,J. \end{cases} \tag{2.3.3}$$

A solution to (2.3.3) can almost always be found when the number of respondents n_r is large, because in this case, the constraints are not too restrictive. If the sample size n is small, it is preferable to have at least $n_r/n = 0.5$ to satisfy the balancing constraints and to find similar donors for each nonrespondent.

Consider the bipartite set of S_r and S_m, $U^* = S_r \times S_m$, of size $n_r \cdot n_m$. The calculation of the final imputation probabilities ψ_{uv} can be viewed as a stratification process. A stratum is assigned to each nonrespondent, such that the population U^* is stratified in n_m strata $U_v^* = \{(u,v)|u \in S_r\}$, for $v \in S_m$. Each stratum corresponds to one nonrespondent and contains the set of n_r possible donors for nonrespondent v. Then, a sample of cells must be selected. Each element u, or possible donor, of a stratum U_v^* has a probability ψ_{uv} of being the selected donor for nonrespondent v. Hence, the inclusion probability of the cell (u,v) is ψ_{uv}, for $(u,v) \in U^*$.

After solving (2.3.3), in most cases, almost all the probabilities ψ_{uv} obtained are equal to either zero or one. This is equivalent to having a stratum of neighbors consisting of a single respondent. In the next section, we adjust the imputation probability calculation process to enable us to select the minimum number of elements in each stratum.

2.3.2 The number of neighbors k

After solving (2.3.3), in most cases, almost all the probabilities ψ_{uv} obtained are equal to either zero or one. However, many researchers encourage considering more than one donor for each non-respondent, for example, as in Jonsson and Wohlin (2004), which adds randomness to the process. This may help to preserve the distribution of the variables and reduce the bias. The constraint that the imputation probabilities need to be smaller than or equal to a quantity k^{-1} can be added to (2.3.3). Thus, at least k respondents will have a probability greater than zero of being a donor for a nonrespondent v, with $0 < k < n_r$ and $v \in S_m$. The program becomes

$$\begin{cases} \underset{\psi_{uv}\in[0,k^{-1}]}{\text{minimize}} & \displaystyle\sum_{v\in S_m}\sum_{u\in S_r} d(u,v)\psi_{uv}, & \\ \text{subject to} & \displaystyle\sum_{u\in S_r}\psi_{uv}=1, & v\in S_m, \\ & \displaystyle\sum_{v\in S_m} d_v r_{vj}\sum_{u\in S_r}\psi_{uv}x_{uj}=\sum_{v\in S_m} d_v r_{vj}x_{vj}, & j=1,\ldots,J. \end{cases} \tag{2.3.4}$$

The number of neighbors k must be chosen well, because it is used to add randomness to the imputation process. A larger k leads to greater variance due to randomness in the method. We recommend choosing k not greater than five although this depends on the size of the data set and the similarities between the responding units.

2.4 Imputation

Once we have calculated the matrix of imputation probabilities $\boldsymbol{\psi}$, we can randomly select the donors. Consider $\boldsymbol{\phi} = (\phi_{uv})$, where $(u,v) \in S_r \times S_m$, the imputation matrix. The element ϕ_{uv} is 1 if unit u is selected to donate its values to unit v, and zero otherwise. Only one donor is selected

per nonrespondent; thus,

$$\sum_{u \in S_r} \phi_{uv} = 1,$$

for each $v \in S_m$. The matrix $\boldsymbol{\phi}$ must satisfy, at best,

$$\sum_{v \in S_m} d_v r_{vj} \sum_{u \in S_r} \phi_{uv} x_{uj} = \sum_{v \in S_m} d_v r_{vj} \sum_{u \in S_r} \psi_{uv} x_{uj},$$

for each variable $j = 1, \ldots, J$. Therefore, a balanced sampling method is used to select the donors, while satisfying the balancing constraints (2.4). To also ensure that only one donor is selected per nonrespondent, the matrix $\boldsymbol{\phi}$ is generated using stratified balanced sampling (Chauvet, 2009; Hasler and Tillé, 2014; Jauslin, Eustache, and Tillé, 2021).

As explained in Section 2.3.1, one cell must be selected in each stratum of cells $U_v^* = \{(u, v) \mid u \in S_r\}$. The sample of cells is selected using stratified balanced sampling. Jauslin, Eustache, and Tillé (2021) propose a method for selecting a stratified balanced sample when the number of strata is large. If the sum of the inclusion probabilities in each stratum is an integer, the method guarantees the selection of a fixed sample size in each stratum. The size of the sample in a stratum is the sum of the inclusion probabilities of the units in this stratum. The matrix $\boldsymbol{\psi}$ is such that $\sum_{u \in S_r} \psi_{uv} = 1$, for any $v \in S_m$, thus, exactly one cell is selected per stratum, that is, one donor is selected per nonrespondent, and Requirement (i) is exactly satisfied. Moreover, by adding balancing vectors, the method can approximately satisfy (2.4) using the cube method (Deville and Tillé, 2004). The balancing variable of each cell $(u, v) \in S_r \times S_m$ is $d_v r_{vj} \psi_{uv} x_{uj}$. Equation (2.4) might only be approximately satisfied because of the complexity of the balancing problem. Therefore, Requirement (iii) is either exactly or approximately fulfilled. Requirement (ii) is also satisfied, because in the matrix $\boldsymbol{\psi}$, only the closest units of each nonrespondent have non-null imputation probabilities.

The imputation of the data set is based on the matrix $\boldsymbol{\phi}$. The missing value of unit v at variable j, such that $r_{vj} = 0$, is imputed randomly as

$$x_{vj}^* = \sum_{u \in S_r} \phi_{uv} x_{uj}. \tag{2.4.5}$$

It is also possible to use a deterministic version of the proposed imputation method. The expectation of ϕ_{uv} is used for $(u, v) \in S_r \times S_m$. Then, the missing value x_{vj} is imputed as

$$x_{vj}^* = \sum_{u \in S_r} \psi_{uv} x_{uj}. \tag{2.4.6}$$

Although this is no longer a donor imputation method, Requirement (iii) is exactly satisfied. In general, the presence of a random component helps to preserve the distribution of the variables, for instance, when estimating a nonlinear estimator as a percentile near or in the tail of the distribution.

2.5 Comparison with FHDI

The FHDI method is reviewed in Yang and Kim (2016), and is popular in practice. Its steps are similar to those of the proposed imputation method, which is a two-phase stratified sampling. First, a set of imputation cells is formed using all observed values for each variable containing nonresponses. For each cell, the imputation weight, called the fractional weight in the FHDI method, is calculated based on the joint probability of the vector of variables $(\mathbf{x}_1, \cdots, \mathbf{x}_J)$. The

calculation of the fractional weights is described in Section 4.1 of Yang and Kim (2016). Second, a hot deck imputation is conducted. Similarly to the proposed imputation method, determining the imputation cells and imputation weights corresponds to stratification, and the hot deck imputation corresponds to stratified sampling. However, although the methods have the same structure, their procedures are different.

FHDI requires discretizing continuous variables to compute the imputation weights. The discretization of each continuous variable is done by dividing its range into a small finite number of segments, as quantiles, for example. This loss of information may become a problem when the number of variables J increases. In addition, the final imputation is a weighted average of the values of the responding units. Thus, the imputed values are not true observed values, but rather a function of several values, and the method is not random. To address this problem, the imputation process described in Section 2.4 replaces the weights ψ_{uv} with the fractional weights. The FHDI method is considered in the simulation study in Section 2.7.

2.6 Properties of the imputed estimator of the total

The proposed imputation method provides a reliable estimation in several different cases. Here, we show that the estimator can be interpreted both as a prediction imputation method and as a reweighting method. Depending on the interpretation, the estimator of the total $\widehat{T}_j$, with the imputed values given in Equation (2.4.5), can be unbiased, under certain assumptions. In the section, we propose three assumptions that imply unbiasedness. The inference is valid when only one of them is satisfied. Some are on the prediction model, and some are on the weights.

Let $\mathrm{E}_p(.)$, $\mathrm{E}_q(.)$ and $\mathrm{E}_{imp}(.)$ denote the expectation with respect to the sampling design, nonresponse mechanism, and random imputation, respectively. The propositions presented in this section hold only when data are missing at random or completely missing at random, in the sense of Rubin (1976).

Proposition 2.6.1. *Consider the notation*

$$\mathbf{x}_i^{(-j)} = (x_{i1}, \ldots, x_{i(j-1)}, x_{i(j+1)}, \ldots, x_{iJ})^\top,$$

for $i \in U$ and $j = 1, \ldots, J$. Suppose further that the context is that of a prediction and assume the following model m:

$$m: \quad x_{ij} = \boldsymbol{\beta}^{(-j)^\top} \mathbf{x}_i^{(-j)} + \varepsilon_i \quad \textit{with} \quad E_m(\varepsilon_i) = 0,$$

where $E_m(.)$ denotes the expectation with respect to the model m. Then, the imputed estimator of the total of the variable j, $\widehat{T}_j$, is unbiased, for $j = 1, \ldots, J$,

$$Bias(\widehat{T}_j) = E_m E_p E_q E_{imp}(\widehat{T}_j - T_j) = 0.$$

The proof is given in the Appendix. Proposition 2.6.1 suggests that if a variable $\mathbf{x}_j$ can be explained by a linear combination of the other variables $\mathbf{x}_{g,g\neq j}$, the estimator $\widehat{T}_j$ will be unbiased.

Proposition 2.6.2. *The estimator of the total can be viewed as an estimator obtained using a reweighting method, such that*

$$\widehat{T}_j = \sum_{u \in S_r} d_u \left(1 + \pi_u \sum_{v \in S_m} d_v \psi_{uv} \right) x_{uj}.$$

When the weight $\left(1+\pi_u \sum_{v\in S_m} d_v \psi_{uv}\right)$ *is a reasonable approximation of the inverse of the probability of the response, that is when*

$$\Pr(u \in S_r|S) \approx \frac{1}{1+\pi_u \sum_{v\in S_m} d_v \psi_{uv}},$$

then the estimator is approximately unbiased,

$$Bias(\widehat{T}_j) = E_p E_q E_{imp}(\widehat{T}_j - T_j) \approx 0.$$

The proof is given in the Appendix. The estimator of the total can be rewritten as a reweighted estimator, such that

$$\widehat{T}_j = \sum_{u\in S_r} d_u w_u x_{uj}.$$

If the weight w_u is equal to the inverse of the probability of the response to variable j, the estimator is unbiased. In other words, the weight w_u compensates for the nonresponse bias, in the same way that the weight d_i compensates for the sampling bias.

Proposition 2.6.3. *The proposed imputation method requires that if* $u \in S_r$ *is the donor for* $v \in S_m$*, then* $u \in knn(v)$*. When*

$$u \in knn(v) \implies (1-r_{vj})(x_{uj}-x_{vj}) = 0,$$

for all $j = 1,\ldots,J$*, then the imputed estimator of the total of the variable* j*,* $\widehat{T}_j$*, is unbiased,*

$$Bias(\widehat{T}_j) = E_p E_q E_{imp}(\widehat{T}_j - T_j) = 0.$$

The proof is given in the Appendix. Proposition 2.6.3 uses the neighborhood principle. Because each donor is selected in the neighborhood of the recipient, the values of the recipient may be, by definition, close to the values of its donor. The closer the values are, the smaller is the bias of the estimator.

2.7 Simulation study

We performed a simulation study to analyze the performance of the proposed imputation methods, using the R package `SwissCheese` (Eustache, Vallée, and Tillé, 2021). We employ an open-source data set from Johnson (1996) that contains 15 variables of morphological data of $n = 250$ men. Only variables with strong correlations are considered: the waist circumference ($\mathbf{x}_1$), the knee circumference ($\mathbf{x}_2$), the chest circumference ($\mathbf{x}_3$), all three in centimeters, the body density in grams per cubic centimeter ($\mathbf{x}_4$), and the percentage of body fat ($\mathbf{x}_5$).

Swiss cheese nonresponses are generated randomly in the data set $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5)$. Nonresponses are generated for the whole data set, such that no variable is fully observed. For each vector $\mathbf{x}_j$ in which we generated missing values, the nonresponse is non-ignorable, because this is the most difficult type of nonresponse to handle. Define the positive values $g_{ij} = x_{ij} - \min(\mathbf{x}_j)$ and a value α_j such that

$$\sum_{i=1}^{n} \min\left[1, \alpha_j\left(g_{ij} + \frac{\sum_{i=1}^{n} g_{ij}}{n^2}\right)\right] = n_r^{(j)},$$

where $n_r^{(j)}$ is the "expected number of units with a missing value at variable j". The expected value for $n_r^{(j)}$ is 113, which gives a proportion of respondents n_r/n of approximately 45%. The

probability p_{ij} that unit i responds to item $j \in \{1,\ldots,5\}$ is

$$p_{ij} = \min\left[1, \alpha_j\left(g_{ij} + \frac{\sum_{i=1}^{n} g_{ij}}{n^2}\right)\right].$$

Missing values are generated randomly using a uniform variable bounded by zero and one. The response indicator r_{ij} is one if unit i responds to item j, and zero otherwise. When $r_{ij} = 0$, the value x_{ij} is missing.

Eight imputation methods were considered to impute the missing values:

- k-nearest neighbor imputation ($k$nn): a missing value for a nonrespondent is imputed as the mean of this variable for the set of k-nearest neighbors;
- Nearest neighbor (nn): the donor of each nonrespondent is its nearest neighbor;
- FHDI: the imputation method proposed by Yang and Kim (2016) and discussed in Section 2.5;
- Sequential regression multiple imputation (SReg): an iterative algorithm that imputes variables one by one using a regression model (Raghunathan et al., 2001; van Buuren, 2018);
- Balanced nearest neighbors (B-nn): the method proposed in Sections 2.3 and 2.4, without constraining a minimum number of neighbors, as in System (2.3.3), with random imputation as in Equation (2.4.5);
- Deterministic balanced nearest neighbors (DB-nn): a deterministic version of the B-nn, as in Equation (2.4.6);
- Balanced k-nearest neighbors (B-$k$nn): the method proposed in Sections 2.3 and 2.4, by constraining the minimum number of neighbors to k, as in System (2.3.4), with random imputation as in Equation (2.4.5);
- Deterministic balanced k-nearest neighbors (DB-$k$nn): a deterministic version of the B-$k$nn, as in Equation (2.4.6).

For each method that uses a number of neighbors k (i.e., $k$nn, FHDI, DB-$k$nn, and B-$k$nn), we use $k = 5$. The sequential regression multiple imputation method is a particular case of the fully conditional specification that imputes multivariate missing data on a variable-by-variable basis (van Buuren et al., 2006). Although the sequential regression multiple imputation is not a donor imputation method, it should work well because of the high correlations between the variables. Based on each imputed data set, we estimate the total of each variable, along with the 50th and the 75th percentiles.

The nonresponse is generated $M_R = 100$ times and, each time, the imputation is repeated $M_I = 100$ times, thus, we create M_R data sets with different nonresponse patterns. For each data set and for each imputation method, we create M_I imputed data sets. Obviously, the M_I imputations for the same nonreponse do not vary for the deterministic methods (i.e., $k$nn, nn, DB-$k$nn, and DB-nn). For each imputation method and parameter, we calculate the Monte Carlo bias of an imputed estimator $\widehat{\theta}$,

$$\text{Bias}_{MC}\left(\widehat{\theta}\right) = \frac{1}{M_R M_I}\sum_{r=1}^{M_R}\sum_{i=1}^{M_I}\left(\widehat{\theta}^{r,i} - \theta\right),$$

where $\widehat{\theta}^{r,i}$ is the value of the imputed estimator of the parameter θ in the simulation (r,i), for $r = 1,\ldots,M_R$ and $i = 1,\ldots,M_I$. We also calculate the Monte Carlo mean squared error (MSE) of

the imputed estimator,

$$\mathrm{MSE}_{MC}\left(\widehat{\theta}\right) = \frac{1}{M_R M_I} \sum_{r=1}^{M_R} \sum_{i=1}^{M_I} \left(\widehat{\theta}^{r,i} - \theta\right)^2.$$

The results for the totals and the percentiles are shown in Tables 1 and 2, respectively, for the eight imputation methods.

For the estimation of the totals, the proposed methods (i.e., B-nn, DB-nn, B-*k*nn, and DB-*k*nn) seem to be equivalent, and outperform the nn and *k*nn imputations in terms of bias and MSE. For the estimation of the percentiles, the proposed methods also outperform than the nn and *k*nn methods. The biases and MSEs of our proposed methods appear to be smaller than those of FHDI for the estimation of the totals. They are similar when estimating a quantile. Globally, the balancing constraints and the donor imputation seem to reduce the bias and MSE of the estimators. The results of our proposed method and those of the SReg imputation are comparable, although the latter is not a donor method. The requirement to use donors is restrictive. Thus, it is promising that our donor methods have almost similar efficiency. Futhermore, the variables are highly correlated, implying that linear regression models are appropriate. SReg is then well suited for the data.

In terms of bias and MSE, the B-nn and DB-nn imputations give similar results, because as expected, almost all the probabilities ψ_{uv} are equal to zero or one, leading to few differences between the two methods. Moreover, they both outperform DB-*k*nn and B-*k*nn. Adding a minimum number $k = 5$ of potential donors to add randomness to the imputation process does not appear to reduce the bias or better preserve the distribution here.

Table 3 shows the bias and MSE of the estimated correlation coefficients between the variables for the B-nn method. The linear relationships between the variables are very well preserved after imputation. We show only the results for the B-nn method, because the other methods yield comparable results.

TABLE 1: Bias and mean squared errors (MSE) with respect to the imputation and the nonresponse mechanisms, of the estimators of the totals, in the case of *k*nn, nn, FHDI, SReg, DB-nn, B-nn, DB-*k*nn and B-*k*nn imputations. The dataset contains Swiss cheese nonresponse, each variable contains approximately 10% of missing values.

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
True value	*9083.35*	*9633.20*	*25165.50*	*263.96*	*4757.90*
Bias					
*k*nn	-61.72	-29.88	-106.23	-0.10	-117.36
nn	-59.85	-27.61	-109.99	-0.05	-121.14
FHDI	-10.25	-13.85	-26.41	-0.02	-11.15
SReg	-5.61	-12.39	-16.13	-0.02	-1.59
DB-nn	-8.79	-12.30	-24.25	-0.03	-5.58
B-nn	-8.45	-11.93	-23.36	-0.02	-5.49
DB-*k*nn	-11.62	-12.79	-29.94	-0.02	-6.60
B-*k*nn	-11.33	-12.25	-29.13	-0.01	-6.36
MSE					
*k*nn	4327.35	1070.79	14137.19	0.03	16193.98
nn	4712.63	1122.08	16226.03	0.04	17816.19
FHDI	250.57	302.01	1417.03	0.00	511.15
SReg	111.90	287.71	920.28	0.00	168.77
DB-nn	197.59	259.34	1675.01	0.01	262.24
B-nn	176.11	242.32	1556.93	0.00	263.62
DB-*k*nn	219.00	240.36	1610.32	0.01	198.89
B-*k*nn	247.59	272.59	1831.69	0.00	370.77

TABLE 2: Pourcentage of bias and mean squared errors (MSE) with respect to the imputation and the nonresponse mechanisms, of estimated 50th and 75th percentiles, in the case of *k*nn, nn, FHDI, SReg, DB-nn, B-nn, DB-*k*nn and B-*k*nn imputations. The dataset contains Swiss cheese nonresponse, each variable contains approximately 10% of missing values.

	P50					P75				
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
True value	*33.28*	*36.92*	*94.25*	*1.04*	*12.42*	*39.04*	*39.87*	*105.30*	*1.07*	*25.20*
Bias										
*k*nn	-15.15	-8.80	-19.11	0.05	-75.75	-43.91	-23.10	-80.67	-0.24	-111.23
nn	-16.77	-7.30	-26.15	0.00	-65.30	-19.25	-10.10	-37.45	-0.04	-55.85
FHDI	0.45	-0.69	-0.78	0.01	-2.34	-6.94	-5.4	-7.36	-0.04	-13.08
SReg	2.37	-2.54	3.14	0.01	-0.98	-4.72	-3.56	3.85	-0.01	-5.17
DB-nn	-0.10	-1.90	-1.55	0.01	-1.20	-1.85	-6.07	-5.88	-0.02	-3.59
B-nn	-0.23	-1.89	-1.62	0.01	-1.34	-1.85	-6.07	-5.97	-0.02	-3.54
DB-*k*nn	1.35	-2.15	1.01	0.01	2.16	-8.46	-7.00	-13.60	0.00	-5.78
B-*k*nn	-0.94	-2.69	-2.84	0.01	-0.34	-3.89	-5.19	-9.92	-0.01	-4.32
MSE										
*k*nn	3.50	1.31	7.90	0.00	67.15	23.07	6.18	76.03	0.00	147.77
nn	3.97	1.12	11.96	0.00	56.67	6.76	1.82	28.27	0.00	42.31
FHDI	0.38	0.41	1.29	0.00	1.89	1.61	0.97	6.12	0.00	5.11
SReg	0.53	0.43	1.37	0.00	0.79	1.04	0.81	5.47	0.00	1.82
DB-nn	0.38	0.46	2.63	0.00	1.21	1.37	0.98	8.32	0.00	2.27
B-nn	0.37	0.48	2.59	0.00	1.23	1.39	0.98	8.43	0.00	2.33
DB-*k*nn	0.36	0.35	1.16	0.00	0.71	1.25	0.88	6.85	0.00	1.57
B-*k*nn	0.53	0.50	2.00	0.00	2.02	1.31	0.91	8.25	0.00	2.62

TABLE 3: Bias and mean squared errors (MSE) with respect to the imputation and the nonresponse mechanisms, of the estimators of the correlation coefficients, in the case of B-nn imputation. The dataset contains Swiss cheese nonresponse, each variable contains approximately 10% of missing values.

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
Bias					
$\mathbf{x}_1$	-	0.0002	-0.0063	-0.0018	-0.0022
$\mathbf{x}_2$		-	0.0001	-0.0058	0.0028
$\mathbf{x}_3$			-	-0.0050	-0.0006
$\mathbf{x}_4$				-	0.0046
$\mathbf{x}_5$					-
MSE					
$\mathbf{x}_1$	-	0.0003	0.0001	0.0001	0.0001
$\mathbf{x}_2$		-	0.0003	0.0006	0.0004
$\mathbf{x}_3$			-	0.0003	0.0001
$\mathbf{x}_4$				-	0.0000
$\mathbf{x}_5$					-

2.8 Discussion

In addition to Properties 2.6.1–2.6.3 on the unbiasedness of the estimated total, the method has two strengths: the possibility of imputing both categorical and continuous variables; and the possibility of forcing the probability ψ_{uv} to be null, if needed, for example, if the survey sampler does not want to allow a respondent u to be the donor of a nonrespondent v.

The variance of estimated parameters is a complex matter when the data sets are imputed, because it needs to consider the variability caused by the sampling design, nonresponses and

the imputation method. Determining an explicit variance estimator requires further investigation, possibly using a pseudo-population bootstrap variance estimator, as described in Chen et al. (2019).

(Eustache, Vallée, and Tillé, 2021) provide a sparse implementation of the methods. The imputation methods can be used in large-scale applications in which both the number of units and the number of variables with missing values are large. With the sparse implementation, the computation of the matrix of imputation probabilities is efficient in terms of computation time.

The choice of the minimum number k of possible donors, as proposed in Section 2.3.2, depends on the data set. The effect of different values of k on total estimators is left to future research.

Acknowledgements

The authors would like to thank the associate editor and the reviewers for their constructive comments.

Appendix

Proof of Property 1. Consider the column vectors of estimated totals

$$\widehat{\mathbf{T}}_{(-j)} = \left(\widehat{T}_1, \ldots, \widehat{T}_{(j-1)}, \widehat{T}_{(j+1)}, \ldots, \widehat{T}_J\right)^\top$$

and of Horvitz-Thompson estimators

$$\widehat{\mathbf{T}}^{HT}_{(-j)} = \left(\widehat{T}^{HT}_1, \ldots, \widehat{T}^{HT}_{(j-1)}, \widehat{T}^{HT}_{(j+1)}, \ldots, \widehat{T}^{HT}_J\right)^\top,$$

with $\widehat{T}^{HT}_j = \sum_{i\in S} d_i x_{ij}$ the Horvitz-Thompson estimator of the total of variable j. We have that

$$\begin{aligned}
\mathrm{E}_m\mathrm{E}_{imp}\left(\widehat{T}_j - \widehat{T}^{HT}_j\right) =& \mathrm{E}_m\left(\sum_{i\in S} r_{ij} d_i x_{ij} + \sum_{v\in S_m}(1-r_{vj})d_v \sum_{u\in S_r}\psi_{uv}x_{uj} - \sum_{i\in S} d_i x_{ij}\right)\\
=& \sum_{i\in S} r_{ij} d_i \boldsymbol{\beta}^{(-j)^\top}\mathbf{x}_i^{(-j)} + \sum_{v\in S_m}(1-r_{vj})d_v \sum_{u\in S_r}\psi_{uv}\boldsymbol{\beta}^{(-j)^\top}\mathbf{x}_u^{(-j)}\\
& - \sum_{i\in S} d_i \boldsymbol{\beta}^{(-j)^\top}\mathbf{x}_i^{(-j)}\\
=& \boldsymbol{\beta}^{(-j)^\top}\left\{\mathrm{E}_{imp}\left(\widehat{\mathbf{T}}_{(-j)}\right) - \widehat{\mathbf{T}}^{HT}_{(-j)}\right\} = 0.
\end{aligned}$$

The last equality comes from Equation (2.3.2). Using the requirements that the data are MAR or CMAR, the different expectations can be reversed to obtain the following development:

$$\begin{aligned}
\mathrm{Bias}\left(\widehat{T}_j\right) &= \mathrm{E}_m\mathrm{E}_p\mathrm{E}_q\mathrm{E}_{imp}(\widehat{T}_j - T_j) = \mathrm{E}_m\mathrm{E}_p\mathrm{E}_q\mathrm{E}_{imp}(\widehat{T}_j - \widehat{T}^{HT}_j + \widehat{T}^{HT}_j - T_j)\\
&= \mathrm{E}_p\mathrm{E}_q\mathrm{E}_m\mathrm{E}_{imp}(\widehat{T}_j - \widehat{T}^{HT}_j) = 0.
\end{aligned}$$

The proof remains the same for each variable $j \in \{1, \ldots, J\}$. ■

Proof of Property 2.

$$\begin{aligned}
\mathrm{E}_{imp}\left(\widehat{T}_j\right) &= \sum_{i\in S} r_{ij} d_i x_{ij} + \sum_{v\in S_m}(1-r_{vj}) d_v \sum_{u\in S_r} \psi_{uv} x_{uj} \\
&= \sum_{i\in S_r} d_i x_{ij} + \sum_{\ell\in S_m} r_{\ell j} d_\ell x_{\ell j} + \sum_{v\in S_m}(1-r_{vj}) d_v \sum_{u\in S_r} \psi_{uv} x_{uj} \\
&= \sum_{i\in S_r} d_i x_{ij} + \sum_{v\in S_m} d_v \sum_{u\in S_r} \psi_{uv} x_{uj} \\
&= \sum_{i\in S_r} d_i \left\{1 + \pi_i \sum_{v\in S_m} d_v \psi_{iv}\right\} x_{ij} \\
&= \sum_{i\in S_r} d_i w_i x_{ij},
\end{aligned}$$

where the third equality comes from Equation (2.3.2). If w_i^{-1} is approximately equal to the true response probability, we have

$$\mathrm{Bias}(\widehat{T}_j) = \mathrm{E}_p\mathrm{E}_q\mathrm{E}_{imp}\left(\widehat{T}_j - T_j\right) = \mathrm{E}_p\mathrm{E}_q\left(\sum_{i\in S_r} d_i w_i x_{ij} - \sum_{i\in S} x_{ij}\right) \approx 0.$$

Indeed, the quantity

$$\sum_{u\in S_r} \frac{d_u x_{uv}}{\Pr(u\in S_r|S)}$$

is an unbiased estimator of T_j if $\Pr(u\in S_r|S) > 0$, for all $u\in S_r$. If the true response probability is exactly w_i^{-1}, $\widehat{T}_j$ is unbiaised, i.e. $\mathrm{Bias}(\widehat{T}_j) = 0$. ■

Proof of Property 3.

$$\begin{aligned}
\mathrm{Bias}\left(\widehat{T}_j\right) &= \mathrm{E}_p\mathrm{E}_q\mathrm{E}_{imp}\left(\widehat{T}_j - T_j\right) \\
&= \mathrm{E}_p\mathrm{E}_q\mathrm{E}_{imp}\left(\sum_{i\in S} r_{ij} d_i x_{ij} + \sum_{v\in S_m}(1-r_{vj}) d_v \sum_{u\in S_r} \phi_{uv} x_{uj} - \sum_{i\in S} x_{ij}\right) \\
&= \mathrm{E}_p\mathrm{E}_q\mathrm{E}_{imp}\left(\sum_{i\in S} r_{ij} d_i x_{ij} + \sum_{v\in S_m}(1-r_{vj}) d_v x_{vj} - \sum_{i\in S} x_{ij}\right) = 0.
\end{aligned}$$

■

Chapter 3

Quasi-Model-Assisted Estimators under Nonresponse in Sample Surveys

Abstract: In the presence of auxiliary information, model-assisted estimators use a working model linking the variable of interest to the auxiliary variables in order to improve the efficiency of the Horvitz-Thompson estimator. In this work, we adapt model-assisted total estimators to missing at random data building on the idea of nonresponse weighting adjustment. We consider nonresponse as a second phase of the survey and reweight the units in model-assisted estimators using the inverse of estimated response probabilities. We show that our proposed estimator can be written as a reweighted estimator with resulting weights calibrated to the total of the auxiliary variables for some working models. We show that our proposed estimator is doubly robust to model misspecification and provide formulae for asymptotic variance and variance estimators. We conduct a simulation study that confirms the performance and robustness to model misspecification of our proposed estimators.

Keywords: auxiliary information, Horvitz-Thompson estimator, missing data, response probabilities, superpopulation model, weighting adjustment.

This chapter is a reprint of article Eustache and Hasler (2022).

3.1 Introduction

In surveys with complete response, the Horvitz-Thompson (HT) estimator is a design-unbiased estimator of population totals (Horvitz and Thompson, 1952). In this article, we focus on estimating the total of a variable of interest. In the presence of auxiliary information, model-assisted estimators can improve the efficiency of the HT estimator by incorporating into the estimator a working model that links the variable of interest and the auxiliary variables. Such estimators are asymptotically unbiased and asymptotically more efficient than the HT estimator regardless of whether the working model is correctly specified or not. The roots of model-assisted is, to the best of our knowldege, the Generalized REGression estimator (GREG) introduced in Särndal (1980).

Nonresponse occurs when the variable of interest is observed only for a part of the sampled units. The aforementioned estimators are unavailable with nonresponse. Estimation in surveys with nonresponse has been widely investigated. The book Särndal and Lundström (2005) is an excellent overview.

One approach to handle nonresponse consists of reweighting the survey respondents to compensate for the nonrespondents. The resulting estimator is called Nonresponse Weighting Adjusted (NWA) estimator, or empirical double expansion estimator, or propensity score adjusted estimator. Overviews and critical reviews of weighting adjustments are presented in Lundström and Särndal (1999), Lee, Rancourt, and Särndal (2002), Brick (2013), and Haziza and Beaumont (2017).

In this article, we propose a quasi-model-assisted estimator adapted for nonresponse. Following Beaumont (2005), we use the prefix "quasi" to indicate that our proposed estimator is not exactly a model-assisted estimator, but rather a blend between model-assisted and NWA estimators. It is based on a working model and a response model. We reweight the respondents in a model-assisted estimator to compensate for the nonrespondents.

We present our quasi-model-assisted estimator for different working models and different approaches to estimating the response model. We show that the proposed estimator can be viewed as a reweighted estimator and that the resulting weights are calibrated to the totals of the auxiliary variables for some working models. We show that our proposed estimator is doubly robust in the sense that it is asymptotically unbiased and asymptotically at least as efficient as the HT estimator, even when one of the two specified models (i.e. the working model or the response model) is misspecified. We provide a formula for the asymptotic variance and a variance estimator of the proposed estimator. We conduct a simulation study showing that our proposed estimator performs well in terms of bias and variance, even when one of the two models is misspecified.

The article is organized as follows. Section 3.2 presents the basic setup. We introduce model-assisted and NWA estimators in Sections 3.3 and 3.4. Our quasi-model-assisted estimator is proposed in Section 3.5. We study different statistical learning techniques as working models in Section 3.6. We develop the asymptotic properties of our proposed estimator in Section 3.7. In section 3.8, we discuss the variance and its estimator. In Section 3.9, two simulation studies, one with simulated data and the other with real data, confirm the performance of our estimator. The main part of this article closes with a short discussion in Section 3.10. Supplementary material is presented in the Appendices.

3.2 Basic setup

We consider a finite population $U = \{1, 2, ..., N\}$. Let $s \subset U$ be a sample of size n selected from U according to a sampling design $p(.)$. The first- and second-order inclusion probabilities are denoted, respectively, by π_k and $\pi_{k\ell} = \text{pr}(k, \ell \in s)$ for generic units k and ℓ. Consider the sample membership indicator a_k of a unit k. We have $\text{pr}(a_k = 1) = \pi_k$, $\text{pr}(a_k = 0) = 1 - \pi_k$, and $\text{E}_p(a_k) = \pi_k$, where subscript p means that the expectation is computed with respect to sampling design $p(.)$. The covariance between the sample membership indicators is $\Delta_{k\ell} = \text{cov}_p(a_k, a_\ell) = \pi_{k\ell} - \pi_k\pi_\ell$.

The goal is to estimate the population total

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

of a variable of interest y with values $\{y_k\}$ known only for those units in the sample. With no additional information, the total t can be estimated by the HT estimator

$$\widehat{t}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k},$$

which is design-unbiased, i.e. $\text{E}_p(\widehat{t}_{HT}) = t$, if $\pi_k > 0$ for all $k \in U$. If, additionally, $\pi_{k\ell} > 0$ for all $k, \ell \in U$, a design-unbiased estimator of the variance of $\widehat{t}_{HT}$ is

$$\widehat{\text{var}}\left(\widehat{t}_{HT}\right) = \sum_{k \in s} \sum_{\ell \in s} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell}. \tag{3.2.1}$$

3.3 Model-assisted estimators

Suppose that a vector of auxiliary variables $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})^\top$ is known for each sampled unit $k \in s$. The idea of model assisted estimation is to postulate a model, sometimes called the *working model* or the *superpopulation model*, that links y_k and $\mathbf{x}_k$. This model is then used to improve the efficiency of the HT estimator while maintaining, or almost, its design unbiasedness.

We assume that the finite population of y_k's conditioned on $\mathbf{x}_k$'s is a realization of an infinite working model ξ, in which

$$\xi : y_k = m(\mathbf{x}_k) + \varepsilon_k,\ k \in U,$$

where $m(.)$ is an unknown function, $\mathrm{E}_\xi(\varepsilon_k) = 0$, $\mathrm{var}_\xi(\varepsilon_k) = \sigma_k^2$, and subscript ξ indicates that the expectation and variance are computed under model ξ.

The model-assisted *difference estimator* of t is

$$\widehat{t}_m = \sum_{k \in U} m(\mathbf{x}_k) + \sum_{k \in s} \frac{y_k - m(\mathbf{x}_k)}{\pi_k}. \tag{3.3.2}$$

Function $m(.)$ is unknown. Say we could estimate it from population values $\{(\mathbf{x}_k, y_k)\}, k \in U$ which would provide $m_U(.)$. Note that $m_U(.)$ is independent from the sample. Replacing $m(.)$ by $m_U(.)$ in the difference estimator yields the *pseudo-generalized difference estimator*

$$\widehat{t}_{m_U} = \sum_{k \in U} m_U(\mathbf{x}_k) + \sum_{k \in s} \frac{y_k - m_U(\mathbf{x}_k)}{\pi_k}. \tag{3.3.3}$$

Breidt and Opsomer (2017) show that both estimators (3.3.2) and (3.3.3) are 1) design unbiased and 2) more efficient than the HT estimator provided that the "residuals" $\{y_k - m(\mathbf{x}_k)\}$ and $\{y_k - m_U(\mathbf{x}_k)\}$ have, respectively, less variability than the "raw values" $\{y_k\}$. This holds regardless of the quality of working model ξ.

The population estimator $m_U(.)$ is generally unknown. It can be estimated by $m_s(.)$ based on the known sample values $\{(\mathbf{x}_k, y_k)\}, k \in s$. Replacing in (3.3.3), we obtain the *model-assisted estimator*

$$\widehat{t}_{m_s} = \sum_{k \in U} m_s(\mathbf{x}_k) + \sum_{k \in s} \frac{y_k - m_s(\mathbf{x}_k)}{\pi_k}. \tag{3.3.4}$$

Breidt and Opsomer (2017) show that, under some regularity conditions and for some specific working models including heteroscedastic multiple regression, linear mixed models, and some statistical learning techniques, the model-assisted estimator $\widehat{t}_{m_s}$ is 1) asymptotically design unbiased and 2) asymptotically more efficient than the HT estimator provided that the "residuals" $\{y_k - m_s(\mathbf{x}_k)\}$ have less variability than the "raw values" $\{y_k\}$. This holds regardless of the quality of working model ξ.

