New methods to handle nonresponse in surveys

PhD Thesis submitted to the Faculty of Science
Institute of Statistics
University of Neuchâtel

For the degree of PhD in Science

by

Caren Hasler

Accepted by the dissertation committee:

Prof. Pascal Felber, jury president, Université de Neuchâtel
Prof. Yves Tillé, thesis director, Université de Neuchâtel
Prof. Anne Ruiz-Gazen, Université de Toulouse 1 Capitole
Prof. David Haziza, Université de Montréal
Prof. Isabel Molina, Universidad Carlos III de Madrid

Thesis defended on August 24, 2015

UNIVERSITÉ DE NEUCHÂTEL

Faculté des sciences
Secrétariat-décanat de Faculté
Rue Emile-Argand 11
2000 Neuchâtel - Suisse
Tél: + 41 (0)32 718 2100
E-mail: secretariat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

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

Madame Caren HASLER

Titre:

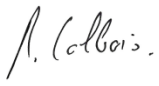
“New methods to handle nonresponse in surveys”

sur le rapport des membres du jury composé comme suit:

- Prof. Yves Tillé, Université de Neuchâtel, directeur de thèse
- Prof. Pascal Felber, Université de Neuchâtel
- Prof. Anne Ruiz-Gazen, Université de Toulouse 1 Capitole, France
- Prof. David Haziza, Université de Montréal, Canada
- Prof. Isabel Molina, Universidad Carlos III de Madrid, Espagne

Neuchâtel, le 25 août 2015

Le Doyen, Prof. B. Colbois



Acknowledgements

I would like to convey my gratitude to my thesis director, Prof. Yves Tillé, for the help and guidance he gave me throughout the realization of this thesis. He has always been available to answer my questions and has provided me with encouraging support. During these past few years, not only I have learned a lot from Pr. Yves Tillé, but also his creativity and intuition have motivated my work. Yves, I'm honored to have worked with you and I'm grateful for all what you have taught me.

The realization of this thesis would not have been possible without the scientific guidance and work of co-authors, Dr. Alina Matei and Prof. Radu V. Craiu, who I would like to thank warmly. Working with them has challenged my ideas and has significantly broadened my knowledge.

I'm thankful to Prof. Pascal Felber to have accepted to act as president of the jury committee, and to Prof. Anne Ruiz-Gazen, to Prof. David Haziza, and to Prof. Isabel Molina to have accepted to be members of the aforementioned committee. I also would like to acknowledge all of them for the time and energy they spent reviewing this work.

This thesis has been financially supported by the Swiss National Science Foundation (project number P1NEP2_151904) and by the Swiss Federal Statistical Office. I'm grateful for their contribution.

I also would like to thank the colleagues and former colleagues of the Institute of statistics for the supportive and warm work environment. Special thanks to my former office mate Alina and to Erika for their friendship, for their understanding, and for the philosophical discussions on statistics and on life in general.

I owe my gratitude to my friends who contributed to my emotional well-being throughout the accomplishment of this work. I feel lucky to have them in my life. I'm particularly thankful to Anaïs, Audrey, Cindy, Jessalynn, Leïla, and Morgane for maintaining close relationships despite the distance that separates us.

My parents and my brother have encouraged and supported me in all what I have undertaken since I was born, including this thesis. Maman, Papa, Pieric, I cannot thank you enough.

Last but not least, I would like to express my profound gratitude to my husband, Michael, whose love and understanding have helped me going through the ups and downs of the completion of this thesis. From the very beginning, he has encouraged me, comforted me, and advised me. Thank you Michael for your support and for being part of my life.

Toronto, June 29, 2015.

Abstract

This document focuses on nonresponse in sample surveys. Mainly, methods to handle nonresponse in complex surveys are proposed. The first chapter of this document introduces concepts and notation of survey sampling and nonresponse. The second chapter proposes an algorithm for stratified balanced sampling for populations with large numbers of strata. The third chapter of this document presents a hot-deck imputation method which combines balanced sampling and a nonparametric approach. This method uses the algorithm presented in the second chapter. The next chapter presents a nonparametric method of imputation for item nonresponse in surveys based on additive regression models. Finally, the fifth chapter proposes three reweighting procedures for handling nonignorable nonresponse in surveys providing that the values of the variable of interest are obtained from a mixture distribution.

Keywords: survey sampling, missing data, imputation, reweighting, nonignorable nonresponse, balanced sampling, stratified sampling.

Resumé

Ce document porte sur la nonréponse dans les enquêtes par échantillonnage. Principalement, des méthodes de traitement de la nonréponse dans des enquêtes complexes sont proposées. Le premier chapitre de ce document introduit des concepts relatifs à l'échantillonnage et à la nonréponse. Le second chapitre propose un algorithme d'échantillonnage équilibré pour des populations hautement stratifiées. Le troisième chapitre de ce document propose une méthode d'imputation par donneur dont la sélection se fait par échantillonnage équilibré combiné à une approche nonparamétrique. Cette méthode nécessite l'utilisation de l'algorithme faisant l'objet du second chapitre. Le chapitre qui suit présente une méthode d'imputation nonparamétrique basée sur les modèles de regression additifs. Finalement, le cinquième chapitre propose trois procédures de repondération pour le traitement de la nonréponse non-ignorable applicable lorsque les valeurs prises par la variable d'intérêt proviennent d'une densité mélange.

Mots-clés: échantillonnage, donnés manquantes, imputation, repondération, nonréponse non-ignorable, échantillonnage équilibré, échantillonnage stratifié.

Contents

List of Figures xiii

List of Tables xv

Introduction 1

1 An introduction to finite population sample surveys and nonresponse 5
- 1.1 General considerations . . . 5
- 1.2 The complete response case . . . 6
- 1.3 Nonresponse in the survey . . . 8
  - 1.3.1 Three types of nonresponse mechanisms . . . 9
  - 1.3.2 Two levels of nonresponse and two handling approaches 9
  - 1.3.3 Reweighting for unit nonresponse . . . 10
  - 1.3.4 Imputation for item nonresponse . . . 13

2 Fast balanced sampling for highly stratified populations 17
- 2.1 Introduction . . . 17
- 2.2 Balanced sampling . . . 19
- 2.3 Chauvet's method for stratified balanced sampling . . . 21
- 2.4 New procedure for highly stratified balanced sampling . . . 23
- 2.5 Case where the sum of the inclusion probabilities is not an integer in each stratum . . . 26
- 2.6 Variance estimation . . . 28
- 2.7 Illustration of the handling of nonresponse . . . 29
  - 2.7.1 Nonresponse and imputation . . . 29
  - 2.7.2 Notation . . . 30
  - 2.7.3 Balanced random imputation to eliminate the imputation variance . . . 31
  - 2.7.4 Stratified balanced sampling for balanced random imputation . . . 32
- 2.8 Simulation study . . . 33
  - 2.8.1 Performance of the proposed algorithm . . . 34
  - 2.8.2 Variance approximation formula and estimator . . . 36
  - 2.8.3 Illustration of the handling of nonresponse . . . 37
- 2.9 Conclusion . . . 38

3 Balanced k-nearest neighbor imputation **41**
3.1 Introduction . . . 41
3.2 Notation and concepts of nonresponse . . . 43
3.3 Methodology for random hot-deck imputation methods . . . 45
3.4 Balanced k-nearest neighbor imputation method . . . 47
3.4.1 Aim of the method . . . 47
3.4.2 Calibration . . . 49
3.4.3 Obtaining the matrix of imputation probabilities $\psi^{[bk]}$ . . . 50
3.4.4 Choice of k and existence of $\psi^{[bk]}$ . . . 51
3.4.5 Stratified balanced sampling . . . 52
3.4.6 Selection of the donors . . . 53
3.5 Approximation of conditional imputation variance . . . 56
3.6 Properties of the imputed total estimator . . . 58
3.6.1 Linear model . . . 58
3.6.2 Response model . . . 58
3.6.3 Neighborhood principle . . . 59
3.6.4 Resistance to model misspecification . . . 59
3.6.5 Asymptotic properties of the total estimator . . . 60
3.7 Simulation study . . . 61
3.7.1 The data . . . 61
3.7.2 Simulation settings . . . 61
3.7.3 Measures of comparison . . . 62
3.7.4 Results of the simulations . . . 63
3.8 Conclusion . . . 65

4 Nonparametric imputation for nonresponse in surveys **67**
4.1 Introduction . . . 67
4.2 Framework . . . 68
4.3 Motivation . . . 70
4.4 Nonparametric tools . . . 71
4.5 The method . . . 73
4.5.1 Estimation and imputation . . . 73
4.5.2 Variance estimation for the imputed total . . . 74
4.6 Simulations . . . 75
4.6.1 Setting 1: simulated data . . . 76
4.6.2 Setting 2: real data . . . 79
4.6.3 Measures of comparison . . . 80
4.6.4 Results of setting 1 . . . 82
4.6.5 Results of setting 2 . . . 83
4.7 Conclusion . . . 84

5 Weighting adjustment for nonignorable nonresponse with a heterogeneous structure of the variable of interest **87**
5.1 Introduction 87
5.2 Framework 89
5.3 Estimating response probabilities 91
5.4 Proposed procedures 93
5.4.1 Reconstruction of latent components 93
5.4.2 The proposed solutions 95
5.5 Variance estimation 97
5.6 Simulations 98
5.7 Application to real data 106
5.8 Conclusion 110

Conclusion **113**

Appendix A Proof of Property 2.1 **117**

Appendix B Proofs of the properties of Chapter 3 **119**

Appendix C Proof of Proposition 3.1 **121**

Bibliography **125**

List of Figures

4.1 Comparison measures of four imputation methods in five populations under SRSWOR. 82
4.2 Comparison measures of four imputation methods in five populations under stratified sampling. 83

5.1 Left panel: scatter plot of the auxiliary variable x against the variable of interest y. Right panel: density estimate of y. 105
5.2 Density estimates of the income from the register (dashed curves) and gaussian densities of three components (solid curves) for sampled units (top panel) and for respondents only (bottom panel). 108

List of Tables

2.1 Correlations between the variables of interest and the balancing variables. 35

2.2 Ratio of the variance of the estimated total of the variables of interest obtained using the new method (Algorithm 2.2) to the variance of the estimated total of the variables of interest obtained using Chauvet's method with step 3 by landing phase by suppression of variables (Algorithm 2.1). 35

2.3 Mean time in seconds and failure rate of selection of a sample with Chauvet's method with step 3 by landing phase by suppression of variables (Algorithm 2.1) and with the new method (Algorithm 2.2) for 25, 50, 100, 250, 500, and 1,000 strata and 1 unit selected in each stratum with equal inclusion probabilities. 36

2.4 Approximated variance, mean of the variance estimator estimated using 10,000 simulations, and variance obtained by 10,000 simulations in the case of the estimation of the total of 4 variables of interest using the new method (Algorithm 2.2). Three cases are considered, namely the selection of samples of size $n = 50, 100, 200$ respectively. 37

2.5 Relative root imputation variance (RRIV) of the imputed estimator for a vector of domain means obtained through balanced random imputation using the new method (Algorithm 2.2). 38

3.1 Monte Carlo relative bias (RB), Monte Carlo relative root mean square error (RRMSE), and Monte Carlo relative root imputation variance (RRIV) for the total estimation, the 10-th percentile estimation, the 90-th percentile estimation, and the variance estimation of the variable of interest $\mathbf{y}$ in Case 1. 64

3.2 Monte Carlo relative bias (RB), Monte Carlo relative root mean square error (RRMSE), and Monte Carlo relative root imputation variance (RRIV) for the total estimation, the 10-th percentile estimation, the 90-th percentile estimation, and the variance estimation of the variable of interest $\mathbf{y}$ in Case 2. 65

3.3 Average over the simulations of the approximated conditional imputation variance of Expression (3.24), Monte Carlo imputation variance of the total and ratio of these two quantities in two different cases. 66

4.1 Average ranks over five populations of each imputation method for each measure of comparison (in absolute value). 84
4.2 Monte Carlo variance of the total, Monte carlo expectation of the bootstrap variance and coverage rate associated with AM imputation for two different sampling designs and five populations. 85
4.3 Comparison measures for four imputation methods for FES data. 85
4.4 Monte Carlo variance of the total, Monte carlo expectation of the bootstrap variance and coverage rate associated with AM imputation for FES data. 86

5.1 Comparison measures of estimators in Setting 1. 102
5.2 Comparison measures of estimators in Setting 2. 103
5.3 Comparison measures of estimators in Setting 3. 105
5.4 Results on SILC 2009 data. 110

Introduction

A *survey* is a statistical method applied to study the characteristics of a population by examining only a part of this one called a *sample*. In contrast with a survey, a *census* is an exhaustive study of the characteristics of a population. *Nonresponse* occurs when the desired information is only observed for a part of the sample and represents one of the sources of error the produced statistics are subject to. Nonresponse has two main consequences on the data. First, because the number of observations is less than initially envisaged, nonresponse increases the variance of estimations. Second, nonresponse introduces a bias in the estimations if the recorded characteristics differ between respondents and nonrespondents.

Because official statistics are used within the decision-making process of authorities, they play an important role in the functioning of our society. The quality of the statistics produced by governmental agencies and other public agencies is therefore of fundamental importance. The control of the different sources of error in the survey contributes to this quality. As source of error, nonresponse is critical to this quest for quality and undeniably has to be given attention.

This document addresses the problem of nonresponse from a data processing and estimation point of view. Typically, we consider the perspective of a statistician who is given a sample survey data file containing missing values with the task of producing point and variance estimation. Imputation and reweighting procedures are proposed. This document does not cover any of the other aspects of nonresponse. For instance, we will not look at factors that influence nonresponse such as: the length or difficulty of items of the questionnaire, the method applied to collect the data, the period the data is collected, the nature of the subject of interest, or the characteristics of the interviewers. Neither will we focus on nonresponse follow-up procedures, which not only increase the response rate but also turn out to be useful when handling nonresponse by helping assess the similarities and dissimilarities between respondents and nonrespondents.

This document is organized as follows. Chapter 1 proposes an overview of nonresponse in finite population sample surveys. It establishes the general framework of the research papers presented in this document. Chapters 2 to 5 are self-contained papers submitted or published in peer-reviewed journals that have been developed in collaboration with different co-authors.

Chapter 2 is a reprint of Hasler and Tillé (2014) and presents an algorithm for stratified balanced sampling which is very fast, regardless of the number of strata. This algorithm turns out to be valuable for the purpose of the imputation method presented in Chapter 3, as well as for many other applications, such as some large-scale surveys. In this chapter, we propose a variance estimator for the total and we illustrate one of the possible applications of the proposed algorithm.

Chapter 3 was co-written with Professor Yves Tillé and proposes a new random hot-deck imputation method. This method combines balanced sampling and a nonparametric approach. This results in an unbiased total estimator under very different models, providing protection against model misspecifiation. Moreover, the proposed method produces negligible imputation variance under specified hypotheses. In this chapter, we also suggest a formula to approximate the imputation variance of the total estimator, we describe the underlying models associated with the proposed method, and we study the asymptotic properties of the total estimator.

Chapter 4 was co-written with Professor Radu V. Craiu and presents a nonparametric method of imputation for item nonresponse in surveys. We consider smoothing splines models within an additive regression framework. This allows us to include a large number of auxiliary variables and to take advantage of some strong auxiliary information available. Because this method is nonparametric, it is very flexible and therefore provides protection against model misspecification in a wide range of problems. This chapter also suggests a bootstrap procedure to estimate the variance of the total.

Chapter 5 was developed in collaboration with Doctor Alina Matei and proposes three reweighting procedures for handling nonignorable nonresponse in surveys. We assume that the values of the variable of interest are sampled from a superpopulation which can be described as a mixture of some hidden components or subpopulations. The response probabilities are modeled through logistic regression; the component structure of the variable of interest is considered in the model. The estimated response probabilities are used in a two-phase estimator of the population mean and an estimator of the variance of the mean is suggested.

This document closes with a general conclusion. Appendices contain technical elements relative to Chapter 2 and to Chapter 3.

1 An introduction to finite population sample surveys and nonresponse

Abstract

This chapter gives a brief overview of nonresponse in finite population sample surveys. It establishes the general framework of the research papers presented in Chapters 3 to 5. After general considerations proposed in Section 1.1, Section 1.2 is devoted to estimation in the complete response case. Section 1.3 addresses nonresponse in the survey and is subdivided as follows. After having defined general concepts and notation of nonresponse, three types of nonresponse mechanisms are described in Section 1.3.1. Then, Section 1.3.2 defines two levels of nonresponse and establishes two general handling approaches: reweighting procedures, further discussed in Section 1.3.3, and imputation, further discussed in Section 1.3.4.

Keywords: nonresponse mechanism, level of nonresponse, reweighting procedure, imputation.

1.1 General considerations

A *sample survey*, often shortened *survey*, is a statistical method applied to study the characteristics of a population by examining only a part of this one called a *sample*. By contrast, a *census survey*, often shortened *census*, is an exhaustive examination of the population. Because surveys require reduced cost and time of work compared to censuses, they represent an interesting option.

This document is concerned with finite population sampling, which studies the selection process of a sample in a population of finite size. In the finite population framework, all the units of the population are identifiable, which is not the case in the more classical infinite population framework. This particularity requires specific estimation tools, some of which being presented in this chapter.

A sampling process can be either probabilistic or non-probabilistic. It is referred to as a *probabilistic sampling* when the units are selected according to a random scheme and it is referred to as a *non-probabilistic sampling* otherwise. Probabilistic sampling is usually preferred by statisticians because, since units

are randomly selected, properties such as the estimated variability or bias are available. With non-probabilistic sampling, however, such properties are unavailable. They are nevertheless circumstances in which probabilistic sampling is inapplicable, which explains the popularity of non-probabilistic sampling. This document focuses on probabilistic sampling and all the tools developed are limited to this context.