To the best of our knowledge, the roots of model-assisted estimation is the GREG estimator in which the working model is linear. It was introduced in Särndal (1980) and further studied in Robinson and Särndal (1983) and Särndal and Wright (1984). Literature on model assisted has flourished in the last three decades. Särndal, Swensson, and Wretman (1992) propose a comprehensive presentation of the model-assisted approach. Särndal and Swensson (1987) study model-assisted estimation with a linear working model at two levels of knowledge for the auxiliary information - sample and population levels. Firth and Bennett (1998) introduce conditions for design consistency of estimators including model-assisted estimators. Section 2.3 of Fuller (2009) presents an overview of the theory of model-assisted estimation. Linear and non-linear models are discussed and some asymptotic results are presented.

More recent works study model-assisted survey estimation with modern and flexible working models. Breidt and Opsomer (2017) provide an overview. To cite just a few of these works, the

working model is estimated with local polynomial regression in Breidt and Opsomer (2000), with penalised splines in Breidt, Claeskens, and Opsomer (2005), with generalized additive models in Opsomer et al. (2007), and with random forests in Dagdoug, Goga, and Haziza (2022). Opsomer et al. (2007) extend the nonparametric model-assisted methodology of Breidt and Opsomer (2000), who considered univariate models and single-phase estimation. Breidt et al. (2007) study model-assisted estimators with semi-parametric working models by incorporating some variables linearly, and others through smooth additive terms. The linear and non-linear parts are estimated simultaneously using a backfitting algorithm for additive models of Hastie and Tibshirani (1990). Cicchitelli and Montanari (2012) present model-assisted estimators where variable space is an auxiliary variable in the working model. This allows the authors to take into account the spacial auto-correlation.

3.4 NWA estimators

In practice, some values $\{y_k\}$ may be missing because they are collected incorrectly or some units refrain from responding. In this work, we will suppose that the values are missing because of the second reason: nonresponse. Let p_k and r_k denote, respectively, the response probability and response indicator to variable y of a unit $k \in U$, with $\text{pr}(r_k = 1) = p_k$ and $\text{pr}(r_k = 0) = 1 - p_k$. Consider $s_r = \{k \in U \mid a_k = 1, r_k = 1\}$, the set of n_r units in s for which variable y is known. The units in s_r are called *respondents*. The process that generates the respondents is called the *response mechanism*.

Nonresponse can be seen as a second phase of the survey as suggested by Hansen and Hurwitz (1946). This second phase consists of selecting a sample of survey respondents from the sampled units based on a response process. In the first phase, a sample s is selected from population U according to a sampling design $p(.)$. In the second phase, a sample s_r is selected from s according to a Poisson sampling design with unknown inclusion probabilities $\{p_k\}$, called the response probabilities (Särndal and Swensson, 1987). It results in a partition of the sample s into two subsamples: the set of respondents and the set of nonrespondents.

In the presence of nonresponse, all of the aforementioned estimators are unavailable. An approach to control nonresponse bias consists of reweighting the survey respondents to compensate for the nonrespondents. The roots of this method is probably Särndal and Swensson (1987), who use the similarity between two-phase sampling and nonresponse to increase the survey weights using the inverse of the response probabilities. If nonresponse is seen as a second phase of the survey, the design weights are multiplied by the inverse of the response probabilities to obtain the *two-phase estimator* or *double expansion estimator*

$$\widehat{t}_{2HT} = \sum_{k \in s_r} \frac{y_k}{\pi_k p_k}.$$

Since the response probabilities are unknown in practice, we may model it using a *response model*. Based on this model, the probabilities are estimated by $\widehat{p}_k$ and used instead of the true response probabilities p_k in the two-phase estimator. This results in the *NWA estimator*, or *empirical double expansion estimator*, or *propensity score adjusted estimator*

$$\widehat{t}_{NWA} = \sum_{k \in s_r} \frac{y_k}{\pi_k \widehat{p}_k}.$$

Since the pioneering work of Särndal and Swensson (1987), many studies on the NWA have been developed and several methods of weighting adjustments - or response probability estimation - have been proposed. We present some of these in the following paragraphs. Overviews and

critical reviews of weighting adjustments are presented in Lundström and Särndal (1999), Lee, Rancourt, and Särndal (2002), Brick (2013), and Haziza and Beaumont (2017).

A first approach to NWA is maximum likelihood. In the context of the Finnish Household budget survey, Ekholm and Laaksonen (1991) model the response probabilities with logistic regression on the explanatory variables and estimate them via maximum likelihood. Kim and Kim (2007) develop the asymptotic properties of the NWA estimator under a general parametric response model on the explanatory variables. The model parameters are estimated via maximum likelihood. They prove that the resulting NWA estimator is generally more efficient than an estimator that uses the true response probabilities. This result is also shown by Beaumont (2005) for a logistic response model.

Another approach to NWA consists of finding weights close to the initial survey weights so that the reweighted estimators of some auxiliary variables respect known (or design-unbiased estimated) population totals. This procedure is called calibration and was first formalized in Deville and Särndal (1992). Särndal and Lundström (2005) provide an overview of the calibration approach to NWA.

Following Lundström and Särndal (1999), two levels of auxiliary information can be considered: 1) Info-U when the population totals of the auxiliary variables are known in addition to the values of the auxiliary variables for all sampled units; 2) Info-S when only the values of these variables for the sampled units are known. At level Info-S, calibration is performed on the design-unbiased HT estimator while at Info-U it can be performed directly on the true totals. Iannacchione, Milne, and Folsom (1991) propose NWA using logistic regression and calibration at Info-S. In Lundström and Särndal (1999), the inverse of the estimated response probabilities are considered as adjustment weights to reduce nonresponse bias at both Info-S and Info-U. Brick and Jones (2008) examines nonresponse bias with different methods of calibration weighting, namely raking and linear.

Kim and Riddles (2012) discuss some asymptotic properties of NWA estimators and derive optimal estimators based on a regression model for the finite population. A parametric model for the response probabilities is assumed. The authors consider a class of consistent estimators of the parameters of the response model that can be written as a solution to an estimating equation. Calibration at Info-S can be viewed as a particular case. Hasler (2023) introduces a common framework in which the asymptotic behavior of the NWA estimator with response probabilities estimated via Maximum likelihood and calibration at both levels, Info-U and Info-S, can be compared. A logistic response model is assumed. Fuller, Loughin, and Baker (1994) propose a reweighting approach based on regression that yields positive and not too extreme weights. An application to the 1987-1988 nation wide Food Consumption Survey is presented.

Under this approach, calibration is performed on the variables used to estimate the response model. Generalized calibration, also called instrumental calibration, allows for these two sets of variables to differ. This approach to NWA was introduced by Deville (1998) and Deville (2002). This method constructs estimators with a variance that is equivalent to that of the HT when applied to the residuals of the regression of the survey variable on the auxiliary variables using the response model variables as instrumental variables. The generalized calibration approach was also considered in Kott (2006), Kott and Chang (2010), and Kott and Liao (2012). Lesage, Haziza, and d'Haultfoeuille (2019) give sufficient conditions for the consistency of the generalized calibration estimator and show how the estimator may have a large bias when some of these conditions are violated. Traditionally, the number of calibration variables is equal to the number of response model variables. Chang and Kott (2008) expand this approach to the case where the number of calibration variables is greater than the number of response model variables.

A nonparametric approach to NWA has also been considered in the literature. For instance, Niyonsenga (1994) and Niyonsenga (1997) considers such an approach when unit and item nonresponse occur together in a survey. Da Silva and Opsomer (2009) use a local polynomial regression to estimate the response probabilities and present the asymptotic properties of the resulting

NWA.

3.5 Quasi-model-assisted estimator

In this paper, we introduce a *quasi-model-assisted estimator* adapted to nonresponse. It is a blend between a model-assisted estimator and a NWA estimator. We borrow the terminology 'quasi-model-assisted' to Beaumont (2005) to indicate that our estimator is not exactly a model-assisted estimator. It is constructed as follows. We replace the estimated function $m_s(\cdot)$, unavailable with nonresponse, by an estimator $m_r(\cdot)$ constructed from the respondents in the model-assisted estimator (3.3.4) and we see nonresponse as a second phase of the survey. This gives the *two-phase model-assisted estimator*

$$\widehat{t}_{m_r,p} = \sum_{k\in U} m_r(\mathbf{x}_k) + \sum_{k\in s_r} \frac{y_k - m_r(\mathbf{x}_k)}{\pi_k p_k}.$$

It is unavailable in practice since it contains the unknown response probabilities $\{p_k\}$. We borrow the idea of the NWA estimation and obtain

$$\widehat{t}_{m_r,\widehat{p}} = \sum_{k\in U} m_r(\mathbf{x}_k) + \sum_{k\in s_r} \frac{y_k - m_r(\mathbf{x}_k)}{\pi_k \widehat{p}_k}. \tag{3.5.5}$$

We call this estimator the *quasi-model-assisted estimator*. It corresponds to a model-assisted estimator where the weights are adjusted for nonresponse. This estimator relies on two models, the working model ξ and the response model. It covers a wide range of estimators depending on the chosen working model ξ and the chosen response model. The first term of estimator (3.5.5) is the population total of the predicted values $\{m_r(\mathbf{x}_k)\}$. For most working models, this requires the values $\{\mathbf{x}_k\}$ to be known for all population units. If these values are known for sampled units only, we may use a HT-type estimator of this sum, see Appendix 3.11.

Some other papers propose estimators that can be defined as quasi-model-assisted. Beaumont (2005) studies an estimator with the same form as the one we propose. He uses calibrated imputation to compensate for nonresponse by modifying the estimated values $\{m_r(\mathbf{x})\}_{k\in S}$, while satisfying some balanced constraints. Kim and Haziza (2014) also propose a quasi-model-assisted estimator. The working model and the response model are estimated simultaneously based on estimating equations. These two estimators previously cited only consider parametric working and response models. Our proposed approach is more general and allows for nonparametric models.

The quasi-model-assisted estimator in (3.5.5) contains two estimators: the response probabilities $\{\widehat{p}_k\}$ and the function $m_r(\cdot)$. Depending on these two choices, we obtain a different estimator. In what follows, we assume, for simplicity, that the response probabilities are parametrically modeled such that $p_k = 1/F(\mathbf{z}_k^\top \boldsymbol{\lambda}_0)$ where $F(.)$ is a known function with unknown parameter $\boldsymbol{\lambda}_0$ and $\mathbf{z}_k$ is a vector of variables observed for both respondents and nonrespondents. The estimated response probabilities are $\widehat{p}_k = 1/F(\mathbf{z}_k^\top \widehat{\boldsymbol{\lambda}})$ for some estimator $\widehat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}_0$. The response probabilities can be estimated using generalized calibration (Deville, 1998), that allows for the variables in the response model to differ from the variables on which we calibrate. This estimator is consistent for the population total under some regularity conditions, see Lesage, Haziza, and d'Haultfoeuille (2019), Section 3. The estimator $\widehat{\boldsymbol{\lambda}}$ is the solution to the estimating equation

$$Q(\boldsymbol{\lambda}) = \sum_{k\in U} \mathbf{x}_k - \sum_{k\in S_r} \frac{\mathbf{x}_k}{\pi_k} F(\mathbf{z}_k^\top \boldsymbol{\lambda}) = 0. \tag{3.5.6}$$

We present quasi-model-assisted estimators with response probabilities estimated via an alternate technique, maximum likelihood, in Appendix 3.12.

Our proposed estimator is generic and reduces to well-known estimators for some choices of response model and working model. If the working model is linear and the response probabilities are estimated via calibration, our proposed estimator reduces to the simple NWA estimator. If the working model is linear and the response probabilities are obtained via generalized calibration, our proposed estimator reduces to the generalized calibration estimator. If the response probabilities are obtained via model calibration where the model calibration includes the working model, then our estimator is a model calibration estimator. If the response probabilities are known and used instead of the estimated ones, then our estimator is a model-assisted estimator for two-phase sampling.

Suppose a particular case where $m_r(\cdot)$ is free from $\widehat{p}_k$. It is for instance the case for the GREG estimator with weights $c_k = 1$ or σ_k^2 and the working models based on statistical learning techniques presented in Section 3.6. Consider that the response probabilities are estimated via generalized calibration where one of the calibration variables $\mathbf{z}_k$ is $m_r(\mathbf{x}_k)$. Because $\mathbf{z}_k$ includes $m_r(\mathbf{x}_k)$, Equation (3.5.6) implies

$$\sum_{k\in U} m_r(\mathbf{x}_k) = \sum_{k\in s_r} \frac{m_r(\mathbf{x}_k)}{\pi_k} F(\mathbf{x}_k^\top \widehat{\boldsymbol{\lambda}}),$$

and the quasi-model-assisted estimator can then be written

$$\widehat{t}_{m_r,\widehat{p}} = \sum_{k\in s_r} \frac{y_k}{\pi_k \widehat{p}_k}.$$

This estimator is a generalized calibration estimator. It is consistent for the population total under some regularity conditions, see Lesage, Haziza, and d'Haultfoeuille (2019), Section 3.

3.6 Statistical Learning Techniques

3.6.1 Generalized Regression

Consider the linear working model

$$\xi : y_k = \mathbf{x}_k^\top \beta + \varepsilon_k,$$

where the ε_k are uncorrelated with mean $E_\xi(\varepsilon_k) = 0$ and variance $\text{var}_\xi(\varepsilon_k) = \sigma_k^2$. The finite population regression coefficient is

$$\mathbf{B}_U = \left(\sum_{k\in U} \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \sum_{k\in U} \mathbf{x}_k y_k.$$

If parameter β is estimated based on s_r we use

$$\mathbf{B}_r = \left(\sum_{k\in s_r} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{c_k}\right)^{-1} \sum_{k\in s_r} \frac{\mathbf{x}_k y_k}{c_k},$$

where $m_r(\mathbf{x}_k) = \mathbf{x}_k^\top \mathbf{B}_r$ and c_k is any of $1, \sigma_k^2, \pi_k \widehat{p}_k, \pi_k \widehat{p}_k \sigma_k^2$.

The quasi-model-assisted estimator can be written in weighted form

$$\begin{aligned}\widehat{t}_{m_r,\widehat{p}} &= \sum_{k\in U}\mathbf{x}_k^\top\mathbf{B}_r + \sum_{k\in s_r}\frac{y_k-\mathbf{x}_k^\top\mathbf{B}_r}{\pi_k\widehat{p}_k}\\ &= \sum_{k\in s_r}\frac{y_k}{\pi_k\widehat{p}_k} + \left(\sum_{k\in U}\mathbf{x}_k - \sum_{k\in s_r}\frac{\mathbf{x}_k}{\pi_k\widehat{p}_k}\right)^\top\left(\sum_{k\in s_r}\frac{\mathbf{x}_k\mathbf{x}_k^\top}{c_k}\right)^{-1}\sum_{k\in s_r}\frac{\mathbf{x}_k y_k}{c_k}\\ &= \sum_{k\in s_r}\left\{\frac{1}{\pi_k\widehat{p}_k} + \left(\mathbf{t}^X-\widehat{\mathbf{t}}^X_{NWA}\right)^\top\left(\sum_{\ell\in s_r}\frac{\mathbf{x}_\ell\mathbf{x}_\ell^\top}{c_\ell}\right)^{-1}\frac{\mathbf{x}_k}{c_k}\right\}y_k\\ &= \sum_{k\in s_r} w_{k,s_r}y_k,\end{aligned}$$

where $\mathbf{t}^X$ is the vector of population totals of the auxiliary variables and $\widehat{\mathbf{t}}^X_{NWA}$ its NWA estimator. The weights w_{k,s_r} are those of the NWA estimator $1/(\pi_k\widehat{p}_k)$ plus a corrective term induced by the working model. The second term cancels when calibration is applied to estimate the response probabilities. The quasi-model-assisted estimator is the NWA estimator in this case. The weights are free from values $\{\mathbf{x}_k\}$ in $U\backslash s_r$ except through the population totals $\mathbf{t}^X$. Only the values $\{\mathbf{x}_k\}$ on s_r and the population totals $\mathbf{t}^X$ are needed to compute the quasi-model-assisted estimator, unless some other values are needed to estimate the response probabilities. The weights are free from $\{y_k\}$. They can therefore be used for several variables of interest provided that they have observed values on s_r. In particular, the weights can be applied to the auxiliary variables. Let $\widehat{\mathbf{t}}^X_{m_r,\widehat{p}}$ be the quasi-model-assisted estimator for the auxiliary variables. It comes

$$\widehat{\mathbf{t}}^X_{m_r,\widehat{p}} = \sum_{k\in s_r}\left\{\frac{1}{\pi_k\widehat{p}_k} + \left(\mathbf{t}^X-\widehat{\mathbf{t}}^X_{NWA}\right)\left(\sum_{k\in s_r}\frac{\mathbf{x}_k\mathbf{x}_k^\top}{c_k}\right)^{-1}\frac{\mathbf{x}_k}{c_k}\right\}\mathbf{x}_k^\top = \mathbf{t}^X.$$

This means that the weights of the quasi-model-assisted estimator are calibrated to the totals of the auxiliary variables when calibration is applied to estimate the response probabilities.

3.6.2 K-Nearest Neighbor

Consider the working model where the prediction for a nonrespondent is obtained by averaging the y-values of the closest respondents. A predicted value $m_r(\mathbf{x}_k)$ is obtained by

$$m_r(\mathbf{x}_k) = \frac{1}{K}\sum_{\ell\in L_k} y_\ell,$$

where $L_k^{(K)} < n_r$ is the set of the K nearest respondents of unit k. The neighborhood is determined based on the auxiliary variables and a distance measure such as the Euclidean distance. Consider $\alpha_{k\ell}$ an indicator that takes value 1 if respondent $\ell\in s_r$ is in the neighborhood $L_k^{(K)}$ of unit $k\in U$. We have $\alpha_{k\ell}=0$ if $\ell\in U\backslash s_r$. A prediction can be written

$$m_r(\mathbf{x}_k) = \frac{1}{K}\sum_{\ell\in s_r}\alpha_{k\ell}y_\ell.$$

The quasi-model-assisted estimator can be written in weighted form

$$\begin{aligned}\widehat{t}_{m_r,\widehat{p}} &= \sum_{k\in U} m_r(\mathbf{x}_k) + \sum_{k\in s_r} \frac{y_k - m_r(\mathbf{x}_k)}{\pi_k \widehat{p}_k} \\ &= \sum_{k\in U} \frac{1}{K} \sum_{\ell\in s_r} \alpha_{k\ell} y_\ell + \sum_{k\in s_r} \frac{y_k}{\pi_k \widehat{p}_k} - \sum_{k\in s_r} \frac{1}{\pi_k \widehat{p}_k K} \sum_{\ell\in s_r} \alpha_{k\ell} y_\ell \\ &= \sum_{\ell\in s_r} \left\{ \frac{1}{\pi_\ell \widehat{p}_\ell} + \frac{1}{K} \left(\sum_{k\in U} \alpha_{k\ell} - \sum_{k\in s_r} \frac{1}{\pi_k \widehat{p}_k} \alpha_{k\ell} \right) \right\} y_\ell .\end{aligned}$$

The weights are the ones of the NWA estimator $1/(\pi_k \widehat{p}_k)$ plus a corrective term induced by the working model. The second term cancels when the response probabilities are calibrated on variables $(\alpha_{1\ell}, \alpha_{2\ell}, \ldots, \alpha_{N\ell})^\top$, $\ell \in s_r$. The quasi-model-assisted estimator is the NWA estimator in this case. The weights depend on the values of the auxiliary variables through the distance measure applied to construct the neighborhoods. They are free from values $\{y_k\}$ and could therefore be used for several variables of interest provided that they have observed values on s_r. In particular, they can be applied to $\{\mathbf{x}_k\}$. This yields

$$\begin{aligned}\widehat{t}^{X}_{m_r,\widehat{p}} &= \sum_{k\in s_r} \frac{\mathbf{x}_k}{\pi_k \widehat{p}_k} + \sum_{k\in U} \frac{1}{K} \sum_{\ell\in s_r} \alpha_{k\ell} \mathbf{x}_\ell - \sum_{k\in s_r} \frac{1}{\pi_k \widehat{p}_k} \frac{1}{K} \sum_{\ell\in s_r} \alpha_{k\ell} \mathbf{x}_\ell \\ &= \sum_{k\in s_r} \frac{\mathbf{x}_k}{\pi_k \widehat{p}_k} + \frac{1}{K} \sum_{k\in U} \left(1 - \frac{a_k r_k}{\pi_k \widehat{p}_k} \right) \sum_{\ell\in s_r} \alpha_{k\ell} \mathbf{x}_\ell .\end{aligned}$$

The weights are calibrated to the totals of the auxiliary variables when

$$K^{-1} \sum_{\ell\in s_r} \alpha_{k\ell} \mathbf{x}_\ell = \mathbf{x}_k$$

for all $k \in U$. This is for instance the case when the neighborhoods $L_k^{(K)}$ are disjoints and have constant values $\{\mathbf{x}_k\}$. In practice, we can reasonably assume that this holds at least approximately for large populations and samples.

3.6.3 Local Polynomial Regression

Local polynomial regression is studied in the context of model-assisted survey estimation in Breidt and Opsomer (2000). Consider a working model in which $\mathbf{x}_k$ is a scalar, i.e. $\mathbf{x}_k = x_k$, $x_k \in \mathbb{R}$. Function $m(\cdot)$ is approximated locally at x_k by q-th order polynomial regression. The model is fitted via weighted least squares with weights based on a kernel function centered at x_k. Breidt and Opsomer (2000) propose and study the model-assisted estimator with a survey weighted estimator of $m(\cdot)$ fitted at the sample level. Adapting their estimator to the context of nonresponse yields

$$m_r(x_k) = \mathbf{e}_1 \cdot \left(\mathbf{X}_{rk}^\top \mathbf{W}_{rk} \mathbf{X}_{rk} \right)^{-1} \mathbf{X}_{rk}^\top \mathbf{W}_{rk} \mathbf{y}_{rk} = \omega_{rk}^\top \mathbf{y}_{rk},$$

where $\mathbf{e}_j$ is a vector with 1 at the j-th coordinate and 0 otherwise,

$$\mathbf{X}_{rk} = \begin{bmatrix} 1 & x_j - x_k & \cdots & (x_j - x_k)^q \end{bmatrix}_{j\in s_r},$$

$$\mathbf{W}_{rk} = \operatorname{diag} \left\{ \frac{1}{k_j h} K \left(\frac{x_j - x_k}{h} \right) \right\}_{j\in s_r},$$

$\omega_{rk}^{\top} = \mathbf{e}_1 \cdot \left(\mathbf{X}_{rk}^{\top}\mathbf{W}_{rk}\mathbf{X}_{rk}\right)^{-1}\mathbf{X}_{rk}^{\top}\mathbf{W}_{rk}$ and

$$\mathbf{y}_{rk}^{\top} = \left[y_j\right]_{j\in s_r},$$

and k_j is either 1 for all $j \in s_r$ or $\pi_j\widehat{p}_j$, $K(\cdot)$ is a continuous kernel function, and h a bandwidth. The quasi-model-assisted estimator can be written in weighted form

$$\begin{aligned}\widehat{t}_{m_r,\widehat{p}} &= \sum_{k\in U} m_r(x_k) + \sum_{k\in s_r}\frac{y_k - m_r(x_k)}{\pi_k\widehat{p}_k}\\ &= \sum_{k\in U}\omega_{rk}^{\top}\mathbf{y}_{rk} + \sum_{k\in s_r}\frac{y_k}{\pi_k\widehat{p}_k} - \sum_{k\in s_r}\frac{1}{\pi_k\widehat{p}_k}\omega_{rk}^{\top}\mathbf{y}_{rk}\\ &= \sum_{k\in s_r}\left\{\frac{1}{\pi_k\widehat{p}_k} + \sum_{\ell\in U}\left(1 - \frac{a_\ell r_\ell}{\pi_\ell\widehat{p}_\ell}\right)\omega_{r\ell}^{\top}\mathbf{e}_k\right\} y_k.\end{aligned}$$

The weights are the one of the NWA estimator $1/(\pi_k\widehat{p}_k)$ plus a corrective term induced by the working model. They are free from values $\{y_k\}$ and could therefore be used for several variables of interest provided that they have observed values on s_r. In particular, they can be applied to $\{x_k\}$. This yields

$$\widehat{t}^{X}_{m_r,\widehat{p}} = \sum_{k\in s_r}\frac{x_k}{\pi_k\widehat{p}_k} + \sum_{\ell\in U}\omega_{r\ell}^{\top}\sum_{k\in s_r}\mathbf{e}_k x_k + \sum_{\ell\in s_r}\frac{\omega_{r\ell}^{\top}}{\pi_\ell\widehat{p}_\ell}\sum_{k\in s_r}\mathbf{e}_k x_k = \sum_{k\in U} x_k,$$

where we used

$$\omega_{r\ell}^{\top}\sum_{k\in s_r}\mathbf{e}_k x_k = x_\ell.$$

The weights are calibrated to the totals of the auxiliary variables.

3.7 Asymptotics and Double Robustness

3.7.1 Double robustness

In this section, we develop the asymptotic properties of the proposed estimator. Estimators adjusted for nonresponse are often based on two models, a response model and a working model, sometimes also called outcome regression model. Such estimators based on two models are doubly robust, or provide double protection against model misspecification, if key properties such as consistency are maintained when one of the two models is misspecified. Double robustness has been studied in the context of sample surveys.

Kott (1994) discusses the double robustness of two estimators: a regression estimator and an imputed estimator, i.e., an HT estimator where the missing values are predicted using a working model. Kott (2006), Chang and Kott (2008), and Kott and Chang (2010) present double robustness of the generalized calibration estimator. Haziza and Rao (2006) show how to obtain imputed values so that the resulting estimator is approximately unbiased under two different approaches to make inference. Kim and Park (2006) present a new ratio imputation method that uses response probability, that is unbiased if one of the two model is correctly specified. Kott and Liao (2012) discuss double robustness of the calibrated estimator with nonlinear weights adjustments. Kim and Haziza (2014) propose a method to compute propensity scores or response probabilities

that leads to doubly robust estimation. The authors also propose a doubly robust variance estimator. Haziza, Chen, and Gao (2022) propose a weighting procedure that yields a doubly robust estimator for variables seen as key survey variables.

3.7.2 Setup

To study the asymptotic properties of the proposed estimator, it is useful to introduce a sampling design $p^*(.)$ that selects the sample s_r directly in population U. The associated first- and second-order inclusion probabilities are, respectively, $\pi_k^* = \pi_k p_k$ and

$$\pi_{k\ell}^* = \begin{cases} \pi_{k\ell} p_k p_\ell, & \text{if } k \neq \ell; \\ \pi_k p_k, & \text{if } k = \ell. \end{cases}$$

The membership indicator of a unit $k \in U$ in the set of respondents s_r is $a_k^* = a_k r_k$. The membership indicator of two different units $k, \ell \in U, k \neq \ell$ in s_r is $a_{k\ell}^* = a_k a_\ell r_k r_\ell$. Given that the nonresponse process is independent from the selected sample, we have $\mathrm{E}_{p^*}(a_k^*) = \mathrm{E}_p \mathrm{E}_q(a_k r_k) = \pi_k p_k$ and $\mathrm{E}_{p^*}(a_{k\ell}^*) = \pi_{k\ell} p_k p_\ell$, where the subscript p^* means that the expectation is computed with respect to the two-phase sampling design $p^*(.)$. The covariance between the membership indicators $\{a_k^*\}$ is

$$\Delta_{k\ell}^* = \begin{cases} \Delta_{k\ell} p_k p_\ell = (\pi_{k\ell} - \pi_k \pi_\ell) p_k p_\ell, & \text{if } k \neq \ell; \\ \pi_k p_k (1 - \pi_k p_k), & \text{if } k = \ell. \end{cases}$$

In what follows, the reference distribution for the convergence of estimators is the distribution induced by the two-phase sampling design $p^*(\cdot)$ unless otherwise specified.

We build on the asymptotic framework of Isaki and Fuller (1982). Consider a sequence U_N of embedded finite populations of size N where N grows to infinity. A sample s_N of size n_N is selected from U_N with sampling design $p_N(\cdot)$. The associated first- and second-order inclusion probabilities are $\pi_{k(N)}$ and $\pi_{k\ell(N)}$, respectively, for some generic units k and ℓ. A subsample s_{rN} is obtained from s_N with Poisson sampling design with unknown inclusion probabilities $p_{k(N)}$.

3.7.3 Conditions

We consider the following common regularity conditions on the sequence of sampling designs.

(A1) $\lim_{N \to +\infty} n_N / N = f \in (0, 1)$,

(A2) There exists $\pi_{min} \in \mathbb{R}$ such that $\pi_{k(N)} > \pi_{min} > 0$ for all $k \in U_N$ and all N,

(A3) There exists $\pi_{2,min} \in \mathbb{R}$ such that $\pi_{k\ell(N)} > \pi_{2,min} > 0$, for all $k, \ell \in U_N$ and all N,

(A4) $\limsup_{N \to +\infty} n_N \max_{\substack{k,\ell \in U_N, \\ k \neq \ell}} \left| \Delta_{k\ell(N)} \right| < +\infty$.

For a sampling design with random sample size, n_N in Assumption (A1) is replaced by the expected sample size. Assumption (A1) states that neither the population nor the sample grow faster than the other one. This assumption is for instance satisfied when the sequence of sampling fractions is constant. This assumption is not satisfied if the sample grows faster than the population, or inversely. Assumption (A2) states that the first order inclusion probabilities are bounded below. This assumption is satisfied for simple random sampling without replacement when Assumption (A1) holds. Assumption (A2) is also satisfied for stratified sampling unless the sampling fraction within some stratum converges to zero. Assumption (A3) states that the second order inclusion probabilities are bounded below. This assumption is satisfied for simple random

sampling without replacement when Assumption (A1) is satisfied and for stratified sampling unless the sampling fraction within some stratum converges to zero. Assumption (A4) states that the sampling designs are not overly dependent. Indeed, we can see

$$n_N \max_{k,\ell \in U_N, k \neq \ell} \left|\Delta_{k\ell(N)}\right|$$

as a measure of dependence between sample indicators. This measure should not grow to infinity. This assumption is satisfied for simple random sampling, Poisson sampling, and any stratified sampling design that is not too highly stratified. This assumption is not satisfied for cluster sampling. Nor is it satisfied for highly stratified sampling for which there is at least one stratum that grows at a slower pace than the overall sample.

We also consider the following condition on the sequence of finite populations.

(A5) The study variable has finite fourth moment and the auxiliary variables has finite first moments, i.e.

$$\limsup_{N \to +\infty} N^{-1} \sum_{k \in U_N} u_k < +\infty.$$

with ξ-probability one where $u_k = (y_k^4, \mathbf{x}_k^\top)^\top$.

Conditions (A1)-(A5) ensure consistency of the HT estimator and its variance estimator in (3.2.1).

We consider the following regularity conditions on the sequence of Poisson sampling designs that generate the sets of respondents.

(A6) $\lim_{N \to +\infty} \sum_{k \in U_N} \pi_{k(N)} p_{k(N)} / n_N = f_r \in (0,1)$,

(A7) There exist $p_{min}, p_{max} \in \mathbb{R}$ such that $0 < p_{min} < p_{k(N)} < p_{max} < 1$, for all $k \in U_N$ and all N,

(A8) The response probabilities are $p_k = 1/F(\mathbf{z}_k^\top \boldsymbol{\lambda}_0)$ for some true unknown parameter vector $\boldsymbol{\lambda}_0$, continuous function F with continuous derivative.

(A9) We have

$$\sum_{k \in U_N} \frac{a_k r_k}{\pi_{k(N)} p_{k(N)}} u_k - \sum_{k \in U_N} u_k = O_{p*}(N n^{-1/2})$$

with ξ-probability one for any characteristic u that is bounded or stochastically bounded.

Assumption (A6) states that the fraction of respondents to sampled units does not increase or decrease as N grows to infinity. Assumption (A7) states that each unit has a strictly positive probability of responding. Assumptions (A1)-(A7) together ensure consistency of the two-phase HT estimator and its variance estimator in (3.2.1) under sampling design $p^*(\cdot)$. Assumption (A8) implies that the data is missing at random (see Rubin, 1976, for a detailed definition) and that the response indicators are independent of one another and of the selected sample s. The values $\{r_k\}$ are obtained from a Poisson sampling design, i.e., the $\{r_k\}$ are generated from independent Bernoulli random variables with $\mathrm{E}_q(r_k \mid s) = \mathrm{E}_q(r_k) = p_k$, where $\mathrm{E}_q(.)$ is the expectation under the nonresponse mechanism. The logistic response model is obtained with $F(x) = 1 + \exp(-x)$. Assumption (A9) states that the respondents first moment estimator of any characteristic that is bounded or stochastically bounded converges to its population first moment.

We also put the following conditions on the estimated response probabilities, the working model, and their estimators.

(A10) $m_r(\mathbf{x}_k) - m_U(\mathbf{x}_k) = O_{p*}(n^{-1/2})$ with ξ-probability one,

(A11) $F(\mathbf{z}_k^\top \widehat{\boldsymbol{\lambda}}) - F(\mathbf{z}_k^\top \boldsymbol{\lambda}_0) = O_{p*}(n^{-1/2})$ with ξ-probability one,

(A12) $\lim_{N\to+\infty} N^{-1}\mathbf{W}^\top \text{diag}(1-p_1, 1-p_2, \ldots, 1-p_N)\mathbf{W} = \mathbf{G}$ where the determinant of $\mathbf{G}$ is strictly positive, $\mathbf{W}^\top = (w_1^\top, w_2^\top, \ldots, w_N^\top)$, and $w_k = (y_k, m_U(\mathbf{x}_k), \mathbf{x}_k^\top)$. If the vector $[m_U(\mathbf{x}_k)]_{k\in U}$ is a linear combination of the columns of $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^\top$, then we set $w_k = (y_k, \mathbf{x}_k^\top)$.

(A13) $m_U(\mathbf{x}_k) - m(\mathbf{x}_k) = O_{\mathbb{P}}(n^{-1/2})$ where subscript $\mathbb{P}$ means that the reference probability is the probability induced jointly by the working model ξ and $p^*(\cdot)$.

(A14) We have

$$\sum_{k\in U_N} \frac{a_k r_k}{\pi_k \widehat{p}_k} \{y_k - m_r(\mathbf{x}_k)\} - \sum_{k\in U_N} \frac{a_k}{\pi_k} \{y_k - m_r(\mathbf{x}_k)\} = O_{p*}(Nn^{-1/2})$$

with ξ-probability one.

In the next section, we show that our proposed estimator is doubly robust. Two scenarios are considered: 1) when the response model is correctly specified and 2) when the working model is correctly specified. Assumptions (A11) and (A12) are needed for the first scenario. Assumptions (A13) and (A14) are needed for the second scenario. Assumption (A10) states that the respondents estimator $m_r(\mathbf{x}_k)$ converges to the population estimator $m_U(\mathbf{x}_k)$. Note that this assumption may be satisfied even if the working model is misspecified. Its validity relies on the selected working model, on the sampling design, and on the response mechanism. This has to be evaluated on a case to case basis. Assumption (A11) states that the inverse of the response probabilities are consistent estimators of the inverse of the true response probabilities. This assumption states that the response model is correctly specified and its validity relies on the chosen function F and chosen estimation technique. This assumption is satisfied for calibration at the population model with the correct function F under (A9) and mild conditions, see Hasler (2023). Assumption (A12) is necessary to ensure that an estimator asymptotically equivalent to ours is well defined, see the lemma below. Assumption (A13) implies that the population estimator $m_U(\mathbf{x}_k)$ is consistent for $m(\mathbf{x}_k)$, i.e. that the working model is correctly specified. The validity of this assumption relies on the selected working model. Assumption (A14) states that the difference between the NWA estimator of and the HT estimator of the residuals $y_k - m_r(\mathbf{x}_k)$ converges to a bound. Two cases in which this assumption holds are: 1) when the $\mathbf{x}_k$ are contained in a compact set and the response probabilities are estimated via calibration at the sample level and 2) when the $\mathbf{x}_k$ are contained in a compact set, the response probabilities are estimated via calibration at the population level, and Assumption (A9) holds. To show that these two cases satisfy Assumption (A14), we write a Taylor expansion of $y_k - m_r(\mathbf{x}_k)$ around a $\mathbf{x}_0$ in the compact set containing all $\mathbf{x}_k$ and use the calibration equations. In what follows, we will omit the subscript N whenever possible to simplify notations.

3.7.4 Asymptotics and Double Robustness

Result 3.7.1 (Response Model Correctly Specified)**.** *Suppose that the sequence of sampling designs, populations, and response mechanisms satisfy Conditions (A1)-(A12). Then*

$$\widehat{t}_{m_r,\widehat{p}} = \widehat{t}_{m_U,p} + O_{p^*}(Nn^{-1/2}), \tag{3.7.7}$$

with ξ-probability one where

$$\widehat{t}_{m_U,p} = \sum_{k\in U} m_U(\mathbf{x}_k) + \sum_{k\in s_r} \frac{y_k - m_U(\mathbf{x}_k)}{\pi_k p_k},$$

is an unbiased estimator of t_y which is at least as efficient as the two-phase HT estimator under sampling design $p^*(\cdot)$.

The proof of Result 3.7.1 is in Appendix 3.13. In the context of model-assisted estimation, $\widehat{t}_{m_U,p}$ is a pseudo-generalized difference estimator under two-phase sampling. It is trivial to show that it is 1) design unbiased and 2) more efficient that the two-phase HT estimator provided that the "residuals" $\{y_k - m_U(\mathbf{x}_k)\}$ have less variability than the "raw values" $\{y_k\}$ (Breidt and Opsomer, 2017, p.192). Hence, the quasi-model-assisted estimator $\widehat{t}_{m_r,\widehat{p}}$ is asymptotically unbiased and asymptotically at least as efficient as the two-phase HT estimator.

Result 3.7.2 (Working Model Correctly Specified)**.** *Suppose that the sequence of sampling designs, populations, and response mechanisms satisfy (A1)-(A10) and (A13)-(A14). Then*

$$\widehat{t}_{m_r,\widehat{p}} = \widehat{t}_m + O_{\mathbb{P}}(Nn^{-1/2}).$$

The proof of Result 3.7.2 is in Appendix 3.13. In the context of model-assisted estimation, $\widehat{t}_m$ is the difference estimator. It is 1) design unbiased 2) more efficient that the two-phase HT estimator provided that the "residuals" $\{y_k - m_U(\mathbf{x}_k)\}$ have less variability than the "raw values" $\{y_k\}$ (Breidt and Opsomer, 2017). Hence, the quasi-model-assisted estimator $\widehat{t}_{m_r,\widehat{p}}$ is asymptotically unbiased and asymptotically at least as efficient as the two-phase HT estimator.

3.8 Variance and Variance Estimation

In this section, we suppose that the same set of auxiliary variables is considered as response model variables, calibration variables, and variables in the working model. That is, $\mathbf{z}_k = \mathbf{x}_k$. The response probabilities are estimated via calibration on the $\mathbf{x}_k$'s. Under nonresponse, we can write the variance of a generic estimator $\widehat{t}_g$ as

$$\operatorname{var}\left(\widehat{t}_g\right) = \operatorname{var}_{sam}\left(\widehat{t}_g\right) + \operatorname{var}_{nr}\left(\widehat{t}_g\right),$$

where the two terms are the sampling variance and the nonresponse variance, respectively, and are given by

$$\operatorname{var}_{sam}\left(\widehat{t}_g\right) = \operatorname{var}_p\left\{\mathrm{E}_q\left(\widehat{t}_g \middle| s\right)\right\},$$

and

$$\operatorname{var}_{nr}\left(\widehat{t}_g\right) = \mathrm{E}_p\left\{\operatorname{var}_q\left(\widehat{t}_g \middle| s\right)\right\}.$$

Using the approximations in Equation (3.7.7) and (3.13.8) and results of Section 7 of Hasler (2023), the variance of the quasi-model-assisted estimator $\widehat{t}_{m_r,\widehat{p}}$ can be approximated by

$$\operatorname{var}\left(\widehat{t}_{m_r,\widehat{p}}\right) \approx \operatorname{var}_{sam}\left(\widehat{t}_{m_U,\widehat{p},\ell}\right) + \operatorname{var}_{nr}\left(\widehat{t}_{m_U,\widehat{p},\ell}\right),$$

where

$$\operatorname{var}_{sam}\left(\widehat{t}_{m_U,\widehat{p},\ell}\right) = \operatorname{var}_p\left[\sum_{k\in s}\frac{1}{\pi_k}\left\{y_k - m_U(\mathbf{x}_k) - \mathbf{x}_k^\top\boldsymbol{\gamma}\right\}\right],$$

$$\operatorname{var}_{nr}\left(\widehat{t}_{m_U,\widehat{p},\ell}\right) = \mathrm{E}_p\left[\sum_{k\in s}\frac{1}{\pi_k^2}\frac{1-p_k}{p_k}\left\{y_k - m_U(\mathbf{x}_k) - \mathbf{x}_k^\top\boldsymbol{\gamma}\right\}^2\right],$$

and

$$\boldsymbol{\gamma} = \left\{\sum_{k\in U}(1-p_k)\mathbf{x}_k\mathbf{x}_k^\top\right\}^{-1}\sum_{k\in U}(1-p_k)\mathbf{x}_k\left\{y_k - m_U(\mathbf{x}_k)\right\}.$$

The first term is the variance of the full sample HT estimator of the differences $\{y_k - m_U(\mathbf{x}_k) - \mathbf{x}_k^\top \boldsymbol{\gamma}\}$. Based on this approximation, a variance estimator is

$$\widehat{\text{var}}\left(\widehat{t}_{m_r,\widehat{p}}\right) = \widehat{\text{var}}_{sam}\left(\widehat{t}_{m_U,\widehat{p},\ell}\right) + \widehat{\text{var}}_{nr}\left(\widehat{t}_{m_U,\widehat{p},\ell}\right),$$

where

$$\widehat{\text{var}}_{sam}\left(\widehat{t}_{m_U,\widehat{p},\ell}\right) = \sum_{k\in s_r} \frac{1-\pi_k}{\pi_k^2}\frac{e_k^2}{\widehat{p}_k} + \sum_{k,\ell\in s_r; k\neq\ell} \frac{\pi_{k\ell}-\pi_k\pi_\ell}{\pi_{k\ell}\pi_k\pi_\ell}\frac{e_k}{\widehat{p}_k}\frac{e_\ell}{\widehat{p}_\ell},$$

$$\widehat{\text{var}}_{nr}\left(\widehat{t}_{m_U,\widehat{p},\ell}\right) = \sum_{k\in s_r} \frac{1}{\pi_k^2}\frac{1-\widehat{p}_k}{\widehat{p}_k^2}e_k^2,$$

$$e_k = y_k - m_r(\mathbf{x}_k) - \mathbf{x}_k^\top \widehat{\boldsymbol{\gamma}},$$

and

$$\widehat{\boldsymbol{\gamma}} = \left(\sum_{k\in s_r} \frac{1}{\pi_k}\frac{1-\widehat{p}_k}{\widehat{p}_k}\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1} \sum_{k\in s_r} \frac{1}{\pi_k}\frac{1-\widehat{p}_k}{\widehat{p}_k}\mathbf{x}_k\left\{y_k - m_r(\mathbf{x}_k)\right\},$$

where we substituted $\widehat{p}_k$ for the unknown p_k and $m_r(\mathbf{x}_k)$ for $m_U(\mathbf{x}_k)$. Note that double robustness of this variance estimator is not guaranteed. It relies on the response model being correctly specified.

3.9 Simulations

3.9.1 Simulated data

Let us consider a population U of size $N = 1000$. For each unit k of U, a vector $\mathbf{x}_k = (x_{k1}, x_{k2})^\top$ is generated from independent and identically distributed random uniform variables with parameters -5 and 5. The goal is to estimate the total t on population U of a survey variable y generated as

$$y_k = 100 + 10.4 \cdot x_{k1} + 9.7 \cdot x_{k2} + \varepsilon_k,$$

where ε_k is the realisation of a normal distribution of mean 0 and variance 1. A unit k of the population has a probability

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

of responding to variable y, with $\boldsymbol{\lambda} = (-0.002, 0.212, 0.231)^\top$. Value $\boldsymbol{\lambda}$ is set so that the expected rate of missing values, i.e. the mean of the $\{p_k\}$ on the population, is 50%.

A comparison between some aforementioned total estimators is performed in different scenarios: when the nonresponse model is correctly versus incorrectly specified, when the working model is correctly versus incorrectly specified. We consider a vector $\widetilde{\mathbf{x}}_k = (\widetilde{x}_{k1}, \widetilde{x}_{k2})^\top$ defined such that

$$\widetilde{x}_{k1} = 10 \cdot \cos(x_{k2}) / \{1 + \exp(x_{k1})\}^{-1} + 10$$

and

$$\widetilde{x}_{k2} = (x_{k2} - x_{k1} + 10) \cdot |x_{k1}|.$$

The correlations between the variable of interest and the other variables are given in Table 1. Vector $\{\mathbf{x}_k\}$ is strongly related to $\{y_k\}$ and $\{p_k\}$. Vector $\{\widetilde{\mathbf{x}}_k\}$ is weakly related to $\{y_k\}$ and $\{p_k\}$. When estimating the models, we use the vector $\widetilde{\mathbf{x}}_k$ instead of $\mathbf{x}_k$ in order to misspecify the models.

Four different scenarios are considered in which different couples of variables are used to fit the response model and the working model. In scenarios 1 and 2, the working model is correctly

TABLE 1: Correlation of the values $\{y_k\}$ of the survey variable with the response probabilities $\{p_k\}$ and $\{x_{k1}\}, \{x_{k2}\}, \{\widetilde{x}_{k1}\}$ and $\{\widetilde{x}_{k2}\}$ of the auxiliary variables on population U.

	$\{x_{k1}\}$	$\{x_{k2}\}$	$\{\widetilde{x}_{k1}\}$	$\{\widetilde{x}_{k2}\}$
$\{y_k\}$	0.712	0.678	0.133	-0.113
$\{p_k\}$	0.670	0.752	-0.402	-0.054

specified, whereas in scenarios 3 and 4 it is incorrectly specified. In scenarios 1 and 3, the response model is correctly specified, whereas in scenarios 2 and 4 it is incorrectly specified. Note that in scenario 1 both models are correctly specified, in scenario 2 and 3 only one of the two models is correctly specified, and in scenario 4 both models are incorrectly specified. Table 2 shows which vector of variables, i.e. $\mathbf{x}_k$ or $\widetilde{\mathbf{x}}_k$, is used to fit the models.

TABLE 2: Couple of variables used to obtain the estimated response probabilities $\widehat{p}_k$ and the estimated function $m_r(.)$ depending on the scenario.

		$\widehat{p}_k$	
		$\{x_{k1}, x_{k2}\}$	$\{\widetilde{x}_{k1}, \widetilde{x}_{k2}\}$
$m_r(.)$	$\{x_{k1}, x_{k2}\}$	Scenario 1	Scenario 2
	$\{\widetilde{x}_{k1}, \widetilde{x}_{k2}\}$	Scenario 3	Scenario 4

We compare five estimators: $\widehat{t}_{HT}$, $\widehat{t}_{NWA}$, and $\widehat{t}_{m_r,\widehat{p}}$, defined in Sections 3.2, 3.4, and 3.5 respectively, the imputed estimator $\widehat{t}_{imp} = \sum_{k\in s_r} y_k/\pi_k + \sum_{k\in s\setminus s_r} m_r(\mathbf{x}_k)/\pi_k$, and the naive estimator $\widehat{t}_{naive} = N n_r^{-1} \cdot \sum_{k\in s_r} y_k$. Estimator $\widehat{t}_{HT}$ is unavailable in practice with nonresponse and serves here as comparison point. Estimators $\widehat{t}_{imp}$ and $\widehat{t}_{m_r,\widehat{p}}$ depend on the estimated function $m_r(.)$. Three different prediction methods are used to obtain $m_r(.)$: generalized regression, local polynomial regression, and K-nearest neighbors with $K = 5$. Estimators $\widehat{t}_{NWA}$ and $\widehat{t}_{m_r,\widehat{p}}$ depend on the estimated response probabilities. The response probabilities are estimated using a logistic model.

We select $I = 10'000$ samples denoted by $s^{(i)}$, $i = 1, \ldots, I$, of size $n = 200$ from population U using simple random sampling without replacement. For each sample, we randomly generate missing values in the survey variable using the response probabilities $\{p_k\}$ and a Poisson sampling design. The expected number of observed values n_r in each sample $s^{(i)}$ is $n/2 = 100$. We can then define the sub-sample $s_r^{(i)} \subset s^{(i)}$ containing the units for which y_k is observed at simulation run i.

In order to evaluate the quality of the nonresponse model and of the working model at simulation run i, $i \in \{1, \ldots, I\}$, two quantities are computed: the mean absolute error of the estimated response probabilities $\{\widehat{p}_k\}$

$$\text{MAE}(\widehat{p}_k) = \frac{1}{n_r} \sum_{k\in s_r^{(i)}} | \widehat{p}_k - p_k |,$$

and the mean relative prediction error

$$\text{MRPE}(m_r(\mathbf{u}_k)) = \frac{1}{N} \frac{\sum_{k\in U} | m_r(\mathbf{u}_k) - y_k |}{\sum_{k\in U} y_k},$$

where $\mathbf{u}_k = \mathbf{x}_k$ or $\mathbf{u}_k = \widetilde{\mathbf{x}}_k$, depending on the scenario. The goodness of fit of the working and response models is assessed by averaging, for each scenario, the MAE and MRPE over the simulation runs. Table 3 contains these averages.

TABLE 3: MAE of the estimated response probabilities $\widehat{p}_k$ and MRPE of $m_r(\mathbf{u}_k)$ of the prediction of the missing values four scenarios for the simulated data.

	Scenario			
	1	2	3	4
MAE($\widehat{p}_k$)				
	0.059	0.175	0.059	0.175
MRPE				
GREG	0.080	0.080	0.366	0.366
poly	0.082	0.082	0.275	0.275
K-nn	0.093	0.093	0.337	0.337

For each samples $s^{(i)}$ and $s_r^{(i)}$, we estimate the population total with the five aforementioned total estimators. For a generic total estimator $\widehat{t}$, we compute the Monte Carlo bias relative to the true total

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

and the Monte Carlo standard deviation relative to the true total

$$\text{RSd}(\widehat{t}) = \frac{\sqrt{(I-1)^{-1}\sum_{i=1}^{I}(\widehat{t}^{(i)} - t)^2}}{t},$$

where $\widehat{t}^{(i)}$ is the value of $\widehat{t}$ obtained at simulation run $i \in \{1, \ldots, I\}$. We compare the total estimators for each scenario. Figure 1 summarizes the results. Detailed results for scenarios 1 to 4 are given in Appendix 3.14 in Tables 7a, 7b, 7c and 7d, respectively.

In scenario 1, both the response and working models fit well the data. Our proposed quasi-model-assisted estimator $\widehat{t}_{m_r,\widehat{p}}$ and $\widehat{t}_{NWA}$ perform the best in this scenario, with a RB close to that of the unbiased estimators $\widehat{t}_{HT}$ and have the lowest RSd. In scenario 2, our proposed estimator $\widehat{t}_{m_r,\widehat{p}}$ shows better results compared to all available estimators, even if the response model fits poorly. It has a RB and a RSd smaller to 1% and 2%, respectively. It shows the best results in term of standard deviation and is more efficient than the NWA and HT estimator. It confirms that the working model allows to improve the efficiency of the total estimator. In scenario 3, the working model is misspecified. The estimator $\widehat{t}_{NWA}$ provides the best results followed by our proposed quasi-model-assisted estimator $\widehat{t}_{m_r,\widehat{p}}$. The reason is that the response model is correctly specified in this scenario. Finally, in scenario 4, both the response model and the working model fit poorly the data. In this case, the performance of $\widehat{t}_{m_r,\widehat{p}}$ is comparable to that of $\widehat{t}_{NWA}$ and $\widehat{t}_{imp}$ that rely on only one of the two models.

The general conclusion of the simulation study is that the proposed estimator $\widehat{t}_{m_r,\widehat{p}}$ globally performs as well or better than estimators $\widehat{t}_{NWA}$, $\widehat{t}_{imp}$, and $\widehat{t}_{naive}$ even when one or both of the working model and the nonresponse model is or are misspecified. Our estimators hence provides protection against model misspecification and greater confidence in the total estimator.

3.9.2 Real data

We carry out a second simulation study on real data. We consider a dataset from the book Särndal (1992), available as dataset MU284 in the R package `sampling`. The population of interest U is composed of $N = 284$ Sweden municipalities. For each municipality $k \in U$, we consider two auxiliary variables: x_{k1}, the population in thousands in 1975, and $\widetilde{x}_{k1}$, the number of Social

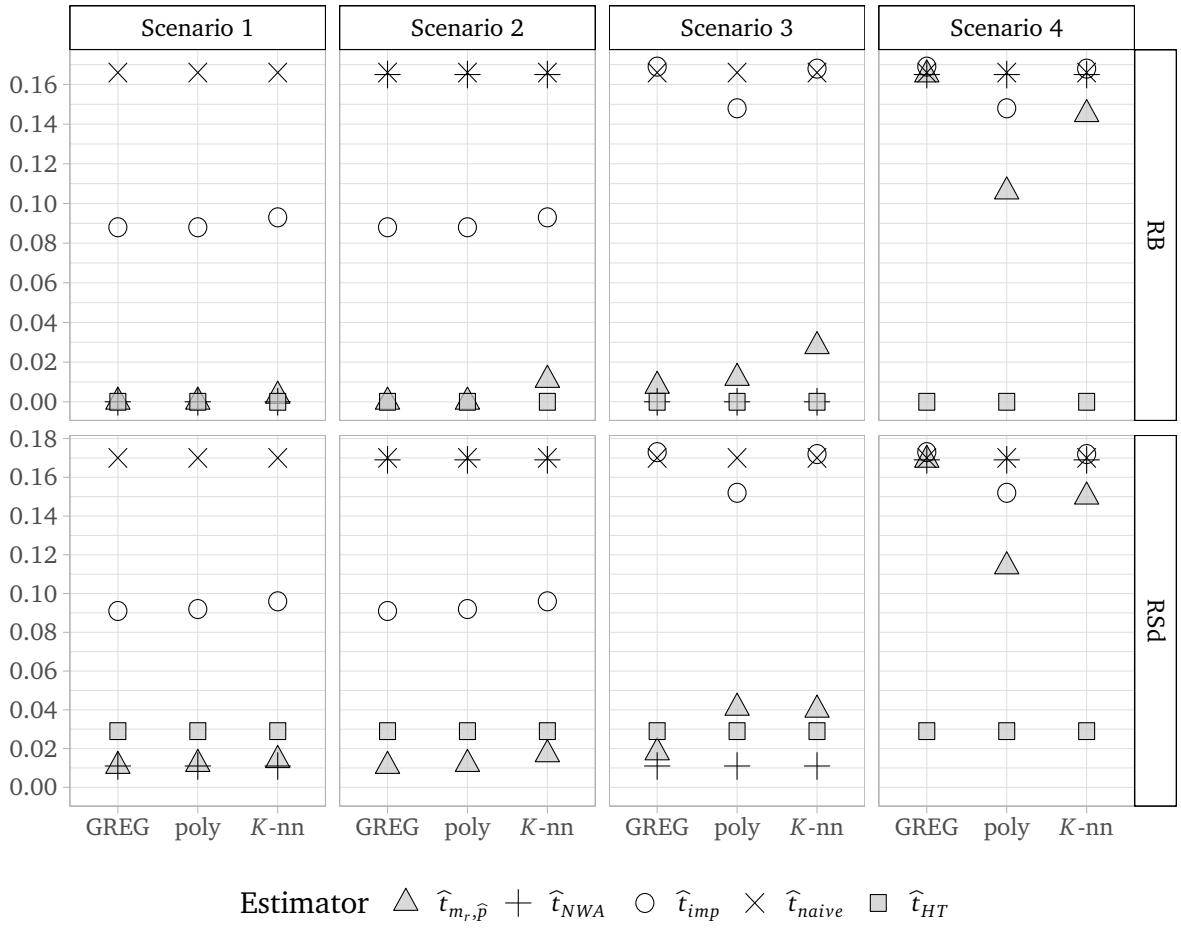

FIGURE 1: Bias and standard deviation of total estimators relative to the true total for the simulated data for scenarios 1 to 4 with three prediction methods: generalized regression (GREG), polynomial regression (poly) and K-nearest neighbors (K-nn) with $K = 5$. Estimator $\widehat{t}_{HT}$ is a comparison point and is unavailable with nonresponse.

Democrat seats on the municipal council. We denote by y the tax revenue variable of a municipality in 1985, in millions of crowns. The aim is to estimate the total of y on population U, i.e., the total municipal tax revenue in Sweden in 1985. We assume that y is subject to nonresponse. The response probability to variable y of a unit $k \in U$ is

$$p_k = \left\{1 + \exp\left[-\boldsymbol{\lambda}^\top (1, x_{k1})^\top\right]\right\}^{-1},$$

with $\boldsymbol{\lambda} = (-0.30, 0.01)^\top$. The expected rate of missing values is 50%.

The correlation between the variable of interest, the response probabilities, and the auxiliary variables is in Table 4. Variable $\{x_{k1}\}$ is strongly related to $\{y_k\}$ and $\{p_k\}$. Variable $\{\widetilde{x}_{k1}\}$ is weakly related to $\{y_k\}$ and $\{p_k\}$. We compare the total estimators in the same four scenarios as in the case of simulated data. That is, we consider variable x_{k1} instead of $\widetilde{x}_{k1}$ when estimating the working model, respectively, response model, in order to obtain a model that does not fit the data well. Table 5 shows which variable, i.e., x_{k1} or $\widetilde{x}_{k1}$, is used to fit the models. The goodness of fit of the working and response models are in Table 6.

We assume a census, i.e. $s = U$ and $\pi_k = 1$ for all $k \in U$. In the simulated data case (see Section 3.9.1), we have selected $I = 10'000$ samples $s^{(i)}$, $i \in \{0, \ldots, I\}$. Following the same steps,

TABLE 4: Correlations of the values $\{y_k\}$ of the survey variable and $\{p_k\}$ of the response probabilities with the auxiliary variables $\{x_{k1}\}$ and $\{\widetilde{x}_{k1}\}$, on population U.

	$\{x_{k1}\}$	$\{\widetilde{x}_{k1}\}$
$\{y_k\}$	0.967	0.401
$\{p_k\}$	0.894	0.637

TABLE 5: Variable used to obtain the estimated response probabilities $\widehat{p}_k$ and the estimated function $m_r(.)$ for four scenarios.

		$\widehat{p}_k$	
		$\{x_{k1}\}$	$\{\widetilde{x}_{k1}\}$
$m_r(.)$	$\{x_{k1}\}$	Scenario 1	Scenario 2
	$\{\widetilde{x}_{k1}\}$	Scenario 3	Scenario 4

TABLE 6: MAE of the estimated response probabilities $\widehat{p}_k$ and MRPE of $m_r(\mathbf{u}_k)$ of the prediction of the missing values four scenarios for the real data.

	Scenario			
	1	2	3	4
MAE($\widehat{p}_k$)	0.030	0.053	0.030	0.053
MRPE				
GREG	0.241	0.241	1.049	1.049
poly	0.108	0.108	0.748	0.748
knn	0.198	0.198	0.809	0.809

the census situation involves that these 10′000 samples are all equal, $s^{(i)} = s$. Similarly as for simulated data, we randomly generate nonresponses in values $\{y_k\}_{k\in U}$ $I = 10'000$ times using the response probabilities $\{p_k\}$ and a poisson sampling design. We obtain the subset $s_r^{(i)} \subset s$ containing the respondents to y_k at simulation run $i \in \{1, \ldots, I\}$.

We compare the total estimators $\widehat{t}_{m_r,\widehat{p}}$, $\widehat{t}_{NWA}$, $\widehat{t}_{imp}$ and $\widehat{t}_{naive}$. We do not consider the HT estimator because it is equivalent to the true total due to the census. We consider the same prediction methods for the working and response models as in Section 3.9.1. For a total estimator $\widehat{t}$, we compute the Monte Carlo bias and standard deviation. The results of the compared total estimators for each scenario are summarized in Figure 2 and detailed in Tables 8a, 8b, 8c, and 8d in the Appendix 3.14.

When the response and working models fit the data well (scenario 1), all estimators except for $\widehat{t}_{naive}$ give good results, with Monte Carlo relative biases and standard deviations below 1%. In scenario 2, the nonresponse model is misspecified. The NWA estimator is much less efficient, with relative bias and standard deviation close to 20%. In this scenario, the relative bias and standard deviations of the proposed estimator $\widehat{t}_{m_r,\widehat{p}}$ remain lower than 8%. In scenario 3, the imputed estimator does not perform well. Its relative standard deviation and bias increase by almost 20% compared to scenarios 1 and 2. This can easily be explained by the fact that the working model fits poorly the data and imputations are therefore far from the true values. Once again, $\widehat{t}_{m_r,\widehat{p}}$ has a relative bias of less than 2% and a relative standard deviation of less than 8%. In scenario 4, both the response model and the working model fit poorly the data. In this case, the performance of $\widehat{t}_{m_r,\widehat{p}}$ is comparable to that of $\widehat{t}_{NWA}$ and $\widehat{t}_{imp}$ which relies on only one of these two models.

The general conclusion of the simulation study on real data is that the proposed estimator $\widehat{t}_{m_r,\widehat{p}}$ globally performs as well or better than estimators $\widehat{t}_{NWA}$, $\widehat{t}_{imp}$, and $\widehat{t}_{naive}$, even when one or both of the working model and response model is or are misspecified. Our estimator hence provides protection against model misspecification and greater confidence in the total estimator.

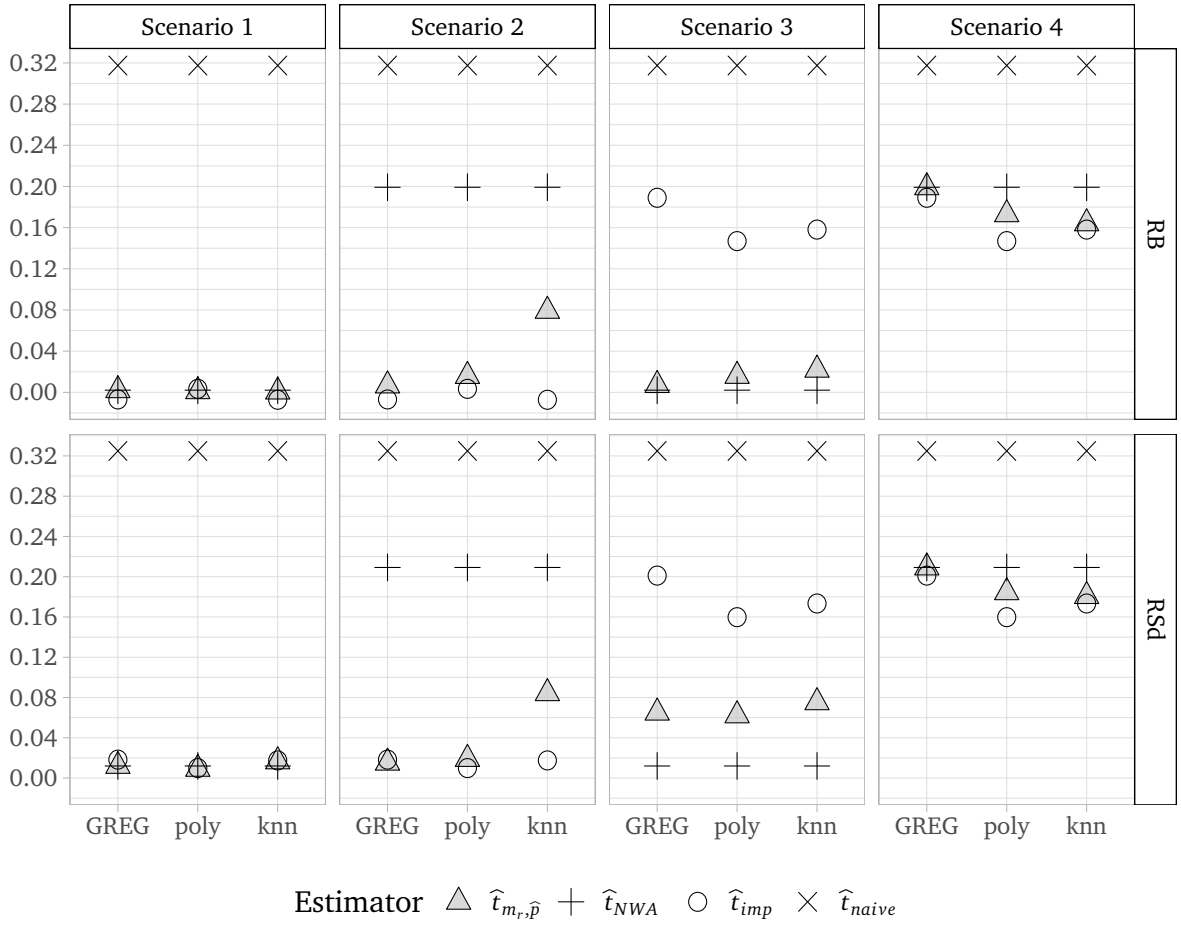

FIGURE 2: Bias and standard deviation of total estimators relative to the true total for the real data and for the scenarios 1 to 4 with three prediction methods: generalized regression (GREG), polynomial regression (poly) and K-nearest neighbors (K-nn) with $K = 5$.

3.10 Discussion

We adapt model-assisted total estimators to missing at random data building on the idea of nonresponse weighting adjustment. We consider nonresponse as a second phase of the survey and reweight the units using the inverse of estimated response probabilities in model-assisted estimators in order to compensate for the nonrespondents. We develop the asymptotic properties of our proposed estimator and show conditions under which it is asymptotically unbiased. Our proposed estimator can be written as a weighted estimator. We show cases in which the resulting weights are calibrated to the total of the auxiliary variables. We conduct a simulation study to empirically study the performance of our estimator. The results of this study confirm that our estimator generally outperforms the competing estimators, even when the underlying models are

misspecified. Further work includes the study of our estimator under other working models as well as the extension to non-missing at random data.

Acknowledgements

This work was partially funded by the Swiss Federal Statistical Office. The views expressed in this paper are those of the authors solely.

Appendices

3.11 Auxiliary Variables Known at the Sample Level Only

The first term of the quasi-model-assisted estimator in (3.5.5) is the population total of the predicted values $m_r(\mathbf{x}_k)$. For most working models, this requires the values $\{\mathbf{x}_k\}$ to be known for all population units. If this population total is unavailable, we may use a Horvitz-Thompson-type estimator of this sum which yields estimator

$$\widehat{t}_{m_r,s,\widehat{p}} = \sum_{k\in s} \frac{m_r(\mathbf{x}_k)}{\pi_k} + \sum_{k\in s_r} \frac{y_k - m_r(\mathbf{x}_k)}{\pi_k \widehat{p}_k}.$$

This estimator is equivalent to that of Kim and Haziza (2014), Equation (3.2). The authors suppose parametric models for the nonresponse model and working model. They estimate the parameter vectors of these two models simultaneously based on a system of estimating equations. This results in doubly robust point and variance estimators. Our approach is different and more general. The working model may be nonparametric and both models may be estimated separately.

3.12 Response Probabilities Estimated via Maximum Likelihood Estimation

A third approach consists of estimating the response probabilities via maximum likelihood. The estimation of parameter vector $\boldsymbol{\lambda}_0$ is then $\widehat{\boldsymbol{\lambda}}$ which is the solution to the estimating equation

$$Q^{mle}(\widehat{\boldsymbol{\lambda}}) = \sum_{k\in s} c_k \left\{ r_k - F^{-1}(\mathbf{x}_k^\top \boldsymbol{\lambda}_0) \right\} \mathbf{x}_k = 0,$$

for some weights c_k, see Kim and Kim (2007). Common choices for the weights are 1 or π_k^{-1}. When $c_k = 1$ usual maximum likelihood estimation is applied. Double robustness of the quasi-model assisted estimator when maximum likelihood is applied may be obtained using arguments similar to those presented in Section 3.7. This goes beyond the scope of this research.

3.13 Proofs of the Results

For the proof of Result 3.7.1, we will need the following Lemma.

Lemma 1. *Suppose that Assumptions (A1)-(A12) are satisfied. Then*

$$\widehat{t}_{m_U,\widehat{p}} = \widehat{t}_{m_U,p} + O_{p^*}(Nn^{-1/2}),$$

with ξ-probability one where

$$\widehat{t}_{m_U,\widehat{p}} = \sum_{k\in U} m_U(\mathbf{x}_k) + \sum_{k\in s_r} \frac{y_k - m_U(\mathbf{x}_k)}{\pi_k \widehat{p}_k}.$$

Proof of Lemma 1. From Hasler (2023), we have

$$\widehat{t}_{m_U,\widehat{p}} = \sum_{k\in U} m_U(\mathbf{x}_k) + \sum_{k\in U}\left[\mathbf{x}_k^\top \boldsymbol{\gamma} + \frac{a_k r_k}{\pi_k p_k}\left\{y_k - m_U(\mathbf{x}_k) - \mathbf{x}_k^\top \boldsymbol{\gamma}\right\}\right] + O_{p^*}(Nn^{-1}),$$

where $\boldsymbol{\gamma} = \left(\sum_{k\in U}(1-p_k)\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{k\in U}(1-p_k)\mathbf{x}_k\{y_k - m_U(\mathbf{x}_k)\}$. Rearranging, we obtain

$$\begin{aligned}\widehat{t}_{m_U,\widehat{p}} =& \sum_{k\in U} m_U(\mathbf{x}_k) + \sum_{k\in U}\frac{a_k r_k}{\pi_k p_k}\{y_k - m_U(\mathbf{x}_k)\} + \sum_{k\in U}\mathbf{x}_k^\top\boldsymbol{\gamma} \\ & - \sum_{k\in U}\frac{a_k r_k}{\pi_k p_k}\mathbf{x}_k^\top\boldsymbol{\gamma} + O_{p^*}(Nn^{-1}) \\ =& \widehat{t}_{m_U,p} + \left(\sum_{k\in U}\mathbf{x}_k^\top - \sum_{k\in U}\frac{a_k r_k}{\pi_k p_k}\mathbf{x}_k^\top\right)\boldsymbol{\gamma} + O_{p^*}(Nn^{-1}).\end{aligned}$$

From Conditions (A7) and (A12), and since $\mathbf{X}$ is of full rank, we have $\boldsymbol{\gamma} = O(1)$. From Conditions (A5) and (A9), we have

$$\sum_{k\in U}\mathbf{x}_k^\top - \sum_{k\in U}\frac{a_k r_k}{\pi_k p_k}\mathbf{x}_k^\top = O_{p^*}(Nn^{-1/2}).$$

We obtain

$$\widehat{t}_{m_U,\widehat{p}} = \widehat{t}_{m_U,p} + O_{p^*}(Nn^{-1/2}).$$

■

Proof of Result 3.7.1. Suppose that the sequence of sampling designs, populations, and response mechanisms satisfy Conditions (A1)-(A12). We have

$$\begin{aligned}\widehat{t}_{m_r,\widehat{p}} - \widehat{t}_{m_U,\widehat{p}} &= \sum_{k\in U}\{m_r(\mathbf{x}_k) - m_U(\mathbf{x}_k)\}\left\{1 - \frac{a_k r_k}{\pi_k}F(\mathbf{x}_k^\top\widehat{\boldsymbol{\lambda}})\right\} \qquad (3.13.8) \\ &= \sum_{k\in U}\{m_r(\mathbf{x}_k) - m_U(\mathbf{x}_k)\}\left\{1 - \frac{a_k r_k}{\pi_k}F(\mathbf{x}_k^\top\boldsymbol{\lambda}_0)\right\} \\ &\quad + \sum_{k\in U}\{m_r(\mathbf{x}_k) - m_U(\mathbf{x}_k)\}\left\{\frac{a_k r_k}{\pi_k}F(\mathbf{x}_k^\top\boldsymbol{\lambda}_0) - \frac{a_k r_k}{\pi_k}F(\mathbf{x}_k^\top\widehat{\boldsymbol{\lambda}})\right\}\end{aligned}$$

The first term is $O_{p^*}(Nn^{-1/2})$ by (A9) and (A10). The second term is $O_{p^*}(Nn^{-1/2})$ by (A7), (A10), and (A11). As a result $\widehat{t}_{m_r,\widehat{p}} = \widehat{t}_{m_U,\widehat{p}} + O_{p^*}(Nn^{-1/2})$. From the Lemma above, we obtain that $\widehat{t}_{m_r,\widehat{p}} = \widehat{t}_{m_U,p} + O_{p^*}(Nn^{-1/2})$. ■

Proof of Result 3.7.2 . We have

$$\widehat{t}_{m_r,\widehat{p}} - \widehat{t}_m = \sum_{k\in U}\{m_r(\mathbf{x}_k) - m(\mathbf{x}_k)\} + \sum_{k\in U}\frac{a_k}{\pi_k}\left(\frac{r_k}{\widehat{p}_k} - 1\right)\{y_k - m_r(\mathbf{x}_k)\} \tag{3.13.9}$$
$$+ \sum_{k\in U}\frac{a_k}{\pi_k}\{m_r(\mathbf{x}_k) - m(\mathbf{x}_k)\}$$

From (A10) and (A13), $m_r(\mathbf{x}_k) - m(\mathbf{x}_k)$ is $O_{\mathbb{P}}(n^{-1/2})$. The first term in (3.13.9) is therefore $O_{\mathbb{P}}(Nn^{-1/2})$. From (A14), the second term is $O_{\mathbb{P}}(Nn^{-1/2})$. From (A2) and since $m_r(\mathbf{x}_k) - m(\mathbf{x}_k)$ is $O_{\mathbb{P}}(n^{-1/2})$, the last term is also $O_{\mathbb{P}}(Nn^{-1/2})$. ■

3.14 Results of the simulations

The results of the simulations for the simulated data are presented in Tables 7.