The central notion of this document is *nonresponse*, which means a failure to obtain responses from the sample; it corresponds to a missing data problem in the survey sampling framework. Despite all the actions taken to increase the response rate, nonresponse impairs most surveys. In contrast to the term *nonresponse*, the term *complete nonresponse* (or *full response*) means a success to obtain all the responses from the sample.

Finally, a notion which plays a central role in the nonresponse framework and which will appear throughout this document is *auxiliary information*. This general notion includes any information not directly linked to the survey. Examples of auxiliary information are, but not limited to: the population total of a variable, the mean in a domain of a variable, or a variable with values known for all the population units or known for all the sampled units. Auxiliary information is used not only at the nonresponse treatment stage of the survey to reduce nonresponse error, but also at the design stage to improve the efficiency of sampling and at the estimation stage to construct accurate estimates. At the nonresponse treatment stage, strong auxiliary information explains the variability in the variable on interest, the variability in the nonresponse process, or, ideally, both simultaneously.

1.2 The complete response case

We consider a population $U = \{1, 2, \ldots, i, \ldots, N\}$ of finite size N, where index i denotes a generic unit of the population. Let y be a variable of interest and let y_i represent the value of the variable interest taken by unit i. The goal of the survey is to estimate a parameter of interest θ, which is a function of the values y_i, $i = 1, \ldots, N$ of the variable of interest. Examples of parameters of interest are, but not limited to: a population mean, a domain mean, a population quantile, a variance, or a regression coefficient. A common parameter of interest is the population total

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

A sample s is randomly selected without replacement from population U using a probability distribution $p(\cdot)$ called *sampling design*. A sampling design is a

probability distribution over all the possible samples in a population, i.e. a function $p(\cdot)$ satisfying

- $p(s) \geq 0$ for all $s \in \mathcal{S}$,
- $\sum_{s \in \mathcal{S}} p(s) = 1$,

where $\mathcal{S}$ is the set of all the possible samples in the population. The sample size, a random quantity, is the number of units contained in sample s. It is denoted by n for convenience but is understood as $n(s)$ as it can differ from one sample to another. The *first order inclusion probability* π_i of unit i is the probability that unit i appears in the selected sample, that is

$$\pi_i = \sum_{\substack{s \in \mathcal{S} \\ s \ni i}} p(s) = \Pr(i \in s) = \Pr(\mathrm{I}_i = 1),$$

where I_i is the sample membership indicator

$$\mathrm{I}_i = \begin{cases} 1 & \text{if unit } i \text{ is selected in the sample,} \\ 0 & \text{otherwise.} \end{cases}$$

It is supposed that a vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})^\top$ of q auxiliary variables is known for each population unit or at least for each sampled unit. The values y_i of the variable of interest are recorded for each sampled unit. In the case of complete response, the Horvitz-Thompson estimator (Horvitz and Thompson, 1952)

$$\widehat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i},$$

is a design-unbiased estimator of the population total Y in the sense that, if $\pi_i > 0$ for all $i \in U$, it satisfies $\mathrm{E}_p(\widehat{Y}) = Y$, where E_p is the expectation with respect to the sampling design.

Before moving to nonresponse in the survey, let us briefly present *two-phase sampling*, a notion which is, as we will mention below, closely related to non-response. Two-phase sampling consists of a double sampling and generalizes the simple one phase sampling described above. In a first phase, an initial sample s_1 is selected from population U with a sampling design $p_1(\cdot)$ and, in a second phase, a subsample s_2 is selected from the first phase sample s_1 with a sampling design $p_2(\cdot|s_1)$. A two-phase sampling design is useful when there is unavailable or limited auxiliary information at the population level. In such circumstances, the first phase sampling is applied and auxiliary information is collected at the first phase sample level. The collected information then enhances the quality of estimates obtained from the second phase sampling. The values of the variable of interest are recorded for the units in the second phase sample s_2. Consider the first and second phase inclusion probabilities

$\pi_{1i} = \Pr(i \in s_1)$ and $\pi_{2i} = \Pr(i \in s_2 | i \in s_1; s_1)$, respectively. Note that the second phase inclusion probabilities are random variables since they depend on the observed values of the first phase sample. The double expansion estimator

$$\widehat{Y}_{DE} = \sum_{i \in s_2} \frac{1}{\pi_{1i}} \frac{1}{\pi_{2i}} y_i \tag{1.1}$$

generalizes the Horvitz-Thompson estimator $\widehat{Y}$ to two-phase sampling. This estimator is design unbiased for the population total Y in the sense that, if $\pi_{1i}, \pi_{2i} > 0$ for all $i \in U$, $\mathrm{E}_p(\widehat{Y}_{DE}) = \mathrm{E}_1\mathrm{E}_2(\widehat{Y}_{DE}|s_1) = Y$, where E_1 and E_2 are the expectations with respect to the first and second phase sampling designs, respectively.

1.3 Nonresponse in the survey

Nonresponse is a failure to obtain responses from the sample, which partitions the sample into two subsets: the set of survey respondents and the set of survey nonrespondents. A sampled unit i is a *survey respondent* (or *respondent*) if its value y_i of the variable of interest is observed; it is a *survey nonrespondent* (or *nonrespondent*) otherwise.

Nonresponse is a random process which represents a second phase of the survey: a sample of survey respondents is randomly selected from the initial sample. For a sampled unit i, the response indicator variable r_i, defined as

$$r_i = \begin{cases} 1 & \text{if unit } i \text{ is a survey respondent,} \\ 0 & \text{otherwise,} \end{cases}$$

is a random quantity to be understood as the second phase sample membership indicator. The *response probability* of a sampled unit i is the probability p_i that the unit is a survey respondent, that is

$$p_i = \mathrm{P}(r_i = 1 | i \in s; s).$$

The *nonresponse mechanism* is the probability distribution $q(\cdot|s)$, where $q(s_r|s)$ is the probability of observing the set $s_r = \{i \in s | r_i = 1\}$ of survey respondents. The nonresponse mechanism describes the process that generates nonresponse. By analogy with two-phase sampling, the response probabilities p_i represent the inclusion probabilities of the second phase, the set s_r of survey respondents is the second phase sample, and the nonresponse mechanism $q(\cdot|s)$ corresponds to the second phase sampling design. The similarity between nonresponse and two-phase sampling is further discussed in Section 1.3.3.

In this document, we suppose that the units respond independently of one

another. In this case, the nonresponse mechanism is described as follows: the response indicator r_i is generated from a Bernoulli random variable with parameter p_i and the set of survey respondents is selected with a conditional poisson sampling design, that is with the following probability distribution

$$q\left(s_r|s\right) = \prod_{i \in s} p_i^{r_i} \left(1 - p_i\right)^{1 - r_i}.$$

1.3.1 Three types of nonresponse mechanisms

Three types of nonresponse mechanisms are distinguished: uniform, ignorable, and non-ignorable. A *uniform nonresponse mechanism* is a nonresponse mechanism where each unit of the population has the same response probability $p_i = p$. It is in this case said that the data is missing completely at random (MCAR).

A nonresponse mechanism is *ignorable* if the response probability p_i does not depend on the variable of interest once having taken into account any appropriate auxiliary information. The response indicator variable satisfies $\Pr(r_i = 1|\mathbf{x}, y) = \Pr(r_i = 1|\mathbf{x})$. In the case of an ignorable nonresponse mechanism, the data is said to be missing at random (MAR, see Rubin, 1976). A uniform nonresponse mechanism is a particular case of ignorable nonresponse mechanism.

Finally, a *non-ignorable nonresponse mechanism* is a nonresponse mechanism which is not ignorable. There is a direct link between the response probability p_i and the variable of interest, and this link still holds once having taken into account any appropriate auxiliary information. The data is in this case said to be not missing at random (NMAR).

Rubin (1976) set the foundation of the concept of ignorability. He defined conditions under which the process that causes the missing data can be ignored. Later, Little (1982) outlined models and provided a basis for the understanding of the effect of non-ignorable nonresponse on the survey estimates. The problems and estimation techniques associated with non-ignorable nonresponse mechanism have then become the subject of much research. Some of the the first research articles that tackled non-ignorable nonresponse mechanisms are Greenlees et al. (1982); Little (1983); Fay (1986).

1.3.2 Two levels of nonresponse and two handling approaches

There are two levels of nonresponse: unit (or total) nonresponse and item (or partial) nonresponse. *Unit* (or *total*) *nonresponse* is a complete lack of information on all the variables of interest for a given unit. *Item* (or *partial*) *nonresponse* is a lack of information on given variables of interest.

The main approaches to handle nonresponse are reweighting and imputation. *Reweighting* procedures consist, firstly, of eliminating the survey nonrespondents from the data file and, secondly, of increasing the initial weights of survey respondents in order to compensate for the eliminated units. See Haziza and Lesage (2015) for a discussion of reweighting procedures for unit nonresponse. *Imputation* procedures consist of creating artificial values to fill in the gaps due to the missing values. See Sande (1981a,b) for discussions of the problems associated with nonresponse and imputation. *Single imputation* means that a single artificial value is created for each missing value and, by contrast, *multiple imputation* (Rubin, 1987) means that at least two artificial values are created for each missing value. In this document, we focus on single imputation. Generally, reweighting is applied to handle unit nonresponse and imputation is applied to handle item nonresponse. Reweighting is further discussed in Section 1.3.3 and single imputation is further discussed in Section 1.3.4.

1.3.3 Reweighting for unit nonresponse

When reweighting for unit nonresponse, two types of approaches are mainly applied: response probabilities modeling or calibration. With *response probabilities modeling*, nonresponse is viewed as a second phase of the survey. Let us suppose for the moment that the response probabilities p_i are known. By analogy with two-phase sampling, the double expansion estimator of Equation (1.1) suggests estimator

$$\widetilde{Y}_{PSA} = \sum_{i \in s_r} \frac{1}{\pi_i} \frac{1}{p_i} y_i.$$

See Särndal and Swensson (1987) for general results for two-phase sampling applied to the case of nonresponse. Similarly to the double expansion estimator, estimator $\widetilde{Y}_{PSA}$ is unbiased for the population total under the assumptions that $\pi_i > 0$ and $p_i > 0$ for all $i \in U$, this latest being somewhat unrealistic since some units are hardcore nonrespondents (Kott, 1994). However, the response probabilities are unknown and a preliminary step consists of estimating them. A model for the response probabilities, called the *nonresponse model*, is supposed. From this model, we obtain estimated response probabilities $\widehat{p}_i$. The estimated response probabilities replace the true response probabilities in estimator $\widetilde{Y}_{PSA}$ and the propensity score adjusted estimator is obtained

$$\widehat{Y}_{PSA} = \sum_{i \in s_r} \frac{1}{\pi_i} \frac{1}{\widehat{p}_i} y_i.$$

Three main estimation techniques applied to estimate the response probabilities are: parametric estimation, nonparametric estimation, and estimation with reweighting classes. With parametric estimation, a parametric nonresponse

model is assumed for the response probabilities

$$p_i = f(\mathbf{x}_i, \boldsymbol{\beta}),$$

where $\mathbf{x}_i$ is a vector of auxiliary variables with values known for every sampled unit and $\boldsymbol{\beta}$ is a vector of parameters. A method of estimation such as maximum likelihood is applied and an estimate $\widehat{\boldsymbol{\beta}}$ of the vector of parameters $\boldsymbol{\beta}$ is obtained. The estimate $\widehat{\boldsymbol{\beta}}$ is plugged in the nonresponse model to obtain estimated response probabilities $\widehat{p}_i = f(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})$. Kim and Kim (2007) showed that, with parametric estimation, the propensity score adjusted estimator $\widehat{Y}_{PSA}$ is generally more efficient than estimator $\widetilde{Y}_{PSA}$, which uses the true response probabilities, provided that the parameters in the nonresponse model are estimated by maximum likelihood. With the second estimation technique, nonparametric estimation, we do not suppose a specific form for function f in the nonresponse model. Rather, general characteristics such as smoothness are assumed. Because more flexible, nonparametric estimation is more robust to model misspecification than parametric estimation and is preferred when there is no prior idea of the form of function f. Nonparametric estimation of the response probabilities for reweighting is considered in Giommi (1987); Niyonsenga (1994, 1997); Da Silva and Opsomer (2006, 2009). Finally, the third estimation technique, estimation with reweighting classes, consists of forming weighting classes based on auxiliary information. The response probabilities are then estimated by the response rate in each weighting class. Särndal and Swensson (1987) refer to these classes to as response homogeneity groups. See Little (1986); Eltinge and Yansaneh (1997); Vartivarian and Little (2002) on the creation of weighting classes and Kott (2012) on the use of the design weights. All three estimation techniques assume a nonresponse model. Misspecification of this model implies a possibly severe bias of the propensity score adjusted estimator $\widehat{Y}_{PSA}$.

With the second approach to reweighting, *reweighting using calibration* (see for instance Folsom and Singh, 2000; Särndal and Lundström, 2005; Särndal, 2007; Kott, 2006), we do not estimate the response probabilities. Rather, the design weights of respondents are adjusted to compensate for nonrespondents by means of calibration. With this approach, one distinguishes between two types of auxiliary variables: auxiliary variables $\mathbf{x}^U$ with values $\mathbf{x}_i^U$ available for each respondent and with known population total $\sum_{i \in U} \mathbf{x}_i^U$, and auxiliary variables $\mathbf{x}^s$ with values $\mathbf{x}_i^s$ available for each respondent and with known estimated total $\sum_{i \in s} d_i \mathbf{x}_i^s$. Information coming from a register is a typical example of the former type and paradata is a typical example of the latter type. The idea of reweighting using calibration is to find calibration weights w_i as close as possible (in an average sense for a given distance) to the initial design

weights $d_i = 1/\pi_i$ while respecting the calibration equation

$$\sum_{i \in s_r} w_i \mathbf{x}_i = \mathbf{X},$$

where

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_i^U \\ \mathbf{x}_i^s \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} \sum_{i \in U} \mathbf{x}_i^U \\ \sum_{i \in s} d_i \mathbf{x}_i^s \end{pmatrix}.$$

Note that this approach does not require the values of the auxiliary variables to be known for each population unit. Whereas calibrating at the sample level (variables $\mathbf{x}^s$ and lower part of the calibration equation) tends to correct the nonresponse error, calibrating at the population level (variables $\mathbf{x}^U$ and upper part of the calibration equation) tends to correct both the sampling error and the nonresponse error. The calibration weights are in the form

$$w_i = d_i F(q_i \mathbf{x}_i^\top \boldsymbol{\lambda}),$$

where F is some function, $\boldsymbol{\lambda}$ is some vector, and q_i is a weight attached to unit i. Several distance functions are proposed in Deville and Särndal (1992) as a means of measuring the distance between the initial design weights $d_i = 1/\pi_i$ and the final weights w_i, each of which providing a particular form for function F (see Deville and Särndal, 1992, p. 378). An estimator of the total is then the *calibrated estimator*

$$\widehat{Y}_C = \sum_{i \in s_r} w_i y_i = \sum_{i \in s_r} \frac{1}{\pi_i} F(q_i \mathbf{x}_i^\top \boldsymbol{\lambda}) y_i.$$

Even though very different in spirit, there is a close parallel between the two aforementioned approaches to reweighting. The calibrated estimator $\widehat{Y}_C$ is indeed a particular case of propensity score adjusted estimator $\widehat{Y}_{PSA}$ where $\widehat{p}_i = F(q_i \mathbf{x}_i^\top \boldsymbol{\lambda})^{-1}$. Hence, the calibrated estimator $\widehat{Y}_C$ can be thought as a propensity score adjusted estimator where the response probabilities are estimated via calibration.

When reweighting using generalized calibration (Deville, 2000, 2002; Kott, 2006) or using generalized raking procedures (see Deville et al., 1993) instead of calibration, one allows for the variables that appear in the final weights to differ from the variables that appear in the calibration equation. One finds calibration weights in the form

$$w_i^G = d_i F(q_i \mathbf{z}_i^\top \boldsymbol{\lambda}),$$

where $\mathbf{z}_i$ is a vector of variables, often called the *intrumental variables*, known

for each survey respondent, while respecting the calibration equation

$$\sum_{i \in s_r} w_i^G \mathbf{x}_i = \mathbf{X}.$$

An estimator of the total is then the *generalized calibration estimator*

$$\widehat{Y}_{GC} = \sum_{i \in s_r} w_i^G y_i = \sum_{i \in s_r} \frac{1}{\pi_i} F(q_i \mathbf{z}_i^\top \boldsymbol{\lambda}) y_i.$$

Similarly to the calibrated estimator, the generalized calibration estimator can be thought as a propensity score adjusted estimator where the response probabilities are estimated nonparametrically via generalized calibration and $\widehat{p}_i = F(q_i \mathbf{z}_i^\top \boldsymbol{\lambda})^{-1}$. Traditionally, vectors $\mathbf{x}_i$ and $\mathbf{z}_i$ are assumed to have the same dimension but a solution was proposed to handle the case of different dimensions (Chang and Kott, 2008). Reweighting using generalized calibration is applied to handle non-ignorable nonresponse (Deville, 2000; Kott and Chang, 2010). The variable of interest is included in the instrumental variables; the effect of the variable of interest on the response probabilities comes into the equation. Particular care should be taken when reweighting for unit nonresponse with generalized calibration: Lesage and Haziza (2015) highlight the risks of bias and variance amplification of the generalized calibration estimator.