TABLE 7: Bias and standard deviation of total estimators relative to the true total for scenarios 1 to 4 with three prediction methods: generalized regression (GREG), polynomial regression (poly) and K-nearest neighbors (K-nn) with $K = 5$. Estimator $\widehat{t}_{HT}$ is a comparison point and is unavailable with nonresponse.

	Estimators				
	$\widehat{t}_{m_r,\widehat{p}}$	$\widehat{t}_{NWA}$	$\widehat{t}_{imp}$	$\widehat{t}_{naive}$	$\widehat{t}_{HT}$
RB					
GREG	<0.001	<0.001	0.088	0.166	<0.001
poly	<0.001	<0.001	0.088	0.166	<0.001
K-nn	0.003	<0.001	0.093	0.166	<0.001
RSd					
GREG	0.011	0.011	0.091	0.170	0.029
poly	0.012	0.011	0.092	0.170	0.029
K-nn	0.014	0.011	0.096	0.170	0.029

(A) Scenario 1

	Estimators				
	$\widehat{t}_{m_r,\widehat{p}}$	$\widehat{t}_{NWA}$	$\widehat{t}_{imp}$	$\widehat{t}_{naive}$	$\widehat{t}_{HT}$
RB					
GREG	<0.001	0.165	0.088	0.166	<0.001
poly	<0.001	0.165	0.088	0.166	<0.001
K-nn	0.011	0.165	0.093	0.166	<0.001
RSd					
GREG	0.011	0.169	0.091	0.170	0.029
poly	0.012	0.169	0.092	0.170	0.029
K-nn	0.017	0.169	0.096	0.170	0.029

(B) Scenario 2

	Estimators				
	$\widehat{t}_{m_r,\widehat{p}}$	$\widehat{t}_{NWA}$	$\widehat{t}_{imp}$	$\widehat{t}_{naive}$	$\widehat{t}_{HT}$
RB					
GREG	0.008	<0.001	0.169	0.166	<0.001
poly	0.012	<0.001	0.148	0.166	<0.001
K-nn	0.028	<0.001	0.168	0.166	<0.001
RSd					
GREG	0.018	0.011	0.173	0.170	0.029
poly	0.041	0.011	0.152	0.170	0.029
K-nn	0.040	0.011	0.172	0.170	0.029

(C) Scenario 3

	Estimators				
	$\widehat{t}_{m_r,\widehat{p}}$	$\widehat{t}_{NWA}$	$\widehat{t}_{imp}$	$\widehat{t}_{naive}$	$\widehat{t}_{HT}$
RB					
GREG	0.165	0.165	0.169	0.166	<0.001
poly	0.106	0.165	0.148	0.166	<0.001
K-nn	0.145	0.165	0.168	0.166	<0.001
RSd					
GREG	0.169	0.169	0.173	0.170	0.029
poly	0.114	0.169	0.152	0.170	0.029
K-nn	0.150	0.169	0.172	0.170	0.029

(D) Scenario 4

The results of the simulations for the real data are presented in Tables 8.

TABLE 8: Bias and standard deviation of the total estimator relative to the true total, for the real data, for scenarios 1 to 4 with three prediction methods: generalized regression (GREG), polynomial regression (poly) and K-nearest neighbors (K-nn) with $K = 5$.

	Estimators			
	$\widehat{t}_{m_r,\widehat{p}}$	$\widehat{t}_{NWA}$	$\widehat{t}_{imp}$	$\widehat{t}_{naive}$
RB				
GREG	0.002	0.002	-0.007	0.318
poly	0.001	0.002	0.003	0.318
K-nn	0.001	0.002	-0.007	0.318
RSd				
GREG	0.012	0.012	0.018	0.325
poly	0.009	0.012	0.010	0.325
K-nn	0.016	0.012	0.018	0.325

(A) Scenario 1

	Estimators			
	$\widehat{t}_{m_r,\widehat{p}}$	$\widehat{t}_{NWA}$	$\widehat{t}_{imp}$	$\widehat{t}_{naive}$
RB				
GREG	0.007	0.199	-0.007	0.318
poly	0.016	0.199	0.003	0.318
K-nn	0.079	0.199	-0.007	0.318
RSd				
GREG	0.015	0.209	0.018	0.325
poly	0.019	0.209	0.010	0.325
K-nn	0.084	0.209	0.018	0.325

(B) Scenario 2

	Estimators			
	$\widehat{t}_{m_r,\widehat{p}}$	$\widehat{t}_{NWA}$	$\widehat{t}_{imp}$	$\widehat{t}_{naive}$
RB				
GREG	0.007	0.002	0.189	0.318
poly	0.016	0.002	0.147	0.318
K-nn	0.022	0.002	0.158	0.318
RSd				
GREG	0.065	0.012	0.201	0.325
poly	0.062	0.012	0.160	0.325
K-nn	0.075	0.012	0.173	0.325

(C) Scenario 3

	Estimators			
	$\widehat{t}_{m_r,\widehat{p}}$	$\widehat{t}_{NWA}$	$\widehat{t}_{imp}$	$\widehat{t}_{naive}$
RB				
GREG	0.199	0.199	0.189	0.318
poly	0.173	0.199	0.147	0.318
K-nn	0.164	0.199	0.158	0.318
RSd				
GREG	0.209	0.209	0.201	0.325
poly	0.184	0.209	0.160	0.325
K-nn	0.181	0.209	0.173	0.325

(D) Scenario 4

Chapter 4

High-dimensional Variance Estimation in Finite Population Sampling

This chapter corresponds to the work in progress Eustache, Dagdoug, and Haziza (2023).

4.1 Introduction

Predictive modeling can be applied at various stages of a survey to enhance the precision of a point estimator and to address the problem of missing values, among others. Using predictive models enables us to establish relationships between a survey variable Y and a set of predictors $X_1, X_2, \ldots, X_p$. For instance, model-assisted estimation procedures make use of a set of predicted values to improve the efficiency of point estimators; e.g., see Särndal (1992) and Breidt and Opsomer (2017). In order to mitigate the potential nonresponse bias caused by item nonresponse, it is common practice to employ some form of imputation, which involves generating a set of predictions to substitute for the missing values; e.g., see Haziza (2009) and Chen and Haziza (2019).

The literature on predictive models for survey data has primarily focused on low-dimensional data settings, where the number of variables p is small relative to the sample size n. Formalized mathematically, it means that $p/n \to 0$. Some notable exceptions include Cardot, Goga, and Shehzad (2017), Ta et al. (2020), Chauvet and Goga (2022) and Dagdoug, Goga, and Haziza (2022). With the advent of big data sets, high-dimensional settings are becoming more prevalent. In this article, a high-dimensional setting refers to a situation where the number of predictors p is of the same order of magnitude as the sample size n so that $p/n \to \kappa^* \in (0,1)$.

High-dimensional linear regression models pose some challenges compared to traditional linear regression models with a smaller number of predictors. In particular, the common variance estimation procedures tend to breakdown when $p/n \to \kappa^* \in (0,1)$. On the one hand, variance estimators based on a first-order Taylor expansion procedure tend to lead to substantial underestimation of the true variance. On the other, resampling procedures such as the jackknife and the bootstrap tend to overestimate the variance of the point estimators; see Wolter (2007) and Mashreghi, Haziza, and Léger (2016) for a treatment of resampling methods in finite population sampling.

In this article, we explain why variance estimators based on a first Taylor and jackknife variance estimators tend to breakdown through a mix of empirical and theoretical investigations. We consider two different setups that involve the customary linear regression model: (1) The model-assisted estimation setup through the use of the generalized regression estimator (see, e.g., Särndal, 1980; Särndal, 1992; Särndal, 2007); (2) The deterministic linear regression imputation setup (e.g., Chen and Haziza, 2019).

We adopt the following notations. Let $U := \{1, 2, ..., N\}$ be a finite population of size N. Our interest lies in estimating the finite population mean

$$\mu_y := \frac{1}{N} \sum_{k \in U} y_k,$$

of a survey variable Y, where y_k denotes the y-value attached to unit k. We select a sample, S, of size n_s and of expected size n, according to a probability sampling design $\mathscr{P}(S \mid \boldsymbol{Z})$, where $\boldsymbol{Z} \in \mathbb{R}^{N\times d}$ denotes the matrix of design information. We restrict our attention to non-informative sampling design; see, e.g., Pfeffermann and Sverchkov (2009). The sample S is fully characterized by the vector of sample selection indicators, $\boldsymbol{I} := [I_1, I_2, \ldots, I_N]^\top$, where $I_k := 1$ if $k \in S$, and $I_k := 0$, otherwise. We denote by $\pi_k := \mathbb{P}(I_k = 1) > 0$ and $\pi_{k\ell} := \mathbb{P}(I_k = 1, I_\ell = 1) > 0$, for $k, \ell \in U$, the first-order and the second-order inclusion probabilities, respectively.

4.2 Linear prediction in survey sampling

We consider the customary linear regression model:

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \epsilon_k, \qquad k \in U,$$

where $\boldsymbol{\beta}$ is a p-vector of unknown coefficients and the errors ϵ_k satisfy $\mathbb{E}[\epsilon_k|\mathbf{x}_k] = 0$, $\mathbb{E}\left[\epsilon_k^2|\mathbf{x}_k\right] := \sigma^2 < \infty$ and are independently and identically distributed. We assume that the intercept is included in the covariates; i.e., the first component of $\mathbf{x}_k$ is 1 for all $k \in U$. Although we assume an homoscedastic variance structure, our results can be easily extended to the case of an heteroscedastic variance structure.

Below, we use the notation $\boldsymbol{y}_S \in \mathbb{R}^{n_s}$ and $\boldsymbol{X}_S \in \mathbb{R}^{n_s\times p}$ to denote the vector of y-values and the design matrix corresponding to the sample, respectively. Also, we use $\boldsymbol{\Pi}_S \in \mathbb{R}^{n_s\times n_s}$ to denote the diagonal matrix, whose k-th diagonal element is π_k.

4.2.1 Model-assisted estimation

In this section, we assume that the observed data are given by

$$\mathscr{D}_{ma} := \{(\mathbf{x}_k, y_k) \ ; \ k \in S\}.$$

In addition, we assume that the vector of population totals,

$$\mathbf{t}_\mathbf{x} := \left[\sum_{k\in U} x_{1k}, \sum_{k\in U} x_{2k}, ..., \sum_{k\in U} x_{pk}\right]^\top,$$

is available from an external source. The Generalized REGression (GREG) estimator of μ_y is given by

$$\widehat{\mu}_{greg} := \frac{1}{N}\left(\sum_{k\in U} \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_S + \sum_{k\in S} \frac{y_k - \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_S}{\pi_k}\right), \tag{4.2.1}$$

where

$$\widehat{\boldsymbol{\beta}}_S = \left(\boldsymbol{X}_S^\top \boldsymbol{\Pi}_S^{-1} \boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S \boldsymbol{\Pi}_S^{-1} \boldsymbol{y}_S$$

is the weighted least squares estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_S := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{k\in S} \frac{\left(y_k - \mathbf{x}_k^\top \boldsymbol{\beta}\right)^2}{\pi_k}.$$

Throughout the paper, we assume that the $p \times p$ matrix, $\boldsymbol{A}_{\Pi S} := \boldsymbol{X}_S^\top \boldsymbol{\Pi}_S^{-1} \boldsymbol{X}_S$, is non-singular. In our setting, the GREG estimator can be written in the so-called projection form:

$$\widehat{\mu}_{greg} = \frac{1}{N} \sum_{k \in U} \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_S$$

since

$$\sum_{k \in S} \pi_k^{-1} \widehat{\epsilon}_{kS} = 0,$$

where $\widehat{\epsilon}_{kS} := y_k - \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_S$ denotes the sample residual attached to unit $k \in S$, Särndal, Swensson, and Wretman (1992, Chapter 6).

We adopt, for distribution of reference of model-assisted estimators, the joint distribution induced by the superpopulation model (4.2) and the sampling design. Consider the following decomposition:

$$\mathbb{V}_{mp}\left(\widehat{\mu}_{greg}\right) = \mathbb{E}_m\left[\mathbb{V}_p\left(\widehat{\mu}_{greg}\right)\right] + \mathbb{V}_m\left(\mathbb{E}_p\left[\widehat{\mu}_{greg}\right]\right), \tag{4.2.2}$$

where the subscripts p and m are used to specify that the variance or the expectation are computed according to the sampling design and the imputation model, respectively.

Using (4.2.2), an estimator of the variance of $\widehat{\mu}_{greg}$ based on a first-order Taylor expansion is given by

$$\widehat{V}_{tay}\left(\widehat{\mu}_{greg}\right) := \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\epsilon}_{kS}}{\pi_k} \frac{\widehat{\epsilon}_{\ell S}}{\pi_\ell} + \frac{\widehat{\sigma}^2}{N},$$

where $\widehat{\sigma}^2$ denotes an unbiased estimator of σ^2. The first term will be the focus of this article.

Särndal, Swensson, and Wretman (1989) advocated for the use of a g-weighted version, which is obtained from (4.2.1) by replacing $\widehat{\epsilon}_{kS}$ with $g_k \times \widehat{\epsilon}_{kS}$, where

$$g_k := 1 + \left(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x},\pi}\right)^\top \boldsymbol{A}_{\Pi S}^{-1} \mathbf{x}_k = \mathbf{t}_{\mathbf{x}}^\top \boldsymbol{A}_{\Pi S}^{-1} \mathbf{x}_k, \qquad k \in S, \tag{4.2.3}$$

is the so-called g-weight attached to unit $k \in S$, with $\widehat{\mathbf{t}}_{\mathbf{x},\pi}$ denoting the Horvitz–Thompson estimator of $\mathbf{t}_{\mathbf{x}}$. The second equality of (4.2.3) is satisfied as the intercept is included in the set of predictors, and the variance structure is homoscedastic. This leads to

$$\widehat{V}_g\left(\widehat{\mu}_{greg}\right) := \frac{1}{N^2} \sum_{k \in U} \sum_{\ell \in U} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{g_k \widehat{\epsilon}_{kS}}{\pi_k} \frac{g_\ell \widehat{\epsilon}_{\ell S}}{\pi_\ell} + \frac{\widehat{\sigma}^2}{N}.$$

Jackknife variance estimation for the GREG estimator has been discussed in Yung and Rao (1996), Duchesne (2000) and Valliant (2002), among others. Here, we consider the generalized jackknife variance estimator of Campbell (1980) and Berger and Skinner (2005). Let $\widetilde{h}_{kk}^{\pi} := \mathbf{x}_k^\top \boldsymbol{A}_{\Pi S}^{-1} d_k \mathbf{x}_k$ be the survey weighted leverage of element $k \in S$, with $d_k = \pi_k^{-1}$. Our next proposition establishes a closed-form for the generalized jackknife variance estimator of the GREG estimator.

Proposition 4.2.1. *An estimator of (4.2.2) based on the generalized jackknife variance estimator of Berger and Skinner (2005) has a closed-formed formula given by*

$$\widehat{V}_{jack}\left(\widehat{\mu}_{greg}\right) = \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{(1-w_k)\, g_k\, \widehat{\epsilon}_{kS}}{\left(1-\widetilde{h}_{kk}^{\pi}\right)\pi_k} \frac{(1-w_\ell)\, g_\ell\, \widehat{\epsilon}_{\ell S}}{\left(1-\widetilde{h}_{\ell\ell}^{\pi}\right)\pi_\ell} + \frac{\widehat{\sigma}^2}{N},$$

where $w_k := (N\pi_k)^{-1}$ *for* $k \in S$.

Proof. See Appendix 4.8. ∎

4.2.2 Deterministic linear regression imputation

Predictions based on a linear regression model are also used in the context of imputation for item nonresponse in surveys. In this context, the survey variable Y is observed only for a subset $S_r \subseteq S$, called the set of respondents to item Y. We denote by $S_m = S - S_r$ the set of nonrespondents to item Y. Let $\boldsymbol{R} := [R_1, R_2, \ldots, R_N]^\top$ the N-vector of response indicators, where $R_k = 1$ if $k \in S_r$, and $R_k = 0$, otherwise. Here, the values of the predictors $X_1, \ldots, X_p$, are assumed to be available for both the respondents and the nonrespondents. We assume that: (i) The data $\{(\mathbf{x}_k, y_k, r_k)\}_{k\in U}$ are identically and independently distributed; (ii) The data are Missing At Random (MAR, Rubin, 1976):

$$\mathbb{P}(R_k = 1|\mathbf{x}_k, y_k) = \mathbb{P}(R_k = 1|\mathbf{x}_k);$$

(iii) The positivity assumption is satisfied; i.e., $\mathbb{P}(R_k = 1|\mathbf{x}_k) > 0$ for all $k \in U$. Available to the imputer are the data

$$\mathscr{D}_{imp} := \{(\mathbf{x}_k, y_k) \ ; \ k \in S_r\} \bigcup \{\mathbf{x}_k \ ; \ k \in S_m\}.$$

Imputation consists of estimating the relationship between Y and $X_1, \ldots, X_p$ based on the respondents and extrapolating this relationship to the set of nonrespondents.

An estimator of μ_y after deterministic linear imputation is given by

$$\widehat{\mu}_{lr} := \frac{1}{\widehat{N}} \left(\sum_{k\in S_r} \frac{y_k}{\pi_k} + \sum_{k\in S_m} \frac{\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_R}{\pi_k} \right) = \frac{1}{\widehat{N}} \sum_{k\in S} \frac{\widetilde{y}_k}{\pi_k},$$

where $\widehat{N} := \sum_{k\in S} \pi_k^{-1}$ and $\widetilde{y}_k := R_k y_k + (1 - R_k)\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_R$ where

$$\widehat{\boldsymbol{\beta}}_R = \left(\sum_{k\in S_r} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\pi_k} \right)^{-1} \sum_{k\in S_r} \frac{\mathbf{x}_k y_k}{\pi_k} = \left(\boldsymbol{X}_R \boldsymbol{\Pi}_R^{-1} \boldsymbol{X}_R^\top \right)^{-1} \boldsymbol{X}_R \boldsymbol{\Pi}_R^{-1} \boldsymbol{y}_R \tag{4.2.4}$$

is the weighted least squares estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}_R := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{k\in S_r} \frac{\left(y_k - \mathbf{x}_k^\top \boldsymbol{\beta}\right)^2}{\pi_k}.$$

In (4.2.4), the quantities $\boldsymbol{X}_R$, $\boldsymbol{\Pi}_R$ and $\boldsymbol{y}_R$ correspond to the counterparts of $\boldsymbol{X}_S$, $\boldsymbol{\Pi}_S$ and $\boldsymbol{y}_S$, respectively, restricted to the set of respondents S_r.

To estimate the variance of $\widehat{\mu}_{lr}$, we consider the reverse framework, originally proposed by Fay (1991) and Shao and Steel (1999); see also Kim and Rao (2009) and Haziza and Vallée (2020). Using this framework, the total variance of $\widehat{\mu}_{lr}$ can be expressed as

$$\mathbb{V}(\widehat{\mu}_{lr}) = \mathrm{E}_m \mathrm{E}_q \mathbb{V}_p(\widehat{\mu}_{lr}) + \mathrm{E}_q \mathbb{V}_m \mathrm{E}_p(\widehat{\mu}_{lr} - \mu_y),$$

where the subscript q denotes the nonresponse mechanism. Let us define $\widehat{h}_{k\ell}^{\pi} := \mathbf{x}_k^\top \boldsymbol{A}_{\Pi R}^{-1} d_k \mathbf{x}_\ell$ and $\widehat{\Gamma}_k := \sum_{\ell\in S_m} \widehat{h}_{k\ell}^{\pi}$, where $\boldsymbol{A}_{\Pi R} := \boldsymbol{X}_R^\top \boldsymbol{\Pi}_R^{-1} \boldsymbol{X}_R$. Again, we assume that the matrix $\boldsymbol{A}_{\Pi R}$ is non-singular.

Proposition 4.2.2. *An estimator of the variance of $\widehat{\mu}_{lr}$ based on a first-order Taylor expansion is given by*

$$\widehat{V}_{tay}(\widehat{\mu}_{lr}) := \frac{1}{\widehat{N}^2} \sum_{k\in S} \sum_{\ell\in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\xi}_k - \widehat{\mu}_{lr}}{\pi_k} \frac{\widehat{\xi}_\ell - \widehat{\mu}_{lr}}{\pi_\ell} + \frac{\sigma^2}{\widehat{N}^2} \sum_{k\in S_r} \frac{1}{\pi_k} \left\{1 - R_k(1 + \widehat{\Gamma}_k)\right\}^2,$$

where

$$\widehat{\xi}_k := \widetilde{y}_k + r_k \widehat{\Gamma}_k \, \widehat{\epsilon}_{kR},$$

with $\widehat{\epsilon}_{kR} = y_k - \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_R$, $k \in S_r$.

The proof of Proposition 4.2.2 is straightforward and is thus omitted.

We now turn to jackknife variance estimation in the context of deterministic linear regression. Berger and Rao (2006) extended the results of Berger and Skinner (2005) to the case of mean and ratio imputations. We extend the results of Berger and Rao (2006) to the more general case of deterministic linear regression imputation, that includes mean and ratio imputations as special cases. Below, we provide an estimator of the total variance in (4.2.2) based on the generalized jackknife of Berger and Rao (2006) in the context of deterministic linear regression imputation and prove a closed-form solution.

Result 4.2.1. *Noting that*

$$\widehat{\mu}_{lr}^{(k)} - \widehat{\mu}_{lr} = \frac{d_k}{\widehat{N} - d_k} \left(\widehat{\mu}_{lr} - \widehat{\xi}_k^{(jack)} \right),$$

where

$$\widehat{\xi}_k^{(jack)} := \widetilde{y}_k + r_k \widehat{\Gamma}_k \frac{\widehat{\epsilon}_{kR}}{1 - \widehat{h}_{kk}^{\pi}},$$

an estimator of $\widehat{V}_{tay}(\widehat{\mu}_{lr})$ *based on the generalized jackknife variance estimator of Berger and Rao (2006) has a closed-form expression given by*

$$\widehat{V}_{jack}(\widehat{\mu}_{lr}) = \frac{1}{\widehat{N}^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\mu}_{lr} - \widehat{\xi}_k^{(jack)}}{\pi_k} \frac{\widehat{\mu}_{lr} - \widehat{\xi}_\ell^{(jack)}}{\pi_\ell} + \frac{\sigma^2}{\widehat{N}^2} \sum_{k \in S_r} \frac{1}{\pi_k} \left\{ 1 - R_k (1 + \widehat{\Gamma}_k) \right\}^2.$$

Proof. See Appendix 4.9. ■

4.3 Behavior of some commonly used variance estimators: Empirical studies

In this section, we present the results of two limited simulation studies. In Section 4.3.1, we examine the empirical performance of the variance estimators in a model-assisted estimation framework (see Section 4.2.1), whereas Section 4.3.2 covers the variance estimators discussed in Section 4.2.2. We denote by $\mathscr{K} = p/n$ the ratio of the number of predictors to the expected sample size. In the case of nonresponse, we have $\mathscr{K} = p/\mathrm{E}[n_r]$ the ratio of the number of predictors to the expected number of respondents.

4.3.1 Model-assisted estimation: the GREG estimator

We generated a finite population U of size $N = 5,000$ consisting of 223 explanatory variables $X_1, \ldots, X_{223}$ and a survey variable Y. The variables $X_1, \ldots, X_{223}$, were generated from a multivariate normal distribution with a mean vector equal to $5 \times \mathbf{1}^\top$ and correlation matrix, whose diagonal elements were equal to 1 and off-diagonal elements equal to 0.3, where $\mathbf{1}$ denotes the vector of ones. Given the X-variables, we generated a survey variable Y according to the linear regression model

$$y_k = 14 - 4x_{1k} + 3x_{2k} + 4x_{3k} + \epsilon_k,$$

where the errors ϵ_k were generated from a normal distribution with mean equal to 0 and variance equal to 25^2. This led to a model R^2 approximately equal to 0.6. In (4.3.1), note that only the first three variables X_1, X_2, and X_3 were used for generating the Y-variable.

From the population, we selected $M = 10,000$ samples, of (expected) sample size $n = 300$, according to two sampling designs: simple random sampling design without replacement and Bernoulli sampling. In each sample, we computed several GREG estimators, $\widehat{\mu}_{greg}$, given by (4.2.1), based on different sets of explanatory variables. In addition to X_1, X_2 and X_3, we included a number of noise variables denoted by p_{noise}. The values for p_{noise} were set to: 0, 20, 40, 60, 80, 100, 120, 140, 160, 180, 200 and 220. This led to 12 estimators, $\widehat{\mu}_{greg}$, of μ_y. To estimate the variance of $\widehat{\mu}_{greg}$, we computed $\widehat{V}_{tay}$ given by (4.2.1) and $\widehat{V}_{jack}$ given by (4.2.1). As a measure of relative bias of a variance estimator, we computed its Monte Carlo percent relative bias (RB). Using the generic notations $\widehat{\mu}$ and $\widehat{V}$ for a point and a variance estimator, respectively, the RB of $\widehat{V}$ is defined as

$$RB(\widehat{V}) := 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{\widehat{V}^{(r)} - V_{MC}(\widehat{\mu})}{V_{MC}(\widehat{\mu})},$$

where $V_{MC}(\widehat{\mu})$ denotes the Monte-Carlo variance of $\widehat{\mu}$ and $\widehat{V}^{(r)}$ denotes the estimator $\widehat{V}$ at the rth iteration, $r = 1, \ldots, 10,000$. The results for simple random sampling without replacement and Bernoulli sampling are shown in Figures 1 and 2, respectively.

From Figures 1 and 2, we note that both $\widehat{V}_{\text{tay}}$ and $\widehat{V}_{\text{jack}}$ performed well for small values of p/n. For instance, for $p/n = 3/223$, which corresponds to the case of $p_{noise} = 0$, the estimator $\widehat{V}_{\text{tay}}$ exhibited a value of RB of about -0.8% for Bernoulli sampling and -2.8% for simple random sampling without replacement. The jackknife variance estimator $\widehat{V}_{\text{jack}}$ showed a bias of about 1.2% for Bernoulli sampling and -0.8% for simple random sampling without replacement. However, for $p/n = 3/83 \approx 0.28$, the RB of $\widehat{V}_{\text{tay}}$ was equal to -30% in the case of Bernoulli sampling and -30.8% in the case of simple random sampling without replacement. The amount of underestimation got worse as p/n increased. On the other hand, the jackknife variance estimator exhibited significant overestimation with values of RB equal to 34.2% in the case of Bernoulli sampling and 33.4% in the case of simple random sampling without replacement. The bias increased as the value of p/n increased.

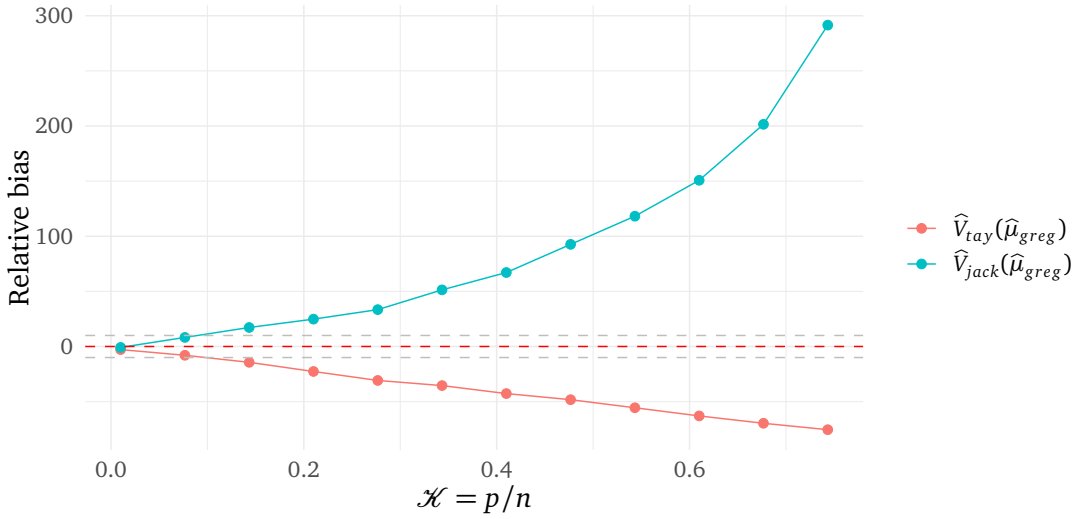

FIGURE 1: Behaviour of two variance estimators of the GREG estimator under simple random sampling without replacement.

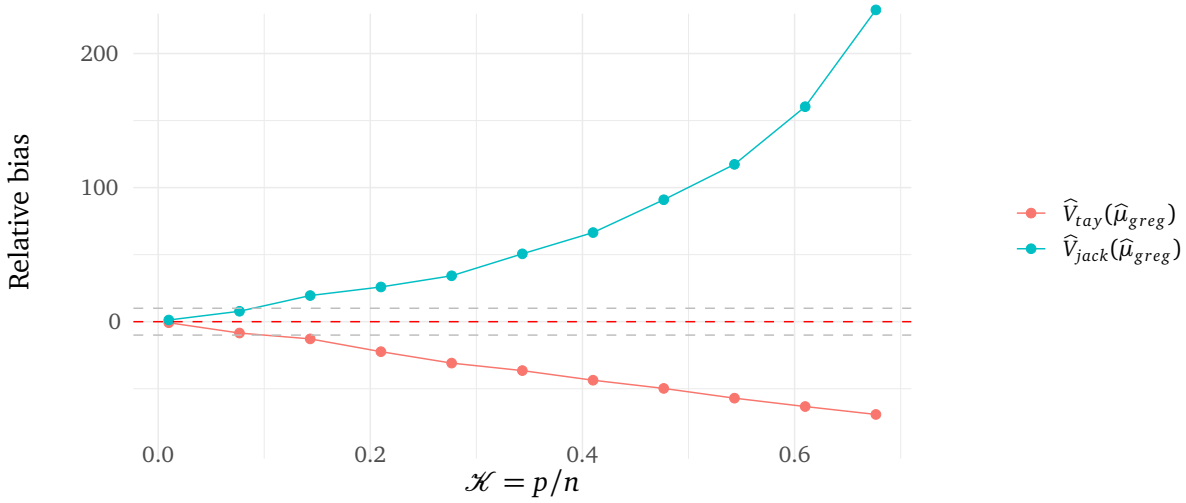

FIGURE 2: Behaviour of two variance estimators of the GREG estimator under Bernoulli sampling.

4.3.2 Deterministic linear regression imputation

We started by generating 5,000 realizations of a vector of explanatory variables, of size 113, from a multivariate normal distribution with a mean vector equal to $5 \times \mathbf{1}^\top$ and correlation matrix, whose diagonal elements were equal to 1 and the off-diagonal elements were equal to 0.3, where $\mathbf{1}$ denotes the vector of ones. We then repeated $R = 10,000$ iterations of the following process:

(i) Given the explanatory variables, we generated the survey variable Y according to the model defined in Section 4.3.1.

(ii) From the finite population of size $N = 5,000$ generated in Step (i), a sample, of (expected) size $n = 300$, was selected according to (1) simple random sampling without replacement and (2) Bernoulli sampling.

(iii) In each sample, the response indicators R_k, $k \in S$, were independently generated according to a Bernoulli distribution with probability

$$p_k = \{1 + \exp(1 + \lambda_1 x_{1k} + \lambda_2 x_{2k} + \lambda_3 x_{3k})\}^{-1},$$

where the values of λ_1-λ_3 were set to obtain an overall response rate of about 50%. Thus, in each sample, the expected number of respondents, $\mathrm{E}(n_r)$, was equal to 150.

(iv) The missing values in each sample were imputed through deterministic linear regression imputation with different subsets of explanatory variables. The first subset of explanatory variables included the variables X_1,X_2 and X_3 only, corresponding to the true model. In addition to X_1,X_2 and X_3, we included a number of noise variables denoted by p_{noise}. This led to 12 sets of explanatory variables of size p, where $p = p_{noise} + 3$. The values for p_{noise} were set to: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 and 110. As a result the ratio $p/\mathrm{E}(n_r)$ ranged from 3/150 to 113/150. Each of the 12 models was fitted on the set of responding units, which led to 12 sets of imputed values.

(v) For each of the 12 sets of imputed values, we computed the imputed estimator $\widehat{\mu}_{lr}$ given by (4.2.2), leading to a set of 12 imputed estimators.

(vi) We estimated the variance of the 12 imputed estimators using two variance estimators: (i) The variance estimator based on a first-order Taylor expansion denoted by $\widehat{V}_{tay}$; see

Section 4.2.2; and (ii) The generalized jackknife variance estimator, denoted by $\widehat{V}_{jack}$; see Section 4.2.2.

As a measure of relative bias of a variance estimator $\widehat{V}$, we computed its Monte Carlo percent relative bias (RB) given by (4.3.1).

From Figures 3 and 4, we note that both $\widehat{V}_{tay}$ and $\widehat{V}_{jack}$ performed well for small values of $p/\mathrm{E}(n_r)$. For instance, for $p/\mathrm{E}(n_r) = 3/113$, which corresponds to the case of $p_{noise} = 0$, the estimator $\widehat{V}_{tay}$ exhibited a value of RB of about -5.8% for both simple random sampling without replacement and Bernoulli sampling. The jackknife variance estimator performed well with values of RB equal to -6.% for simple random sampling without replacement and -3.9% for Bernoulli sampling. However, for larger values of $p/\mathrm{E}(n_r)$ both variance estimators did not perform well. For instance, for $p/\mathrm{E}(n_r) = 33/113 \approx 0.29$, the estimator $\widehat{V}_{tay}$ underestimated the true variance with values of RB equal to -17.2% for simple random sampling without replacement and -17.7% for Bernoulli sampling. On the other hand, the estimator $\widehat{V}_{jack}$ was 17.2% too large for simple random sampling without replacement and 20.3% too large for Bernoulli sampling.

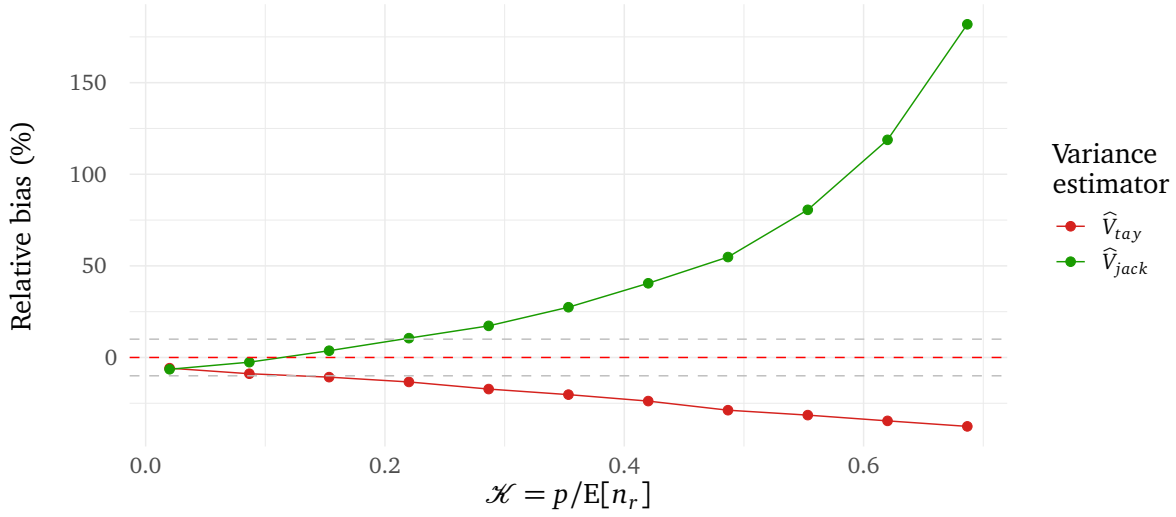

FIGURE 3: Behaviour of two variance estimators of the linear imputed estimator under simple random sampling without replacement.

4.3.3 Explaining the behavior of classical variance estimators

In the context of both model-assisted estimation and deterministic linear regression imputation, the customary variance estimators based on a first-order Taylor expansion and the generalized jackknife variance estimator tend to breakdown when $p/n \to \kappa^* \in (0,1)$ (or $p/\mathrm{E}[n_r] \to \kappa^* \in (0,1)$). In this section, we explain why this is the case. For simplicity, we confine to the case of model-assisted estimation under simple random sampling without replacement. Arguments similar to the ones below can also be used to explain the behavior of classical variance estimators under deterministic linear regression imputation.

The variance estimators based on a first-order Taylor expansion given by (4.2.1) involves the sample residuals $\widehat{\epsilon}_{kS}$. It turns out that, in a high-dimensional setting, the distribution of the sample residuals $\widehat{\epsilon}_{kS}$ is not a good approximation of the distribution of the errors ϵ_k in (4.2). In particular, we have

$$\mathbb{V}_m(\widehat{\epsilon}_{kS}) = \sigma^2(1-\widetilde{h}_{kk}),$$

where $\widetilde{h}_{kk}$ denotes the kth diagonal element of the hat matrix $\boldsymbol{X}_S^\top \left(\boldsymbol{X}_S\boldsymbol{X}_S^\top\right)^{-1}\boldsymbol{X}_S$. The validity of classical variance estimators relies on the assumption that $\widetilde{h}_{kk} \to 0$ as n and N go to infinity. In a

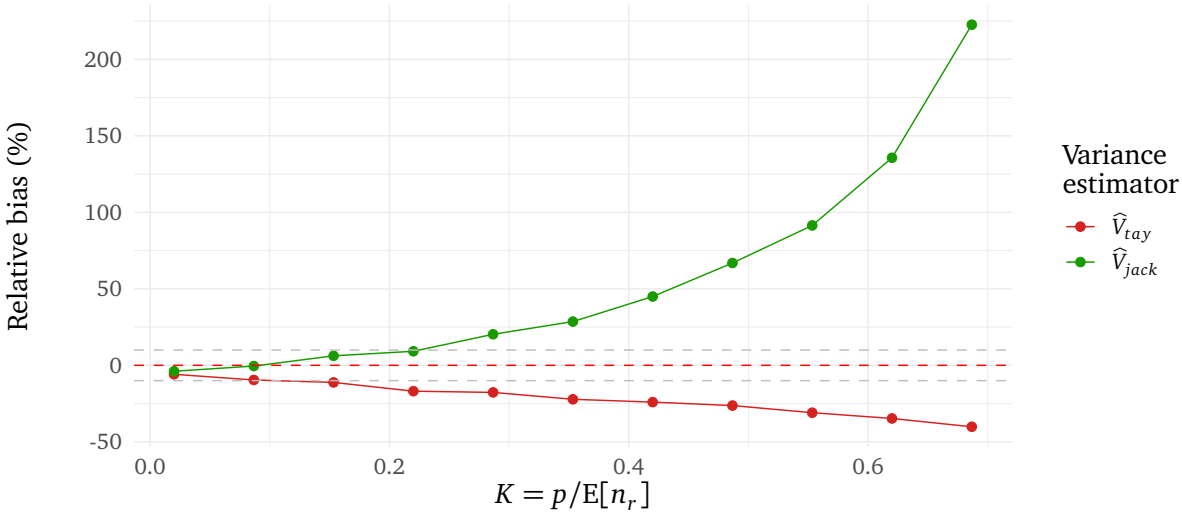

FIGURE 4: Behaviour of two variance estimators of the linear imputed estimator under Bernoulli sampling.

high-dimensional setting, this assumption no longer holds, as it can be shown that

$$\widetilde{h}_{kk} = \frac{p}{n} + o_p(1),$$

for a wide class of distributions for the design matrix $\boldsymbol{X}_S$ (e.g., the multivariate normal distribution); see e.g., El Karoui and Purdom (2018), Pajor and Pastur (2009) and Karoui and Koesters (2011) for a discussion. As a result, the variance of the sample residuals $\widehat{\epsilon}_{kS}$ is approximately equal to $\sigma^2(1-p/n)$, which can be considerably smaller than σ^2 for large values of p/n. This, in turn, explains why the variance estimator based on a first-order Taylor expansion tends to underestimate the true variance of $\widehat{\mu}_{greg}$ for large values of p/n.

Turning to generalized jackknife variance estimators, we note from (4.2.1) that it involves the residuals $\widehat{\epsilon}_{kS}^{(k)} = \widehat{\epsilon}_{kS}/(1-\widetilde{h}_{kk})$. Since

$$\mathbb{V}_m(\widehat{\epsilon}_{kS}^{(k)}) = \frac{\sigma^2}{1-\widetilde{h}_{kk}} \simeq \frac{\sigma^2}{1-\frac{p}{n}},$$

the variance of $\widehat{\epsilon}_{kS}^{(k)}$ may be considerably larger than σ^2 for large values of p/n. As a result, the generalized jackknife variance estimator tends to overestimate the true variance of $\widehat{\mu}_{greg}$ for large values of p/n.

4.4 Bias: Model-assisted estimation

In the previous section, we illustrated that commonly used variance estimators for the GREG may present important biases in high-dimensional scenarios. In this section, we analyze theoretically the reasons of these biases. For simplicity, we restrict our analysis to the case of Bernoulli sampling. However, it is essential to note that our results likely remain good approximations with other sampling designs. We consider the framework introduced in Isaki and Fuller (1982) for our asymptotic results. An increasing sequence of finite populations $\{U_\nu\}_{\nu\in\mathbb{N}}$ of respective sizes $\{N_\nu\}_{\nu\in\mathbb{N}}$ is considered. We emphasize that the populations are assumed to be embedded, that is, $U_\nu \subset U_{\nu+1}$, for all $\nu \in \mathbb{N}$. In each population U_ν, a sample S_ν is selected using a sample

design $\mathscr{P}_\nu(\,\cdot\,,\,\mathbf{Z}_\nu)$ with purpose of estimating $\mu_{y,\nu}$. The first and second order inclusion probabilities of $\mathscr{P}_\nu(\,\cdot\,,\,\mathbf{Z}_\nu)$ are denoted by $\{\pi_{k,\nu}\}_{k\in U_\nu}$ and $\{\pi_{k\ell,\nu}\}_{k\neq\ell\in U_\nu}$, respectively. For simplicity, in what follows, we omit the index ν whenever no confusion arises. For two sequences $\{a_\nu\}_{\nu\in\mathbb{N}}$ and $\{b_\nu\}_{\nu\in\mathbb{N}}$, we write $a_\nu \simeq b_\nu$ to express that they share the same limit, i.e., $\lim_{\nu\to\infty} a_\nu/b_\nu = 1$. We extend this meaning to sequences of random variables where the limit is to be understood in the probability sense. Moreover, in this article, asymptotic order notations are to be understood in a high-dimensional asymptotic framework in which $\lim_{\nu\to\infty} p_\nu/n_\nu = \kappa^* \in (0;1)$.

In this article, the inference is made conditionally to the predictors. They may either be fixed or originated from a random sample, in which case we assume that the conditions on the design that we impose below hold almost surely.

(H1) The predictors are such that, for all k,

$$\widetilde{h}_{kk} = \frac{p_\nu}{n_\nu} + o(1).$$

Assumption (H1) should hold when the data originated from a random sample. We refer to Portnoy (1987) for more details on the matter, including a proof when $\boldsymbol{X}_U$ is a Gaussian data matrix, where $\boldsymbol{X}_U$ is the design matrix corresponding to the population.

Result 4.4.1. *Consider a Bernoulli sampling design and denote by* $\mathbb{V}_m\big(\widehat{\mu}_{greg,\nu}\big)$ *the variance of* $\widehat{\mu}_{greg,\nu}$ *with respect to the superpopulation model.*

i) If the superpopulation model is linear, the superpopulation model variance $\mathbb{V}_m\big(\widehat{\mu}_{greg,\nu}\big)$ *is unbiased for the unconditional variance* $\mathbb{V}\big(\widehat{\mu}_{greg,\nu}\big)$*, that is,*

$$\mathbb{E}_p\big[\mathbb{V}_m\big(\widehat{\mu}_{greg,\nu}\big)\big] = \mathbb{V}\big(\widehat{\mu}_{greg,\nu}\big).$$

Moreover, if, for an arbitrary k, $\lim_{\nu\to\infty} \mathbb{V}_p\big(g_{k,\nu}\big) = 0$*, then* $\mathbb{V}_m\big(\widehat{\mu}_{greg,\nu}\big)$ *and* $\mathbb{V}\big(\widehat{\mu}_{greg,\nu}\big)$ *are asymptotically equivalent, that is,*

$$\frac{\mathbb{V}_m\big(\widehat{\mu}_{greg,\nu}\big)}{\mathbb{V}\big(\widehat{\mu}_{greg,\nu}\big)} \xrightarrow{\mathbb{P}} 1.$$

ii) The biases of $\widehat{V}_{tay,\nu}\big(\widehat{\mu}_{greg,\nu}\big)$*,* $\widehat{V}_{g,\nu}\big(\widehat{\mu}_{greg,\nu}\big)$ *and* $\widehat{V}_{jack,\nu}\big(\widehat{\mu}_{greg,\nu}\big)$ *can be expressed as follows:*

$$\frac{\mathbb{E}_m\big[\widehat{V}_{tay,\nu}\big(\widehat{\mu}_{greg,\nu}\big)\big]}{\mathbb{V}_m\big(\widehat{\mu}_{greg,\nu}\big)} = \frac{N}{\sum_{k\in U} g_k}\frac{n_s}{n}\left(1+\pi\left\{\frac{n}{n_s}-1\right\}-\mathscr{K}(1-\pi)\right),$$

$$\frac{\mathbb{E}_m\big[\widehat{V}_{g,\nu}\big(\widehat{\mu}_{greg,\nu}\big)\big]}{\mathbb{V}_m\big(\widehat{\mu}_{greg,\nu}\big)} \simeq (1-\pi)(1-\mathscr{K}) + \pi\times\frac{N}{\sum_{k\in U} g_k},$$

$$\frac{\mathbb{E}_m\big[\widehat{V}_{jack,\nu}\big(\widehat{\mu}_{greg,\nu}\big)\big]}{\mathbb{V}_m\big(\widehat{\mu}_{greg,\nu}\big)} = \frac{1-\pi}{1-\mathscr{K}} + \pi\times\frac{N}{\sum_{k\in U} g_k}.$$

Part i) of Result 4.4.1 shows that the model variance of the GREG estimator is unbiased and consistent for the unconditional variance. This holds only under the crucial assumption that the regression function is linear in the covariates. It is worth noting that Part i) also holds for general sampling designs, provided that appropriate assumptions are made on higher-order inclusion

probabilities. These properties allow for simpler investigations of the properties of variance estimators since the model variance is easily tractable. Part ii) of Result 4.4.1 reveals the reasons of the under-estimation of the Taylor and g-weighted estimators and the over-estimation of the Jackknife estimator. It is worth noting that the biases are computable and can be used in practice to construct debiased variance estimators.

Remark 4.4.1. *The behavior of g-weights $\{w_k\}_{k\in U}$ strongly influences the high-dimensional behavior of the variance estimators. If p_ν is either fixed or increases slowly with respect to n_ν, it can be shown (under appropriate assumptions) that, uniformly in k,*

$$g_k \xrightarrow{\mathbb{P}} 1.$$

This behavior, however, may not hold for general covariates settings within a high-dimensional framework in which $\lim_{\nu\to\infty} p_\nu/n_\nu > 0$. The assumption in Result 4.4.1 that

$$\lim_{\nu\to\infty} \mathbb{V}_p\left(g_{k,\nu}\right) = 0$$

is much weaker as it only requires that this limit is degenerate.

Corollary 4.4.1. *Consider a Bernoulli sampling design with an asymptotically negligible sampling fraction. Then,*

$$\frac{\mathbb{E}_m\left[\widehat{V}_{tay,\nu}\left(\widehat{\mu}_{greg,\nu}\right)\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg,\nu}\right)} \simeq \frac{N}{\sum_{k\in U} g_k}(1-\mathscr{K}),$$

$$\frac{\mathbb{E}_m\left[\widehat{V}_{g,\nu}\left(\widehat{\mu}_{greg,\nu}\right)\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg,\nu}\right)} \simeq (1-\mathscr{K}),$$

$$\frac{\mathbb{E}_m\left[\widehat{V}_{jack,\nu}\left(\widehat{\mu}_{greg,\nu}\right)\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg,\nu}\right)} \simeq \frac{1}{1-\mathscr{K}}.$$

In the case of negligible sampling fractions, the biases of Taylor, g-weighted, and jackknife variance estimators are greatly simplified. Interestingly, the bias of the jackknife matches the bias found in El Karoui and Purdom (2018) for the jackknife estimation of the variance of a linear model. This reveals that, even in moderate dimensions, traditional variance estimators underestimate or underestimate the true variance and may lead to unsatisfactory inference (e.g., invalid confidence intervals).

4.5 Bias: Deterministic linear regression imputation

In this section, we study the behavior of variance estimators in the context of deterministic linear regression imputation in high-dimensional situations.

Result 4.5.1. *Under Bernoulli sampling and Assumption (H1), the bias of* $\widehat{V}_{tay,v}\left(\widehat{\mu}_{lr}\right)$ *and* $\widehat{V}_{jack,v}\left(\widehat{\mu}_{lr,v}\right)$ *can be expressed as follows*

$$\frac{\mathrm{E}_m\left[\widehat{V}_{tay,v}(\widehat{\mu}_{lr,v})\right]}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{lr,v})\right]} \simeq 1 + \frac{-\mathscr{K}\widehat{A}_n - \frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k}{\frac{1}{\sigma^2}\sum_{k\in S}B_k + n_s - 1 + \frac{1}{1-\pi_v}\widehat{A}_n},$$

$$\frac{\mathrm{E}_m\left[\widehat{V}_{jack,v}(\widehat{\mu}_{lr,v})\right]}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{lr,v})\right]} \simeq 1 + \frac{\frac{\mathscr{K}}{1-\mathscr{K}}(\widehat{A}_n - n_s\mathscr{K} + 2n_s) - \frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k}{\frac{1}{\sigma^2}\sum_{k\in S}B_k + n_s - 1 + \frac{1}{1-\pi_v}\widehat{A}_n},$$

where

$$B_k = \frac{1}{n_s}\sum_{\ell\in S}\left\{(\mathbf{x}_\ell^\top\boldsymbol{\beta})^2 - \mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta}\right\},$$

and

$$\widehat{A}_n = \sum_{k\in S_r}(\widehat{\Gamma}_k + 1)^2 - n_s.$$

Corollary 4.5.1. *Consider a Bernoulli sampling design with an asymptotically negligible sampling fraction. Then,*

$$\frac{\mathrm{E}_m\left[\widehat{V}_{tay,v}(\widehat{\mu}_{lr,v})\right]}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{lr,v})\right]} \leq \frac{\frac{1}{\sigma^2}\sum_{k\in S}B_k + n_r + (1-\mathscr{K})\left(2n_m + \sum_{k\in S_r}\Gamma_k^2\right)}{\frac{1}{\sigma^2}\sum_{k\in S}B_k + n_r + 2n_m + \sum_{k\in S_r}\Gamma_k^2} \leqslant 1, \tag{4.5.5}$$

$$\frac{\mathrm{E}_m\left[\widehat{V}_{jack,v}(\widehat{\mu}_{lr,v})\right]}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{lr,v})\right]} \leq \frac{\frac{1}{\sigma^2}\sum_{k\in S}B_k + n_r + (\mathscr{K}+2)\,n_m + \frac{1}{1-\mathscr{K}}\sum_{k\in S_r}\Gamma_k^2}{\frac{1}{\sigma^2}\sum_{k\in S}B_k + n_r + 2n_m + \sum_{k\in S_r}\Gamma_k^2} \geqslant 1. \tag{4.5.6}$$

Result 4.5.1 describes the bias of the Taylor and the jackknife variance estimator. Because the covariates are treated as fixed in these bias statements, these biases are fairly difficult to interpret. Nonetheless, we note that, as a consequence of Cauchy-Schwartz inequality,

$$\begin{aligned} & \frac{1}{n_s}\left(\sum_{k\in S}\mathbf{x}_k^\top\boldsymbol{\beta}\right)^2 \leqslant \sum_{k\in S}\left(\mathbf{x}_k^\top\boldsymbol{\beta}\right)^2 \\ \Leftrightarrow \quad & \sum_{k\in S}\left(\mathbf{x}_k^\top\boldsymbol{\beta}\right)^2 - \frac{1}{n_s}\left(\sum_{k\in S}\mathbf{x}_k^\top\boldsymbol{\beta}\right)^2 = \sum_{k\in S}B_k \geqslant 0. \end{aligned}$$

As Corollary 4.5.1 reveals, these bias terms can be greatly simplified in case of negligible sampling fractions. It follows from (4.5.5) that the Taylor variance estimator underestimates the variance of $\widehat{\mu}_{lr,v}$, and from (4.5.6) that the jackknife variance estimator overestimates it. Also, $\sum_{k\in S}B_k = 0$ if $\boldsymbol{\beta} = \mathbf{0}_p$, in which case the biases further simplify and inequalities in (4.5.5) and (4.5.6) become strict.

4.6 Empirical behavior of bias-adjusted estimators

The derivation of the high-dimensional asymptotic biases, such as those presented in Results 4.4.1 and 4.5.1 suggest using bias-adjusted estimators. To be more precise, using the generic notation

$\widehat{V}$ for a variance estimator of $\widehat{\mu}_{greg}$, the adjusted variance estimator of $\mathbb{V}\left(\widehat{\mu}_{greg}\right)$ is defined as

$$\widehat{V}^{(adj)}(\widehat{\mu}_{greg}) := \widehat{V}(\widehat{\mu}_{greg}) \times \frac{\mathbb{V}_m(\widehat{\mu}_{greg})}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{greg})\right]}.$$

In case of model-assisted estimators, the quantities

$$\frac{\mathbb{V}_m(\widehat{\mu}_{greg})}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{greg})\right]}$$

are known in practice and can be computed from the observed data. These estimators are expected to be unbiased, independently of the dimension; indeed,

$$\begin{aligned}\mathbb{E}\left[\widehat{V}^{(adj)}(\widehat{\mu}_{greg})\right] &= \mathbb{E}_p\left[\mathbb{E}_m\left[\widehat{V}(\widehat{\mu}_{greg})\right]\frac{\mathbb{V}_m(\widehat{\mu}_{greg})}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{greg})\right]}\right] \\ &= \mathbb{E}_p\left[\mathbb{V}_m(\widehat{\mu}_{greg})\right] \\ &= \mathbb{V}(\widehat{\mu}_{greg}).\end{aligned}$$

In this section, we are interested in investigating the empirical behaviors of bias-adjusted estimators for the Taylor, the g-weighted and the jackknife variance estimators. We considered the settings described in Section 4.3.1. In addition to the RB of the variance estimators, we computed the RB of their adjusted versions. The results for Bernoulli and simple random sampling without replacement are in Tables 1 and 2, respectively.

TABLE 1: Relative bias for variance estimators and adjusted variance estimators of Taylor, jackknife and Taylor g-weight, for the GREG estimator, for Bernoulli sampling.

		RB						
p	$\mathscr{K}$	$\mathbb{V}_m$	$\widehat{V}_{tay}$	$\widehat{V}_g$	$\widehat{V}_{jack}$	$\widehat{V}_{tay}^{(adj)}$	$\widehat{V}_g^{(adj)}$	$\widehat{V}_{jack}^{(adj)}$
3	0.01	3.21	8.94	4.67	7.35	5.71	5.75	6.42
23	0.08	9.32	-9.03	-2.14	14.77	6.03	6.03	6.86
43	0.14	9.84	-20.83	-8.74	23.92	6.72	6.65	7.72
63	0.21	9.82	-31.83	-15.79	33.83	6.82	6.75	7.91
83	0.28	6.65	-43.99	-25.14	41.54	3.77	3.73	5.01
103	0.34	6.01	-53.48	-32.38	54.97	3.24	3.25	4.86
123	0.41	7.17	-61.84	-39.02	73.64	4.05	3.86	5.81
143	0.48	5.00	-70.09	-47.09	92.58	2.17	1.80	4.20
163	0.54	6.23	-76.64	-53.61	124.01	3.33	2.65	5.58
183	0.61	5.55	-82.84	-60.89	163.98	2.62	1.98	5.57
203	0.68	4.05	-88.06	-68.33	221.38	1.39	0.47	4.99

In these tables, the quantity $\mathbb{V}_m := \mathbb{V}_m(\widehat{\mu}_{greg})$ denotes the model variance of $\widehat{\mu}_{greg}$.
We observe that, as stated by Result 4.4.1 part i), $\mathbb{V}_m$ exhibited negligible biases across all scenarios of $\mathscr{K}$ included in our simulations for both simple random sampling without replacement and Bernoulli sampling. We now restrict temporarily our attention to Table 1 for Bernoulli sampling. As discussed previously in Section 4.3, on the one hand, the Taylor and the g-weighted variance estimators showed negligible biases when $\mathscr{K}$ was negligible but exhibited significant negative biases for moderate to high-dimensional scenarios, with up −88.06% and −68.06%. Similarly,

TABLE 2: Relative bias for variance estimators and adjusted variance estimators of Taylor, jackknife and Taylor g-weight, for the GREG estimator for simple random sampling without replacement.

		RB						
p	$\mathscr{K}$	$\mathbb{V}_m$	$\widehat{V}_{tay}$	$\widehat{V}_g$	$\widehat{V}_{jack}$	$\widehat{V}_{tay}^{(adj)}$	$\widehat{V}_g^{(adj)}$	$\widehat{V}_{jack}^{(adj)}$
3	0.01	5.42	4.23	5.16	7.91	6.22	6.22	6.95
23	0.08	6.38	-7.90	-1.15	16.06	7.08	7.04	8.09
43	0.14	4.06	-21.71	-10.18	21.71	4.87	4.81	5.92
63	0.21	3.50	-33.12	-17.63	30.78	4.32	4.20	5.65
83	0.28	3.49	-43.20	-24.41	42.50	4.61	4.41	6.06
103	0.34	4.35	-52.33	-30.93	57.56	5.30	5.06	7.07
123	0.41	2.28	-61.86	-39.14	71.54	3.13	3.03	5.28
143	0.48	4.59	-68.75	-44.77	97.60	5.66	5.38	8.08
163	0.54	3.51	-76.02	-52.30	124.30	4.54	4.28	7.50
183	0.61	1.46	-82.49	-60.18	157.45	2.32	1.92	5.76
203	0.68	-1.72	-88.05	-68.16	201.01	-1.21	-1.68	2.83

the jackknife variance estimator had negligible biases when $\mathscr{K}$ was negligible but presented significant positive biases otherwise. On the other hand, the three bias-adjusted variance estimators tracked closely the estimator $\mathbb{V}_m$ and, as a result, exhibited negligible biases in low, moderate, and high-dimensional scenarios. We finally note that, even though our bias statements detailed in Result 4.4.1 were derived under the assumption of Bernoulli sampling, all three bias-adjusted estimators also had negligible biases in all scenarios of simple random sampling without replacement, as shown in Table 2.

4.7 Final remarks

In this article, we considered the problem of variance estimation for model-assisted and imputed estimators based on linear regression with a large number of covariates. We have shown that customary variance estimators such as those based on Taylor expansion, their g-weighted versions, or based on jackknife exhibit large biases when p/n is not negligible. The Taylor and its g-weighted version are biased negatively, and the jackknife is biased positively. We proved these biases do not vanish, even asymptotically, unless $\lim_{v\to\infty} p_v/n_v = 0$. For model-assisted variance estimators, we obtained closed-form, computable expressions for the biases of both Taylor, g-weights, and the jackknife variance estimators. These expressions can be used to define bias-adjusted variance estimators, which are unbiased independently of p/n; the simulations presented in Section 4.6 confirmed their good behavior empirically. Unfortunately, for imputed estimators, the biases depend on unknown quantities, and defining a bias-adjusted estimator requires an intermediate estimation step. This is beyond the scope of this article and is a topic currently under investigation. Similarly, unpublished simulation results suggest that the bootstrap variance estimator is also biased positively in high dimension; this is consistent with recent results obtained by El Karoui and Purdom (2018) and Zhao and Candes (2022). This is a topic of future research.

Appendices

4.8 Proof of Proposition 2.1.

The generalized jackknife variance estimator of $\widehat{\mu}_{greg}$ is defined in Berger and Skinner (2005) as

$$\widehat{V}_{jack}\left(\widehat{\mu}_{greg}\right) := \sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}z_k z_\ell, \tag{4.8.7}$$

where

$$z_k := (1-w_k)\left(\widehat{\mu}_{greg} - \widehat{\mu}_{greg}^{(k)}\right). \tag{4.8.8}$$

In (4.8.8), $\widehat{\mu}_{greg}^{(k)}$ denotes the GREG estimator as defined in (4.2.1), computed by deleting the element k of the sample (but not of the population). More generally, in what follows, we use the subscript (k) to denote any statistic (or, set) computed after deleting element k.

To derive a closed-form formula for (4.8.7), it suffices to find a closed-form formula for $\widehat{\mu}_{greg}^{(k)}$. In Duchesne (2000), a formula is given when the inclusion probabilities are reweighted after deletion of an element, a practice that we do not do with the generalized jackknife variance estimator. Below, we give a simple proof relying on the following lemma.

Lemma 2. *The following relation holds:*

$$\widehat{\boldsymbol{\beta}}^{(k)} = \widehat{\boldsymbol{\beta}} - \frac{\boldsymbol{A}_{\Pi S}^{-1}d_k\mathbf{x}_k\widehat{\epsilon}_{kS}}{1-\widetilde{h}_{kk}^{\pi}}.$$

Lemma 2 is a weighted extension of the well-known closed-form formula for leave-one-out linear regression. We let $\widehat{t}_{greg} := N\widehat{\mu}_{greg}$ be the usual total GREG estimator. We begin by noting that if the full-sample GREG estimator can be written in projection form, then the leave-one-out version of the estimator can too. It follows that

$$\begin{aligned}\widehat{t}_{greg}^{(k)} &= \boldsymbol{t}_x^\top\widehat{\boldsymbol{\beta}}^{(k)}\\ &= \boldsymbol{t}_x^\top\left(\widehat{\boldsymbol{\beta}} - \frac{\boldsymbol{A}_{\Pi S}^{-1}d_k\mathbf{x}_k\widehat{\epsilon}_{kS}}{1-\widetilde{h}_{kk}^{\pi}}\right)\\ &= \widehat{t}_{greg} - \frac{d_k g_k\widehat{\epsilon}_{ks}}{1-\widetilde{h}_{kk}^{\pi}}.\end{aligned}$$

Now,

$$\widehat{\mu}_{greg}^{(k)} := \frac{\widehat{t}_{greg}^{(k)}}{N} = \widehat{\mu}_{greg} - \frac{d_k g_k\widehat{\epsilon}_{ks}}{N\left(1-\widetilde{h}_{kk}^{\pi}\right)},$$

from which it follows that

$$\widehat{\mu}_{greg}^{(k)} - \widehat{\mu}_{greg} = -\frac{1}{N}\frac{d_k g_k\widehat{\epsilon}_{ks}}{1-\widetilde{h}_{kk}^{\pi}}.$$

Finally,

$$z_k = \frac{1}{N}\frac{(1-w_k)\,g_k\widehat{\epsilon}_{ks}}{\pi_k\left(1-\widetilde{h}_{kk}^{\pi}\right)},$$

which concludes the proof.

4.9 Proof of Result 2.1.

The effect of deleting a respondent or nonrespondent is different; we treat these cases separately. For $k \in S_m$, write

$$\begin{aligned}\widehat{\mu}_{lr}^{(k)} &= \frac{1}{\widehat{N}-d_k}\left(\sum_{\ell\in S_r}\frac{y_\ell}{\pi_\ell}+\sum_{j\in S_r^{(k)}}\frac{\mathbf{x}_\ell^\top\widehat{\boldsymbol{\beta}}_R}{\pi_\ell}\right)\\ &= \frac{1}{\widehat{N}-d_k}\left(\sum_{\ell\in S_r}\frac{y_\ell}{\pi_\ell}+\sum_{\ell\in S_m}\frac{\mathbf{x}_\ell^\top\widehat{\boldsymbol{\beta}}_R}{\pi_\ell}-\frac{\mathbf{x}_k^\top\widehat{\boldsymbol{\beta}}_R}{\pi_k}\right)\\ &= \frac{\widehat{N}}{\widehat{N}-d_k}\left(\widehat{\mu}_{lr}-\frac{\mathbf{x}_k^\top\widehat{\boldsymbol{\beta}}_R}{\widehat{N}\pi_k}\right).\end{aligned}$$

Similarly, for $k \in S_r$, we have

$$\begin{aligned}\widehat{\mu}_{lr}^{(k)} &= \frac{1}{\widehat{N}-d_k}\left(\sum_{\ell\in S_r^{(k)}}\frac{y_\ell}{\pi_\ell}+\sum_{\ell\in S_m}\frac{\mathbf{x}_\ell^\top\widehat{\boldsymbol{\beta}}_R^{(k)}}{\pi_\ell}\right)\\ &= \frac{1}{\widehat{N}-d_k}\left\{\sum_{\ell\in S_r}\frac{y_\ell}{\pi_\ell}-\frac{y_k}{\pi_k}+\sum_{\ell\in S_m}\frac{\mathbf{x}_\ell^\top}{\pi_\ell}\left(\widehat{\boldsymbol{\beta}}_R-\frac{\mathbf{A}_{\Pi R}^{-1}\mathbf{x}_k\widehat{\epsilon}_{Rk}}{\pi_k(1-\widehat{h}_{kk}^{\pi})}\right)\right\}\\ &= \frac{\widehat{N}}{\widehat{N}-d_k}\left[\widehat{\mu}_{lr}-\frac{1}{\widehat{N}\pi_k}\left\{y_k+\sum_{\ell\in S_m}\frac{\mathbf{x}_\ell^\top}{\pi_\ell}\left(\frac{d_k\mathbf{A}_{\Pi R}^{-1}\mathbf{x}_k\widehat{\epsilon}_{Rk}}{1-\widehat{h}_{kk}^{\pi}}\right)\right\}\right]\end{aligned}$$

Thus, introducing response indicators, we obtain for an arbitrary element $k \in S$,

$$\widehat{\mu}_{lr}^{(k)} = \frac{\widehat{N}}{\widehat{N}-d_k}\left(\widehat{\mu}_{lr}-\frac{1}{\widehat{N}\pi_k}\widehat{\xi}_k^{(jack)}\right)$$

from which it follows that

$$\widehat{\mu}_{lr}^{(k)}-\widehat{\mu}_{lr} = \frac{d_k}{\widehat{N}-d_k}\left(\widehat{\mu}_{lr}-\widehat{\xi}_k^{(jack)}\right).$$

4.10 Proof of Corollary 2.1.

The Generalized Jackknife variance estimator proposed in Berger and Rao (2006) is

$$\widehat{V}_{jack}(\widehat{\mu}_{lr}) := \sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}e_k e_\ell, \tag{4.10.9}$$

where

$$e_k := (1-w_k)\left(\widehat{\mu}_{lr}-\widehat{\mu}_{lr}^{(k)}\right).$$

Using Result 4.2.1, a closed-form formula of e_k is given by

$$e_k = \frac{d_k}{\widehat{N}}\left(\widehat{\mu}_{lr} - \widehat{\xi}_k^{(jack)}\right).$$

Replacing e_k by its closed formula in (4.10.9) leads to the result.

4.11 Proof of Result 4.1

Statement i)

Consider the following variance decomposition:

$$\mathbb{V}\left(\widehat{\mu}_{greg}\right) = \mathbb{E}_p\left[\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)\right] + \mathbb{V}_p\left(\mathbb{E}_m\left[\widehat{\mu}_{greg}\right]\right).$$

Note that,

$$\mathbb{E}_m\left[\widehat{\mu}_{greg}\right] = \mathbb{E}_m\left[\frac{1}{N}\sum_{k\in U}\mathbf{x}_k^\top\widehat{\boldsymbol{\beta}}_S\right] = \frac{1}{N}\sum_{k\in U}\mathbf{x}_k^\top\boldsymbol{\beta}.$$

In particular, observe that this quantity is independent of S, so that

$$\mathbb{V}\left(\widehat{\mu}_{greg}\right) = \mathbb{E}_p\left[\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)\right],$$

when the model is correctly specified. This establishes unbiasedness. For the asymptotic equivalence, we show that

$$\lim_{\nu\to\infty} n_\nu^2 \times \mathbb{E}\left[\left\{\mathbb{V}\left(\widehat{\mu}_{greg}\right) - \mathbb{V}_m\left(\widehat{\mu}_{greg}\right)\right\}^2\right] = 0.$$

We begin with the following decomposition:

$$\mathbb{V}\left(\widehat{\mu}_{greg}\right) - \mathbb{V}_m\left(\widehat{\mu}_{greg}\right) = \frac{\sigma^2}{\pi N_\nu^2}\sum_{k\in U_\nu}\left(g_{k,\nu} - \mathbb{E}_p\left[g_{k,\nu}\right]\right).$$

It thus follows that

$$\begin{aligned} n_\nu^2 \times \mathbb{E}_p\left[\left\{\mathbb{V}\left(\widehat{\mu}_{greg}\right) - \mathbb{V}_m\left(\widehat{\mu}_{greg}\right)\right\}^2\right] &= \frac{n_\nu^2\sigma^4}{\pi^2 N_\nu^4}\mathbb{E}_p\left[\left\{\sum_{k\in U_\nu}\left(g_{k,\nu} - \mathbb{E}_p\left[g_{k,\nu}\right]\right)\right\}^2\right] \\ &\leqslant \frac{n_\nu^2\sigma^4}{\pi^2 N_\nu^3}\sum_{k\in U_\nu}\mathbb{E}_p\left[\left(g_{k,\nu} - \mathbb{E}_p\left[g_{k,\nu}\right]\right)^2\right]. \end{aligned}$$

By symmetry, we obtain

$$n_\nu^2 \times \mathbb{E}_p\left[\left\{\mathbb{V}\left(\widehat{\mu}_{greg}\right) - \mathbb{V}_m\left(\widehat{\mu}_{greg}\right)\right\}^2\right] \leqslant \frac{n_\nu^2\sigma^4}{\pi^2 N_\nu^2}\mathbb{V}_p\left(g_{1,\nu}\right) = o(1),$$

by assumption.