1.3.4 Imputation for item nonresponse

Reweighting is usually avoided to handle item nonresponse because it requires many sets of weights for some large number of variables. Rather, imputation is preferred. Imputation is simple in the sense that it generates a complete data file which is available to estimate all parameters of interest. However, particular attention should be paid when handling a data file containing imputed values. Such values are artificial and considering them as observed produces invalid variance estimators and invalid inference. There is an extensive literature devoted to variance estimation and inference with imputed data, see for instance Rao (1990); Särndal (1990); Rao and Shao (1992); Lee et al. (1994); Rao and Sitter (1995); Fay (1996); Rao (1996); Shao and Sitter (1996); Kim (2001); Brick et al. (2004); Haziza and Rao (2006). An effective imputation method should impute consistent values; some procedures consist of imputing manually based on logical rules or automatically based on a systematic approach (Fellegi and Holt, 1976).

Imputation methods are classified into two groups: deterministic imputation methods and random imputation methods. For fixed sample and fixed set of survey respondents, *deterministic imputation methods* produce the same imputed values if the imputation is repeated in the same sample and in the

same set of survey respondents. Examples of deterministic imputation are: ratio imputation (see David and Sukhatme, 1974; Rao and Sitter, 1995; Shao, 2000; Kim and Park, 2006), regression imputation, respondent mean imputation, nearest neighbor imputation (studied in Rancourt et al., 1994; Chen and Shao, 1997, 2000, 2001; Shao and Wang, 2008; Shao, 2009), predictive mean matching (Little, 1988), and auxiliary value imputation (Beaumont et al., 2011). *Random imputation methods* have, as indicated by their name, a random component. As a result, they produce, for fixed sample and fixed set of survey respondents, different imputed values if the imputation is repeated. Examples of random imputation methods are: random hot-deck imputation, imputation with added residuals (Chen et al., 2000; Chauvet et al., 2011b), and fractional hot-deck imputation (Kim and Fuller, 2004). Unlike deterministic imputation methods, random imputation methods tend to preserve the distribution of the variable being imputed at the expense of an additional variance in estimations which is called the *imputation variance*. Many authors have been interested in minimizing the imputation variance, see for instance Kalton and Kish (1981, 1984); Chen et al. (2000); Kim and Fuller (2004); Fuller and Kim (2005); Chauvet et al. (2011b).

Alternatively, imputation methods can be classified as either donor or predicted value. *Donor imputation methods* replace the missing value of a survey nonrespondent with an observed value. The unit providing the value is called a *donor* and the unit receiving the value is called a *recipient*. Examples of donor imputation methods are: random hot-deck imputation, nearest neighbor imputation, and previous value imputation. *Predicted value imputation methods* use functions of the observed values of survey respondents to predict the missing values of survey nonrespondents. Examples of predicted value imputation methods are: ratio imputation, regression imputation, and respondent mean imputation.

In Section 1.3.3, when reweighting using response probabilities modeling, we assumed a model for the response probabilities, the nonresponse model. When imputing, we rather assume a model for the variable of interest, the *imputation model* (Kalton and Kasprzyk, 1986; Särndal, 1992). From this model, an imputed value y_i^* is obtained for each survey nonrespondent i and the *imputed estimator*

$$\widehat{Y}_I = \sum_{i \in s} \frac{1}{\pi_i} \left[y_i r_i + y_i^* \left(1 - r_i \right) \right]$$

is considered. Every imputation method assumes an imputation model, either defined clearly or underlain. Misspecification of this model implies a possibly severe bias of estimator $\widehat{Y}_I$.

Parametric and nonparametric imputation models are considered. In a parametric framework, we consider a general imputation model

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i,$$

where f is a specified function, $\mathbf{x}_i$ is a vector of auxiliary variables with values known for every sampled unit, $\boldsymbol{\beta}$ is a vector of parameters, and ε_i are zero-mean independent errors with variance σ_i^2. For deterministic imputation, an imputed value y_i^* for a survey nonrespondent i is obtained as follows:

$$y_i^* = f(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_r),$$

where $\widehat{\boldsymbol{\beta}}_r$ is an estimate of $\boldsymbol{\beta}$ based on the observed values y_i and $\mathbf{x}_i$ of survey respondents. For random imputation, a random residual ε_i^* is added, i.e. $y_i^* = f(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}_r) + \varepsilon_i^*$. Some examples of imputation methods which assume a parametric imputation model are: regression imputation, mean imputation, and ratio imputation. In a nonparametric framework, we do not specify the form of function f or the structure of the variance σ_i^2. Some examples of imputation methods which assume a nonparametric imputation model are: nearest neighbor imputation and predictive mean matching.

2 Fast balanced sampling for highly stratified populations

Abstract

Balanced sampling is a very efficient sampling design when the variable of interest is correlated to the auxiliary variables on which the sample is balanced. A procedure to select balanced samples in a stratified population has previously been proposed. Unfortunately, this procedure becomes very slow as the number of strata increases and it even fails to select samples for some large numbers of strata. A new algorithm to select balanced samples in a stratified population is proposed. This new procedure is much faster than the existing one when the number of strata is large. Furthermore, this new procedure makes it possible to select samples for some large numbers of strata, which was impossible with the existing method. Balanced sampling can then be applied on a highly stratified population when only a few units are selected in each stratum. Finally, this algorithm turns out to be valuable for many applications as, for instance, for the handling of nonresponse.[1]

Keywords: balanced sampling, stratified sampling, cube method, unequal probability sampling, auxiliary information.

2.1 Introduction

Auxiliary information is a central point in survey statistics. It is widely-used in a large set of sampling designs. For instance, auxiliary information can be used to select stratified samples; it can also be used to define sampling designs with unequal probabilities. Regardless of the way auxiliary information is used, the main goal is to improve the quality of the estimates.

A stratified sampling design consists of dividing the population into subgroups (the *strata*) and of selecting samples in each stratum. Auxiliary information must be available to define the strata. The way the population has to be stratified is not always clear. A lot of research has been conducted on this topic. Neyman (1934) looked into optimum allocation. A method for the iterative improvement of the points of stratification was given and illustrated in Dalenius and Hodges (1959). Bülher and Deutler (1975) presented a method

[1] This article is a reprint of Hasler, C. and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics & Data Analysis*, 74:81–94. http://dx.doi.org/10.1016/j.csda.2013.12.005

to determine a global optimal solution by linear programming whereas Lavallée and Hidiroglou (1988) tackled the issue of stratification of a highly skewed population. Díaz-García and Garay-Tápia (2007) considered the allocation problem in stratified surveys as a problem of stochastic programming. Stratified sampling designs have the interesting property of reducing the variance of the Horvitz-Thompson estimator compared to unstratified sampling designs if the values of the variable of interest are somewhat homogenous inside the strata.

A balanced sampling design consists of selecting samples in such a way that the Horvitz-Thompson estimator for some auxiliary variables matches the population total. These auxiliary variables are called the *balancing variables*. Deville et al. (1988) described a method to obtain balanced samples and, later, the cube method (Deville and Tillé, 2004) was proposed for the same purpose. Some methods have been proposed for the computing of optimal inclusion probabilities for balanced sampling as for instance those given in Tillé and Favre (2005), Nedyalkova and Tillé (2008), and Chauvet et al. (2011a). A balanced sampling design is a very efficient sampling design when the variable of interest is correlated to the balancing variables.

In the presence of auxiliary variables correlated to the variable of interest and in the presence of strata, it is thus very useful to select samples applying a procedure which produces both stratified and balanced samples. Brewer (1999), indeed, showed that balanced sampling inside the strata can considerably improve the robustness and efficiency of some estimates. Chauvet (2009) proposed a stratified balanced sampling procedure: his algorithm selects samples which are approximately balanced in each stratum, balanced across the entire population and such that the sample size is fixed in each stratum. Unfortunately, Chauvet's procedure can be slow when the number of strata is large. In this paper, a new algorithm for stratified balanced sampling is proposed. This algorithm is much faster than Chauvet's algorithm when the number of strata is large.

The proposed algorithm turns out to be valuable for many applications, namely the selection of balanced samples in highly stratified populations when only a few units are selected in each stratum. For example, the proposed algorithm could improve the quality of estimates produced by some large-scale surveys. Indeed, in some large-scale multistage surveys, only one or two primary sampling units or first-stage units are selected in each stratum and the number of strata can be very large. Besides, the proposed method can also be used to treat nonresponse. Stratified sampling has long been used for the purpose of imputation. For instance, Kalton and Kish (1984) had already proposed selecting stratified sample of respondents to act as donors in order to reduce imputation variance. This idea can be extended by using the proposed method

for stratified balanced sampling. Indeed, Chauvet et al. (2011b) proposed a class of imputation methods that they called balanced random imputation and which use balanced sampling. This class of method is constructed such that the imputation variance is eliminated. Furthermore, the imputed values can be obtained through stratified balanced sampling. In this framework, however, the considered number of strata may be very large, hence the proposed method for stratified balanced sampling turns out to be useful in this context.

The paper is organized as follows. In Section 2.2, notions and concepts of balanced sampling are reviewed. Then in Section 2.3, Chauvet's method is described. The new method is presented in Section 2.4. A solution to apply the new method in cases where the sum of the inclusion probabilities is not an integer in each stratum is given in Section 2.5. Section 2.6 focusses on estimation of the variance of the Horvitz-Thompson estimator whereas Section 2.7 presents a possible application of the new method to the handling of nonresponse. Brief simulation studies were conducted to test the performance of the new sampling algorithm, to test the accuracy of the proposed formulas for the variance, and to illustrate the application of the new sampling algorithm in the context of handling of nonresponse. The results of these studies are given in Section 2.8. Finally, Section 2.9 closes the paper with concluding remarks.

2.2 Balanced sampling

Consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ of size N in which the aim is to select a random sample S, i.e. a subset of the population randomly selected. A sampling design $p(\cdot)$ assigns to each subset $s \subset U$ a probability $p(s)$ of being selected with

$$\sum_{s \subset U} p(s) = 1.$$

The inclusion probability π_k is the probability of selecting a particular unit k. The aim is to estimate a total

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

for some variable of interest y. If $\pi_k > 0$ for all $k \in U$, then the Horvitz and Thompson (1952) estimator given by

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

is unbiased for t_y.

Consider now that a column vector $\mathbf{x}_k \in \mathbb{R}^q$ of auxiliary variables is available

for all the units $k \in U$. A sampling design $p(\cdot)$ with inclusion probabilities π_k is said to be balanced on $\mathbf{x}_k$ if

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

for every subset $s \subset U$ such that $p(s) > 0$. In many cases, it is not possible to find a subset $s \subset U$ satisfying exactly equation (2.1). As a result, a sampling design $p(\cdot)$ can often not be exactly balanced. This problem is referred to as a *rounding problem*. Consider the sample membership indicators $\mathbf{s} = (s_1 \ \ldots \ s_k \ \ldots \ s_N)^\top$ where

$$s_k = \begin{cases} 1 & \text{if } k \in S, \\ 0 & \text{if } k \notin S. \end{cases}$$

When a rounding problem is encountered, it is not possible to find a vector $\mathbf{s}$ of zeros and ones that exactly satisfies the equation

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

Deville and Tillé (2004) proposed the cube method, which allows for the selection of balanced samples. The cube method is an algorithm composed of two phases: the *flight phase* and the *landing phase*. In what follows, the results given by the two phases of the algorithm are presented. The aim is not to describe the cube method in detail but only the outputs of both phases.

- The flight phase provides a vector of random variables $\boldsymbol{\phi} = (\phi_1 \ \ldots \ \phi_k \ \ldots \ \phi_N)^\top$, with $0 \leq \phi_k \leq 1$, such that
  (i) $\mathrm{E}(\phi_k) = \pi_k$ for all $k \in U$,
  (ii) $\sum_{k \in U} \frac{\mathbf{x}_k}{\pi_k} \phi_k = \sum_{k \in U} \mathbf{x}_k$,
  (iii) $\#\{k | 0 < \phi_k < 1\} \leq q$, where q is the dimension of $\mathbf{x}_k$.
  A unit k with $\phi_k = 1$ is selected in the sample and a unit k with $\phi_k = 0$ is definitely rejected. Whether there is a rounding problem or not, the equation in *(ii)* is exactly satisfied. In the presence of a rounding problem and as explained at the end of the previous paragraph, it is, however, not possible to find a vector $\boldsymbol{\phi}$ of zeros and ones which is a solution to the equation in *(ii)*. In that case, some ϕ_k's are not integers and some units are not yet selected or rejected at the end of the flight phase. It is possible to show, as stated in *(iii)*, that the number of non-integer ϕ_k's is at most q. In other words, at most q units are not yet selected or rejected at the end of the flight phase. Chauvet and Tillé (2006) proposed a fast algorithm for the flight phase. In what follows, the flight phases are carried out by means of this algorithm.

- The landing phase is used to deal with the rounding problem. Its main idea is to relax the balancing constraint in order to address the problem of the units that have not yet been rejected or selected at the end of the flight phase. The landing phase provides a vector $\mathbf{s} = (s_1 \ \dots \ s_k \ \dots \ s_N)^\top$ of sample membership indicator such that

  (i) $\mathrm{E}(s_k|\boldsymbol{\phi}) = \phi_k$ for all $k \in U$,

  (ii) $\sum_{k\in U} \frac{\mathbf{x}_k}{\pi_k} s_k \approx \sum_{k\in U} \mathbf{x}_k$.

  A unit k with $s_k = 1$ is selected in the sample and a unit k with $s_k = 0$ is rejected. At the end of the landing phase, every unit has been selected or rejected. Deville and Tillé (2004) have proposed two ways of running the landing phase: by linear programming or by suppression of variables. The landing phase by linear programming consists of solving a linear programming problem through the simplex algorithm. The list of all possible samples from a population of size q, where q is the dimension of $\mathbf{x}_k$, must be generated and this can be impossible when q exceeds a limit. Therefore, the landing phase by linear programming cannot be applied when the number of auxiliary variables q exceeds this limit, which is generally 20.

When one of the variables in $\mathbf{x}_k$ is equal or proportional to π_k, the balancing constraint (2.1) implies that

$$\sum_{k\in s} \frac{\pi_k}{\pi_k} = \sum_{k\in U} \pi_k \Leftrightarrow n(s) = \sum_{k\in U} \pi_k,$$

where $n(s)$ is the size of the subset $s \subset U$. This means that the sampling design has a fixed sample size. This equality can only be exactly satisfied if the sum of the inclusion probabilities is an integer. If the sum of the inclusion probabilities is not an integer, then the cube method usually selects a sample whose size is the smallest integer larger than this sum or the largest integer smaller than this sum.

2.3 Chauvet's method for stratified balanced sampling

The population is presumed to be partitioned into H nonoverlapping strata $U_1, \dots, U_h, \dots, U_H$. Let $\mathbb{1}(k \in U_h)$ be the stratum membership indicator that takes value 1 if unit k belongs to stratum h and 0 otherwise. A stratified balanced sampling design $p(\cdot)$ is a sampling design which is balanced on $\mathbf{x}_k$ in each stratum, i.e.

$$\sum_{k\in s\cap U_h} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U_h} \mathbf{x}_k,$$

for each $s \subset U$ with $p(s) > 0$ and for each $h = 1, \dots, H$.

Suppose that the goal is to balance on $\mathbf{x}_k \in \mathbb{R}^q$ such that none of the auxiliary

variables is proportional to π_k. Chauvet's method (Chauvet, 2009) is presented in Algorithm 2.1. The main idea of this method is to first run flight phases independently inside the strata. This ensures that the samples are as balanced as possible within the strata. Next, in a second step, a general flight phase is run on all the units of the population that have not yet been selected or rejected at the end of the first step. It results in samples that are as balanced as possible across the entire population. Finally, a third step is carried out to handle the case of units that have not yet been rejected or selected at the end of the second step. Originally in Chauvet (2009), the third step consisted of unequal probability sampling whereas the third step of the new procedure presented in Section 2.4 consists of a landing phase by suppression of variables. A landing phase by suppression of variables takes the balancing constraint into account and therefore provides more accurate estimates than unequal probability sampling. Henceforth, for a fair comparison between methods with respect to the accuracy of the estimates and as pointed out by one of the referees, the third step of Chauvet's algorithm has here been modified onto a landing phase by suppression of variables.

Algorithm 2.1 Chauvet stratified balanced sampling with step 3 by landing phase by suppression of variables

Step 1: Carry out a flight phase, with balancing variables $(\pi_k \quad \mathbf{x}_k^\top)^\top$ and inclusion probabilities π_k independently in each stratum U_h.
Step 1 provides vector $\boldsymbol{\phi}$ of ϕ_k's.

Step 2: Carry out a flight phase, with balancing variables

$$\left(\phi_k \mathbb{1}(k \in U_1) \ \ldots \ \phi_k \mathbb{1}(k \in U_h) \ \ldots \ \phi_k \mathbb{1}(k \in U_H) \quad \frac{\phi_k \mathbf{x}_k^\top}{\pi_k}\right)^\top$$

and inclusion probabilities ϕ_k on the set of units with non-integer ϕ_k, i.e. on the units that are not yet selected or rejected at the end of step 1.
Step 2 provides vector $\boldsymbol{\psi}$ of ψ_k's.

Step 3: Do a landing phase with inclusion probabilities ψ_k and balancing variables

$$\left(\psi_k \mathbb{1}(k \in U_1) \ \ldots \ \psi_k \mathbb{1}(k \in U_h) \ \ldots \ \psi_k \mathbb{1}(k \in U_H) \quad \frac{\psi_k \mathbf{x}_k^\top}{\pi_k}\right)^\top$$

on the set of units with non-integer ψ_k. Use the landing phase by suppression of variables.

In step 1 of Algorithm 2.1, $q+1$ balancing variables are considered in each flight phase. Therefore at most $q+1$ units in each stratum are not yet selected or rejected at the end of step 1. As a result, step 2 concerns at most $(q+1)H$ units. In step 2, $H+q$ balancing variables are considered in the flight phase.