Statement ii)

Taylor: Write $\widehat{\mu}_{greg}$ as $\widehat{\mu}_{greg} = \boldsymbol{\mu}_x^\top \widehat{\boldsymbol{\beta}}_S$ with $\boldsymbol{\mu}_x := \boldsymbol{t}_x/N$. It follows that

$$\mathbb{V}_m\left(\widehat{\mu}_{greg}\right) = \mathbb{V}_m\left(\boldsymbol{\mu}_x^\top \widehat{\boldsymbol{\beta}}_S\right) = \boldsymbol{\mu}_x^\top \mathbb{V}_m\left(\widehat{\boldsymbol{\beta}}_S\right)\boldsymbol{\mu}_x = \frac{\sigma^2}{\pi}\boldsymbol{\mu}_x^\top \boldsymbol{A}_{\Pi S}^{-1}\boldsymbol{\mu}_x = \frac{\sigma^2}{\pi N^2}\sum_{k\in U} g_k.$$

Under a Bernoulli sampling design, the Taylor variance estimator reduces to

$$\widehat{V}_{tay}\left(\widehat{\mu}_{greg}\right) = \frac{1}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\widehat{\epsilon}_{kS}^2 + \frac{\widehat{\sigma}^2}{N}.$$

Thus,

$$\begin{aligned}
\mathbb{E}_m\left[\widehat{V}_{tay}\left(\widehat{\mu}_{greg}\right)\right] &= \frac{1}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\sigma^2\left(1-\widetilde{h}_{kk}\right) + \frac{\sigma^2}{N} \\
&= \frac{\sigma^2}{N^2}\frac{1-\pi}{\pi^2}\left(n_s - p\right) + \frac{\sigma^2}{N} \\
&= \frac{\sigma^2 n_s}{N^2}\frac{1-\pi}{\pi^2}\left(1-\mathscr{K}\right) + \frac{\sigma^2}{N}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\frac{\mathbb{E}_m\left[\widehat{V}_{tay}\left(\widehat{\mu}_{greg}\right)\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} &= \frac{n_s\left(1-\pi\right)\pi^{-1}\left(1-\mathscr{K}\right)}{\sum_{k\in U} g_k} + \frac{\pi N}{\sum_{k\in U} g_k} \\
&= \frac{n_s}{n}\times\left(1-\pi\right)\left(1-\mathscr{K}\right)\times\frac{N}{\sum_{k\in U} g_k} + \frac{\pi N}{\sum_{k\in U} g_k} \\
&= \frac{N}{\sum_{k\in U} g_k}\frac{n_s}{n}\left\{\left(1-\pi\right)\left(1-\mathscr{K}\right) + \frac{n}{n_s}\pi\right\} \\
&= \frac{N}{\sum_{k\in U} g_k}\frac{n_s}{n}\left\{1 + \pi_v\left(\frac{n}{n_s}-1\right) - \mathscr{K}\left(1-\pi\right)\right\}.
\end{aligned}$$

G-weighted: We begin by noting the following fact:

$$\sum_{k\in S} g_k^2 = N^2\pi^2\sigma^{-2}\mathbb{V}_m\left(\widehat{\mu}_{greg}\right). \tag{4.11.10}$$

To prove (4.11.10), write

$$\begin{aligned}
\sum_{k\in S} g_k^2 &= \sum_{k\in S} \boldsymbol{t}_{\mathbf{x}}\boldsymbol{A}_{\Pi S}^{-1}\mathbf{x}_k\mathbf{x}_k\boldsymbol{A}_{\Pi S}^{-1}\boldsymbol{t}_{\mathbf{x}} \\
&= \pi\times\boldsymbol{t}_{\mathbf{x}}\boldsymbol{A}_{\Pi S}^{-1}\sum_{k\in S}\frac{\mathbf{x}_k\mathbf{x}_k}{\pi}\boldsymbol{A}_{\Pi S}^{-1}\boldsymbol{t}_{\mathbf{x}} \\
&= \pi\times\boldsymbol{t}_{\mathbf{x}}\boldsymbol{A}_{\Pi S}^{-1}\boldsymbol{t}_{\mathbf{x}} \\
&= N^2\pi^2\sigma^{-2}\times\frac{\sigma^2}{N^2\pi}\boldsymbol{t}_{\mathbf{x}}\boldsymbol{A}_{\Pi S}^{-1}\boldsymbol{t}_{\mathbf{x}} \\
&= N^2\pi^2\sigma^{-2}\mathbb{V}_m\left(\widehat{\mu}_{greg}\right).
\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}_m\left[\widehat{V}_g\left(\widehat{\mu}_{greg}\right)\right] &= \frac{1}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\sigma^2\left(1-\widetilde{h}_{kk}\right)g_k^2+\frac{\sigma^2}{N}\\ &\simeq \frac{1-\mathscr{K}}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\sigma^2 g_k^2+\frac{\sigma^2}{N}\\ &= (1-\mathscr{K})(1-\pi)\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)+\frac{\sigma^2}{N}.\end{aligned}$$

Thus,

$$\begin{aligned}\frac{\mathbb{E}_m\left[\widehat{V}_g\left(\widehat{\mu}_{greg}\right)\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)} &\simeq \frac{(1-\mathscr{K})(1-\pi)\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)}+\frac{\pi N}{\sum_{k\in U}g_k}\\ &= (1-\mathscr{K})(1-\pi)+\frac{\pi N}{\sum_{k\in U}g_k}.\end{aligned}$$

Jackknife: In Bernoulli sampling, the Jackknife variance estimator reduces to

$$\widehat{V}_{jack}\left(\widehat{\mu}_{greg}\right)-\frac{\widehat{\sigma}^2}{N}=\frac{(n-1)^2}{n^2N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\frac{g_k^2\widehat{\epsilon}_{ks}^2}{\left(1-\widetilde{h}_{kk}\right)^2}\simeq\frac{1}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\frac{g_k^2\,\widehat{\epsilon}_{ks}^2}{\left(1-\widetilde{h}_{kk}\right)^2},$$

so that

$$\mathbb{E}_m\left[\widehat{V}_{jack}\left(\widehat{\mu}_{greg}\right)\right]-\frac{\sigma^2}{N}=\frac{(n-1)^2}{n^2N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\frac{g_k^2\widehat{\epsilon}_{ks}^2}{\left(1-\widetilde{h}_{kk}\right)^2}\simeq\frac{\sigma^2}{N^2}\frac{1-\pi}{\pi^2}\sum_{k\in S}\frac{g_k^2}{\left(1-\widetilde{h}_{kk}\right)}.$$

Using (4.11.10), we get

$$\mathbb{E}_m\left[\widehat{V}_{jack}\left(\widehat{\mu}_{greg}\right)\right]\simeq\frac{1-\pi}{1-\mathscr{K}}\mathbb{V}_m\left\{\widehat{V}_{jack}\left(\widehat{\mu}_{greg}\right)\right\}+\frac{\sigma^2}{N}.$$

It follows that

$$\frac{\mathbb{E}_m\left[\widehat{V}_{jack}\left(\widehat{\mu}_{greg}\right)\right]}{\mathbb{V}_m\left(\widehat{\mu}_{greg}\right)}\simeq\frac{1-\pi}{1-\mathscr{K}}+\frac{\pi N}{\sum_{k\in U}g_k}.$$

4.12 Proof of Result 5.1.

We first obtain an expression of the variance of $\widehat{\mu}_{lr}$ using the method of Särndal (1992). The total variance of $\widehat{\mu}_{lr}$ can be decomposed as

$$\mathbb{V}(\widehat{\mu}_{lr})=\mathbb{V}_{sam}(\widehat{\mu}_{lr})+\mathbb{V}_{nr}(\widehat{\mu}_{lr})+\mathbb{V}_{mix}(\widehat{\mu}_{lr}),$$

where

$$\mathbb{V}_{sam}(\widehat{\mu}_{lr})=\mathrm{E}_m\mathbb{V}_p(\widehat{\mu}_H),$$

$$\mathbb{V}_{nr}(\widehat{\mu}_{lr})=\mathrm{E}_q\mathrm{E}_p\mathbb{V}_m(\widehat{\mu}_{lr}-\widehat{\mu}_H)$$

and

$$\mathbb{V}_{\mathrm{mix}}(\widehat{\mu}_{lr})=2\mathrm{E}_p\mathrm{E}_q\mathrm{Cov}_m\left\{\widehat{\mu}_H-\mu_y,\widehat{\mu}_{lr}-\widehat{\mu}_H\right\}$$

where $\widehat{\mu}_H = \widehat{N}^{-1}\sum_{k\in S}\pi_k^{-1}y_k$. In the case of linear regression imputation and Bernoulli sampling, it can be shown that $\mathbb{V}_{mix}(\widehat{\mu}_{lr}) = 0$. Now, using a first-order Taylor expansion, a full sample estimator of $\mathbb{V}_{sam}(\widehat{\mu}_{lr})$ is given by

$$\widehat{V}_{sam}(\widehat{\mu}_{lr}) = \frac{1-\pi}{n_s^2}\sum_{k\in S}(y_k - \widehat{\mu}_H)^2.$$

After some straightforward algebra, we obtain

$$\begin{aligned}
\mathrm{E}_m\left[\widehat{V}_{sam}(\widehat{\mu}_{lr})\right] &= \frac{1-\pi}{n_s^2}\sum_{k\in S}\mathrm{E}_m\left[(y_k-\widehat{\mu}_H)^2\right] \\
&= \frac{1-\pi}{n_s^2}\left\{\sum_{k\in S}(\mathbf{x}_k^\top\boldsymbol{\beta})^2 + n_s\sigma^2 - \frac{2}{n}\sum_{k\in S}\sum_{\ell\in S}\mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta} - 2\sigma^2\right. \\
&\left. + \frac{1}{n}\sum_{k\in S}\sum_{\ell\in S}\mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta} + \sigma^2\right\} \\
&= \frac{1-\pi}{n_s^2}\sigma^2\left\{\frac{1}{\sigma^2}\sum_{k\in S}B_k + n_s - 1\right\},
\end{aligned}$$

where

$$B_k = \frac{1}{n_s}\sum_{\ell\in S}\left\{(\mathbf{x}_\ell^\top\boldsymbol{\beta})^2 - \mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta}\right\}.$$

We now turn to the nonresponse component. It can be shown that

$$\mathbb{V}_{nr} = \mathrm{E}_q\mathrm{E}_p\left[\frac{\sigma^2}{(\widehat{N}\pi)^2}\sum_{k\in S}\left\{R_k(1+\widehat{\Gamma}_k)-1\right\}^2\right].$$

Assuming σ^2 is known, the above quantity can be estimated by

$$\widehat{V}_{nr} = \frac{\sigma^2}{n_s^2}\widehat{A}_n,$$

where

$$\widehat{A}_n = \sum_{k\in S_r}(1+\widehat{\Gamma}_k)^2 - n_s.$$

Adding (4.12) and (4.12), we obtain the estimator of $\mathrm{E}_m[\mathbb{V}(\widehat{\mu}_{lr})]$

$$\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{lr})\right] = \mathrm{E}_m\left(\widehat{V}_{sam} + \widehat{V}_{nr}\right) = \frac{1-\pi}{n_s^2}\sigma^2\left\{\frac{1}{\sigma^2}\sum_{k\in S}B_k + (n_s-1) - \frac{1}{1-\pi}\widehat{A}_n\right\}.$$

Next, we establish part (i) of Result (4.5.1). We consider the variance decomposition of Shao and Steel (1999)

$$\mathbb{V}(\widehat{\mu}_{lr}) = \mathrm{E}_q\mathrm{E}_m\mathbb{V}_p(\widehat{\mu}_{lr}) + \mathrm{E}_q\mathbb{V}_m\mathrm{E}_p(\widehat{\mu}_{lr}). \tag{4.12.11}$$

Using a first-order Taylor expansion, it can be shown that an estimator of the first term of the variance decomposition (4.12.11) is given by

$$\widehat{V}_1(\widehat{\mu}_{lr}) = \frac{1-\pi}{n_s^2}\sum_{k\in S}\left(\widehat{\xi}_k - \widehat{\mu}_{lr}\right)^2.$$

After some tedious but somewhat straightforward algebra, we can show that

$$\mathrm{E}_m\left[\widehat{V}_1\right] = \frac{1-\pi}{n_s^2}\left[\sum_{k\in S} B_k + \sigma^2\left\{\sum_{k\in S}\widehat{h}_{kk} + \sum_{k\in S_r}(1+\widehat{\Gamma}_k)^2(1-\widehat{h}_{kk}) - 1 - \frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k\right\}\right],$$

where

$$B_k = \frac{1}{n_s}\sum_{\ell\in S}\left\{(\mathbf{x}_\ell^\top\boldsymbol{\beta})^2 - \mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta}\right\}.$$

Using a first-order Taylor expansion, it can be shown that an estimator of the second term of the decomposition (4.12.11) is given by

$$\widehat{V}_2(\widehat{\mu}_{lr}) = \frac{\pi\sigma^2}{n_s^2}\widehat{A}_n.$$

An estimator of the total variance $\mathbb{V}(\widehat{\mu}_{lr})$ based on a first-order Taylor expansion is obtained by summing (4.12) and (4.12), which leads to

$$\widehat{V}_{tay}(\widehat{\mu}_{lr}) = \widehat{V}_1(\widehat{\mu}_{lr}) + \widehat{V}_2(\widehat{\mu}_{lr}).$$

It follows that

$$\begin{aligned}\mathrm{E}_m\left[\widehat{V}_{tay}(\widehat{\mu}_{lr})\right] = \frac{1-\pi}{n_s^2}\sigma^2\Bigg\{&\frac{1}{\sigma^2}\sum_{k\in S}B_k + \sum_{k\in S}\widehat{h}_{kk} + \sum_{k\in S_r}(1+\widehat{\Gamma}_k)^2(1-\widehat{h}_{kk})\\ &-1-\frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k + \frac{\pi}{1-\pi}\widehat{A}_n\Bigg\}.\end{aligned}$$

It follows that

$$\frac{\mathrm{E}_m\left[\widehat{V}_{tay}(\widehat{\mu}_{lr})\right]}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{lr})\right]} - 1 \approx \frac{\sum_{k\in S}\widehat{h}_{kk} + \sum_{k\in S_r}(1+\widehat{\Gamma}_k)^2(1-\widehat{h}_{kk}) - n_s - \frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k - \widehat{A}_n}{\frac{1}{\sigma^2}\sum_{k\in S}B_k + n_s - 1 + \frac{1}{1-\pi}\widehat{A}_n}.$$

The simplification $\widehat{h}_{kk} \approx \mathscr{K} = p/\mathrm{E}[n_r]$ gives

$$\begin{aligned}\frac{\mathrm{E}_m\left[\widehat{V}_{tay}(\widehat{\mu}_{lr})\right]}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{lr})\right]} &\approx 1 + \frac{(1-\mathscr{K})\left\{\sum_{k\in S_r}(1+\widehat{\Gamma}_k)^2 - n_s\right\} - \frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k - \widehat{A}_n}{\frac{1}{\sigma^2}\sum_{k\in S}B_k + n_s - 1 + \frac{1}{1-\pi}\widehat{A}_n}\\ &= 1 + \frac{-\mathscr{K}\widehat{A}_n - \frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k}{\frac{1}{\sigma^2}\sum_{k\in S}B_k + n_s - 1 + \frac{1}{1-\pi}\widehat{A}_n}.\end{aligned}$$

Jackknife: For a Bernoulli sampling, the expression of the generalized jackknife variance estimator of Berger and Rao (2006) is

$$\widehat{V}_{jack}(\widehat{\mu}_{lr}) = \frac{1-\pi}{n_s^2}\sum_{k\in S}\left(\widehat{\xi}_k^{(jack)} - \widehat{\mu}_{lr}\right)^2 + \frac{\pi\sigma^2}{n_s^2}\widehat{A}_n.$$

It follows that

$$\mathrm{E}_m\left[\widehat{V}_{jack}(\widehat{\mu}_{lr})\right] - \frac{\pi\sigma^2}{n_s^2}\widehat{A}_n = \frac{1-\pi}{n_s^2}\sum_{k\in S}\mathrm{E}_m\left[\left(\widehat{\xi}_k^{(jack)} - \widehat{\mu}_{lr}\right)^2\right].$$

By linearity,

$$\mathrm{E}_m\left[\left(\widehat{\xi}_k^{(jack)}-\widehat{\mu}_{lr}\right)^2\right]=\mathrm{E}_m\left[\left(\widehat{\xi}_k^{(jack)}\right)^2\right]-2\mathrm{E}_m\left[\widehat{\xi}_k^{(jack)}\widehat{\mu}_{lr}\right]+\mathrm{E}_m\left[\widehat{\mu}_{lr}^2\right],$$

so we may treat these terms separately. We have

$$\mathrm{E}_m\left[\left(\widehat{\xi}_k^{(jack)}\right)^2\right]=(\mathbf{x}_k^\top\boldsymbol{\beta})^2+\sigma^2\widehat{h}_{kk}+\sigma^2R_k(1+\widehat{\Gamma}_k)^2\frac{1}{1-\widehat{h}_{kk}}.$$

Based on the results of the previous part, we have

$$\mathrm{E}_m\left[\widehat{\xi}_k^{(jack)}\widehat{\mu}_{lr}\right]=\frac{1}{n_s}\sum_{\ell\in S}\mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta}+\frac{\sigma^2}{n_s}(1+\widehat{\Gamma}_k)$$

and

$$\mathrm{E}_m\left[\widehat{\mu}_{lr}^2\right]=\frac{1}{n_s^2}\sum_{\ell\in S}\sum_{k\in S}\mathbf{x}_\ell^\top\boldsymbol{\beta}\mathbf{x}_k^\top\boldsymbol{\beta}+\frac{\sigma^2}{n_s}\left(1+\frac{1}{n_s}\sum_{\ell\in S}\widehat{\Gamma}_\ell\right).$$

Then, finally we obtain

$$\begin{aligned}\sum_{k\in S}\mathrm{E}_m\left[\left(\widehat{\xi}_k^{(jack)}-\widehat{\mu}_{lr}\right)^2\right]&=\sum_{k\in S}\left\{(\mathbf{x}_k^\top\boldsymbol{\beta})^2+\sigma^2\widehat{h}_{kk}+\sigma^2R_k(1+\widehat{\Gamma}_k)^2\frac{1}{1-\widehat{h}_{kk}}\right.\\&\left.-\frac{2}{n_s}\sum_{\ell\in S}\mathbf{x}_k^\top\boldsymbol{\beta}\mathbf{x}_\ell^\top\boldsymbol{\beta}-\frac{2\sigma^2}{n_s}(1+\widehat{\Gamma}_k)+\frac{1}{n_s^2}\sum_{\ell\in S}\sum_{k\in S}\mathbf{x}_\ell^\top\boldsymbol{\beta}\mathbf{x}_k^\top\boldsymbol{\beta}+\frac{\sigma^2}{n_s}\left(1+\frac{1}{n_s}\sum_{\ell\in S}\widehat{\Gamma}_\ell\right)\right\}\\&=\sigma^2\left\{\frac{1}{\sigma^2}\sum_{k\in S}B_k+\sum_{k\in S}\widehat{h}_{kk}+\sum_{k\in S_r}(1+\widehat{\Gamma}_k)^2\frac{1}{1-\widehat{h}_{kk}}-1-\frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k\right\}.\end{aligned}$$

$$\begin{aligned}\mathrm{E}_m\left[\widehat{V}_{jack}(\widehat{\mu}_{lr})\right]=&\frac{1-\pi}{n_s^2}\sigma^2\left\{\frac{1}{\sigma^2}\sum_{k\in S}B_k+\sum_{k\in S}\widehat{h}_{kk}+\sum_{k\in S_r}(1+\widehat{\Gamma}_k)^2\frac{1}{1-\widehat{h}_{kk}}-1-\frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k\right.\\&\left.+\frac{\pi}{1-\pi}\widehat{A}_n\right\}\end{aligned}$$

It follows that

$$\frac{\mathrm{E}_m\left[\widehat{V}_{jack}(\widehat{\mu}_{lr})\right]}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{lr,v})\right]}=1+\frac{\sum_{k\in S}\widehat{h}_{kk}+\sum_{k\in S_r}(1+\widehat{\Gamma}_k)^2\frac{1}{1-\widehat{h}_{kk}}-n_s-\frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k-\widehat{A}_n}{\frac{1}{\sigma^2}\sum_{k\in S}B_k+n_s-1+\frac{1}{1-\pi}\widehat{A}_n}$$

The simplification $\widehat{h}_{kk}\approx\mathscr{K}=p/\mathrm{E}[n_r]$ gives

$$\begin{aligned}\frac{\mathrm{E}_m\left[\widehat{V}_{jack}(\widehat{\mu}_{lr})\right]}{\mathrm{E}_m\left[\widehat{V}(\widehat{\mu}_{lr,v})\right]}&\approx 1+\frac{\frac{1}{1-\mathscr{K}}\sum_{k\in S_r}(1+\widehat{\Gamma}_k)^2-n_s+n_s\mathscr{K}-\frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k-\widehat{A}_n}{\frac{1}{\sigma^2}\sum_{k\in S}B_k+n_s-1+\frac{1}{1-\pi}\widehat{A}_n}\\&=1+\frac{\frac{\mathscr{K}}{1-\mathscr{K}}\left(\widehat{A}_n+2n_s-n_sK\right)-\frac{1}{n_s}\sum_{k\in S}\widehat{\Gamma}_k}{\frac{1}{\sigma^2}\sum_{k\in S}B_k+n_s-1+\frac{1}{1-\pi}\widehat{A}_n}.\end{aligned}$$

Chapter 5

Spatiotemporal Sampling With Spatial Spreading and Rotation of Units in Time

Abstract: When the sampled population belongs to a metric space, the selection of neighboring units will imply often similarities in the collected data due to their geographical proximity. In order to estimate parameters such as means or totals, it is therefore more efficient to select samples that are well distributed in space. Often, the interest lies not only in estimating a parameter at one point in time, but rather in estimating it at several points and studying its evolution. Because of the temporal autocorrelation of successive values from the same unit, a system of temporal rotation of the units in the samples must be provided. In other words, this type of problem forces us to consider two types of autocorrelation: spatial and temporal. In this article, we propose two new spatiotemporal sampling methods for equal or unequal inclusion probabilities. Systematic sampling is used to promote a rotation of the selection of the same unit over time, and thus address temporal spread. Both methods select samples that are well distributed in space at each sampling time. They differ by the fact that these samples are of random size for the first one, while for the second one, more complex, their sizes are controlled. Thus, the first method is called spatiotemporal sampling with random sample sizes (SPAR) and the second, spatiotemporal sampling with fixed sample sizes (SPAF). Simulations show that our methods outperform and generalize existing methods.

Keywords: balanced sampling, Cube method, Pivotal method, Spatial balance, Systematic sampling.

This chapter is a reprint of article Eustache, Jauslin, and Tillé (2022).

5.1 Introduction

Sampling is almost always done to estimate unknown population parameters, for instance a total. When spatial data are considered, information from two neighboring units are generally very similar. In this case, the selection of close units thus provides less information than the observation of spatially well-distributed units, and a less efficient estimator.

In addition, many applications require not only the selection of a spread sample at a given time, but also a rotation system of the selected units over time. For example, in some environmental monitoring such as that described in Tillé and Ecker (2013), a different part of the population is visited each year. Similarly, in the new census techniques applied in France and Italy, rotation groups of small municipalities are formed and one is selected each year. In these applications, if each annual sample is well spread, the gain in accuracy is likely to be significant. Therefore, the samples must remain spatially spread at each sampling time.

A large number of methods have been proposed to select spread samples. In one dimension, Quenouille (1949b) has shown that systematic sampling is the optimal design for obtaining the most spread sample with equal inclusion probabilities. A first family of methods consists in transforming a multi-dimensional problem into a one-dimensional one in order to apply a systematic design. This is the case of the stratified sampling method by generalized random tessellation developed by Stevens Jr. and Olsen (1999), Stevens and Olsen (2003), and Stevens and Olsen (2004), who use a quadrant-recursive partition of the unit square to map to a one-dimensional problem. Dickson and Tillé (2016) have used the travelling salesman problem to reduce the sampling problem to one dimension.

Another family of methods introduces a repulsion in the selection of neighboring units. Grafström (2011) has proposed the method called "Spatially correlated Poisson sampling" which generates a strong negative correlation between inclusion probabilities of close units using sampling weights. Grafström, Lundström, and Schelin (2012) generalize the pivotal method proposed by Deville and Tillé (1998) to the selection of spread samples. This method has been modified by Grafström and Tillé (2013) to obtain samples that are both spread and balanced on totals of known auxiliary variables. Another generalization enables its application to continuous populations (Grafström and Matei, 2018). Grafström and Lundström (2013) recommend the use of spatially balanced samples on variables that are not geographic coordinates. Because of their similarities, groups of neighboring units can be seen as strata. Thus, the selection of spread samples can be compared to a multidimensional stratification, as in the method proposed by Jauslin and Tillé (2020).

Information from a unit collected at close sampling times will probably be similar. Zhao and Grafström (2020) propose a method of spatiotemporal sampling to improve estimators of change by selecting positively coordinated samples, i.e. by maximizing their overlap. If the goal is not to estimate the evolution of a parameter, it is preferable to do the opposite: select samples with negative coordination. Indeed, selecting samples that overlap as little as possible reduces redundant information. An appropriate rotation of the units must then be planned. This problem is complex when inclusion probabilities are unequal. Several solutions have been proposed in Deville and Tillé (2000) or Rivest and Ebouele (2020) but these methods do not take into account spatial autocorrelations. The selection of several spread samples from the same population over time becomes much more complex. Some solutions have already been proposed. Khavarzadeh, Mohammadzadeh, and Mateu (2018) divide the space into primary units that are chosen with a balanced design and the units are selected to maximize the spread. Wang and Zhu (2019) propose a spatiotemporal sampling method based on a consecutive application of the local pivotal method. However, it cannot be applied when inclusion probabilities vary over time and does not allow to select a unit more than once over time.

In this paper, a set of new solutions for spatiotemporal sampling generalizing the method of Wang and Zhu (2019) is proposed. These methods can be applied to any temporal matrix of equal or unequal inclusion probabilities. If the inclusion probabilities allow the same unit to be selected several times, an appropriate rotation of the selected units over time is provided. The two proposed methods select spatiotemporal samples that are temporally and spatially spread. In other words, temporal spread means that a unit should not be selected at close sampling times, and spatial spread means that at each sampling time, the selected sample will be well distributed in space. The methods differ in that one produces random size samples at each sampling time while the other, much more complex, produces fixed size samples. Their effectiveness was demonstrated by a set of simulations on spatial biological data using the R package `SpotSampling` from Eustache, Jauslin, and Tillé (2020). This package enables to apply the methods proposed in this paper.

5.2 Spreading in the context of spatial statistic trinity

In sampling theory, Hájek (1981) defines a pairwise strategy consisting of a design and an estimator. In survey sampling, we can also distinguish between a design-based approach and a model-based approach, depending on whether the inference is conducted according to the model or according to the design that generates the population (see among others Valliant, Dorfman, and Royall, 2000, for the model-based approach, and Tillé, 2020; Lohr, 2021, for the design-based approach). Wang, Gao, and Stein (2020) define the trinity of spatial statistics as the triplet composed of a population, a sampling design and an estimator. In this section, we will justify the interest of our method in this context.

Consider a finite population U of units denoted by $k \in \{1, \ldots, N\}$. A sample $s \subset U$ is selected by means of a sampling design $p(.)$ such that

$$p(s) \geq 0, \text{ for all } s \in U \text{ and } \sum_{s \subset U} p(s) = 1.$$

A variable a_k has a Bernoulli distribution and takes the value 1 if unit k is in the sample and 0 otherwise. The first and second order inclusion probability are respectively

$$\pi_k = \sum_{s \ni k} p(s) = \mathrm{E}_p(a_k) \text{ and } \pi_{k\ell} = \sum_{s \ni k,\ell} p(s) = \mathrm{E}_p(a_k a_\ell), \text{ for all } k, \ell \in U.$$

where $\mathrm{E}_p(.)$ is the expectation under the sampling design. Moreover, define $\Delta_{k\ell} = \mathrm{cov}_p(a_k, a_\ell) = \pi_{k\ell} - \pi_k \pi_\ell$ as the covariance under the sampling design between a_k and a_ℓ, with $\mathrm{cov}_p(.,.)$ the covariance under the sampling design. In order to estimate a total

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

of a variable of interest y_k, $k \in U$, the Horvitz-Thompson estimator (Horvitz and Thompson, 1952)

$$\widehat{t}_y = \sum_{k \in s} \frac{y_k}{\pi_k}$$

gives the simplest unbiased estimator provided that all first order inclusion probabilities are not null.

Furthermore, suppose that the population is governed by a model M. As in Grafström and Tillé (2013), we can consider the general linear model, with possible heteroscedasticity and autocorrelation,

$$y_k = \boldsymbol{x}_k^\top \beta + \varepsilon_k, \text{ for all } k \in U, \tag{5.2.1}$$

where $\boldsymbol{x}_k$ is a column vector of the values taken by p auxiliary variables on unit k and $\beta \in \mathbb{R}^p$ is the vector of regression coefficients. Moreover, the ε_k is a random variable such that $\mathrm{E}_M(\varepsilon_k) = 0$, $\mathrm{var}_M(\varepsilon_k) = \sigma_k^2$, for all $k \in U$, and

$$\mathrm{cov}_M(\varepsilon_k, \varepsilon_\ell) = \sigma_k \sigma_\ell \rho_{k\ell}, \text{ with } k \neq \ell \in U,$$

where $\mathrm{E}_M(.)$, $\mathrm{var}_M(.)$ and $\mathrm{cov}_M(.,.)$ respectively denote the expectation, variance and covariance under model M. The spatial heterogeneity studied in Wang, Zhang, and Fu (2016) is a particular case of model (5.2.1) when vector $\boldsymbol{x}_k$ contains the indicator variables of the strata and the σ_k^2 are equal within a stratum but can be unequal from one stratum to another. Usually the closer the units are, the more correlated they are. The $\rho_{k\ell}$ are thus supposed to be decreasing in function of a distance that can be computed between k and ℓ. For instance, the correlations could be written as $\rho_{k\ell} = \rho^{d(k,\ell)}$, where $d(k,\ell)$ is a distance between units k and ℓ.

Isaki and Fuller (1982) define the anticipated variance as

$$\text{Avar}(\widehat{Y}) = \text{E}_M E_p(\widehat{Y} - Y)^2.$$

The anticipated variance allows to evaluate the precision of an estimator under a given design and a model. It thus allows to conceive the best design according to a superpopulation model that would have generated the population. Nedyalkova and Tillé (2008) compute the anticipated variance for a very general class of linear estimators. Grafström and Tillé (2013) prove that, under model (5.2.1), the anticipated variance of the Horitz-Thompson estimator can be shown to be

$$\text{Avar}(\widehat{Y}) = \text{E}_p \left[\left(\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} - \sum_{k \in U} \mathbf{x}_k \right)^\top \beta \right]^2 + \sum_{k \in U} \sum_{\ell \in U} \sigma_k \sigma_\ell \rho_{k\ell} \frac{\Delta_{k\ell}}{\pi_k \pi_\ell}.$$

Grafström and Tillé (2013) also showed that the design that minimizes the anticipated variance should be balanced on $\boldsymbol{x}_k$ in the sense defined by Deville and Tillé (2004). Furthermore, it must have unequal inclusion probabilities proportional to σ_k and must be as spread as possible in the space making the quantity $\rho_{k\ell}\Delta_{k\ell}$ small. Neyman optimal allocation (Neyman, 1934) is also a particular case of this result. Variance estimators for spread samples are proposed in Grafström, Lundström, and Schelin (2012).

Grafström and Lundström (2013) has also shown that spread samples are automatically well balanced on the variables used to compute the distance. The spread samples can then be considered approximately stratified over any compact set of units in space. Furthermore, Grafström, Saarela, and Ene (2014) has shown, through a set of simulations, that a spread sample improves not only the accuracy of the Horvitz-Thompson estimator but also that of the nearest neighbor estimator. A rigorous proof of this result is given in Fattorini et al. (2021). There are thus multiple applications where it is interesting to select samples spread out with unequal probabilities. As we have seen, several methods exist to select such samples. However, the selection of several spread samples with unequal probabilities that are negatively coordinated from the same population is a problem that is not yet solved. We propose a solution in the following sections.

5.3 Spatiotemporal sampling notations and requirements

Suppose that each unit k of the population U belongs to a metric space of dimension $r \geq 2$ and the spatial coordinates of each unit are known. Consider also $T \in \mathbb{N}$ different moments spaced out in time. For example, these T times may correspond to years or months. The selection of a spatiotemporal sample must satisfy given inclusion probabilities which can be equal or unequal. Let π_k^t be the probability that unit $k \in U$ is selected at time $t \in \{1, \ldots, T\}$. Let $\boldsymbol{\Pi}$ denote the $N \times T$ matrix of temporal inclusion probabilities:

$$\boldsymbol{\Pi} = \begin{pmatrix} \pi_1^1 & \cdots & \pi_1^t & \cdots & \pi_1^T \\ \vdots & & \vdots & & \vdots \\ \pi_k^1 & \cdots & \pi_k^t & \cdots & \pi_k^T \\ \vdots & & \vdots & & \vdots \\ \pi_N^1 & \cdots & \pi_N^t & \cdots & \pi_N^T \end{pmatrix}.$$

The tth column of matrix $\boldsymbol{\Pi}$ is denoted by $\boldsymbol{\pi}^t = (\pi_1^t, \ldots, \pi_k^t, \ldots, \pi_N^t)^\top$ and contains the inclusion probabilities of all the units at sampling time t. The kth row of matrix $\boldsymbol{\Pi}$ is denoted by $\boldsymbol{\pi}_k = (\pi_k^1, \ldots, \pi_k^t, \ldots, \pi_k^T)$ and contains the inclusion probabilities of k at each sampling time. The sum

of the tth column of $\boldsymbol{\Pi}$ is denoted by $\psi^t = \sum_{k \in U} \pi_k^t$ and the sum of the kth row by $\psi_k = \sum_{t=1}^{T} \pi_k^t$. The sums ψ^t and ψ_k are not necessarily integer.

The aim is to generate a matrix of indicator random variables a_k^t that are equal to 1 if plot k is selected in the sample at sampling time t and 0 otherwise. Matrix $\mathbf{A}$ is the $N \times T$ sampling indicator matrix:

$$\mathbf{A} = \begin{pmatrix} a_1^1 & \cdots & a_1^t & \cdots & a_1^T \\ \vdots & & \vdots & & \vdots \\ a_k^1 & \cdots & a_k^t & \cdots & a_k^T \\ \vdots & & \vdots & & \vdots \\ a_N^1 & \cdots & a_N^t & \cdots & a_N^T \end{pmatrix}.$$

The tth column of matrix $\mathbf{A}$ is denoted by $\mathbf{a}^t = (a_1^t, \ldots, a_k^t, \ldots, a_N^t)^\top$ and corresponds to the cross-sectional sample at time t. The kth row of matrix $\mathbf{A}$ is denoted by $\mathbf{a}_k = (a_k^1, \ldots, a_k^t, \ldots, a_k^T)$ and corresponds to the longitudinal sample of k. Let also $n^t = \sum_{k \in U} a_k^t$ be the number of units selected at the tth sampling time and $n_k = \sum_{t=1}^{T} a_k^t$ be the number of times that unit k is selected during the T times.

The objective is to select a spatiotemporal sample $\mathbf{A}$, which best meets the following three requirements:

(i) The sampling design satisfies the inclusion probabilities given in $\boldsymbol{\Pi}$, i.e. $\mathrm{E}_p(\mathbf{A}) = \boldsymbol{\Pi}$.

(ii) The longitudinal sample $\mathbf{a}_k = (a_k^1, \ldots, a_k^t, \ldots, a_k^T)$ is as spread over time as possible, for all $k \in U$, in the sense that once a unit has been selected, it should remain out of the following samples as long as possible.

(iii) The cross-sectional sample $\mathbf{a}^t = (a_1^t, \ldots, a_k^t, \ldots, a_N^t)^\top$ is as spread in space as possible, for all $t \in \{1, \ldots, T\}$, in the sense that we avoid selecting geographically neighboring units.

Requirement (i) is equivalent to have $\mathrm{E}_p(a_k^t) = \pi_k^t$, for each element a_k^t of matrix $\mathbf{A}$ and implies $\mathrm{E}_p(n^t) = \psi^t$ and $\mathrm{E}_p(n_k) = \psi_k$.

A longitudinal sample corresponds to select or not the same unit at T different times. The same variable measured at several different times on a unit k is positively autocorrelated over time. For this reason, the objective is to obtain a sample $\mathbf{a}_k$ as spread as possible if the vector of inclusion probabilities $\boldsymbol{\pi}_k$ allows to select the unit k more than once. By spreading each sample $\mathbf{a}_k$ (requirement (ii)), once a unit k is selected, it remains out of the sample as long as possible, depending on the vector of inclusion probabilities $\boldsymbol{\pi}_k$. This generates an appropriate rotation of the units selected in the cross-sectional samples $\mathbf{a}^t$ and minimizes the overlap between successive samples.

If the units are geolocated, spatial autocorrelation must be taken into account. By selecting a spread sample based on the spatial coordinates at each sampling time t, the accuracy of the estimate should be better than with unspread samples. Requirement (iii) prevents the selection of similar units at the same time.

Finding a method to meet all of these requirements is not straightforward. In the following sections, we proceed step-by-step, first explaining how the rotation of units in time is optimized (ii), and then describing the two spatiotemporal methods.

5.4 Method of Wang and Zhu

The new methods proposed in this paper are based on an existing method, developed by Wang and Zhu (2019), for the problem of spatiotemporal sampling. It based on consecutive applications of the local pivotal method developed by Grafström, Lundström, and Schelin (2012). The

local pivotal method generalizes the pivotal method, a sampling method without replacement described in Deville and Tillé (1998), to the selection of spread samples (see Appendix A and Appendix B).

The method of Wang and Zhu (2019) is described in Algorithm 1 and consists of two steps. First a spatially spread set of units, denoted by $G \subset U$, is selected using the local pivotal method. Then, samples $\mathbf{a}^1, \ldots, \mathbf{a}^T$ are selected from G without replacement. The method satisfies the constraints (i)-(iii), but can only be applied under two conditions:

(*i*) The columns of matrix $\mathbf{\Pi}$ are proportional.

(*ii*) The sums of the rows of $\mathbf{\Pi}$ are equal or smaller than one, i.e. $\psi_k \leq 1$ for all $k \in U$.

Algorithm 1 Wang and Zhu method

1. Select an initial set from U, denoted by G, by the local pivotal method with probabilities $\pi_k^\circ = L \sum_{t=1}^{T} \pi_k^t$, where $L \geq 1$ is a predefined value.
2. For $t = 1, \ldots, T$, repeat the following steps.
   (a) Select a sample $\mathbf{a}^t$ from G of size n^t by the local pivotal method with equal inclusion probabilities.
   (b) Update G by $G^* = G \backslash \{\mathbf{a}^1 \cup \cdots \cup \mathbf{a}^t\}$.

In the step 1 of Algorithm 1, probabilities of π_k° must remain smaller than one and L must not be too large to have a good spatial balance. The authors recommend to take $L \leq \min\{2, \min_{k \in U}(\psi_k)^{-1}\}$.

In real sampling problems, the assumptions on $\mathbf{\Pi}$ are not always satisfied. Indeed, inclusion probabilities are not necessarily proportional as in condition (*i*) of Wang and Zhu, especially if they are based on a variable that changes over time. Condition (*ii*) of Wang and Zhu is also very restrictive. In practical problems, the sums of the rows of $\mathbf{\Pi}$ could be larger that one. In this case, a unit would be selected several times during the period using a rotation scheme. With Wang and Zhu's method, each unit can only be selected once for the entire time period. The new methods proposed in this paper are not restricted to these two requirements.