It may be impossible to carry out the flight phase of step 2 if the considered design is highly stratified (i.e. if H is very large). Indeed, the fast algorithm for the flight phase proposed by Chauvet and Tillé (2006) requires the use of a matrix that is equal in size to the number of balancing variables times the number of balancing variables plus one. However, this approach can only be used with matrices of a limited size. This limit depends on the computer. If a highly stratified design is considered, the flight phase of step 2 requires the use of a huge matrix and it may be impossible to carry it out. Henceforth, Algorithm 2.1 is likely to fail for highly stratified designs.

2.4 New procedure for highly stratified balanced sampling

In this section, it is supposed that the sum of the inclusion probabilities in each stratum

$$\sum_{k \in U_h} \pi_k = n_h,$$

is an integer. This hypothesis will be relaxed in Section 2.5 but will considerably simplify the complexity of the proposed algorithm. The main idea of the proposed method described in Algorithm 2.2, is to first run a flight phase independently in each stratum. Then, in a second step, U_1 and U_2 are merged and a flight phase is run. Next, U_1 and U_2 are merged with U_3 and a flight phase is run again and so on. Finally, a landing phase by suppression of variables is carried out in a third step.

This alternative implementation might look like a simple variant but it actually offers major advantages. An important advantage is that it greatly reduces computation time when some large numbers of strata are considered. This reduction in computation time is explained in what follows. In step 2 of Algorithm 2.2, the flight phases are carried out with the balancing variables $\mathbf{z}_k^{(j)}$. Consider matrix $\mathbf{Z}^{(j)}$ whose rows are the $\mathbf{z}_k^{(j)^\top}$ restricted to the k with non-integer $\phi_k^{(j-1)}$, i.e. the k such that $0 < \phi_k^{(j-1)} < 1$.

Property 2.1 *With Algorithm 2.2, for* $j = 2, \ldots, H$

(i) $$\#\left\{ k \in \bigcup_{i=1}^{j} U_i \,\middle|\, 0 < \phi_k^{(j)} < 1 \right\} \leq 2q + 2,$$

(ii) the number of non-null columns of matrix $\mathbf{Z}^{(j)}$ *is less than or equal to* $2q+2$*, where a null column is a column that contains only zeros.*

The proof is given in Appendix A and requires that the sum of the inclusion probabilities is an integer in each stratum. In light of Property 2.1, it appears that the flight phase must never be applied on a matrix of balancing variables with more than $2q + 2$ columns with Algorithm 2.2 because the null columns can be removed. However the flight phase could be applied on a matrix of balancing variables with up to $q + H$ columns with Algorithm 2.1. This size

Algorithm 2.2 New procedure for highly stratified balanced sampling

Step 1: Carry out a flight phase, with balancing variables $(\pi_k \quad \mathbf{x}_k^\top)^\top$ and inclusion probabilities π_k independently in each stratum U_h.
Step 1 provides vector $\boldsymbol{\phi}^{(1)}$ of $\phi_k^{(1)}$'s.

Step 2: For $j = 2$ to H:

- Carry out a flight phase on the union of strata $U_1, \ldots, U_j$, with balancing variables

$$\mathbf{z}_k^{(j)} = \left(\phi_k^{(j-1)} \mathbb{1}(k \in U_1) \ \ldots \ \phi_k^{(j-1)} \mathbb{1}(k \in U_j) \quad \frac{\phi_k^{(j-1)} \mathbf{x}_k^\top}{\pi_k} \right)^\top$$

  and inclusion probabilities $\phi_k^{(j-1)}$ on the set of units with non-integer $\phi_k^{(j-1)}$. The flight phase provides a vector $\boldsymbol{\phi}^{(j)}$ of $\phi_k^{(j)}$'s for units with non-integer $\phi_k^{(j-1)}$.
- Set $\phi_k^{(j)} = \phi_k^{(j-1)}$ for units with integer $\phi_k^{(j-1)}$.

Step 2 provides vector $\boldsymbol{\phi}^{(H)}$ of $\phi_k^{(H)}$'s.

Step 3: Do a landing phase with inclusion probabilities $\phi_k^{(H)}$ and balancing variables

$$\mathbf{z}_k^{(H+1)} = \left(\phi_k^{(H)} \mathbb{1}(k \in U_1) \ \ldots \ \phi_k^{(H)} \mathbb{1}(k \in U_H) \quad \frac{\phi_k^{(H)} \mathbf{x}_k^\top}{\pi_k} \right)^\top$$

on the set of units with non-integer $\phi_k^{(H)}$. Use the landing phase by suppression of variables.

difference of the matrices considered in the flight phases affects the execution time of the algorithms. Indeed, even if Algorithm 2.2 requires us to run $2H - 1$ flight phases against only $H + 1$ for Algorithm 2.1, Algorithm 2.2 becomes much faster than Algorithm 2.1 as H increases. Even more interesting is the fact that Algorithm 2.2 is much more resistent to numerical instability than Algorithm 2.1 thanks to the reduction in size stated above. Indeed, numerical instability increases when the dimension of the matrices to deal with increases. In step 2 of Algorithm 2.1, flight phases operate with matrices of up to $(q + H) \times (q + H + 1)$ in size whereas flight phases operate with matrices of up to $(2q + 2) \times (2q + 3)$ in size in step 2 of Algorithm 2.2. As this dimension depends on H for Algorithm 2.1, numerical instability increases as H increases. This is not the case with Algorithm 2.2.

Another advantage of the proposed method is that a landing phase can be applied in the last step even if the population is highly stratified. Indeed, at the last loop of step 2, we have

$$\# \left\{ k \in U \,\middle|\, 0 < \phi_k^{(H)} < 1 \right\} \leq 2q + 2.$$

This implies that the last step concerns at most $2q + 2$ units. This quantity is independent of the number of strata H. Consequently, a landing phase can be applied in step 3 regardless of the number of strata. Therefore, the balancing can be taken into consideration in step 3 of Algorithm 2.2 even if the population is highly stratified. As far as the last step of Algorithm 2.1 is concerned, a landing phase may not be applied for highly stratified populations as the number of units considered can reach $q + H$. The landing phase of step 3 of Algorithm 2.2 must be done by suppression of variables in order to ensure fixed size sampling inside the strata. Indeed, steps 1 and 2 of Algorithm 2.2 consist of flight phases. The balancing equations of the carried out flight phases imply that

$$\sum_{k \in U_h} \phi_k^{(j)} = n_h,$$

for each $j = 1, \ldots, H$ and each $h = 1, \ldots, H$. In particular, the following equation

$$\sum_{k \in U_h} \phi_k^{(H)} = n_h, \tag{2.2}$$

is satisfied for each $h = 1, \ldots, H$. In step 3 of Algorithm 2.2, a landing phase is carried out. The aim is to derive a sample $\mathbf{s}$ of s_k's. The balancing equations linked to the first H balancing variables $\phi_k^{(H)} \mathbb{1}(k \in U_h)$, $h = 1, \ldots, H$ simplify to

$$\sum_{k \in U_h} s_k = \sum_{k \in U_h} \phi_k^{(H)}, \tag{2.3}$$

for $h = 1, \ldots, H$. Combining Equation (2.2) together with Equation (2.3) leads to

$$\sum_{k \in U_h} s_k = n_h, \tag{2.4}$$

for $h = 1, \ldots, H$. As n_h is in this section supposed to be an integer, equation (2.4) can always be satisfied; all that is required is to select n_h units in each stratum U_h. The landing phase by suppression of variables consists of alternate dropping the last balancing variables and running a flight phase again until the remaining constraints are exactly satisfied. As explained above, the constraints linked to the first balancing variables $\phi_k^{(H)} \mathbb{1}(k \in U_h)$, $h = 1, \ldots, H$, can always be satisfied. As the landing phase is carried out by suppression of variables in step 3 of Algorithm 2.2, only the last q variables

$$\frac{\phi_k^{(H)} \mathbf{x}_k^\top}{\pi_k}$$

are suppressed and fixed size sampling inside the strata is ensured.

Furthermore, the selected sample $\mathbf{s} = (s_1 \ \ldots \ s_k \ \ldots \ s_N)^\top$ satisfies $\mathrm{E}(s_k) = \pi_k$. To summarize, the sampling design associated with Algorithm 2.2 is balanced,

can be highly stratified and ensures fixed size sampling within the strata. Selection of samples with highly stratified designs becomes tractable with this new procedure.

Parallel computing can be used to slightly speed up both Algorithm 2.1 and Algorithm 2.2. Firstly, it is conceivable to carry out the flight phases of steps 1 of both Algorithms in parallel. It is also possible to adapt step 2 of Algorithm 2.2 to use parallel computing. Indeed, even though it is impossible to roughly apply parallel computing as iterative procedures are involved, step 2 can be adapted for it as follows. The procedure proposed in step 2 of Algorithm 2.2 can be applied in parallel on non-overlapping groups of strata first. Then some of these groups can be gathered and the same procedure can be used, and so on.

Finally, Algorithm 2.1 and Algorithm 2.2 can both be applied if the number of balancing variables q exceeds the size of a stratum. However, the balancing does not perform in such a stratum. Indeed, the random vector provided by the flight phase of step 1 in such a stratum match the initial inclusion probabilities. It means that none of the units of this stratum are selected or rejected yet at the end of the flight phase (except the one with integer inclusion probabilities). In steps 2, the same phenomena can occur, depending on whether the number of balancing variables still exceeds the number of units involved in the flight phases. Steps 3 can be applied if the number of balancing variables q exceeds the size of a stratum.

2.5 Case where the sum of the inclusion probabilities is not an integer in each stratum

As explained before, great advantages of Algorithm 2.2 can be gained from the fact that the sum of the inclusion probabilities is an integer in each stratum. However, most of the stratified designs used in practice can show a sum of the inclusion probabilities which is not an integer in each stratum. Stratification with proportional allocation and stratification with optimal allocation are, among others, such designs. In what follows, a procedure to extend the use of Algorithm 2.2 to the case in which the sum of the inclusion probabilities is not an integer in each stratum is presented. It thus becomes possible to apply the new algorithm regardless of which stratified design is used.

The goal of this section is to introduce a procedure to round the sum of the inclusion probabilities in each stratum. This is a typical problem of rounding of allocations. This topic has already been widely explored and several procedures already exist such as those implemented in the R package stratification by Baillargeon and Rivest (2011) or those presented by Wright (2012). We propose a new random procedure of rounding that agrees with the balancing in

the sense that it does not overly unbalance the totals of the auxiliary variables. Hence, the proposed rounding procedure consists of randomly rounding the sum of the inclusion probabilities in each stratum to the smallest integer larger than this sum or the largest integer smaller than this sum while taking into account constraints in relation to the balancing and sample size.

Let $\lfloor \cdot \rfloor$ denote the floor function. Consider

$$n_h = \sum_{k \in U_h} \pi_k,$$

and $p_h = n_h - \lfloor n_h \rfloor$. We have $0 \leq p_h \leq 1$ for all $h = 1, \ldots, H$. Since

$$\sum_{h=1}^{H} n_h = n$$

is an integer,

$$m = \sum_{h=1}^{H} p_h$$

is an integer as well.

The main idea of the proposed procedure is to select a sample of strata in which the number of selected units will be rounded up. Let $\boldsymbol{J} = (J_1 \ \ldots \ J_h \ \ldots \ J_H)$ denote a vector of sample membership indicators, where

$$J_h = \begin{cases} 1 & \text{if stratum } U_h \text{ is selected in the sample of strata,} \\ 0 & \text{otherwise.} \end{cases}$$

The probability that a stratum U_h is selected is p_h, or equivalently $\mathrm{E}(J_h) = p_h$. The rounded sample size in strata h is then $n_h^* = \lfloor n_h \rfloor + J_h$ for $h = 1, \ldots, H$. It means that the sample size in a stratum U_h is the smaller integer larger than n_h if the stratum is selected in the sample of strata and the larger integer smaller than n_h if the stratum is not selected in the sample of strata.

Constraints are imposed on the sample of strata $\boldsymbol{J}$. First, it is selected such that the total number of selected units remains the same despite the change in sample size in some strata, i.e.

$$\sum_{h=1}^{H} n_h^* = \sum_{h=1}^{H} n_h.$$

This last Equation is equivalent to

$$\sum_{h=1}^{H} \lfloor n_h \rfloor + J_h = \sum_{h=1}^{H} n_h. \tag{2.5}$$

Moreover, as explained before, this rounding of the sample size must not overly unbalance the totals of the auxiliary variables, which is formalized as

$$\sum_{h=1}^{H} \frac{\lfloor n_h \rfloor + J_h}{n_h} \sum_{k \in U_h} \mathbf{x}_k = \sum_{h=1}^{H} \sum_{k \in U_h} \mathbf{x}_k. \tag{2.6}$$

Considering Equation (2.5) together with Equation (2.6) leads to

$$\sum_{h=1}^{H} \frac{\lfloor n_h \rfloor + J_h}{n_h} \sum_{k \in U_h} \begin{pmatrix} \pi_k \\ \mathbf{x}_k \end{pmatrix} = \sum_{h=1}^{H} \sum_{k \in U_h} \begin{pmatrix} \pi_k \\ \mathbf{x}_k \end{pmatrix},$$

or equivalently

$$\sum_{h=1}^{H} \frac{J_h}{n_h} \sum_{k \in U_h} \begin{pmatrix} \pi_k \\ \mathbf{x}_k \end{pmatrix} = \sum_{h=1}^{H} \frac{p_h}{n_h} \sum_{k \in U_h} \begin{pmatrix} \pi_k \\ \mathbf{x}_k \end{pmatrix}.$$

The last equation can be rewritten

$$\sum_{h=1}^{H} \frac{J_h}{p_h} \mathbf{V}_h = \sum_{h=1}^{H} \mathbf{V}_h, \tag{2.7}$$

where

$$\mathbf{V}_h = \frac{p_h}{n_h} \sum_{k \in U_h} \begin{pmatrix} \pi_k \\ \mathbf{x}_k \end{pmatrix} = \begin{pmatrix} p_h \\ \frac{p_h}{n_h} \sum_{k \in U_h} \mathbf{x}_k \end{pmatrix}.$$

Expression (2.7) is a usual system of balancing equations that can be solved by the cube method. The sample of strata $\boldsymbol{J}$ is therefore obtained by balanced sampling.

The inclusion probabilities π_k must then be slightly modified in new probabilities π_k^* in such a way that

$$\sum_{k \in U_h} \pi_k^* = n_h^* = \lfloor n_h \rfloor + J_h,$$

and that $\mathrm{E}(\pi_k^*) = \pi_k$. This modification is not trivial with unequal inclusion probabilities. Several solutions exist and are discussed in Grafström et al. (2012). Once the new inclusion probabilities are computed, their sums are integers in the strata and Algorithm 2.2 can be used.

2.6 Variance estimation

The variance can be approximated with the method proposed by Deville and Tillé (2005). The same method was considered in Chauvet (2009). Set

$$\mathbf{z}_k = \begin{pmatrix} \pi_k \mathbb{1}(k \in U_1) & \pi_k \mathbb{1}(k \in U_2) & \dots & \pi_k \mathbb{1}(k \in U_H) & \mathbf{x}_k^\top \end{pmatrix}^\top.$$

An approximation of the variance of the total estimator $\widehat{t}_y$ is

$$\mathrm{var}_{\mathrm{app}}\left(\widehat{t}_y\right) = \sum_{k\in U} b_k \left(\frac{y_k}{\pi_k} - \boldsymbol{\beta}^\top \frac{\mathbf{z}_k}{\pi_k}\right)^2, \tag{2.8}$$

where

$$b_k = \pi_k\left(1-\pi_k\right)\frac{N}{N-(H+q)} \quad \text{and} \quad \boldsymbol{\beta} = \left(\sum_{\ell\in U} b_\ell \frac{\mathbf{z}_\ell}{\pi_\ell}\frac{\mathbf{z}_\ell}{\pi_\ell}^\top\right)^{-1} \sum_{\ell\in U} b_\ell \frac{\mathbf{z}_\ell}{\pi_\ell}\frac{y_\ell}{\pi_\ell}.$$

Various definitions of b_k's, and thus various approximations of the variance, are given in Deville and Tillé (2005). An estimator of the approximated variance (2.8) is

$$\widehat{\mathrm{var}}\left(\widehat{t}_y\right) = \sum_{k\in S} c_k \left(\frac{y_k}{\pi_k} - \widehat{\boldsymbol{\beta}}^\top \frac{\mathbf{z}_k}{\pi_k}\right)^2, \tag{2.9}$$

where

$$c_k = \left(1-\pi_k\right)\frac{n}{n-(H+q)} \quad \text{and} \quad \widehat{\boldsymbol{\beta}} = \left(\sum_{\ell\in S} c_\ell \frac{\mathbf{z}_\ell}{\pi_\ell}\frac{\mathbf{z}_\ell}{\pi_\ell}^\top\right)^{-1} \sum_{\ell\in S} c_\ell \frac{\mathbf{z}_\ell}{\pi_\ell}\frac{y_\ell}{\pi_\ell}.$$

The performance of the approximated variance and that of the variance estimator provided above are tested in Section 2.8. Nevertheless, the variance estimator (2.9) is intractable if the number of balancing variables exceeds the sample size, which means here if $H+q>n$. Indeed, the matrix $\sum_{\ell\in S} c_\ell \frac{\mathbf{z}_\ell}{\pi_\ell}\frac{\mathbf{z}_\ell}{\pi_\ell}^\top$ is in this case not invertible. However, it is possible in this case to estimate the variance using a collapsed stratum procedure (see Wolter, 1985, p. 50–57). Hence, the H strata are combined into G groups such that $G+q\leq n$ and the procedure given above to estimate the variance is applied considering the G groups instead of the H strata.