5.5 Preliminary step to spatiotemporal sampling: a two-phase sampling approach

Spatiotemporal sampling methods can begin by the selection of an initial spread set of units, as in the method of Wang and Zhu. This allows to obtain a better spreading of the cross-sectional samples $\mathbf{a}^t$. This first sampling phase consists in selecting a first well-spread set U' of N' units, $N' \leq N$, which will then be considered for the spatiotemporal design. Thus, the inclusion probabilities used for this design become conditional on the first sampling phase.

All units $k \in \{U \backslash U'\}$ will therefore be permanently excluded from all the cross-sectional samples, which means that their inclusion probability will be $\pi_k^t = 0$ for each sampling time t. This first sampling phase is a generalization of the first step of the Wang and Zhu method without restricting $\mathbf{\Pi}$ to condition (*i*). Concerning the second condition (*ii*), this first phase should not be applied if $\mathbf{\Pi}$ does not satisfy it. Indeed, if we have $\psi_k \geq 1$ for a unit k, this unit must be selected in at least one cross-sectional sample $\mathbf{a}^t$, and then can not be completely excluded.

This selection is made using the local cube method (Grafström and Tillé, 2013). The local cube method is based on two methods: the cube method, that allows to select balanced samples on totals of auxiliary variables (Deville and Tillé, 2004), and the local pivotal method. Similarly to the cube method, the local cube method is divided in two phases: the *flight phase* and the *landing phase* (see Appendix B). More precisely, this preliminary step uses only the flight phase

of the local cube method (see Appendix C). Algorithm 2 describes the main steps of the selection of this initial spread set.

Algorithm 2 Preliminary step: selection of an initial spread set

1. Compute inclusion probabilities $\boldsymbol{\pi}^\circ = (\pi_1^\circ, \ldots, \pi_k^\circ, \ldots \pi_N^\circ)^\top$ such that $\pi_k^\circ = \min(L \sum_{t=1}^T \pi_k^t, 1)$.
2. Run the flight phase of the local cube method with $\boldsymbol{\pi}^\circ$ as inclusion probabilities and column of $\mathbf{\Pi}$ as balancing constraints. A vector $\boldsymbol{\pi}^\bullet = (\pi_1^\bullet, \ldots, \pi_k^\bullet, \ldots \pi_N^\bullet)^\top$ of inclusion probabilities is obtained.
3. Update matrix $\mathbf{\Pi}$ with $\mathbf{\Pi}^\bullet$ such that $\mathbf{\Pi}^\bullet = \text{diag}(\boldsymbol{\pi}^\bullet)\,\text{diag}(\boldsymbol{\pi}^\circ)^{-1}\mathbf{\Pi}$.

The same recommendation as in the Wang and Zhu method is applied to the choice of L, i.e. $L \leq \min\{2, \min_{k\in U}(\psi_k)^{-1}\}$.

Proposition 5.5.1. *During the process of Algorithm 2, the sums of the columns of matrix* $\mathbf{\Pi}$ *are equal to that of matrix* $\mathbf{\Pi}^\bullet$.

Proof. After step 2 of Algorithm 2, because of the flight phase of the local cube method, many lines of $\mathbf{\Pi}^\bullet$ may contain only zeros and we have $\mathbf{\Pi}^\top \boldsymbol{\pi}^\bullet = \mathbf{\Pi}^\top \boldsymbol{\pi}^\circ$. This allows to deduce that $\mathbf{\Pi}^{\bullet\top}\mathbf{1}_N = \mathbf{\Pi}^\top \text{diag}(\boldsymbol{\pi}^\circ)^{-1}\boldsymbol{\pi}^\bullet\mathbf{1}_N = \mathbf{\Pi}^\top \text{diag}(\boldsymbol{\pi}^\circ)^{-1}\boldsymbol{\pi}^\circ\mathbf{1}_N = \mathbf{\Pi}^\top\mathbf{1}_N$, so the sums of the columns of $\mathbf{\Pi}^\bullet$ is exactly the same as the ones of $\mathbf{\Pi}$, with $\mathbf{1}_N$ a column vector composed of N ones. ■

The Proposition 5.5.1 will allow us to obtain the fixed sample size based on the original matrix $\mathbf{\Pi}$ during the second sampling phase, i.e. the spatiotemporal sampling methods presented below. The sums of the rows of $\mathbf{\Pi}^\bullet$ are either equal to zero or not greater than $1/L$.

Matrix $\mathbf{\Pi}$ can be singular. In this case, it is more efficient to remove the columns of $\mathbf{\Pi}$ that are linearly dependent on the others. Indeed, this reduces the number of balancing constraints. If all the columns of matrix $\mathbf{\Pi}$ are proportional (i.e. linearly dependent), as in condition (i) of the Wang and Zhu method, only one column can be used and the local cube method is reduced to the local pivotal method. This shows that this first phase generalizes Wang and Zhu's.

5.6 Temporal spreading

As explained in Section 5.3, samples $\mathbf{a}^t$ must be as spread over time as possible while $\mathbf{a}_k$ must be as spread as possible in the space of dimension $r \geq 2$. These two spreads are difficult to manage simultaneously. The method used to address the first problem of temporal spreading is presented in this section.

The first constraint to be managed is the temporal spreading. In other words, each sample $\mathbf{a}_k$ must be spread in a one-dimensional space corresponding to the T different times. To that end, the longitudinal samples $\mathbf{a}_k$ are generated using systematic sampling, $k \in U$. Systematic sampling for unequal probabilities was proposed by Madow (1949) (see also Iachan, 1982; Iachan, 1983; Bellhouse, 1988; Bellhouse and Sutradhar, 1988). This sampling method selects a sample according to a random starting point but with a fixed, periodic interval based on the inclusion probabilities (see Appendix D). In one dimension, Quenouille (1949b) and Bellhouse (1977) proved that systematic sampling is the best design to obtain the most spread sample when the inclusion probabilities are equal. Therefore, systematic sampling is a good way to select each longitudinal sample $\mathbf{a}_k$ meeting requirement (ii).

For each vector of inclusion probabilities over time $\boldsymbol{\pi}_k$, all possible systematic samples are computed. Let $h(k)$ denote the number of possible systematic samples with non-zero probabilities for k. This number $h(k)$ is not greater than T, if ψ_k is integer, and is not greater than $(T+1)$ otherwise (Pea, Qualité, and Tillé, 2007). Let also $H = \sum_{k\in U} h(k)$ be the total number of longitudinal samples. In addition to taking into account autocorrelation as explained above, the

advantage of systematic sampling is that the total number H of possible samples is relatively small compared to other designs for which $H = N \times 2^T$. This makes it possible to describe all possible systematic samples in a simple way.

Consider $\mathbf{S}_k$ the matrix containing in rows the $h(k)$ possible longitudinal samples of a unit k such that $\mathbf{S}_k = (\mathbf{s}_{k,1}^\top, \ldots, \mathbf{s}_{k,i}^\top, \ldots, \mathbf{s}_{k,h(k)}^\top)^\top$, where $\mathbf{s}_{k,i} = (s_{k,i}^1, \ldots, s_{k,i}^t, \ldots, s_{k,i}^T)$ is the ith possible systematic sample. Consider also $\mathbf{p}_k = (p_{k,1}, \ldots, p_{k,i}, \ldots, p_{k,h(k)})^\top$ the probabilities of selecting the samples of $\mathbf{S}_k$. We have $\mathbf{S}_k^\top \mathbf{p}_k = \pi_k$. All matrices $\mathbf{S}_k$ and vectors $\mathbf{p}_k$ are respectively concatenated in a matrix $\mathbf{S}$ and a vector $\mathbf{p}$, such that $\mathbf{S} = (\mathbf{S}_1^\top, \ldots, \mathbf{S}_k^\top, \ldots, \mathbf{S}_N^\top)^\top$ and $\mathbf{p} = (\mathbf{p}_1^\top, \ldots, \mathbf{p}_k^\top, \ldots, \mathbf{p}_N^\top)^\top$. The rows of matrix $\mathbf{S}$ thus contains H longitudinal samples of size T. Vector $\mathbf{p}$ contains the selection probabilities of the H longitudinal samples in $\mathbf{S}$.

Example 5.6.1. *Consider the matrix of inclusion probabilities* $\mathbf{\Pi}$ *with* $N = 3$ *and* $T = 4$

$$\mathbf{\Pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} = \begin{pmatrix} 0.4 & 0.6 & 0.2 & 0.8 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.9 & 0.3 & 0.7 \end{pmatrix}.$$

The longitudinal sampling designs using systematic sampling are computed:

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_1 \\ \hline \mathbf{S}_2 \\ \hline \mathbf{S}_3 \end{pmatrix} = \left(\begin{array}{c} \mathbf{s}_{1,1} \\ \mathbf{s}_{1,2} \\ \mathbf{s}_{1,3} \\ \hline \mathbf{s}_{2,1} \\ \mathbf{s}_{2,2} \\ \hline \mathbf{s}_{3,1} \\ \mathbf{s}_{3,2} \\ \mathbf{s}_{3,3} \end{array}\right) = \left(\begin{array}{cccc} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array}\right) \text{ and } \mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \hline \mathbf{p}_2 \\ \hline \mathbf{p}_3 \end{pmatrix} = \left(\begin{array}{c} p_{1,1} \\ p_{1,2} \\ p_{1,3} \\ \hline p_{2,1} \\ p_{2,2} \\ \hline p_{3,1} \\ p_{3,2} \\ p_{3,3} \end{array}\right) = \left(\begin{array}{c} 0.2 \\ 0.2 \\ 0.6 \\ \hline 0.5 \\ 0.5 \\ \hline 0.1 \\ 0.2 \\ 0.7 \end{array}\right),$$

and the number of samples in $\mathbf{S}$ *are* $H = h(1) + h(2) + h(3) = 3 + 2 + 3 = 8$.

Once all longitudinal sampling designs have been computed, a systematic sample must be choosen for each unit $k \in U$. Let $\mathbf{q} \in \{0,1\}^H$ denote the vector of indicators, with the same dimension as $\mathbf{p}$, that indicates which systematic sample is definitively selected. Each element $q_{k,i}$ is a Bernoulli random variable that is equal to 1 if the ith longitudinal sample of the unit k is selected and 0 otherwise. Each realization of $\mathbf{q}$ corresponds to the selection of the longitudinal samples and we have $\mathbf{S}_k^\top \mathbf{q}_k = \mathbf{a}_k$. The aim is to define a method to select exactly one longitudinal sample for each unit k, this implies

$$\sum_{i=1}^{h(k)} q_{k,i} = \sum_{i=1}^{h(k)} p_{k,i} = 1, k \in U.$$

Vector $\mathbf{a}_k$ will be a systematic sample, the temporal spreading is therefore guaranteed. Some constraints must also be applied on $\mathbf{q}$ to ensure the spatial spreading of cross-sectional samples $\mathbf{a}^t$. If unit k is in the neighborhood of ℓ, the idea is to choose systematic samples that do not select k and ℓ at the same time. In other words, it could be better to have as less as possible $a_k^t = a_\ell^t$. In the next section, two spatiotemporal sampling methods are explained. They use different methods to randomly select $\mathbf{a}_k$ in each submatrix $\mathbf{S}_k$ while satisfying the spatial spreading at each sampling time.

5.7 Spatiotemporal sampling

The constraint of knowing the size of the sample selected before sampling is often required. In the context of spatiotemporal sampling, this constraint must be taken into account for T samples. This makes it much more complicated to satisfy. To obtain fixed sample sizes, vector $\mathbf{q}$ must satisfy the balancing equation

$$\mathbf{S}^\top \mathbf{q} = \mathbf{S}^\top \mathbf{p} = (n^1, \ldots, n^t, \ldots, n^T)^\top.$$

In this section, we propose two different spatiotemporal sampling methods that select spatially well-distributed samples at the sampling times. It is important to note that the methods differ in that one generates random sample size at each sampling time while the other generates fixed sample size. The second method is much more complex because of the complexity of satisfying Equation 5.7 while still considering the requirements (i)-(iii) presented above. Both can be considered as generalizations of Wang and Zhu's method.

5.7.1 Spatiotemporal sampling with random sample sizes (SPAR)

The *spatiotemporal sampling with random sample sizes* (SPAR) method is a spatiotemporal sampling method with random sample size at each sampling time. The sizes of the samples $\mathbf{a}^t$ are fixed only if the inclusion probabilities in $\boldsymbol{\pi}^t$ are all equal. In this restricted case, this method is the same as that of Wang and Zhu. In the other cases, this method has only fixed sample size at the sampling time $t = 1$.

The SPAR method is described in Algorithm 3.

Algorithm 3 SPAR Sampling

For $t = 1$ to T, apply the following instructions:

1. Apply a spread sampling method on vector of inclusion probabilities $\boldsymbol{\pi}^t$ of matrix $\boldsymbol{\Pi}$ and obtain vector $\mathbf{a}^t = (a_1^t, \ldots, a_k^t, \ldots, a_N^t)^\top$.
2. For each unit k, update the probabilities $\mathbf{p}_k = (p_{k,1}, \ldots, p_{k,j}, \ldots, p_{k,h(k)})^\top$ of the systematic sampling design as follows:

   - If $a_k^t = 1$,

$$p_{k,j}^* = \begin{cases} p_{k,j}\left(\sum_{j=1}^{h(k)} p_{k,j} s_{k,j}^t\right)^{-1} & \text{if } s_{k,j}^t = 1, \\ 0 & \text{if } s_{k,j}^t = 0. \end{cases}$$

   - If $a_k^t = 0$,

$$p_{k,j}^* = \begin{cases} p_{k,j}\left(\sum_{j=1}^{h(k)} p_{k,j}(1 - s_{k,j}^t)\right)^{-1} & \text{if } s_{k,j}^t = 0, \\ 0 & \text{if } s_{k,j}^t = 1. \end{cases}$$

3. Update matrix $\boldsymbol{\Pi}$ by $\boldsymbol{\Pi}^*$ such that the kth row of $\boldsymbol{\Pi}^*$ is $\boldsymbol{\pi}_k^* = \mathbf{S}_k^\top \mathbf{p}_k^*$, with $k = 1, \ldots, N$.

The spread of the selected units is managed by recursively applying a spread sampling method at each sampling time. This spread sampling method used could be for instance the local pivotal method, the generalized random tessellation stratified method (Stevens Jr. and Olsen, 1999; Stevens and Olsen, 2003; Stevens and Olsen, 2004), , the wave sampling method (Jauslin and Tillé, 2020) or even the travelling salesman problem-systematic method (Dickson and Tillé, 2016).

By updating $\mathbf{p}$ values to 0, some systematic samples are excluded at each iteration depending on the results of the spread sampling. For example, at time $t = 1$, if unit $k = 1$ is selected in sample $\mathbf{a}^1$, any systematic samples of unit k for which the first element is 0 are excluded. The

latest update of vector $\mathbf{p}$ in Algorithm 3 is only composed of 0s and 1s and corresponds to the vector $\mathbf{q}$. This sampling satisfies inclusion probabilities, requirement (i) is met. This vector $\mathbf{q}$ also satisfies requirements (ii) due to the systematic sampling and (iii) due to the first step, but does not have a fixed sample size.

5.7.2 Spatiotemporal sampling with fixed sample sizes (SPAF)

The *spatiotemporal sampling with fixed sample sizes* (SPAF) method allows to select spatiotemporal samples that are spatially, temporally spread, while controlling the size of cross-sectional samples. The procedure is based on the local pivotal method. Let $(\mathbf{S}_k, \mathbf{p}_k)$ denote the couple containing the longitudinal sampling design of a unit k and $\mathfrak{I}(\mathbf{M})$ be the image of a matrix $\mathbf{M}$. The steps of the SPAF method are described in Algorithm 4.

Each of the T sampling times are processed one by one. Two units k and ℓ in U are considered at main stages of Algorithm 4. The main idea is to recursively update the vectors $\mathbf{p}_k$ and $\mathbf{p}_\ell$ such that there is a repulsion in the selection of k and ℓ at the same sampling time. All the couples of units (k, ℓ) are treated in a particular order. Indeed, they are sorted in increasing order according to the Euclidean distance between their coordinates in the metric space to which they belong. Nearby units are thus favoured because they are treated first, and they will have less chances to be selected at the same sampling time. The update of $\mathbf{p}_k$ and $\mathbf{p}_\ell$ is similar to that of the local pivot method. The particularity here is that the update focuses on the decision making for the sampling time considered at this iteration. During the procedure, the $\mathbf{p}_k$ vectors are updated and some of their values are potentially set to 0. The probabilities updated to 0 imply the exclusion of the corresponding samples in the $\mathbf{S}_k$ matrices. These samples should therefore not be taken into account for the rest of the procedure, so sub-couple $(\widetilde{\mathbf{S}}_k, \widetilde{\mathbf{p}}_k)$, containing only systematic sampling with non-null probabilities of being selected in $\mathbf{p}_k$, must be defined for each k. These samples must not be taken into account for the rest of the procedure. It is thus necessary to define for each k a sub-couple $(\widetilde{\mathbf{S}}_k, \widetilde{\mathbf{p}}_k)$, containing only systematic samples having non-zero probabilities in $\mathbf{p}_k$ to be selected.

Proposition 5.7.1. *Consider the procedure described in Algorithm 4. Throughout this process, the following propositions are satisfied:*

(i) The vectors $\mathbf{b}_k$ and $\mathbf{b}_\ell$ exist.

(ii) If the sums of the rows of $\mathbf{S}_k$ are all integer, the vectors $\mathbf{q} \in \{0,1\}^H$ and $\mathbf{p} \in [0,1]^H$ satisfy

$$\sum_{i=1}^{h(k)} q_{k,i} = \sum_{i=1}^{h(k)} p_{k,i} = 1, k \in U,$$

(iii) $\mathrm{E}(\mathbf{A}) = \mathbf{\Pi}$,

(iv) The units selected in each cross-sectionnal sample $\mathbf{a}^t$, $t \in \{1, \ldots, T\}$, are well spread in the metric space in which they belong,

(v) The sum of the vector $\boldsymbol{\pi}$ remains unchanged after each of its updates:

$$\sum_{k=1}^{n} \pi_k^* = \sum_{k=1}^{n} \pi_k.$$

Proof. *(i)* The existence of $\mathbf{b}_k$ and $\mathbf{b}_\ell$ is guaranteed because $\mathbf{u}_p \in \{\mathfrak{I}(\mathbf{U}_k) \cap \mathfrak{I}(\mathbf{U}_\ell)\}$ can always be written as a linear combination of the samples contained in rows of $\mathbf{S}_k$ and $\mathbf{S}_\ell$

Algorithm 4 SPAF sampling

(i) Define the vector $\mathbf{d}_t \in \mathbb{R}^T$ such that its tth element is equal to 1 and the others to 0, with $t \in \{1, \ldots, T\}$.

(ii) Consider all pairs of units $(k, \ell) \in (U \times U)$.

(iii) For each sampling time $t = 1$ to T, apply the following steps to each pair of units (k, ℓ) from the closest to the farthest in terms of Euclidean distance between their spatial coordinates.

1. Define the subdesign $(\widetilde{\mathbf{S}}_k, \widetilde{\mathbf{p}}_k)$ of $(\mathbf{S}_k, \mathbf{p}_k)$ such that $(\widetilde{\mathbf{S}}_k, \widetilde{\mathbf{p}}_k)$ contains only the $\widetilde{h}(k)$ samples of $(\mathbf{S}_k, \mathbf{p}_k)$ with non-null probabilities. Compute subdesigns $(\widetilde{\mathbf{S}}_k, \widetilde{\mathbf{p}}_k)$ and $(\widetilde{\mathbf{S}}_\ell, \widetilde{\mathbf{p}}_\ell)$.
2. Compute matrices $\mathbf{U}_k = \widetilde{\mathbf{S}}_k^\top - \boldsymbol{\pi}_k \mathbf{1}_{\widetilde{h}(k)}^\top$ and $\mathbf{U}_\ell = \widetilde{\mathbf{S}}_\ell^\top - \boldsymbol{\pi}_\ell \mathbf{1}_{\widetilde{h}(\ell)}^\top$.
3. While $\pi_k^t \notin \{0, 1\}$ and $\{\mathfrak{I}(\mathbf{U}_k) \cap \mathfrak{I}(\mathbf{U}_\ell)\} \neq \{\mathbf{0}\}$, repeat the following instructions:
   (a) Let $\mathbf{u}_p \in \mathbb{R}^T$ be the orthogonal projection of $\mathbf{d}_t$ on the set $\{\mathfrak{I}(\mathbf{U}_k) \cap \mathfrak{I}(\mathbf{U}_\ell)\}$. If $\mathbf{u}_p$ is null, move on to another couple of units.
   (b) Find two vectors $\mathbf{b}_k \in \mathbb{R}^{\widetilde{h}(k)}$ and $\mathbf{b}_\ell \in \mathbb{R}^{\widetilde{h}(\ell)}$ such that $\widetilde{\mathbf{S}}_k^\top \mathbf{b}_k = \mathbf{u}_p$ and $\widetilde{\mathbf{S}}_\ell^\top \mathbf{b}_\ell = \mathbf{u}_p$.
   (c) Compute the largest values for γ_{k1}, γ_{k2}, $\gamma_{\ell 1}$ and $\gamma_{\ell 2}$ that satisfy $\mathbf{0} \leq \widetilde{\mathbf{p}}_k + \gamma_{k1}\mathbf{b}_k \leq \mathbf{1}$, $\mathbf{0} \leq \widetilde{\mathbf{p}}_k - \gamma_{k2}\mathbf{b}_k \leq \mathbf{1}$, $\mathbf{0} \leq \widetilde{\mathbf{p}}_\ell + \gamma_{\ell 1}\mathbf{b}_\ell \leq \mathbf{1}$ and $\mathbf{0} \leq \widetilde{\mathbf{p}}_\ell - \gamma_{\ell 2}\mathbf{b}_\ell \leq \mathbf{1}$.
   (d) Compute $\lambda_1 = \min(\gamma_{k1}, \gamma_{\ell 2})$ and $\lambda_2 = \min(\gamma_{k2}, \gamma_{\ell 1})$.
   (e) Update randomly vectors $\widetilde{\mathbf{p}}_k$ and $\widetilde{\mathbf{p}}_\ell$ such that

$$\begin{cases} \widetilde{\mathbf{p}}_k \leftarrow \widetilde{\mathbf{p}}_k + \lambda_1 \mathbf{b}_k \\ \widetilde{\mathbf{p}}_\ell \leftarrow \widetilde{\mathbf{p}}_\ell - \lambda_1 \mathbf{b}_\ell \end{cases} \text{ with probability } \lambda_2/(\lambda_1 + \lambda_2),$$

   or

$$\begin{cases} \widetilde{\mathbf{p}}_k \leftarrow \widetilde{\mathbf{p}}_k - \lambda_2 \mathbf{b}_k \\ \widetilde{\mathbf{p}}_\ell \leftarrow \widetilde{\mathbf{p}}_\ell + \lambda_2 \mathbf{b}_\ell \end{cases} \text{ with probability } \lambda_1/(\lambda_1 + \lambda_2).$$

   (f) Update non-null probabilities in $\mathbf{p}_k$ and $\mathbf{p}_\ell$ with the updated values of $\widetilde{\mathbf{p}}_k$ and $\widetilde{\mathbf{p}}_\ell$.
   (g) If there is at least one null value in vector $\widetilde{\mathbf{p}}_k$, remove them from $\widetilde{\mathbf{p}}_k$ and remove also their corresponding longitudinal samples in matrix $\widetilde{\mathbf{S}}_k$. Do the same for unit ℓ to considered only samples with positive selection probabilities in $\widetilde{\mathbf{S}}_\ell$ and $\widetilde{\mathbf{p}}_\ell$.
   (h) Update the inclusion probabilities $\boldsymbol{\pi}_k$ and $\boldsymbol{\pi}_\ell$ because $\widetilde{\mathbf{S}}_k$, $\widetilde{\mathbf{S}}_\ell$, $\widetilde{\mathbf{p}}_k$ and $\widetilde{\mathbf{p}}_\ell$ could be modified in the previous step.
   (i) Recompute also matrices $\mathbf{U}_k$ and $\mathbf{U}_\ell$.
4. Update $\mathbf{p}$ with $\mathbf{p}^*$ such that $\mathbf{p}_k^*$ and $\mathbf{p}_\ell^*$ are the updated $\mathbf{p}_k$ and $\mathbf{p}_\ell$.
5. Update $\boldsymbol{\Pi}$ with $\boldsymbol{\Pi}^*$ where the kth and ℓth rows of $\boldsymbol{\Pi}^*$ are respectively $\boldsymbol{\pi}_k^* = \mathbf{S}_k^\top \mathbf{p}_k^*$ and $\boldsymbol{\pi}_\ell^* = \mathbf{S}_\ell^\top \mathbf{p}_\ell^*$.

(iv) If $\boldsymbol{\Pi}$ still contains values not equal to zero or one, apply the following steps:

1. Consider the population of size H of the systematic samples contains in $\mathbf{S}$.
2. Stratify the population in N strata so that each stratum contains the subgroup of systematic samples of a unit $k \in U$, i.e. $\mathbf{S}_k$.
3. Apply a stratified balanced sampling on the population of systematic samples using vector $\mathbf{p}$ as inclusion probability, columns of $\mathbf{S}$ as balancing variables and strata defined in the previous step.

(ii) Throughout algorithm 4, $\mathbf{p}$ is iteratively modified in order to finaly obtain $\mathbf{q}$. Only the non-zeros values of $\mathbf{p}$ are considered at each step, so a subvector $\widetilde{\mathbf{p}}_k$ of $\mathbf{p}_k$ containing only these relevant values is defined for each unit k. For a unit k, $\widetilde{\mathbf{p}}_k$ is updated by $(\widetilde{\mathbf{p}}_k + \lambda \mathbf{b}_k)$, with $\lambda \in \mathbb{R}$. If $\mathbf{b}_k$ is centered, the proposition *(i)* is immediately proven. Since $\widetilde{\mathbf{S}}_k^\top \mathbf{b}_k = \mathbf{u}_p$ and $\mathbf{1}_T^\top \mathbf{u}_p = 0$, we have $\mathbf{1}_T^\top \widetilde{\mathbf{S}}_k^\top \mathbf{b}_k = 0$. In the case where the sums of the rows of $\mathbf{S}_k$ are all equal because $\psi_k \in \mathbb{N}$, this implies $\mathbf{1}_T^\top \widetilde{\mathbf{S}}_k^\top = n_k \mathbf{1}_{\widetilde{h}(k)}^\top$ and then $n_k \mathbf{1}_{\widetilde{h}(k)}^\top = 0$. The vector $\mathbf{b}_k$ is thus centered and the proposition *(ii)* is proven.

(iii) For a unit k, the expectation of π_k under the random of the update, i.e. the expectation of π_k^*, is:

$$\mathrm{E}_p(\pi_k^*) = \frac{\lambda_2}{\lambda_1+\lambda_2}\widetilde{\mathbf{S}}_k^\top(\widetilde{\mathbf{p}}_k + \lambda_1\mathbf{b}_k) + \frac{\lambda_1}{\lambda_1+\lambda_2}\widetilde{\mathbf{S}}_k^\top(\widetilde{\mathbf{p}}_k - \lambda_2\mathbf{b}_k) = \pi_k.$$

(iv) The update is made such that if a probability π_k^t is increased by the update, the corresponding π_ℓ^t will be decreased and reciprocally. There is a repulsion in the selection of the neighboring units k and ℓ at the same sampling time t, as in the local pivotal method. This allows to obtain a spread sample.

(v) At each iteration, only the inclusion probability vectors $\boldsymbol{\pi}_k$ and $\boldsymbol{\pi}_\ell$ of two units k and ℓ are updated by $\boldsymbol{\pi}_k^*$ and $\boldsymbol{\pi}_\ell^*$. So the proposition *(v)* is proven if $(\boldsymbol{\pi}_k + \boldsymbol{\pi}_\ell) = (\boldsymbol{\pi}_k^* + \boldsymbol{\pi}_\ell^*)$. The sum $(\boldsymbol{\pi}_k^* + \boldsymbol{\pi}_\ell^*)$ can be computed:

$$\begin{aligned}
\boldsymbol{\pi}_k^* + \boldsymbol{\pi}_\ell^* &= \widetilde{\mathbf{S}}_k^\top(\widetilde{\mathbf{p}}_k + \lambda\mathbf{b}_k) + \widetilde{\mathbf{S}}_\ell^\top(\widetilde{\mathbf{p}}_\ell - \lambda\mathbf{b}_\ell) \\
&= \widetilde{\mathbf{S}}_k^\top\widetilde{\mathbf{p}}_k + \lambda\widetilde{\mathbf{S}}_k^\top\mathbf{b}_k + \widetilde{\mathbf{S}}_\ell^\top\widetilde{\mathbf{p}}_\ell - \lambda\widetilde{\mathbf{S}}_\ell^\top\mathbf{b}_\ell \\
&= \boldsymbol{\pi}_k + \lambda\mathbf{u}_p + \boldsymbol{\pi}_\ell - \lambda\mathbf{u}_p = \boldsymbol{\pi}_k + \boldsymbol{\pi}_\ell,
\end{aligned}$$

with $\lambda = \lambda_1$ or $\lambda = -\lambda_2$.

■

Only one longitudinal sample $\mathbf{a}_k$ must be selected among $\mathbf{S}_k$ for each unit k (Equation 5.6). This is satisfied by keeping the sum of the components of $\mathbf{p}_k$ and $\mathbf{p}_\ell$ equal to one at each step, as in the Proposition 5.7.1 (ii).

This is satisfied only under the condition that the sums of the rows of $\boldsymbol{\Pi}$ are integer, i.e. $\psi_k \in \mathbb{N}$ for all $k \in U$. During each stage, if at least one of the two units k and ℓ does not satisfy this condition, one can simply solve the problem by adding phantom sampling times to vectors of inclusion probabilities $\boldsymbol{\pi}_k$ and $\boldsymbol{\pi}_\ell$, as proposed by Grafström et al. (2012), to sum to an integer.

This trick allows to apply the algorithm, without restriction on the $\boldsymbol{\Pi}$ matrix. Propositions 5.7.1 (iii) and 5.7.1 (iv) correspond to requirements (i) and (iii) respectively. Then, the size of the cross-sectional samples is fixed at each sampling time t, this can be deduced from the Proposition 5.7.1 (iv).

5.8 Simulations

First of all, the interest of this section is to compare our method to other existing methods in different situations. However, when the probability matrix $\boldsymbol{\Pi}$ is totally unequal, no method exists to select spatiotemporal samples. In this case, the general dispersion of each sample selected using the methods SPAR and SPAF must be evaluated and criticized, without any means of comparison.

5.8.1 Spreading measures

A commonly used spatial balanced index has been developed by Stevens and Olsen (2004). This index is based on the partition of space into Voronoï polygons and is particularly effective for comparing the spatial spread of different samples of the same population. Let $s^t = \{i \in U \mid a_i^t = 1\}$ be the set of selected units index at time t in a cross-sectionnal sample $\mathbf{a}^t$, $t \in \{1,\dots,T\}$. The Voronoï polygon of the ith selected unit in $\mathbf{a}_t$ is defined as the polygon which includes all non-selected units that are closest to i than all other selected units, with $i \in s^t$. Let z_i be the sum of the inclusion probabilities of units included into the ith Voronoï polygon. Grafström, Lundström, and Schelin (2012) assert that a best spatially balanced sample is one with $z_i = 1$ for all selected unit i and suggest that the variance $B(s^t) = 1/n_t \sum_{i\in s^t}(z_i - 1)^2$ can represent a measure of spatial

balance for a sample of size $n_t = \sum_{k \in U} a_k^t$. The smaller its value, the better $\mathbf{a}^t$ is spatially balanced. Since this measure depends on the spatial pattern of the population, this allows to compare the spreading of different samples from the same population. It is very useful in determining which sampling design best selects well-spread samples at each sampling time.

Moran's I index is a measure of spatial autocorrelation proposed by Moran (1950). It is based on the fact that the level of spatial autocorrelation of an indicator random variable as $\mathbf{a}^t$ shows its level of spatial spreading. Because of the limitations of this index, Tillé et al. (2018) have developed a normalized version of Moran's I index: the I_B index. This new index version can take values from -1 (perfect spatial balance) to 1 (maximum concentration) and a neutral value 0. It allows to evaluate the spatial spreading and the spatial balance of a sample.

5.8.2 Biological data

To evaluate our methods, the *Centre Suisse de Cartographie de la Faune* (CSCF) provided us with a spatial biological data set. These data list *odonata* (i.e. dragonflies and damselflies) species observed in land squares in Switzerland between 1840 and 2020. Data includes 1400 land squares with an area of 1 km^2 and 83 different *odonata* species located in the Swiss cantons of Fribourg, Neuchâtel and Vaud. Figure 1 represents a map with the 1400 land squares.

We focused on a sampling design to study the rare species. The importance of a square is related to the number of rare species observed there. Let $\mathbf{M}$ denote the matrix that contains in rows the 1400 land squares and in columns the 83 species. Matrix $\mathbf{M}$ is composed of 0s and 1s that specify if a species has already been observed within a square. Consider column vector $\mathbf{g} \in \mathbb{N}^{83}$ that contains the inverse of the species occurrence rate. In other words, g_i is equal to the inverse of the sum of the ith column of $\mathbf{M}$, $i = 1, \ldots, 83$. Consider also vector $\mathbf{c} \in \mathbb{R}^{1400}$, such that $\mathbf{c} = \mathbf{g}^\top \mathbf{M}$, containing a square importance measure based on rare species.

5.8.3 Results

To evaluate our sampling methods, we consider the problem of selecting a spatiotemporal sample of land squares that are both spread and with fixed size at each sampling time. We considered $T = 3$ sampling times. The simulations were run using the '`SpotSampling`' R package (Eustache, Jauslin, and Tillé, 2020).

For the first scenario of simulations, we considered equal inclusion probabilities at each sampling time, with $n^1 = 200$, $n^2 = 250$ and $n^3 = 300$. The columns of $\boldsymbol{\Pi}$ are then proportional and the sum of its rows are all equal to $15/28$. By taking this structure of inclusion probabilities, the method of Wang and Zhu can be applied and then compared to our methods. For the second scenario, inclusion probabilities are chosen totally unequal, with the idea of increasing emphasis on rare species over time. For the first sampling time, inclusion probabilities in $\boldsymbol{\pi}^1$ are all equal. For the second one, $\boldsymbol{\pi}^2$ is proportional to the number of species in land squares. Next, vector $\boldsymbol{\pi}^3$ of inclusion probabilities of the last sampling time is proportional to vector $\mathbf{c}$ to give more importance to squares potentially containing rare species. Sample sizes are $n^1 = n^2 = n^3 = 250$.

Figure 1 shows a spatiotemporal sample selected with the SPAF method using equal inclusion probabilities from the first scenario of the simulations. Land squares filled in light grey represent the initial spread set selected using the preliminary step described in Section 5.5. This first step can be applied because the sums of the rows of $\boldsymbol{\Pi}$ are not greater than 1. This initial set is represented in Figure 1 at the top left. Land squares definitively selected with the SPAF sampling are filled in black. Figure 1 at top right, bottom left and bottom right respectively represent the selected samples at sampling times 1, 2 and 3.