2.7 Illustration of the handling of nonresponse

2.7.1 Nonresponse and imputation

Imputation is a process that consists of replacing a missing value with a substituted one. It is especially used to compensate for item nonresponse. Imputation methods can be classified into two groups: deterministic and random. Deterministic methods are adequate for the purpose of totals estimation but they often fail to estimate quantiles because they disturb the distribution of the imputed variable. Random methods, on the other hand, are often appropriate for the aim of totals and quantiles estimation as they tend to preserve the distribution of the imputed variable. Unfortunately, the randomness of the imputation adds an additional amount of variance to the estimators. This

additional amount of variance is called *imputation variance*. Random imputation methods that produce the least possible imputation variance are therefore effective methods of handling item nonresponse when the aim is to estimate totals as well as quantiles. Random hot-deck imputation is the process that consists of replacing a missing value with an observed value extracted from the same survey and selected at random.

2.7.2 Notation

A finite population $U = \{1, \ldots, k, \ldots, N\}$ of size N is considered. In a first phase, a random sample S of size n is drawn with a given sampling design $p(\cdot)$. For each $k \in U$, consider the first order inclusion probability $\pi_k = \Pr(k \in S)$ and let $d_k = 1/\pi_k$ denote its Horvitz-Thompson weight (Horvitz and Thompson, 1952). It is supposed in this part that the vector of q auxiliary variables $\mathbf{x}_k$ is observed for each sampled unit $k \in S$. However, the values of the variable of interest y_k are potentially missing for some $k \in S$. Nonresponse can be viewed as a second phase of the sampling process. A subset $S_r \subset S$ of units k with observed y_k is indeed obtained from S with a usually unknown conditional distribution $q(S_r|S)$. Let S_m denote the complement of S_r in S, i.e. the subset of S containing the units k with missing y_k (the nonrespondents). For $k \in S$, let r_k be the response indicator variable

$$r_k = \begin{cases} 1 & \text{if } k \in S_r, \\ 0 & \text{otherwise.} \end{cases}$$

Imputation can be viewed as a third phase of the sampling process. Imputed values y_k^*, $k \in S_m$, are indeed drawn with a conditional distribution

$$I(y_k^*|S, S_r).$$

Suppose the aim is to estimate the regression coefficient

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

In the case of complete response, the estimator

$$\widehat{\theta}_N = \left(\sum_{k \in S} d_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S} d_k \mathbf{x}_k y_k,$$

is adequate. In the presence of nonresponse, this estimator is intractable and the imputed estimator

$$\widehat{\theta}_I = \left(\sum_{k \in S} d_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \left(\sum_{k \in S} d_k r_k \mathbf{x}_k y_k + \sum_{k \in S} d_k (1 - r_k) \mathbf{x}_k y_k^* \right),$$

can be used. The total variance of $\widehat{\theta}_I$ can be expressed as follows

$$\mathrm{Var}\left(\widehat{\theta}_I\right) = \mathrm{Var}_p \mathrm{E}_q \mathrm{E}_I \left(\widehat{\theta}_I\right) + \mathrm{E}_p \mathrm{Var}_q \mathrm{E}_I \left(\widehat{\theta}_I\right) + \mathrm{E}_p \mathrm{E}_q \mathrm{Var}_I \left(\widehat{\theta}_I\right), \quad (2.10)$$

where the subscripts p, q and I indicate the expectations and variances with respect to the sampling mechanism, with respect to the nonresponse mechanism, and with respect to the imputation mechanism, respectively. The first term in Expression (2.10) represents the sampling variance, the second term represents the nonresponse variance and the last term represents the imputation variance.

2.7.3 Balanced random imputation to eliminate the imputation variance

Chauvet et al. (2011b) proposed a class of random imputation methods which they called balanced random imputation. The proposed method consists of randomly selecting residuals while satisfying given constraints. It eliminates the imputation variance while preserving the distribution of the variable being imputed. An application of the new stratified balanced sampling procedure (Algorithm 2.2) for the purpose of balanced random imputation is provided here. For reasons of simplicity, the particular case of random hot-deck imputation in the context of the estimation of a domain means vector is considered. Algorithm 2.2 is however adaptive to the whole class of random imputation methods proposed in Chauvet et al. (2011b).

Suppose that $\mathbf{x}_k$ is a vector of H domain indicators and that the aim is to estimate the vector of domains means

$$\theta_N = \left(\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_H \right)^\top,$$

for some variable of interest y. A random sample S is therefore selected. Suppose that the vector of q auxiliary variables $\mathbf{x}_k$ is observed for each sampled unit $k \in S$ and that the value of the variable of interest y_k is missing for some sampled units $k \in S$. The imputed estimator $\widehat{\theta}_I$ is in this case

$$\widehat{\theta}_I = \left(\frac{\sum_{k \in D_h} d_k r_k y_k + \sum_{k \in D_h} d_k (1 - r_k) y_k^*}{\sum_{k \in D_h} d_k} \right)_{1 \leq h \leq H},$$

where D_h, $h = 1, \ldots, H$, represent the H domains considered. Random hot-deck imputation is then used to compensate for nonresponse. Survey

weighted imputation is considered, which means that the survey weights d_k are considered in the imputation process.

In what follows, it is explained how the method presented in Chauvet et al. (2011b) proceeds in this particular framework to select imputed values such that the imputation variance of the estimator $\widehat{\theta}_I$ is eliminated. The imputation is here explained by the following imputation model

$$\mathrm{m} : y_k = \beta + \sigma\varepsilon_k,$$

where β and σ are unknown parameters and ε_k are independent and identically distributed random variables with mean 0 and variance 1. For $i \in S_m$, the imputed value is given by

$$y_i^* = \overline{y}_r + \widehat{\sigma}\varepsilon_i^*,$$

where $\overline{y}_r$ is the estimated mean value over the respondents of the variable of interest, i.e.

$$\overline{y}_r = \left(\sum_{k\in S} d_k r_k\right)^{-1} \sum_{k\in S} d_k r_k y_k,$$

$\widehat{\sigma}$ is an estimator of σ, and the ε_i^*, $i \in S_m$, are selected independently and with replacement from $E_r = \{\widetilde{e}_j = \widehat{\sigma}^{-1}(y_j - \overline{y}_r); j \in S_r\}$ with probabilities

$$\widetilde{w}_j = \Pr\left(\varepsilon_i^* = \widetilde{e}_j\right) = \frac{d_j}{\sum_{\ell\in S} d_\ell r_\ell}.$$

In order to eliminate the imputation variance of the imputed estimator $\widehat{\theta}_I$, it is proposed in Chauvet et al. (2011b) to select the residuals ε_i^* such that

$$\frac{\sum_{i\in D_h} d_i(1-r_i)\widehat{\sigma}\varepsilon_i^*}{\sum_{i\in D_h} d_i} = 0, \tag{2.11}$$

for each $h = 1, \ldots, H$. The aim of the method proposed in Chauvet et al. (2011b) is therefore to select residuals ε_i^* for $i \in S_m$ with replacement in E_r while respecting Equation (2.11). This is a problem of balanced sampling with replacement. As explained in Chauvet et al. (2011b), it can alternatively be viewed as a problem of balanced sampling without replacement within a population of cells. This idea is used in the following section to explain how Algorithm 2.2 can be applied to select ε_i^* having the properties stated above.

2.7.4 Stratified balanced sampling for balanced random imputation

One of the possible applications of Algorithm 2.2 is the selection of residuals ε_i^*, $i \in S_m$, for balanced random imputation. Consider the population of

cells $U^* = \{(i,j) \in S_m \times S_r\}$. Moreover, consider $\psi_{ij} = \widetilde{w}_j$ the inclusion probability attach to each population unit $(i,j) \in U^*$, $c_{ij} = d_i \psi_{ij} \widetilde{e}_j$ an auxiliary variable attached to each population unit $(i,j) \in U^*$, and $U_h^* = \{(h,j); j \in S_r\}$ a strata defined for each $h \in S_m$. A solution to the balanced sampling with replacement problem stated above is given by stratified balanced sampling without replacement as follows. Select a random sample S^* with a stratified sampling design $p(\cdot)$ with inclusion probabilities ψ_{ij} balanced on c_{ij} and set $\varepsilon_i^* = \widetilde{e}_j$ for $i \in S_m$ and $j \in S_r$ if unit (i,j) is selected in the sample. This procedure indeed gives a solution to the balanced sampling with replacement problem stated above because

$$\Pr\left(\varepsilon_i^* = \widetilde{e}_j\right) = \Pr\left\{(i,j) \in S^*\right\} = \psi_{ij} = \widetilde{w}_j,$$

and for each $s^* \subset U^*$ with $p(s^*) > 0$

$$\sum_{(i,j) \in s^* \cap U_h^*} \frac{c_{ij}}{\psi_{ij}} = \sum_{(i,j) \in U_h^*} c_{ij} \quad \text{for all } h \in S_m,$$

which implies that the residuals ε_i^* for $i \in S_m$ are selected such that Equation (2.11) is satisfied. However, it is often not possible to select samples such that Equation (2.11) is exactly satisfied but only approximately satisfied. As a result, the imputation variance of $\widehat{\theta}_I$ is not completely eliminated but is relatively small.

As previously shown, stratified balanced sampling can be used for the purpose of balanced random imputation. In this context, a stratum U_i^* is attached to each nonrespondent $i \in S_m$. The number of strata considered in the stratified balanced sampling hence matches the number of nonrespondents. It may therefore be very large. For instance, in Statistics on Income and Living Conditions (SILC) in Switzerland in 2009, more than 1,800 persons did not indicate their income as they had been asked to. Therefore, approximately 1,800 strata would be required to carry out balanced random imputation through stratified balanced sampling. In this context, the new algorithm (Algorithm 2.2) clearly has an edge over Algorithm 2.1 because it is much faster when the number of strata is large and the selection of samples becomes tractable for some highly stratified cases that could not be handled using Algorithm 2.1.

2.8 Simulation study

Brief simulation studies are conducted to test the performance of the new sampling algorithm, to test the accuracy of the proposed formulas for variance and to illustrate the application of the new sampling algorithm in the context of the handling of nonresponse.

2.8.1 Performance of the proposed algorithm

The simulations conducted in Chauvet (2009) are extended. First, a population of size 1,000 is generated and is partitioned into 25 strata of equal size. Four balancing variables and four variables of interest are considered. The four balancing variables x_1, x_2, x_3, and x_4 are generated using independent gamma distributions with parameters 4 and 25. The four variables of interest are generated as follows

$$
\begin{aligned}
y_1 &= 20\alpha + \varepsilon_1 \\
y_2 &= 500 + 5x_1 + 5x_2 + \varepsilon_2 \\
y_3 &= 500 + 100x_1 + 100x_2 + 100x_3 + 100x_4 + \varepsilon_3 \\
y_4 &= 500 + 200x_1 + 100x_2 + 100x_3 + 50x_4 + \varepsilon_4
\end{aligned}
$$

where ε_i, $i = 1, \ldots, 4$ are normally distributed with mean 0 and standard deviation respectively 120 (i = 1), 270 (i = 2), and 1,000 ($i = 3, 4$). The variable α indicates the strata. Its first 40 coordinates are 1, its 40 following coordinates are 2, and so on up to 25. The aim is to estimate the population total of the variables of interest. The following cases are considered:

Case 1: Only two balancing variables (x_1 and x_2) are considered and a sample of size $n = 25$ is selected with equal inclusion probabilities.

Case 2: Only two balancing variables (x_1 and x_2) are considered and a sample of size $n = 50$ is selected with equal inclusion probabilities.

Case 3: The four balancing variables (x_1, x_2, x_3, and x_4) are considered and a sample of size $n = 25$ is selected with equal inclusion probabilities.

Case 4: The four balancing variables (x_1, x_2, x_3, and x_4) are considered and a sample of size $n = 50$ is selected with equal inclusion probabilities.

In each case, a sample is selected using the new method (Algorithm 2.1) and another one using Chauvet's method with step 3 by landing phase by suppression of variables (Algorithm 2.2). For each sample, the total of the four variables of interest is estimated. The variance of the estimated total of the variables of interest is then computed conducting 10,000 simulations. In order to compare the results, the ratio of the variance of the estimated total of the variables of interest obtained using the new method (Algorithm 2.2) to the variance of the estimated total of the variables of interest obtained using Chauvet's method with step 3 by landing phase by suppression of variables (Algorithm 2.1) is computed. Table 2.1 presents the correlation between the variables of interest and the balancing variables. The results of the simulations are presented in Table 2.2.

Tab. 2.1.: Correlations between the variables of interest and the balancing variables.

Auxiliary variables	Variables of interest			
	y_1	y_2	y_3	y_4
x_1	0.053	0.562	0.492	0.791
x_2	0.035	0.566	0.508	0.432
x_3	-0.020	-0.007	0.517	0.409
x_4	0.022	0.010	0.503	0.203

Tab. 2.2.: Ratio of the variance of the estimated total of the variables of interest obtained using the new method (Algorithm 2.2) to the variance of the estimated total of the variables of interest obtained using Chauvet's method with step 3 by landing phase by suppression of variables (Algorithm 2.1).

	Variables of interest			
	y_1	y_2	y_3	y_4
Case 1	0.974	0.985	0.991	1.019
Case 2	1.024	0.992	0.991	0.970
Case 3	1.016	0.980	1.047	1.065
Case 4	0.977	0.990	1.010	1.040

In order to compare the execution time of both algorithms, a population of size 10,000, and the same balancing variables x_1 and x_2 as above are considered. The population is respectively partitioned into 25, 50, 100, 250, 500, and 1,000 strata of equal size. Samples of one unit per stratum balanced on the two balancing variables are selected with equal inclusion probabilities using Chauvet's method with step 3 by landing phase by suppression of variables (Algorithm 2.1) and the new method (Algorithm 2.2). For each scenario, the mean time in seconds of selection of a sample and the failure rate of selection of a sample are observed for each method. One hundred samples are selected with each of the two methods to obtain these observations. The results are presented in Table 2.3.

Table 2.2 shows that the new method (Algorithm 2.2) produces similar results in term of variance of the estimated total as Chauvet's method with step 3 by landing phase by suppression of variables (Algorithm 2.1). The new method has, however, the advantage over Chauvet's method with step 3 by landing phase by suppression of variables. Indeed, an important gain in execution time arises if the new method is applied and selection of samples with highly stratified designs becomes tractable, as confirmed by Table 2.3. The original

Tab. 2.3.: Mean time in seconds and failure rate of selection of a sample with Chauvet's method with step 3 by landing phase by suppression of variables (Algorithm 2.1) and with the new method (Algorithm 2.2) for 25, 50, 100, 250, 500, and 1,000 strata and 1 unit selected in each stratum with equal inclusion probabilities.

	Algorithm 2.1		Algorithm 2.2	
Number of strata	Mean time	Failure rate	Mean time	Failure rate
25	1.883	0.00	1.941	0.00
50	2.066	0.00	2.108	0.00
100	2.805	0.00	2.289	0.00
250	22.451	0.00	2.959	0.00
500	387.126	0.03	4.039	0.00
1000	9770.745	0.13	6.656	0.00

Chauvet's method with step 3 by unequal probability sampling would, however, perform less well in term of variance than the two methods with step 3 by landing phase by suppression of variables considered here. Indeed, a third step by unequal probability sampling would not take the balancing into account, which would result in a greater variance in the estimations.

2.8.2 Variance approximation formula and estimator

In order to study the performance of the proposed variance approximation formula and its estimator, the same balancing variables x_1 and x_2 and the same variables of interest y_1 to y_4 are considered. The population of size 1,000 is partitioned into 25 equal size strata. Three scenarios are considered, namely the selection using the new method (Algorithm 2.2) of respectively 2 units, 4 units, and 8 units per stratum balanced on the two balancing variables. As above, the units are selected with equal inclusion probabilities. This results in samples of size 50, 100, and 200 respectively. For each scenario, the approximated variance is computed using formula (2.8). The simulation variance of the total estimator $\widehat{t}_y$ and the mean of the variance estimator (2.9) are estimated drawing 10,000 samples for each scenario. The results are presented in Table 2.4.

The mean variance estimator almost matches the approximated variance. The estimator (2.9) is an almost unbiased estimator of the approximated variance (2.8). They are both close to the variance obtained by simulation. However, they tend to slightly underestimate the variance. The gap between the approximated variance and the simulated variance is due to the fact that the formula proposed by Deville and Tillé (2005) does not include the variance induced by the landing phase.

Tab. 2.4.: Approximated variance, mean of the variance estimator estimated using 10,000 simulations, and variance obtained by 10,000 simulations in the case of the estimation of the total of 4 variables of interest using the new method (Algorithm 2.2). Three cases are considered, namely the selection of samples of size $n = 50, 100, 200$ respectively.

		Variables of interest			
n		y_1 $\times 10^7$	y_2 $\times 10^8$	y_3 $\times 10^{11}$	y_4 $\times 10^{11}$
50	Approximated var.	27.53	12.99	9.44	6.29
	Mean var. estimator	27.23	12.95	9.58	6.39
	Simulation var.	28.62	14.59	9.67	7.18
100	Approximated var.	13.04	6.15	4.48	2.99
	Mean var. estimator	13.02	6.14	4.48	2.99
	Simulation var.	13.15	6.54	4.58	3.24
200	Approximated var.	5.80	2.74	1.99	1.32
	Mean var. estimator	5.79	2.74	1.99	1.33
	Simulation var.	5.66	2.90	2.01	1.40

2.8.3 Illustration of the handling of nonresponse

An illustration of the use of the new method (Algorithm 2.2) in the context of nonresponse is shown here. Ilocos data set available in the R package ineq by Zeileis (2013) is considered. The data shows household income in a region of the Philippines called Ilocos and comes from two Philippines' National Statistics Office surveys. The data coming from the 1998 Annual Poverty Indicators Survey are considered here. The sample size is 632. Five domains D_h for $h = 1, \ldots, 5$ are created by grouping households by family size (variable AP.family.size). Each domain, except the last one, refers to 2 consecutive family sizes. The first domain D_1 therefore contains the households whose family size lies in $\{1, 2\}$, the second domain D_2 contains the households whose family size lies in $\{3, 4\}$, and so on until the fourth domain. The fifth domain contains the households whose family size exceeds 8. The variable of interest y is the income (variable AP.income) and $\mathbf{x}_k$ is the vector of domain indicators. A respondents set is created by generating a response indicator vector $\mathbf{r} = (r_k)$, $k \in S$. For $k \in S$, the component r_k is generated from a Bernoulli random variable with parameters 0.55, 0.60, 0.65, 0.70, or 0.75 if the unit $k \in S$ belongs to domain D_1, D_2, D_3, D_4, or D_5 respectively (uniform nonresponse mechanism inside the domains). This results in an overall mean response rate of 0.60.

Then 1,000 hot-deck and survey weighted imputations are conducted by the method proposed in Chauvet et al. (2011b) and presented in Section 2.7.3.

The new method (Algorithm 2.2) is applied as explained in Section 2.7.4 to obtain the imputed values. For each imputation, the imputed estimator

$$\widehat{\theta}_I = \left(\frac{\sum_{k \in D_h} d_k r_k y_k + \sum_{k \in D_h} d_k (1 - r_k) y_k^*}{\sum_{k \in D_h} d_k} \right)_{1 \leq h \leq 5},$$

for the vector of domain means $\theta = \left(\overline{Y}_1, \ldots, \overline{Y}_5 \right)$ is computed. Let $\widehat{\theta}_I^i$ be the imputed estimate of simulation i. To check that the imputation variance of $\widehat{\theta}_I$ is eliminated (or almost), the vector of relative root imputation variances (RRIV) defined as

$$\text{RRIV}\left(\widehat{\theta}_I\right) = \sqrt{\frac{\frac{1}{999} \sum_{i=1}^{1000} \left(\widehat{\theta}_I^i - \frac{1}{1000} \sum_{i=1}^{1000} \widehat{\theta}_I^i \right)^2}{\widehat{\theta}}},$$

where

$$\widehat{\theta} = \left(\frac{\sum_{k \in D_h} d_k y_k}{\sum_{k \in D_h} d_k} \right)_{1 \leq h \leq 5},$$

is computed. Table 2.5 presents the results. It shows that the RRIV is almost eliminated through balanced random imputation with the new method (Algorithm 2.2).

Tab. 2.5.: Relative root imputation variance (RRIV) of the imputed estimator for a vector of domain means obtained through balanced random imputation using the new method (Algorithm 2.2).

Domain	RRIV
1	$4.60 \cdot 10^{-07}$
2	$7.58 \cdot 10^{-08}$
3	$4.66 \cdot 10^{-08}$
4	$1.21 \cdot 10^{-07}$
5	$2.22 \cdot 10^{-07}$

2.9 Conclusion

In this paper, a new algorithm for stratified balanced sampling has been proposed. This algorithm selects samples which are approximately balanced in each stratum, balanced across the entire population and such that a fixed number of units is selected in each stratum. It is faster and more resistent to numerical instability than the previous methods proposed in this context. Moreover, this new algorithm greatly reduces the number of variables considered in the balancing procedures. Therefore, it makes it possible to select

stratified balanced samples in some highly stratified populations that could not be handled using existing methods. A variance approximation formula for the total and its estimator have been proposed. A possible application of the new method to the handling of nonresponse has been provided. Finally, results of a simulation study have confirmed the performance of the proposed method, the accuracy of the formula for the variance approximation and its estimator, and the usefulness of the method for the handling of nonresponse.

Acknowledgements

The authors wish to thank the associate editor and the three reviewers for their useful and constructive comments and suggestions, which helped to considerably improve this manuscript. This research was supported by the Swiss federal statistical office.

Balanced k-nearest neighbor imputation

3

Abstract

In order to overcome the problem of item nonresponse, random imputation methods are often used because they tend to preserve the distribution of the imputed variable. Among the random imputation methods, the random hot-deck has the interesting property of imputing observed values. A new random hot-deck imputation method is proposed. The key innovation of this method is that the selection of donors is viewed as a sampling problem and uses calibration and balanced sampling. This approach makes it possible to select donors such that if the auxiliary variables were imputed, their estimated totals would not change. As a consequence, very accurate and stable totals estimations can be obtained. Moreover, the method is based on a nonparametric procedure. Donors are selected in neighborhoods of recipients. In this way, the missing value of a recipient is replaced with an observed value of a similar unit. This new approach is very flexible and can greatly improve the quality of estimations. Also, this method is unbiased under very different models and is thus resistant to model misspecification. Finally, the new method makes it possible to introduce edit rules while imputing.[1]

Keywords: balanced sampling, calibration, missing data, nonresponse.

3.1 Introduction

Nonresponse is an important problem in surveys. Indeed, the error caused by nonresponse on estimates can be more severe than the error caused by the sampling design. Nonresponse arises when a sampled unit does not respond to one or more items of a survey. One differentiates item nonresponse (a sampled unit does not respond to a particular question) from unit nonresponse (a sampled unit does not respond to the entire survey). Reweighting procedures are often used to deal with unit nonresponse whereas imputation methods are used to treat item nonresponse. Imputation denotes a procedure to replace a missing value with a substituted one.

Imputation methods are classified as either deterministic or random. Deterministic refers to imputation methods that yield the same imputed values if the imputation is repeated. Deterministic imputation methods include ratio imputation, regression imputation, respondent mean imputation, and nearest neighbor imputation. Deterministic imputation methods produce good totals

[1] This chapter is a working paper co-written with Professor Yves Tillé.

estimations. Nevertheless they often fail to estimate quantiles. Random imputation refers to methods that yield different imputed values if the imputation is repeated. Random imputation methods include among others multiple imputation presented in Rubin (1987), imputation with added residuals considered in Chauvet et al. (2011b), and random k-nearest neighbor imputation (kNNI) (see among others Dahl, 2007). Unlike deterministic imputation methods, random imputation methods offer the advantage of tending to preserve the distribution of the imputed variable. Nevertheless such methods imply the presence of an additional amount of variance due to the randomness of imputation, which is called *imputation variance*. Many authors have been interested in minimizing imputation variance. For instance, Kalton and Kish (1981, 1984) proposed two ways to reduce imputation variance. The first way consists of selecting donors among the respondents without replacement rather than with replacement. The second way consists of first constructing strata using the respondents' values of the variable of interest and selecting proportionate stratified samples to act as donors. To the same end, Chen et al. (2000) proposed adjustment of the imputed values; Kim and Fuller (2004) and Fuller and Kim (2005) used fractional hot-deck imputation, and Chauvet et al. (2011b) introduced a class of balanced random imputation methods that consists of randomly selecting residuals while satisfying given constraints.

Imputation methods can alternatively be classified as either donor or predicted value. Donor imputation methods replace the missing value of a nonrespondent with the observed value of a respondent. The unit providing the value is called a *donor* and the unit receiving the value is called a *recipient*. A hot-deck method is a donor imputation method where a missing value is replaced with an observed value extracted from the same survey. Such a method is particularly of interest because it imputes feasible and observed values. The reader can, for instance, refer to Andridge and Little (2010) for a review of hot-deck imputation. In contrast, predicted value imputation methods use function of the respondents values to predict the missing values.

In this paper, a new method of random hot-deck imputation is proposed: the balanced k-nearest neighbor imputation method (bkNNI). The main feature of this method is that the selection of donors is viewed as a sampling problem and uses calibration and balanced sampling. This makes it possible to select donors such that if the auxiliary variables were imputed, their estimated totals would not change. Moreover, this method is based on a nonparametric procedure. Indeed, it provides donors selected in neighborhoods of recipients. In this way, the gap due to one unit's missing value is filled with a similar unit's observed value. The novelty of the proposed method lies not only in the fact that the selection of donors uses balanced sampling but also in the fact that this is paired with a nonparametric selection of donors. Considered together in the

same procedure, these two features imply a robustness in terms of model misspecification. Moreover, this method uses a methodology that makes it possible to take edit rules into account while imputing. The proposed method is also particularly effective, produces negligible imputation variance and a quasi-null bias in specified cases.

This paper is organized as follows. In Section 3.2, notation and concepts of nonresponse are reviewed. A methodology for random hot-deck imputation methods is introduced in Section 3.3. Section 3.4 focusses on the presentation of the bkNNI. Section 3.5 introduces a formula to approximate the conditional imputation variance of the total when the new method is applied. Section 3.6 describes the models underlain by the method and studies the asymptotic properties of the total estimator. Then, in Section 3.7, the performance of the new imputation method and the accuracy of the proposed estimator for imputation variance are tested through a simulation study. A short discussion concludes the paper in Section 3.8.

3.2 Notation and concepts of nonresponse

Consider a finite population $U = \{1, 2, \ldots, i, \ldots, N\}$ and suppose that the target is the variable of interest $\mathbf{y} = (y_1, y_2, \ldots, y_i, \ldots, y_N)^\top$. In a first phase, a random sample S of size n is drawn from U with a given sampling design $p(\cdot)$ where $p(s) = \Pr(S = s)$ for $s \subset U$. Let $\pi_i = \Pr(i \in S)$ denote the first order inclusion probability of unit i and let $d_i = 1/\pi_i$ denote its Horvitz-Thompson weight (Horvitz and Thompson, 1952). If a census is considered, the inclusion probabilities and the design weights are equal to 1. A vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iQ})^\top$ of Q auxiliary variables is assumed to be known for each unit i in the sample S. In what follows, it is supposed that one of the auxiliary variables is constant. In a second phase, a subset of respondents $S_r \subset S$ is obtained from S with a usually unknown conditional distribution $q(S_r|S)$. The values y_i of the variable of interest are known for the units of S_r only. Let $S_m = S \setminus S_r$ denote the complement of S_r in S, i.e. the subsample of S containing the units with missing data (the nonrespondents). The respective sizes of these subsets are n_r and n_m with $n_r + n_m = n$. For $i \in S$, let r_i be the response indicator variable

$$r_i = \begin{cases} 1 & \text{if unit } i \text{ belongs to } S_r, \\ 0 & \text{otherwise.} \end{cases}$$

It is supposed that the units respond independently from each other. For each unit $i \in S$, r_i is therefore generated from a Bernoulli random variable with parameter $\theta_i = \Pr(i \in S_r | i \in S)$. The parameter θ_i represents the response propensity of unit i and is usually unknown. Hence, the conditional distribution

$q(S_r|S)$ is a Poisson sampling design, i.e.

$$q(S_r|S) = \prod_{i \in S_r} \theta_i \prod_{i \in S_m} (1 - \theta_i).$$

Three types of nonresponse mechanisms exist: uniform, ignorable, and non-ignorable. A uniform nonresponse mechanism is a nonresponse mechanism where each unit of the population has the same response propensity, i.e. $\theta_i = \theta$ for each $i \in U$. It is in this case said that the data is missing completely at random (MCAR). An ignorable nonresponse mechanism (Rubin, 1976) is a nonresponse mechanism where the response propensity θ_i does not depend on the variable of interest once the auxiliary variables have been taken into account. In the case of an ignorable nonresponse mechanism, the data is said to be missing at random (MAR). Finally, a non-ignorable nonresponse mechanism is a nonresponse mechanism where the response propensity θ_i depends on the variable of interest. The data is in this case said to be not missing at random (NMAR). In a third phase, nonresponse can be corrected through imputation. Imputed values y_j^*, $j \in S_m$ are drawn with a conditional distribution

$$I\left(y_j^* | S, S_r\right).$$

The aim is to estimate the population total

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

of the variable of interest $\mathbf{y}$. In the case of complete response, the estimator

$$\widehat{Y} = \sum_{i \in S} d_i y_i,$$

is a design-unbiased estimator of Y. In the presence of nonresponse, the previous estimator is intractable and the imputed estimator

$$\widehat{Y}_I = \sum_{i \in S_r} d_i y_i + \sum_{j \in S_m} d_j y_j^*,$$

is used. Moreover, consider

$$\begin{aligned}
\mathbf{X} &= \sum_{i \in U} \mathbf{x}_i, \\
\widehat{\mathbf{X}} &= \sum_{i \in S} d_i \mathbf{x}_i, \\
\widehat{\mathbf{X}}_I &= \sum_{i \in S_r} d_i \mathbf{x}_i + \sum_{j \in S_m} d_j \mathbf{x}_j^*,
\end{aligned}$$

where $\mathbf{x}_j^*$ represents the imputed value we would have obtained for $j \in S_m$

if we were to impute the auxiliary variables. For instance, suppose hot-deck imputation is used. For each nonrespondent $j \in S_m$, a donor $i \in S_r$ is chosen. The imputed value y_j^* for unit $j \in S_m$ is therefore the donor's observed value y_i. In this case, the imputed value $\mathbf{x}_j^*$ is the observed value $\mathbf{x}_i$ of the same donor as the one used to obtain y_j^*.

In the presented framework, the bias and variance of an imputed estimator $\widehat{\theta}_I$ (for a total or another statistic θ) are given by

$$\begin{aligned} \text{Bias}\left(\widehat{\theta}_I\right) &= \text{E}_p\text{E}_q\text{E}_I\left(\widehat{\theta}_I - \theta\right), \\ \text{Var}\left(\widehat{\theta}_I\right) &= \text{Var}_p\text{E}_q\text{E}_I\left(\widehat{\theta}_I\right) + \text{E}_p\text{Var}_q\text{E}_I\left(\widehat{\theta}_I\right) + \text{E}_p\text{E}_q\text{Var}_I\left(\widehat{\theta}_I\right), \end{aligned} \quad (3.1)$$

where the subscripts p, q and I indicate respectively the expectations and variances with regards to the sampling mechanism, with regards to the nonresponse mechanism, and with regards to the imputation mechanism. The first term in Expression (3.1) represents the sampling variance, the second term represents the nonresponse variance and the last term represents the imputation variance.

3.3 Methodology for random hot-deck imputation methods

In this section, we propose an original formalization for random hot-deck donor imputation. The proposed method is presented by means of this formalization, but this last one can be used for any random hot-deck method. Random hot-deck imputation consists of replacing a missing value with an observed value extracted from the same survey. For each nonrespondent, a donor is hence randomly chosen among the respondents. Consequently, a random hot-deck imputation can be achieved through the realization of a random matrix $\boldsymbol{\phi} = (\phi_{ij})$, $(i,j) \in S_r \times S_m$ such that

$$\phi_{ij} = \begin{cases} 1 & \text{if nonrespondent } j \text{ is imputed by respondent } i, \\ 0 & \text{otherwise,} \end{cases}$$

which can be rewritten

$$\phi_{ij} = \mathbb{1}_{y_j^* = y_i}. \quad (3.2)$$

As exactly one donor must be selected for each nonrespondent, $\boldsymbol{\phi}$ must satisfy

$$\sum_{i \in S_r} \phi_{ij} = 1, \quad \text{for each } j \in S_m. \quad (3.3)$$

It is here considered that a respondent can be used to impute several nonrespondents. Therefore, no conditions are imposed on

$$\sum_{j \in S_m} \phi_{ij},$$

for $i \in S_r$. Taking the conditional expectation of Equation (3.2) generates a matrix of imputation probabilities $\boldsymbol{\psi} = (\psi_{ij})$, $(i,j) \in S_r \times S_m$

$$\psi_{ij} = \mathrm{E}_I(\phi_{ij}) = \mathrm{E}_I\left(\mathbb{1}_{y_j^*=y_i}\right) = \Pr\left(y_j^* = y_i | S, S_r\right). \tag{3.4}$$

By definition, $\boldsymbol{\psi}$ satisfies

$$\sum_{i \in S_r} \psi_{ij} = 1 \quad \text{for each } j \in S_m, \tag{3.5}$$

$$\psi_{ij} \geq 0 \quad \text{for each } (i,j) \in S_r \times S_m. \tag{3.6}$$

The considered methodology for random hot-deck donor imputation is therefore operated in two stages. In the first stage, the matrix of imputation probabilities $\boldsymbol{\psi}$ is defined. Then, in the second stage, a realization of the matrix of imputation $\boldsymbol{\phi}$ is carried out. This methodology has, among other things, the interesting property to make it possible to implement edit rules in the imputation process to correct the data at the record level. Indeed, suppose that the value of the variable of interest y_i of a respondent $i \in S_r$ is, for some reason, inconsistent with a nonrespondent $j \in S_m$. To take this inconsistency into account, it is sufficient to set a zero in coefficient ψ_{ij} of the matrix of imputation probabilities $\boldsymbol{\psi}$. In this way, indeed, unit $i \in S_r$ is removed from the set of possible donors for nonrespondent $j \in S_m$. Note that it might then be required to rescale the column of matrix $\boldsymbol{\psi}$ corresponding to nonrespondent j in order to ensure that this one still sums up to 1. Two examples of application of this methodology are provided below.

Example 1: simple random imputation method (with replacement)

Simple random imputation consists of selecting, for each nonrespondent, a donor among the respondents. The donors are selected randomly, with replacement (a donor can be used several times), and with equal probabilities. This results in the matrix of imputation probabilities $\boldsymbol{\psi}^{[srs]} = \left(\psi_{ij}^{[srs]}\right)$, $(i,j) \in S_r \times S_m$ with

$$\psi_{ij}^{[srs]} = \frac{1}{n_r}.$$

Notation $[srs]$ in superscript of the matrix $\boldsymbol{\psi}$ means that the matrix linked to simple random imputation is considered.