We performed 10'000 simulations. For each compared method evaluated, the average values of the spread measures I_B and B during the simulations are calculated for each sampling time t. The results are summarized in Table 1. In the first scenario, SPAR, SPAF and Wang and Zhu

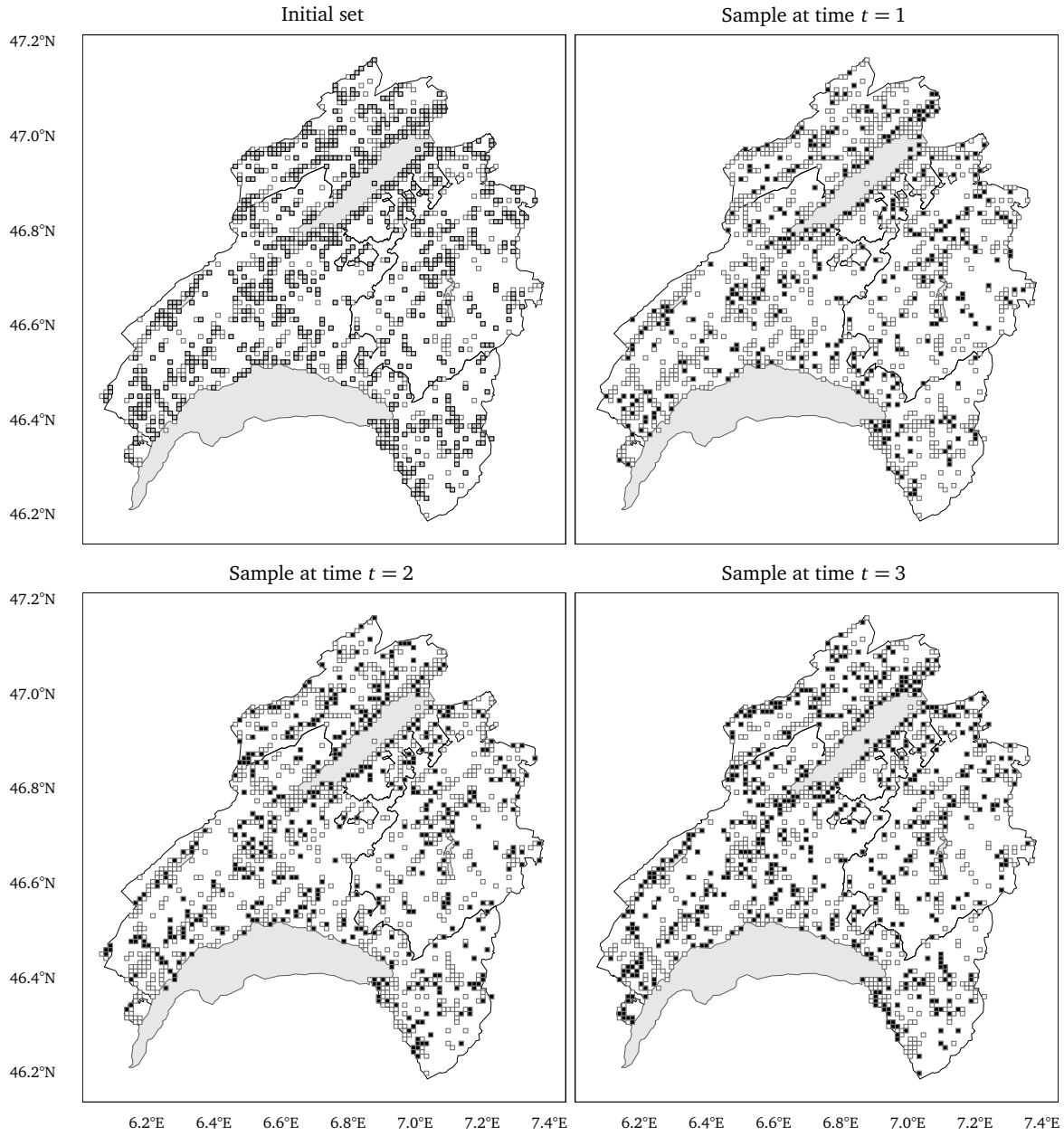

FIGURE 1: Figures showing a spatiotemporal sample selected with the SPAF sampling method. The sampling population are composed of 1400 land squares located in the Swiss cantons of Fribourg, Neuchâtel and Vaud. Three sampling times are considered with sample size respectively equal to 200, 250 and 300. Inclusion probabilities are equal at each sampling time. An initial spread set is selected before applying the SPAF method. Land squares that are not excluded from the population during this preselection are filled in light grey in figure at top left. Land squares definitively selected with the SPAF sampling are filled in black. Figures at top right, bottom left and bottom right respectively represent the selected samples at sampling times 1, 2 and 3.

methods are comparable in terms of I_B and B measures. For the second scenario, spreading measures show that samples selected with the SPAF sampling method are well spread. However,

spreading of samples selected with SPAR is better than those selected with SPAF. This is easily explained by the fact that SPAR generates samples with random size.

TABLE 1: Spreading measures of spatiotemporal samples based on 10'000 simulations on the spatial biological data from CSCF. The WZ method cannot be applied on totally unequal inclusion probabilities.

	Sampling design				
	Equal probabilities			Unequal probabilities	
	SPAR	SPAF	WZ	SPAR	SPAF
I_B					
$t=1$	-0.388	-0.374	-0.389	-0.386	-0.360
$t=2$	-0.393	-0.378	-0.395	-0.264	-0.233
$t=3$	-0.387	-0.357	-0.386	-0.309	-0.273
B					
$t=1$	0.116	0.123	0.115	0.130	0.140
$t=2$	0.128	0.135	0.127	0.151	0.165
$t=3$	0.138	0.147	0.138	0.145	0.158

SPAR, spatiotemporal sampling with random sample sizes; SPAF, spatiotemporal sampling with fixed sample sizes; WZ, Wang and Zhu method.

5.9 Conclusion

The selection of spatiotemporal samples is a complex problem particularly when one wants to impose simultaneously constraints of temporal and spatial spreading. However, these constraints are important to optimize the collection of information.

In this paper, we have solved the problem in the most general case, i.e. when inclusion probabilities are unequal and variable over time. Two spatiotemporal sampling methods are described. They provide a random spatiotemporal sample that is well spread in time and in space. The first one, called SPAR, select a sample at each sampling time that is of random size while the second one, the SPAF method, also allows to control their sizes. a random spatiotemporal sample with random size and spread at each sampling time. The second one, the SPAF method gives a random spatiotemporal sample containing longitudinal samples of fixed size and well spatially spread.

The proposed sampling methods are evaluated on spatial biological data given by the *Centre Suisse de Cartographie de la Faune*. Simulations show that samples selected with SPAR and SPAF are well spread in space. The SPAF method is the first spatiotemporal sampling method that allows to consider unequal and time-varying probabilities. All of these results indicate that the spot method is very efficient to select a well spread sample. These methods can be very easily used by means of the '`SpotSampling`' R package (Eustache, Jauslin, and Tillé, 2020).

Acknowledgements

The authors warmly thank members of the *Centre Suisse de Cartographie de la Faune* for the trust they have given them by making available a set of data to evaluate our method. They are grateful for the time spent by Lionel Qualité to make relevant corrections and remarks on this paper. They also thank Zhonglei Wang and Zhengyuan Zhu who shared their R code, which made it possible to compare their methods.

Appendix A: The local pivotal method

Consider a vector of inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$. The local pivotal method is described in Algorithm 5.

Algorithm 5 Local pivotal method

Repeat the following steps until all elements in $\boldsymbol{\pi}$ are equal to zero or one.

1. Select two neighboring units $k \in U$ and $\ell \in U$ that still have non-integer inclusion probabilities π_k and π_ℓ.
2. Compute $\lambda_1 = \min(\pi_k + \pi_\ell, 1)$ and $\lambda_2 = \max(0, \pi_k + \pi_\ell - 1)$.
3. Update the probabilities π_k and π_ℓ such that

$$\begin{cases} \pi_k \leftarrow \lambda_1, \pi_\ell \leftarrow \lambda_2 & \text{with probability } (\pi_k - \lambda_2)/(\lambda_1 - \lambda_2), \\ \pi_k \leftarrow \lambda_2, \pi_\ell \leftarrow \lambda_1 & \text{with probability } (\lambda_1 - \pi_k)/(\lambda_1 - \lambda_2). \end{cases}$$

In Algorithm 5, the definition of the neighbourhood of unit k may be based on spatial coordinates. Several variants of the method exists but they only differ by the way of selecting the two neighboring units from spatial coordinates. At each step, if π_k is increased, π_ℓ is decreased and reciprocally. This repulsion in the selection of neighboring units allows to obtain a spread sample at the end of the algorithm.

Appendix B: The cube method

Consider a vector of inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$. The cube method allows to generate a random vector $\mathbf{a} = (a_1, \ldots, a_N)^\top$ of Bernoulli variables such that $\mathrm{E}_p(\mathbf{a}) = \boldsymbol{\pi}$ and which is balanced on the totals of auxiliary variables. A sample $\mathbf{a}$ is said to be balanced on the J auxiliary variables if it satisfies equation

$$\mathbf{M}^\top \mathbf{a} = \mathbf{M}^\top \boldsymbol{\pi}, \tag{5.9.2}$$

where $\mathbf{M} = (\mathbf{x}_1/\pi_1, \ldots, \mathbf{x}_N/\pi_N)^\top$. Expression (5.9.2) is called the system of balancing equations. Sometimes, this equality cannot be exactly satisfied.

The cube method is divided into two phases: the flight phase and the landing phase. The flight phase of the cube method is a random walk from $\boldsymbol{\pi}$ to $\boldsymbol{\pi}^*$ such that $\mathrm{E}_p(\boldsymbol{\pi}^*) = \boldsymbol{\pi}$, $\mathbf{M}^\top \boldsymbol{\pi}^* = \mathbf{M}^\top \boldsymbol{\pi}$ and $\#\{k \mid \boldsymbol{\pi}^* \notin \{0,1\}\} \leq J$. So, there remains at most J (i.e. the number of columns of $\mathbf{M}$) non-integer values in $\boldsymbol{\pi}^*$ at the end of the flight phase. The landing phase consists of rounding to 0 or 1 probabilities of the remaining units. If there are not a lot of remaining units, a solution satisfying exactly the balancing constraints can be found using linear programming. Otherwise, the constraints must be relaxed. One possibility consists of removing balancing variables one by one until a sample satisfying remaining balancing constraints can be selected. This landing phase by suppression of variables requires a priority order on the variables.

Appendix C: The flight phase of the local cube method

Consider a vector of inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$. The flight phase of the local cube method is described in Algorithm 6.

At each step, only $(J+1)$ neighboring units of k are considered. At the end of the flight phase of the cube method, the updated vector of $\boldsymbol{\pi}$ contains mainly 0s and 1s, except for at most J components.

Algorithm 6 Flight phase of the local cube method

Repeat the following steps until there remains less than $(J+1)$ non-integer values in $\boldsymbol{\pi}$.

1. Select $(J+1)$ neighboring units with a non-integer inclusion probability in $\boldsymbol{\pi}$.
2. Apply the flight phase of the cube method only on these units and update $\boldsymbol{\pi}$.

Appendix D: Systematic sampling

Consider a vector of inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$ and suppose that π sums to an integer number n, i.e. $\psi = n$, with $n \in \mathbb{N}$. The usual systematic method is described in Algorithm 7.

Algorithm 7 Systematic sampling

1. Compute the cumulative inclusion probabilities $V_k = \sum_{j \leq k} \pi_j$ with $k = 1, \ldots, N$ and $V^0 = 0$.
2. Generate a uniform continuous random variable u on interval [0,1].
3. Next, for $i = 1, \ldots, n$, select the units $k(i)$ such that $V_{k(i)-1} \leq u + i - 1 < V_{k(i)}$.

Let $V_k = \sum_{j \leq k} \pi_j$ denote the cumulative inclusion probability, with $k = 1, \ldots, N$. Define $v_j \in [0, 1]$ such that $V_j \bmod 1 = v_j$, for $j = 0, \ldots, N-1$. Let also $v_{(j)}$ be the v_js sorted by increasing order with $v_{(N)} = 1$. Each interval $[v_{(k-1)}, v_{(k)}[$ corresponds to the selection of a unique sample and the length of this interval is the probability of selecting this sample, for $t = 1, \ldots, T$. The probability associated to each sample is thus $(v_{(k)} - v_{(k-1)})$.

General Conclusion

This manuscript is mainly devoted to the topic of nonresponse in sample surveys. I was delighted to work on this interesting subject, which is inevitable when working with data. Nonresponse makes it difficult or impossible to use conventional survey methods. We have worked on the adaptation and development of methods while trying to focus on the situations and problems that statisticians are usually confronted with. We study various survey operations that can be affected by nonresponse.

The Swiss Federal Office requires nonresponses to be imputed with observed values in order to avoid implausible results. In the case of nonresponses appearing in several survey variables, the best way to proceed is to impute all the nonresponses of one unit by real values of only one other unit, and not several, to avoid impossible combinations of values. In Chapter 2, we have developed a new imputation method that meets this requirement.

In the case of nonresponse, it is possible to impute but it is also possible to use the nonresponse weighted adjustment to compensate for these unavailable values. When auxiliary information is available, model-assisted estimators are well used for total estimates. In Chapter 3, we have proposed an estimator mixing the model-assisted estimator and the nonresponse weighted adjusted estimator. This estimator adapts the idea of model-assisted estimators to data containing nonresponse.

We have also raised a problem with the efficiency of variance estimators using resampling methods in high-dimensional data sets in Chapter 4. We show an overestimation of the variance estimate of total estimators for resampling-based estimation methods. We have proposed a correction for the existing estimators in order to reduce this overestimation.

At a time when data is ubiquitous in most businesses and new items, I have enjoyed studying nonresponse in different forms and trying to improve existing tools. Nonresponse is not easy to control and can take many different forms, which is why it needs to be fully understood and studied before it can be treated. In Chapter 4, we have highlighted a problem in the presence of nonresponse: when the ratio between the number of fully-observed units and the number of survey variables approaches one. In fact, this problem is also present in the other methods introduced. In Chapter 3, if the set of fully-observed units is limited, it may be difficult to associate a fully-observed unit with each unit containing missing values. In Chapter 4, a model is postulated and estimated between the survey and the auxiliary variables. If the observed data in the survey variable is small, it may be easy to misspecify the linkage model. Note that in this case, the problem is more related to the number of fully-observed units than to the ratio.

Chapter 5 of this manuscript introduces a nonresponse sampling method. We have presented a spatiotemporal sampling method. The two sources of autocorrelation - spatial and temporal - lead to a problem that is difficult to solve in order to avoid information redundancy. Research into this method was also motivated by a real problem, as many institutions require samples to be selected from the same population each year, for example.

All the simulation studies supporting the works in this manuscript were carried out using R software. Methods of Chapter 2 and 5 are respectively available in the R packages Eustache, Jauslin, and Tillé (2020) and Eustache, Vallée, and Tillé (2021).

Bibliography

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

Beaumont, J.-F. (2005). “Calibrated imputation in surveys under a quasi-model-assisted approach”. In: *Journal of the Royal Statistical Society. Series B* 67, pp. 445–458.

Bellhouse, D. R. (1977). “Some optimal designs for sampling in two dimensions”. In: *Biome-trika* 64.3, pp. 605–611.

— (1988). “Systematic sampling”. In: *Sampling*. Vol. 6. Handbook of Statistics. Elsevier, pp. 125–145.

Bellhouse, D. R. and B. C. Sutradhar (1988). “Variance Estimation for Systematic Sampling When Autocorrelation Is Present”. In: *The Statistician: Journal of the Institute of Statisticians* 37, pp. 327–332.

Berger, Y. G. and J. N. K. Rao (2006). “Adjusted jackknife for imputation under unequal probability sampling without replacement”. In: *Journal of the Royal Statistical Society* B68, pp. 531–547.

Berger, Y. G. and C. J. Skinner (2005). “A jackknife variance estimator for unequal probability sampling”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1, pp. 79–89.

Breidt, F. J., G. Claeskens, and J. D. Opsomer (2005). “Model-assisted estimation for complex surveys using penalized splines”. In: *Biometrika* 92, pp. 831–846.

Breidt, F. J. and J. D. Opsomer (2000). “Local polynomial regression estimators in survey sampling”. In: *Annals of Statistics* 28, pp. 1026–1053.

Breidt, F. J. and Jean Opsomer (2017). “Model-assisted survey estimation with modern prediction techniques”. In: *Statistical Science* 32, pp. 190–205.

Breidt, F. J. et al. (2007). “Semiparametric model-assisted estimation for natural resource surveys”. In: *Survey methodology* 33, pp. 35–44.

Brewer, K. R. W. (1963). “Ratio estimation in finite populations: Some results deductible from the assumption of an underlying stochastic process”. In: *Australian Journal of Statistics* 5, pp. 93–105.

Brick, J. Michael (2013). “Unit Nonresponse and Weighting Adjustments: A Critical Review”. In: *Journal of Official Statistics* 29.3, pp. 329–353.

Brick, J. Michael and Michael E. Jones (2008). “Propensity to respond and nonresponse bias”. In: *Metron* 66.1, pp. 51–73.

Campbell, C. (1980). “A different view of finite population estimation”. In: *Proceedings of the Survey Research Methods Section of the American Statistical Accociation*. Baltimore, pp. 319–24.

Cardot, H., C. Goga, and M.-A. Shehzad (2017). “Calibration and Partial Calibration on Principal Components when the Number of Auxiliary Variables is large”. In: *Statistica Sinica* 27.243-260.

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

Chang, T. and P. S. Kott (2008). “Using calibration weighting to adjust for nonresponse under a plausible model”. In: *Biometrika* 95, pp. 555–571.

Chauvet, G. (2009). “Stratified Balanced Sampling”. In: *Survey Methodology* 35, pp. 115–119.

Chauvet, G., J.-C. Deville, and D. Haziza (2011). “On balanced random imputation in surveys”. In: *Biometrika* 98.2, pp. 459–471.

Chauvet, Guillaume and Camelia Goga (2022). “Asymptotic efficiency of the calibration estimator in a high-dimensional data setting”. In: *Journal of Statistical Planning and Inference* 217, pp. 177–187.

Chen, Hua Yun (2010). “Compatibility of conditionally specified models”. In: *Statistics & Probability Letters* 80.7-8, pp. 670–677.

Chen, S et al. (2019). “Pseudo-population bootstrap methods for imputed survey data”. In: *Biometrika* 106.2, pp. 369–384.

Chen, Sixia and David Haziza (2019). “Recent Developments in Dealing with Item Non-response in Surveys: A Critical Review”. In: *International Statistical Review* 87, S192–S218.

Cicchitelli, Giuseppe and Giorgio E. Montanari (2012). “Model-Assisted Estimation of a Spatial Population Mean”. In: *International Statistical Review* 80.1, pp. 111–126.

Da Silva, D. N. and J. D. Opsomer (2009). “Nonparametric propensity weighting for survey nonresponse through local polynomial regression”. In: *Survey Methodology* 35.2, pp. 165–176.

Dagdoug, Mehdi, Camelia Goga, and David Haziza (2022). “Model-Assisted Estimation Through Random Forests in Finite Population Sampling”. In: *To appear in the Journal of the American Statistical Association*, pp. 1–18.

Deville, J.-C. (1998). “La correction de la non-réponse par calage ou par échantillonnage équilibré”. In: *Recueil de la Section des méthodes d’enquêtes des communications présentées au 26ème congrès de la Société Statistique du Canada*. Sherbrooke, pp. 103–110.

— (2002). “La correction de la nonréponse par calage généralisé”. In: *Actes des Journées de Méthodologie Statistique*. Paris: Insee-Méthodes.

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

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

— (2000). “Selection of several unequal probability samples from the same population”. In: *Journal of Statistical Planning and Inference* 86, pp. 215–227.

— (2004). “Efficient balanced sampling: The cube method”. In: *Biometrika* 91, pp. 893–912.

Dickson, Maria Michela and Yves Tillé (2016). “Ordered spatial sampling by means of the traveling salesman problem”. In: *Computational Statistics* 31.4, pp. 1359–1372.

Duchesne, Pierre (2000). “A note on jackknife variance estimation for the general regression estimator”. In: *Journal of Official Statistics* 16.2, p. 133.

Durbin, J. (1959). “A note on the application of Quenouille’s method of bias reduction to the estimation of ratio”. In: *Biometrika* 46, pp. 477–480.

Ekholm, A. and S. Laaksonen (1991). “Weighting via response modeling in the Finish Household Budget Survey”. In: *Journal of Official Statistics* 3, pp. 325–337.

El Karoui, Noureddine and Elizabeth Purdom (2018). “Can we trust the bootstrap in high-dimensions? The case of linear models”. In: *The Journal of Machine Learning Research* 19.1, pp. 170–235.

Eustache, Esther, Mehdi Dagdoug, and David Haziza (2023). “High-dimensional variance estimation in finite population sampling”. Work in progress.

Eustache, Esther and Caren Hasler (2022). “Quasi-Model-Assisted Estimators under Nonresponse in Sample Surveys”. Submitted for publication.

Eustache, Esther, Raphaël Jauslin, and Yves Tillé (2020). *The R Package ‘SpotSampling’*. Vienna: CRAN project.

Eustache, Esther, Raphaël Jauslin, and Yves Tillé (2022). “Spatiotemporal sampling with spatial spreading and rotation of units in time”. In: *Spatial Statistics* 47, p. 100613.

Eustache, Esther, Audrey-Anne Vallée, and Yves Tillé (2021). *The SwissCheese R Package*. R package version beta. URL: `https://github.com/EstherEustache/SwissCheese`.

— (2024). “Balanced Donor Imputation Handling Swiss Cheese Nonresponse”. In: *To appear in Statistica Sinica*, pp. 1–39.

Fattorini, Lorenzo et al. (2021). "Design-based properties of the nearest neighbour spatial interpolator and its bootstrap mean squared error estimator". In: *Biometrics* 113, pp. 463–475.

Fay, R. E. (1991). "A Design-Based Perspective on Missing Data Variance". In: *Proceedings of the 1991 Annual Research Conference*. U.S. Census Bureau, pp. 429–440.

Firth, D. and K. E. Bennett (1998). "Robust Models in Probability Sampling". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 60.1, pp. 3–21. ISSN: 13697412, 14679868.

Fuller, W. A., M. M. Loughin, and H. D. Baker (1994). "Regression weighting in the presence of nonresponse with application to the 1987–1988 nationwide Food Consumption Survey". In: *Survey Methodology* 20, pp. 75–85.

Fuller, Wayne A. (2009). *Sampling statistics*. John Wiley & Sons.

Fuller, Wayne A. and Jae Kwang Kim (2005). "Hot Deck Imputation for the Response Model". In: *Survey Methodology* 31, pp. 139–149.

Graf, M. (2011). "Use of survey weights for the analysis of compositional data". In: *Compositional Data Analysis: Theory and Applications*. Ed. by V. Pawlowsky-Glahn and A. Buccianti. Chichester: Wiley.

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

Grafström, A., N. L. P. Lundström, and L. Schelin (2012). "Spatially balanced sampling through the Pivotal method". In: *Biometrics* 68.2, pp. 514–520.

Grafström, A. and Y. Tillé (2013). "Doubly Balanced Spatial Sampling with Spreading and Restitution of Auxiliary Totals". In: *Environmetrics* 14.2, pp. 120–131.

Grafström, A et al. (2012). "Size constrained unequal probability sampling with a non-integer sum of inclusion probabilities". In: *Electronic Journal of Statistics* 6, pp. 1477–1489.

Grafström, Anton and Niklas L. P. Lundström (2013). "Why well spread probability samples are balanced?" In: *Open Journal of Statistics* 3.1, pp. 36–41.

Grafström, Anton and Alina Matei (2018). "Spatially balanced sampling of continuous populations". In: *Scandinavian Journal of Statistics* 45.3, pp. 792–805.

Grafström, Anton, Svetlana Saarela, and Liviu Theodor Ene (2014). "Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space". In: *Canadian Journal of Forest Research* 44.10, pp. 1156–1164.

Hájek, J. (1971). "Discussion of an essay on the logical foundations of survey sampling, part on by D. Basu". In: *Foundations of Statistical Inference*. Ed. by V. P. Godambe and D. A. Sprott. Toronto, Canada: Holt, Rinehart, Winston, p. 326.

— (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.

Hansen, M. H. and W. N. Hurwitz (1946). "The problem of non-response in sample surveys". In: *Journal of the American Statistical Association* 41, pp. 517–529.

Hasler, Caren (2023). *Inference from Sampling with Response Probabilities Estimated via Calibration*. Tech. rep. University of Neuchâtel.

Hasler, Caren, Radu V. Craiu, and Louis-Paul Rivest (2018). "Vine Copulas for Imputation of Monotone Non-response". In: *International Statistical Review* 86.3, pp. 488–511.

Hasler, Caren and Yves Tillé (2014). "Fast balanced sampling for highly stratified population". In: *Computational Statistics & Data Analysis* 74, pp. 81–94. ISSN: 0167-9473.

— (2016). "Balanced k-nearest neighbour imputation". In: *Statistics* 50.6, pp. 1310–1331.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized additive Models*. Boca Raton: Chapman & Hall/CRC.

Haziza, D. and J. N. K. Rao (2006). "A Nonresponse Model Approach to Inference under Imputation for Missing Survey Data". In: *Survey Methodologyh* 32.1, pp. 59–71.

Haziza, David (2009). "Imputation and Inference in the Presence of Missing Data". In: Handbook of Statistics 29. Ed. by C.R. Rao, pp. 215–246. ISSN: 0169-7161.

Haziza, David and Jean-François Beaumont (2017). "Construction of Weights in Surveys: A Review". In: *Statistical Science* 32.2, pp. 206 –226.

Haziza, David, Sixia Chen, and Yimeng Gao (2022). *Targeting key survey variables at the nonresponse treatment stage*. Tech. rep. Univeristy of Ottawa, University of Oklahoma, University of Montreal.

Haziza, David and Audrey-Anne Vallée (2020). "Variance estimation procedures in the presence of singly imputed survey data: a critical review". In: *Japanese Journal of Statistics and Data Science* 3.2, pp. 583–623.

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

Iachan, R. (1982). "Systematic sampling: A critical review". In: *International Statistical Review* 50, pp. 293–303.

— (1983). "Asymptotic theory of systematic sampling". In: *Annals of Statistics* 11, pp. 959–969.

Iannacchione, V. G., J. G. Milne, and R. E. Folsom (1991). "Response probability weight adjustments using logistic regression". In: *In Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 637–642.

Im, Jongho, In Ho Cho, and Jae Kwang Kim (2018). "FHDI: An R Package for Fractional Hot Deck Imputation". In: *The R Journal* 10.1, pp. 140–154.

Isaki, C. T. and W. A. Fuller (1982). "Survey Design under the Regression Superpopulation Model". In: *Journal of the American Statistical Association* 77, pp. 89–96.

Jauslin, R. and Y. Tillé (2020). "Spatial Spread Sampling Using Weakly Associated Vectors". In: *Journal of Agricultural, Biological, and Environmental Statistics* 25.3, pp. 431–451.

Jauslin, Raphaël, Esther Eustache, and Yves Tillé (July 2021). "Enhanced cube implementation for highly stratified population". In: *Japanese Journal of Statistics and Data Science* 4.2, pp. 783–795.

Johnson, Roger W. (1996). "Fitting Percentage of Body Fat to Simple Body Measurements". In: *Journal of Statistics Education* 4.1, pp. 1–8.

Jonsson, Per and Claes Wohlin (2004). "An evaluation of *k*-nearest neighbour imputation using Likert data". In: *Proceedings of the 10th International Symposium on Software Metrics*. Chicago, pp. 108–118.

Judkins, D. R. (1997). "Imputing for swiss cheese patterns of missing data". In: *Proceedings of Statistics Canada Symposium*. Statistics Canada, p. 97.

Karoui, Noureddine El and Holger Koesters (2011). "Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods". In: *arXiv: Statistics Theory*.

Khavarzadeh, Ramin, Mohsen Mohammadzadeh, and Jorge Mateu (2018). "A simple two-step method for spatio-temporal design-based balanced sampling". In: *Stochastic environmental research and risk assessment* 32.2, pp. 457–468.

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

Kim, J. K. and J.J. Kim (2007). "Nonresponse weighting adjustment using estimated response probability". In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 35.4, pp. 501–514.

Kim, J. K. and H. Park (2006). "Imputation using response probability". In: *Canadian Journal of Statistics* 34.1, pp. 171–182.

Kim, J. K. and J.N.K. Rao (2009). "A unified approach to linearization variance estimation from survey data after imputation for item nonresponse". In: *Biometrika* 96.4, pp. 917–932.

Kim, J. K. and M. K. Riddles (2012). "Some theory for propensity-score-adjustment estimators in survey sampling". In: *Survey Methodology* 38.2, pp. 157–165.

Kim, Jae Kwang and David Haziza (2014). "Doubly Robust Inference With Missing Data in Survey Sampling". In: *Statistica Sinica* 24.1, pp. 375–394. ISSN: 10170405, 19968507.

Kott, P. S. (1994). “A note on handling nonresponse in surveys”. In: *Journal of the American Statistical Association* 89.426, pp. 693–696.

— (2006). “Using calibration weighting to adjust for nonresponse and coverage errors”. In: *Survey Methodology* 32.2, pp. 133–142.

Kott, P. S. and T. Chang (2010). “Using calibration weighting to adjust for nonignorable unit nonresponse”. In: *Journal of the American Statistical Association* 105.491, pp. 1265–1275. ISSN: 0162-1459.

Kott, Phillip S. and Dan Liao (2012). “Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine”. In: *Survey Research Methods* 6.2, pp. 105–111.

Lee, H., E. Rancourt, and C.-E. Särndal (2002). “Variance estimation from survey data under single imputation”. In: *In Survey Nonresponse*. Ed. by R. M. Groves et al. New York: Wiley, pp. 315–328.

Lee, Kok-Huat (1973). “Variance Estimation in Stratified Sampling”. In: *Journal of the American Statistical Association* 68.342, pp. 336–342. ISSN: 01621459.

Lesage, Éric, David Haziza, and Xavier d’Haultfoeuille (2019). “A Cautionary Tale on Instrumental Calibration for the Treatment of Nonignorable Unit Nonresponse in Surveys”. In: *Journal of the American Statistical Association* 114.526, pp. 906–915.

Lohr, Sharon L (2021). *Sampling: design and analysis*. Chapman and Hall/CRC.

Lundström, S. and C.-E. Särndal (1999). “Calibration as a standard method for treatment of nonresponse”. In: *Journal of Official Statistics* 15, pp. 305–327.

Madow, W. G. (1949). “On the theory of systematic sampling, II”. In: *Annals of Mathematical Statistics* 20, pp. 333–354.

Mashreghi, Zeinab, David Haziza, and Christian Léger (2016). “A survey of bootstrap methods in finite population sampling”. In: *Statistics Surveys* 10.none.

Moran, Patrick A. P. (1950). “Notes on continuous stochastic phenomena”. In: *Biometrika* 37.1/2, pp. 17–23.

Murray, Jared and Jerome P. Reiter (2014). “Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models With Local Dependence”. In: *Journal of the American Statistical Association* 111, pp. 1466–1479.

Nedyalkova, Desislava and Yves Tillé (2008). “Optimal sampling and estimation strategies under linear model”. In: *Biometrika* 95, pp. 521–537.

Neyman, J. (1934). “On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection”. In: *Journal of the Royal Statistical Society* 97, pp. 558–606.

Niyonsenga, T. (1994). “Nonparametric estimation of response probabilities in sampling theory”. In: *Survey Methodology* 20.2, pp. 177–184.

— (1997). “Response probability estimation”. In: *Journal of Statistical Planning and Inference* 59, pp. 111–126.

Opsomer, Jean et al. (Feb. 2007). “Model-Assisted Estimation of Forest Resources With Generalized Additive Models”. In: *Journal of the American Statistical Association* 102, pp. 400–409.

Pajor, Alain and Leonid Pastur (2009). “On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution”. In: *Studia Mathematica* 195.1, pp. 11–29.

Pea, Johan, Lionel Qualité, and Yves Tillé (2007). “Systematic sampling is a minimal support design”. In: *Computational Statistics & Data Analysis* 51, pp. 5591–5602.

Pfeffermann, Danny and Michail Sverchkov (2009). “Inference under informative sampling”. In: *Handbook of statistics*. Vol. 29. Elsevier, pp. 455–487.

Portnoy, Stephen (1987). “A central limit theorem applicable to robust regression estimators”. In: *Journal of multivariate analysis* 22.1, pp. 24–50.

Quenouille, M. H. (1949a). “Approximation tests of correlation in time series”. In: *Journal of the Royal Statistical Society* B11, pp. 18–84.

Quenouille, Maurice H. (1949b). “Problems in plane sampling”. In: *The Annals of Mathematical Statistics* 20, pp. 355–375.

Raghunathan, Trivellore E. et al. (2001). “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models”. In: *Survey Methodology* 27.1, pp. 85–95.

Rao, J. N. K. and J. Shao (1992). “Jackknife variance estimation with survey data under hot-deck imputation”. In: *Biometrika* 79, pp. 811–822.

Rivest, Louis-Paul and Sergio Ewane Ebouele (2020). “Sampling a two dimensional matrix”. In: *Computational Statistics & Data Analysis* 149, p. 106971.

Robinson, P. M. and C.-E. Särndal (1983). “Asymptotic properties of the generalized regression estimator in probability sampling”. In: *Sankhyā* B45, pp. 240–248.

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

Rubin, D. B. (1976). “Inference and missing data”. In: *Biometrika* 63, pp. 581–592.

Sang, Hejian, Jae Kwang Kim, and Danhyang Lee (2022). “Semiparametric Fractional Imputation Using Gaussian Mixture Models for Handling Multivariate Missing Data”. In: *Journal of the American Statistical Association* 117, pp. 654–663.

Särndal, C.-E. (1978). “Design-based and model-based inference in survey sampling”. In: *Scandinavian Journal of Statistics* 5, pp. 27–52.

— (1980). “On π-inverse weighting versus best linear unbiased weighting in probability sampling”. In: *Biometrika* 67, pp. 639–650.

— (1992). “Methods for estimating the precision of survey estimates when imputation has been used”. In: *Survey Methodology* 18.2, pp. 241–252.

— (2007). “The calibration approach in survey theory and practice”. In: *Survey Methodology* 33.2, pp. 99–119.

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

Särndal, C.-E. and B. Swensson (1987). “A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse”. In: *International Statistical Review* 55.3, pp. 279–294.

Särndal, C.-E., B. Swensson, and J. H. Wretman (1989). “The weighted residual technique for estimating the variance of the general regression estimator of the finite population total”. In: *Biometrika* 76, pp. 527–537.

— (1992). *Model Assisted Survey Sampling*. New York: Springer.

Särndal, C.-E. and R. L. Wright (1984). “Cosmetic form of estimators in survey sampling”. In: *Scandinavian Journal of Statistics* 11, pp. 146–156.

Shao, J. and P. Steel (1999). “Variance estimation for survey data with composite imputation and nonneglible sampling fractions”. In: *Journal of the American Statistical Association* 94, pp. 254–265.

Stekhoven, Daniel J. and Peter Bühlmann (Oct. 2011). “MissForest—non-parametric missing value imputation for mixed-type data”. In: *Bioinformatics* 28.1, pp. 112–118.

Stevens, Don L. and Anthony R. Olsen (2004). “Spatially Balanced Sampling of Natural Resources”. In: *Journal of the American Statistical Association* 99.465, pp. 262–278.

Stevens, Don L. Jr. and Anthony R. Olsen (2003). “Variance Estimation for Spatially Balanced Samples of Environmental Resources”. In: *Environmetrics* 14.6, pp. 593–610.

Stevens Jr., Don L. and Anthony R. Olsen (1999). “Spatially restricted surveys over time for aquatic resources”. In: *Journal of Agricultural, Biological, and Environmental Statistics* 4, pp. 415–428.

Ta, Tram et al. (2020). “Generalized Regression Estimators with High-Dimensional Covariates”. In: *Statistica Sinica* 30.3, pp. 1135–1154.

Tillé, Yves (2020). *Sampling and Estimation From Finite Populations*. Hoboken: Wiley.

Tillé, Yves and Klaus Ecker (2013). “Complex national sampling design for long-term monitoring of protected dry grasslands in Switzerland”. English. In: *Environmental and Ecological Statistics* 21, pp. 1–24. ISSN: 1352-8505.

Tillé, Yves et al. (2018). “Measuring the spatial balance of a sample: A new measure based on the Moran’s I index”. In: *Spatial Statistics* 23, pp. 182–192.

Tukey, J. W. (1958). “Bias and confidence in not quiet large samples”. In: *Annals of Mathematical Statistics* 29, p. 614.

Valliant, R., A. H. Dorfman, and R. M. Royall (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

Valliant, Richard (2002). “Variance estimation for the general regression estimator”. In: *Survey methodology* 28.1, pp. 103–108.

van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC, Boca Raton.

van Buuren, S. et al. (2006). “Fully conditional specification in multivariate imputation”. In: *Journal of Statistical Computation and Simulation* 76.12, pp. 1049–1064.

Wang, Jin-Feng, Tong-Lin Zhang, and Bo-Jie Fu (2016). “A measure of spatial stratified heterogeneity”. In: *Ecological Indicators* 67, pp. 250–256.

Wang, Jinfeng, Bingbo Gao, and Alfred Stein (2020). “The spatial statistic trinity: A generic framework for spatial sampling and inference”. In: *Environmental Modelling & Software* 134, p. 104835.

Wang, Zhonglei and Zhengyuan Zhu (2019). “Spatiotemporal Balanced Sampling Design for Longitudinal Area Surveys”. In: *Journal of Agricultural, Biological and Environmental Statistics* 24.2, pp. 245–263.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. Second. New York: Springer.

Yang, Shu and Jae Kwang Kim (2016). “Fractional Imputation in Survey Sampling: A Comparative Review”. In: *Statistical Science* 31.3, pp. 415 –432.

Yung, W and JNK Rao (1996). “Jackknife linearization variance estimators under stratified multistage sampling”. In: *Survey Methodology* 22, pp. 23–32.

Zhao, Qian and Emmanuel J Candes (2022). “An Adaptively Resized Parametric Bootstrap for Inference in High-dimensional Generalized Linear Models”. In: *arXiv preprint arXiv:2208.08944*.

Zhao, Xin and Anton Grafström (2020). “A sample coordination method to monitor totals of environmental variables”. In: *Environmetrics* 31.6.