Example 2: random k-nearest neighbor imputation method

The *k-nearest neighbors* of a nonrespondent unit $j \in S_m$ are here defined as its k most similar respondents units $i \in S_r$, i.e.

$$\mathrm{knn}(j) = \{i \in S_r | \mathrm{rank}\left(d(i,j)\right) \leq k\},$$

where $d(\cdot,\cdot)$ is for instance the Mahalanobis distance defined through the nonconstant auxiliary variables

$$d(i,j) = \left\{ (\mathbf{x}_i - \mathbf{x}_j)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right\}^{1/2},$$

where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of the nonconstant auxiliary variables. If the values of the auxiliary variables are known at the sample level only, $\boldsymbol{\Sigma}$ must be estimated.

The random k-nearest neighbor imputation method (kNNI) consists of replacing the missing value of a unit $j \in S_m$ with the value of one of its k-nearest neighbors. The donors are randomly selected with equal probabilities. This results in the matrix of imputation probabilities $\boldsymbol{\psi}^{[k]} = \left(\psi_{ij}^{[k]}\right)$, $(i,j) \in S_r \times S_m$ with

$$\psi_{ij}^{[k]} = \begin{cases} \frac{1}{k} & \text{if } i \text{ belongs to the } k \text{ nearest neighbors of } j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

Notation $[k]$ in superscript of the matrix $\boldsymbol{\psi}$ means that the matrix linked to kNNI is considered. The matrix of imputation probabilities related to the kNNI is a matrix containing exactly k non-null coefficients in each column and all these non-null coefficients are equal to $1/k$. This particular matrix of imputation probabilities is the starting point of the method proposed in this paper.

3.4 Balanced k-nearest neighbor imputation method

3.4.1 Aim of the method

Random imputation methods show the nice feature that they tend to preserve the distribution of the variable being imputed. This is often not the case for deterministic imputation methods. In return, this randomness implies the undesirable presence of imputation variance. Hence, a random imputation method that makes it possible to keep the imputation variance relatively small is of great interest. Moreover, donor imputation methods, such as hot-deck, have the advantage of imputing feasible and observed values.

An imputation method can be based on either a parametric procedure or a nonparametric procedure. Parametric procedures rely on strong modeling assumptions. In this case, one can assume that the distribution of the variable of interest is known, but the parameters of this distribution must be estimated. Practically, this kind of hypotheses is satisfied in only a few cases. On the other hand, nonparametric procedures typically rely on weaker assumptions. Hence, such procedures are usually much more flexible than parametric procedures. An imputation method based on a nonparametric procedure therefore makes

it possible to handle a wider range of data without violating its underlying hypotheses than an imputation method based on a parametric procedure.

The proposed method shows the nice features stated above. First, it is a random hot-deck imputation method. It therefore tends to preserve the distribution of the variable being imputed and it imputes observed and feasible values. Moreover, even though it is random, the proposed method makes it possible to control the imputation variance of the total estimator. Indeed, the imputation process conserves the estimator of the total of the auxiliary variables. If the auxiliary variables suffered from nonresponse, their imputed total estimator would match their total estimator under complete response. The imputation mechanism relative to bkNNI is therefore such that conditionally on the sampling mechanism and on the nonresponse mechanism

$$\widehat{\mathbf{X}}_I = \widehat{\mathbf{X}}. \tag{3.8}$$

If a linear model fits well the relation between the variable of interest and the auxiliary variables, the imputation variance of the total estimator can in this way be reduced. As a result, the proposed method makes it possible to impute randomly while keeping imputation variance of the total estimator relatively small. Then, the proposed method is based on a nonparametric procedure. Indeed, the donors are chosen in neighborhoods of recipients. For each nonrespondent, a donor is randomly selected among its k-nearest neighbors. As nonparametric, this procedure is very adaptive and makes it possible to impute values that are close to the unobserved missing value for a wide range of data. Moreover and as stated above, the imputation process conserves the estimator of the total of the auxiliary variables. If a linear model fits well the relation between the variable of interest and the auxiliary variables, this property implies that the imputed total estimator for the variable being imputed is close to the total estimator that would have been obtained under complete response (Horvitz-Thompson estimator). As this last one is unbiased, the obtained imputed total estimator is nearly unbiased in the case specified above. Details regarding the properties of the total estimator are given in Section 3.6. Last, the proposed method used the methodology proposed in Section 3.3. This makes it possible for the user to take into account edit rules directly while imputing as explained in Section 3.3.

In what follows, it is explained how donors can be obtained. Notation $[bk]$ in the superscript of the matrices $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ means that the matrices linked to the bkNNI are considered. The method proceeds in two steps. The first step consists of obtaining the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$ whereas the second step consists of generating a realization of the matrix of imputation $\boldsymbol{\phi}^{[bk]}$. The aim is that the imputation mechanism satisfies Equation (3.8). With this

aim, the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$ is, in the first step, constructed such that Equation

$$\mathrm{E}_I\left(\widehat{\mathbf{X}}_I\right) = \widehat{\mathbf{X}} \tag{3.9}$$

is satisfied and the matrix of imputation probabilities $\boldsymbol{\phi}^{[bk]}$ is, in the second step, generated such that Equation

$$\widehat{\mathbf{X}}_I = \mathrm{E}_I\left(\widehat{\mathbf{X}}_I\right). \tag{3.10}$$

is satisfied. These two steps are presented in Section 3.4.3 and in Section 3.4.6 respectively. Equation (3.9) together with Equation (3.10) lead to Equation (3.8). As a result, the imputation mechanism obtained through the two steps described above satisfies Equation (3.8).

3.4.2 Calibration

The aim of this Section is to briefly describe calibration (Deville and Särndal, 1992) which is the main tool used in Section 3.4.3 to obtain the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$. Suppose a vector of Q auxiliary variables $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iQ})^\top$ is known for each unit of the population U. The aim of calibration is to find calibration weights w_i for $i \in S$ as close as possible to the initial design weights $d_i = 1/\pi_i$ (in an average sense for a given distance) while respecting the calibration equation

$$\sum_{i \in S} w_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i = \mathbf{X}.$$

Several distance functions are proposed in Deville and Särndal (1992) as a means of measuring the distance between the initial design weights $d_i = 1/\pi_i$ and the final weights w_i. Each distance provides a particular form for the final weights w_i. The raking method is the calibration obtained considering the distance function $G(\cdot, \cdot)$ given by

$$G(w_i, d_i) = w_i \log\left(\frac{w_i}{d_i}\right) - w_i + d_i.$$

This leads to the final weights $w_i = d_i \exp\left(\boldsymbol{\lambda}^\top \mathbf{x}_i\right)$ where the vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_Q)^\top$ is the solution to the calibration equation

$$\sum_{i \in S} d_i \exp\left(\boldsymbol{\lambda}^\top \mathbf{x}_i\right) \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i.$$

From Deville and Särndal (1992), if a solution $\boldsymbol{\lambda}$ to this calibration problem exists, then it is unique.

3.4.3 Obtaining the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$

In this Section, a procedure to obtain the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$ is presented. This matrix must satisfy

$$\psi_{ij}^{[bk]} \neq 0 \quad \text{only if} \quad i \in \text{knn}(j),$$

because it is imposed that donors are chosen in neighborhoods of recipients. Moreover, as stated in Section 3.3, a matrix of imputation must satisfy equations (3.5) and (3.6). Finally, as stated in Section 3.4.1, in order to select donors such that Equation (3.8) is satisfied, it is imposed on $\boldsymbol{\psi}^{[bk]}$ to satisfy Equation (3.9), i.e

$$\mathrm{E}_I\left(\widehat{\mathbf{X}}_I\right) = \widehat{\mathbf{X}}. \tag{3.11}$$

This constraint states that conditionally on the sampling mechanism and the nonresponse mechanism, the imputation mechanism provides an unbiased imputed total estimator for the auxiliary variables. Equation (3.11) is equivalent to

$$\sum_{j\in S_m} d_j \sum_{i\in S_r} \psi_{ij}^{[bk]} \mathbf{x}_i = \sum_{j\in S_m} d_j \mathbf{x}_j.$$

Hence, the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]} = \left(\psi_{ij}^{[bk]}\right)$ must satisfy simultaneously

$$\psi_{ij}^{[bk]} \neq 0 \quad \text{only if} \quad i \in \text{knn}(j), \tag{3.12}$$

$$\sum_{i\in S_r} \psi_{ij}^{[bk]} = 1 \quad \text{for each } j \in S_m, \tag{3.13}$$

$$\psi_{ij}^{[bk]} \geq 0 \quad \text{for each } (i,j) \in S_r \times S_m, \tag{3.14}$$

$$\sum_{j\in S_m} d_j \sum_{i\in S_r} \psi_{ij}^{[bk]} \mathbf{x}_i = \sum_{j\in S_m} d_j \mathbf{x}_j. \tag{3.15}$$

The existence of a matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$ satisfying the above conditions and the choice of the number of nearest neighbors k are discussed in Section 3.4.4. Algorithm 3.1 presents a procedure to obtain the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$. The main idea of Algorithm 3.1 is to find a matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$ close to the matrix of imputation probabilities relative to the kNNI $\boldsymbol{\psi}^{[k]}$, while satisfying equations (3.12), (3.13), (3.14), and (3.15). This Algorithm initializes with the matrix $\boldsymbol{\psi}^{[k]}$. Throughout all the steps, a null coefficient remains null which implies that equation (3.12) is satisfied. Thereafter, calibrations and normalizations are alternated. Calibrations provide matrices $\boldsymbol{\psi}(2\ell)$ for $\ell \geq 1$ satisfying equations (3.14) and (3.15). However, these matrices $\boldsymbol{\psi}(2\ell)$ are not matrices of imputation probabilities because they do not satisfy (3.13). Normalizations provide matrices $\boldsymbol{\psi}(2\ell+1)$

for $\ell \geq 1$ satisfying equations (3.13) and (3.14) but not necessarily satisfying equation (3.15). Iterations stop when the matrix $\boldsymbol{\psi}(2\ell+1)$ obtained by normalization approximately satisfies equation (3.15). The matrix $\boldsymbol{\psi}^{[bk]}$ is the last $\boldsymbol{\psi}(2\ell+1)$ considered. Hence, $\boldsymbol{\psi}^{[bk]}$ satisfies equations (3.12), (3.13), (3.14), and (3.15) simultaneously.

Algorithm 3.1 Procedure to obtain the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$

Step 1: *Initialization*
Set $\boldsymbol{\psi}(1) = \boldsymbol{\psi}^{[k]}$, the matrix of imputation probabilities relative to the kNNI defined in Expression (3.7).

Step 2: *Iterations*
Repeat for $\ell = 1, 2, \ldots$

- *Calibration*
  For $i \in S_r$, consider the initial weights $\tilde{d}_i = \sum_{j \in S_m} d_j \psi(2\ell-1)_{ij}$ and obtain the calibrated weights $w_i = \tilde{d}_i \exp\left(\boldsymbol{\lambda}^\top \mathbf{x}_i\right)$ by means of the raking method. The calibration equation is

$$\sum_{i \in S_r} w_i \mathbf{x}_i = \sum_{j \in S_m} d_j \mathbf{x}_j.$$

- Let $\boldsymbol{\psi}(2\ell)$ be the matrix defined as $\psi(2\ell)_{ij} = \psi(2\ell-1)_{ij} \exp\left(\boldsymbol{\lambda}^\top \mathbf{x}_i\right)$.
- *Normalization*
  Let $\boldsymbol{\psi}(2\ell+1)$ be the matrix defined as $\psi(2\ell+1)_{ij} = \frac{\psi(2\ell)_{ij}}{\sum_{l \in S_r} \psi(2\ell)_{lj}}$.
- *Stop criterion*
  If

$$\max_{1 \leq q \leq Q} \left| \frac{\sum_{j \in S_m} \sum_{i \in S_r} d_j \psi(2\ell+1)_{ij} x_{iq} - \sum_{j \in S_m} d_j x_{jq}}{\sum_{j \in S_m} d_j x_{jq}} \right| \leq tol$$

  for a small fixed error tolerance tol, then
  - Set $\boldsymbol{\psi}^{[bk]} = \boldsymbol{\psi}(2\ell+1)$,
  - Stop.

3.4.4 Choice of k and existence of $\boldsymbol{\psi}^{[bk]}$

In this Section, the choice of the number of nearest neighbors k and existence of a matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$ having the required properties are discussed.

The number of nearest neighbors k is relevant when constructing the matrix $\boldsymbol{\psi}^{[bk]}$ in Section 3.4.3. Indeed, this matrix must contain at most $k \cdot n_m$ non-null coefficients as confirmed by Equation (3.12). Moreover, the coefficients of this matrix must satisfy Equation (3.13) and Equation (3.15). These two equations together form a system of $n_m + q$ constraints. Hence, when constructing the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$, the aim is to find $k \cdot n_m$ unknown coefficients that satisfy $n_m + q$ linear constraints. As a result, a necessary condition to find a matrix $\boldsymbol{\psi}^{[bk]}$ satisfying Equations (3.12), (3.13) and (3.15)

is

$$k \geq \frac{n_m + q}{n_m}. \tag{3.16}$$

However, this condition is not sufficient to ensure that a solution $\boldsymbol{\psi}^{[bk]}$ exists and elements external to the choice of k have an effect. Such an element is the configuration of the nonrespondents. Indeed, Equation (3.14) and Equation (3.13) imply together that all the coefficients $\psi_{ij}^{[bk]}$ must lie between 0 and 1. Consider the extreme case in which all the units with the largest values of one auxiliary variable are nonrespondents. For this auxiliary variable, and as all the coefficients $\psi_{ij}^{[bk]}$ must lie between 0 and 1, the small values of respondents do not make it possible to compensate the large missing values of the nonrespondents. Hence, a matrix $\boldsymbol{\psi}^{[bk]}$ with the properties stated above might not exist, whatever the value of k is. Condition (3.16) is therefore necessary but not sufficient for a solution to exist. Note that, if a solution exists, then it is unique. Indeed, the solution to the calibration problem in Algorithm 3.1 is unique (see last sentence of Section 3.4.2).

The choice of k can have an impact on the bias of the total estimator. Indeed, the smaller k is, the closer the values of the auxiliary variables are in a neighborhood of k units and therefore the closer the values of the variable of interest y tend to be in the same neighborhood. As a result and from Property 3.3 below, under the hypothesis that the values of variables y are similar in a neighborhood, the smaller k is, the smaller the bias of the total estimator tends to be. Hence, with the aim of controlling the bias, k should be chosen as small as possible. However, note that this choice has no impact on the bias of the total estimator when the hypotheses of Property 3.1 or those of Property 3.2 are satisfied. Indeed, the bias of the total estimator are in these cases null whatever the value of k is. Moreover, the larger k is, the larger the imputation variance of the total is.

For these reasons, we suggest to choose the smallest value k for which a solution for the computation of matrix $\boldsymbol{\psi}^{[bk]}$ exists. Thus, we propose to fix the value of k as follows. First, choose a value of k that satisfies Equation (3.16) and that is relatively small. Then, apply Algorithm 3.1 and see if this one finds a solution, i.e. returns a matrix $\boldsymbol{\psi}^{[bk]}$ satisfying the required conditions. If this is the case, the user can then apply Algorithm 3.2 in order to select the donors. Otherwise, we propose gradually increasing the value of k and repeating this procedure until a solution is found.

3.4.5 Stratified balanced sampling

The aim of this Section is to briefly describe stratified balanced sampling. This represents the main tool used in Section 3.4.6 to obtain the matrix of imputation $\boldsymbol{\phi}^{[bk]}$. Suppose a vector of Q auxiliary variables $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iQ})^\top$ is

known for each unit of the population U. A sampling design $p(\cdot)$ is said to be balanced on these auxiliary variables if

$$\sum_{i \in s} \frac{\mathbf{x}_i}{\pi_i} = \sum_{i \in U} \mathbf{x}_i,$$

for each $s \subset U$ with $p(s) > 0$. This last equation can be rewritten

$$\sum_{i \in U} \frac{\mathbf{x}_i}{\pi_i} \mathbb{1}_{i \in s} = \sum_{i \in U} \mathbf{x}_i,$$

where $\mathbb{1}_{i \in s}$ is the indicator function which takes value 1 if unit i belongs to s and 0 otherwise. The cube method (Deville and Tillé, 2004) is a method for balanced sampling. It allows a balanced sample to be selected while satisfying the inclusion probabilities in the sense that $\mathrm{E}\left(\mathbb{1}_{i \in S}\right) = \pi_i$ for each $i \in U$. Suppose moreover that the population U is partitioned into H nonoverlapping strata $U_1, U_2, \ldots, U_H$. A stratified balanced sampling design is a sampling design balanced in each stratum, i.e.

$$\sum_{i \in U_h} \frac{\mathbf{x}_i}{\pi_i} \mathbb{1}_{i \in s} = \sum_{i \in U_h} \mathbf{x}_i \quad \text{for each } 1 \leq h \leq H, \tag{3.17}$$

for each $s \subset U$ with $p(s) > 0$. Chauvet (2009) and Hasler and Tillé (2014) proposed methods for stratified balanced sampling which are based on the cube method. The algorithm proposed in Hasler and Tillé (2014) is particularly fast and is applicable when the number of strata is very large. The samples selected with these methods are approximately balanced in each stratum and approximately balanced in the overall population while satisfying the inclusion probabilities in the sense that

$$\mathrm{E}\left(\mathbb{1}_{i \in S}\right) = \pi_i \quad \text{for each } i \in U. \tag{3.18}$$

Remark 3.1 *If the sum of the inclusion probabilities*

$$n_h = \sum_{i \in U_h} \pi_i$$

is integer in each stratum U_h, $1 \leq h \leq H$, the algorithm proposed in Hasler and Tillé (2014) selects exactly n_h units in each stratum U_h, $1 \leq h \leq H$.

3.4.6 Selection of the donors

In this Section, a procedure to obtain the matrix of imputation $\boldsymbol{\phi}^{[bk]}$ from the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$ is presented. The main idea of this procedure is to select a donor among the respondents for each nonrespondent while satisfying constraints. The key feature of the method is that the selection

of donors is viewed as a sampling problem and the constraints are satisfied through stratified balanced sampling, where only one unit is selected in each stratum.

As stated in Section 3.4.1, in order to select donors such that Equation (3.8) is satisfied, it is imposed on $\boldsymbol{\phi}^{[bk]}$ to satisfy Equation (3.10), i.e

$$\widehat{\mathbf{X}}_I = \mathrm{E}_I\left(\widehat{\mathbf{X}}_I\right). \tag{3.19}$$

This constraint states that the imputation variance of the auxiliary variables cancels out. A sufficient condition for Equation (3.19) to hold is

$$\sum_{i\in S_r} d_j \phi_{ij}^{[bk]} \mathbf{x}_i = \sum_{i\in S_r} d_j \psi_{ij}^{[bk]} \mathbf{x}_i \quad \text{for each } j \in S_m,$$

which can be rewritten

$$\sum_{i\in S_r} \frac{d_j \psi_{ij}^{[bk]} \mathbf{x}_i}{\psi_{ij}^{[bk]}} \phi_{ij}^{[bk]} = \sum_{i\in S_r} d_j \psi_{ij}^{[bk]} \mathbf{x}_i \quad \text{for each } j \in S_m.$$

Moreover, matrices $\boldsymbol{\psi}^{[bk]}$ and $\boldsymbol{\phi}^{[bk]}$ must be linked by Equation (3.4) , i.e.

$$\psi_{ij}^{[bk]} = \mathrm{E}_I\left(\phi_{ij}^{[bk]}\right) \text{ for each } (i,j) \in S_r \times S_m.$$

The aim is therefore to generate matrix $\boldsymbol{\phi}^{[bk]}$ such that

$$\sum_{i\in S_r} \frac{d_j \psi_{ij}^{[bk]} \mathbf{x}_i}{\psi_{ij}^{[bk]}} \phi_{ij}^{[bk]} = \sum_{i\in S_r} d_j \psi_{ij}^{[bk]} \mathbf{x}_i \quad \text{for each } j \in S_m, \tag{3.20}$$

$$\mathrm{E}_I\left(\phi_{ij}^{[bk]}\right) = \psi_{ij}^{[bk]} \quad \text{for each } (i,j) \in S_r \times S_m. \tag{3.21}$$

A matrix $\boldsymbol{\phi}^{[bk]}$ satisfying exactly the above equations often does not exist. However, when generating this matrix with the procedure presented below, the constraints are relaxed until a solution is found. See Remark 3.3. A solution to this problem was proposed in Chauvet et al. (2011b). A slight modification of that is presented here. Consider the population of cells

$$\dot{U} = \{(i,j) | i \in S_r, j \in S_m\}.$$

This population is partitioned into n_m strata $\dot{U}_j$, $j \in S_m$ where

$$\dot{U}_j = \{(i,j) | i \in S_r\}.$$

Each stratum corresponds to one nonrespondent. Then, exactly one unit will be selected in each stratum, providing in this way exactly one donor for each nonrespondent. For each unit $(i,j) \in \dot{U}$, consider the initial inclusion

probability

$$\dot{\pi}_{(i,j)} = \psi_{ij}^{[bk]},$$

and the auxiliary variables

$$\dot{\mathbf{x}}_{(i,j)} = d_j \psi_{ij}^{[bk]} \mathbf{x}_i.$$

Moreover, as $\phi_{ij}^{[bk]}$ is 1 if respondent i is the donor for nonrespondent j and 0 otherwise, consider

$$\mathbb{1}_{(i,j)\in S} = \phi_{ij}^{[bk]}.$$

It means that respondent $i \in S_r$ is used to impute the missing value of nonrespondent $j \in S_m$ if unit (i,j) is selected in the sample. The problem formed by equations (3.20) and (3.21) can be rewritten as follows

$$\sum_{(i,j)\in \dot{U}_j} \frac{\dot{\mathbf{x}}_{(i,j)}}{\dot{\pi}_{(i,j)}} \mathbb{1}_{(i,j)\in S} = \sum_{(i,j)\in \dot{U}_j} \dot{\mathbf{x}}_{(i,j)} \quad \text{for each } j \in S_m, \tag{3.22}$$

$$\mathrm{E}_I\left(\mathbb{1}_{(i,j)\in S}\right) = \dot{\pi}_{(i,j)} \quad \text{for each } (i,j) \in \dot{U}. \tag{3.23}$$

This is a typical problem of stratified balanced sampling where only one unit is selected in each stratum because each respondent receives exactly one value. Equations (3.22) and (3.23) correspond respectively to equations (3.17) and (3.18). The procedure to obtain matrix $\boldsymbol{\phi}^{[bk]}$ therefore uses stratified balanced sampling and is presented in Algorithm 3.2. The first step of the algorithm consists of selecting a stratified balanced sample in the cell population $\dot{U} = \{(i,j) | i \in S_r, j \in S_m\}$ such that equations (3.22) and (3.23) are satisfied. In a second and last step, matrix $\boldsymbol{\phi}^{[bk]}$ is obtained by setting $\phi_{ij}^{[bk]} = 1$ if the cell (i,j) has been selected in the sample and 0 otherwise.

Remark 3.2 *For each $j \in S_m$, the following equation is satisfied*

$$\sum_{(i,j)\in \dot{U}_j} \dot{\pi}_{(i,j)} = \sum_{i\in S_r} \psi_{ij}^{[bk]} = 1.$$

This means that the sum of the inclusion probabilities is equal to 1 in each stratum considered in the stratified balanced sampling problem of Algorithm 3.2. Therefore, as the procedure for balanced stratified sampling proposed in Hasler and Tillé (2014) is applied in Algorithm 3.2 and from Remark 3.1, exactly 1 unit is selected in each stratum, i.e

$$\sum_{(i,j)\in \dot{U}_j} \mathbb{1}_{(i,j)\in S} = 1 \quad \textit{for each } j \in S_m.$$

Moreover, as

$$\sum_{(i,j)\in\dot{U}_j} \mathbb{1}_{(i,j)\in S} = \sum_{i\in S_r} \phi_{ij}^{[bk]} \quad \textit{for each } j \in S_m,$$

the matrix $\boldsymbol{\phi}^{[bk]}$ satisfies Equation (3.3). This means that Algorithm 3.2 provides exactly one donor for each nonrespondent $j \in S_m$.

Remark 3.3 *It is often not possible to select samples such that the balancing equations are exactly satisfied. As a result, Algorithm 3.2 often generates a matrix $\boldsymbol{\phi}^{[bk]}$ such that Equation (3.22) is only approximately satisfied. Equivalently, donors are often selected such that Equation (3.19) is only approximately satisfied.*

Algorithm 3.2 Procedure to obtain the matrix of imputation $\boldsymbol{\phi}^{[bk]}$

Step 1: *Stratified balanced sampling*

Select a stratified balanced sample S in the cells population $\dot{U} = \{(i,j)|i \in S_r, j \in S_m\}$ with the method proposed in Hasler and Tillé (2014) such that

$$\sum_{(i,j)\in\dot{U}_j} \frac{\dot{\mathbf{x}}_{(i,j)}}{\dot{\pi}_{(i,j)}} \mathbb{1}_{(i,j)\in S} = \sum_{(i,j)\in\dot{U}_j} \dot{\mathbf{x}}_{(i,j)} \quad \text{for each } j \in S_m,$$

$$\mathrm{E}_I\left(\mathbb{1}_{(i,j)\in S}\right) = \dot{\pi}_{(i,j)} \quad \text{for each } (i,j) \in \dot{U},$$

where

- $\dot{\mathbf{x}}_{(i,j)} = d_j \psi_{ij}^{[bk]} \mathbf{x}_i$ is the vector of balancing variables linked to unit $(i,j) \in S_r \times S_m$,
- $\dot{\pi}_{(i,j)} = \psi_{ij}^{[bk]}$ is the inclusion probability attached to unit $(i,j) \in S_r \times S_m$,
- $\dot{U}_j = \{(i,j)|i \in S_r\}$ is the stratum attached to unit $j \in S_m$.

Step 2: *Matrix of imputation*

Let $\boldsymbol{\phi}^{[bk]}$ be the matrix defined as $\phi_{ij}^{[bk]} = \mathbb{1}_{(i,j)\in S}$.

3.5 Approximation of conditional imputation variance

A procedure to approximate the imputation variance conditional on the design and on the nonresponse mechanism of the total $\mathrm{Var}_I(\widehat{Y}_I)$ is described in this Section. For the sake of brevity, we refer to the latter variance as conditional imputation variance, because it is conditional to the sampling design and the nonresponse mechanism. The proposed procedure relies on the idea that, in the new method, the selection of donors is viewed as a sampling problem and imputation is achieved through stratified balanced sampling. However, all the values of the variable of interest are known prior to selecting the sample of imputed values. Indeed, donors are selected among respondents and the values of the variable of interest are fully observed for these. Deville and Tillé (2005) proposed an approximation formula for the variance of a total under balanced sampling that can be used in this framework. Based on this approximation formula, we propose the following formula for the conditional imputation

variance.

$$\mathrm{Var}_I^{app}(\widehat{Y}_I) = \sum_{i \in S_r} \sum_{\substack{j \in S_m \\ \psi_{ij}^{[bk]} \neq 0}} c_{ij} d_j^2 \left(y_i - \mathbf{b}^\top \mathbf{x}_i\right)^2, \tag{3.24}$$

$$c_{ij} = \psi_{ij}^{[bk]} \left(1 - \psi_{ij}^{[bk]}\right) \frac{n_m k}{n_m k - q},$$

$$\mathbf{b} = \left(\sum_{i \in S_r} \sum_{\substack{j \in S_m \\ \psi_{ij}^{[bk]} \neq 0}} c_{ij} d_j^2 \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i \in S_r} \sum_{\substack{j \in S_m \\ \psi_{ij}^{[bk]} \neq 0}} c_{ij} d_j^2 \mathbf{x}_i y_i.$$

Notice that only a single respondents set is necessary to approximate the conditional imputation variance. However, it can underestimate this one. Indeed, this formula comes from the variance of the total for balanced sampling. When applying balanced sampling, it is often not possible to exactly satisfy the balancing constraints, which is referred to as a *rounding problem*. A sample can thus be only approximately balanced (see Deville and Tillé, 2004). Indeed, the balancing constraints must be relaxed in order to make it possible to select a sample. Hence, the variance of the total when balanced sampling is applied can be broken down into two terms: a first term derived under the hypothesis that the balancing equations are perfectly satisfied and a second term due to the rounding problem (see Deville and Tillé, 2005). As the other approximations and estimators of the variance proposed in this framework, Formula (3.24) does not capture the part of the variance due to the rounding problem and can therefore underestimate the conditional imputation variance.

The stronger the linear relation between the variable of interest and the auxiliary variables is, the more formula (3.24) tends to underestimates the conditional imputation variance. To understand the reason for this, suppose that there is a strict linear relation between the variable of interest and the auxiliary variables. In this case, if the auxiliary variables are perfectly balanced, so is also the variable of interest. As a result, the term in the variance due to the balancing itself is null. Hence, the variance is only due to the rounding problem. In this case, an estimator that captures only the term due to the balancing therefore captures 0% of the actual variance. Then, as the linear relation between the variable of interest and the auxiliary variables weakens, the variable of interest becomes less well balanced. This implies that the variance due to the balancing increases, and, therefore, that the part of the actual variance that is returned by an estimator that captures only the variance due to this one increases.

3.6 Properties of the imputed total estimator

The performance of the imputed total estimator under the proposed imputation method relies on two underlying models: the linear model and the response model. Moreover, as donors are chosen in neighborhoods of recipients, the performance of the imputed total estimator depends on a third principle, which is the neighborhood principle. These models and principles are here detailed. Moreover, the asymptotic properties of the imputed total estimator are studied.

3.6.1 Linear model

Property 3.1

1. *Suppose the data is MAR (or MCAR). Consider the linear model*

$$m: \quad y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i \quad \textit{with} \quad \mathrm{E}_m(\varepsilon_i) = 0,$$

where $\mathrm{E}_m(.)$ denotes the expectation with respect to the model m. If the model m holds, then the bkNNI provides an unbiased imputed total estimator $\widehat{Y}_I$ in the sense that

$$\mathrm{Bias}\left(\widehat{Y}_I\right) = \mathrm{E}_m\mathrm{E}_p\mathrm{E}_q\mathrm{E}_I\left(\widehat{Y}_I - Y\right) = 0.$$

2. *Moreover, if the relation between the variable of interest and the auxiliary variables is strictly linear, i.e.*

$$y_i = \boldsymbol{\beta}^\top \mathbf{x}_i,$$

then the bkNNI provides an imputed total estimator $\widehat{Y}_I$ with a quasi-null imputation variance, i.e.

$$\mathrm{Var}_{imp} = \mathrm{E}_p\mathrm{E}_q\mathrm{Var}_I\left(\widehat{Y}_I\right) \approx 0,$$

regardless of the nonresponse process (MCAR, MAR or NMAR).

The proof is given in Appendix B. It results that if the linear model m reasonably fits the population data, then the bkNNI provides an almost unbiased imputed total estimator $\widehat{Y}_I$ with a small imputation variance.

3.6.2 Response model

Property 3.2 *Let $\boldsymbol{\psi}^{[bk]} = \left(\psi_{ij}^{[bk]}\right), (i,j) \in S_r \times S_m$, be the matrix of imputation probabilities relative to the bkNNI. If*

$$\theta_i = \frac{1}{1 + \sum_{j \in S_m} \frac{d_j}{d_i} \psi_{ij}^{[bk]}},$$

perfectly fits the true response probability of each unit $i \in S_r$, then the bkNNI provides an unbiased imputed total estimator $\widehat{Y}_I$, i.e.

$$\text{Bias}\left(\widehat{Y}_I\right) = \text{E}_p\text{E}_q\text{E}_I\left(\widehat{Y}_I - Y\right) = 0.$$

The proof is given in Appendix B. It results that if the model stated above estimates the response probabilities for $i \in S_r$ reasonably well, then the imputed estimator $\widehat{Y}_I$ is an almost unbiased estimator for Y.

This result can be interpreted in the following way. Respondent $i \in S_r$ acts as a donor a certain number of times in such a way that in expectation its weight is equal to

$$d_i + \sum_{j \in S_m} d_j \psi_{ij}^{[bk]}.$$

If θ_i is the response probability of unit $i \in S_r$, then the bias due to nonresponse is compensated. This can happen when the auxiliary variables $\mathbf{x}_i$ can describe the nonresponse mechanism, because the weights $\psi_{ij}^{[bk]}$ are obtained by calibration on the estimated totals of these variables.

3.6.3 Neighborhood principle

If none of the two previous models hold, a third principle can correct the situation, namely the neighborhood principle. The neighborhood principle states that neighboring units (i.e units showing close auxiliary values) show close y values.

Property 3.3 *Consider $(i, j) \in S_r \times S_m$. If the implication*

$$i \in knn(j) \quad \Rightarrow \quad y_i - y_j = 0$$

holds, then the bkNNI provides an unbiased imputed total estimator $\widehat{Y}_I$, i.e.

$$\text{Bias}\left(\widehat{Y}_I\right) = \text{E}_p\text{E}_q\text{E}_I\left(\widehat{Y}_I - Y\right) = 0.$$

The proof is given in Appendix B. It results that if neighboring units (i.e units showing close auxiliary values) have y values which are close, then the imputed estimator $\widehat{Y}_I$ is an almost unbiased estimator for Y.

3.6.4 Resistance to model misspecification

The new method is resistant to model misspecification in terms of bias of the imputed total estimator $\widehat{Y}_I$. It is indeed sufficient that only one of the three models or principle stated above holds to obtain an unbiased imputed total estimator $\widehat{Y}_I$. However, a unique model provides an imputed total estimator

$\widehat{Y}_I$ with a quasi-null imputation variance, namely the strictly linear model $y_i = \boldsymbol{\beta}^\top \mathbf{x}_i$.

3.6.5 Asymptotic properties of the total estimator

We now study the asymptotic properties of the total estimator under bkNNI. Suppose that there is a sequence of finite populations indexed by ℓ such that the population size N_ℓ and the sample size n_ℓ tend to $+\infty$ as $\ell \to +\infty$. Thereafter, index ℓ is omitted in order to make the notation less cluttered but the asymptotic results and convergences are understood to be as $\ell \to +\infty$. The following assumptions are considered:

(A1): $\pi_{ij} - \pi_i \pi_j = O\left(\frac{n}{N^2}\right)$ for each $i, j \in U$, $i \neq j$.

(A2): $d_i = O\left(\frac{N}{n}\right)$ for each $i \in U$.

(A3): The data is MAR.

(A4): The following model holds:

$$\text{m:} \quad y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i,$$

with $\mathrm{E}_m(\varepsilon_i) = 0$, $\mathrm{E}_m(\varepsilon_i \varepsilon_j) = \sigma^2 < +\infty$ if $i = j$ and 0 otherwise and where $\mathrm{E}_m(\cdot)$ denotes the expectation with respect to the model m.

(A5): The imputation design is exactly balanced, i.e.

$$\sum_{i \in S_r} \phi_{ij}^{[bk]} \mathbf{x}_i = \sum_{i \in S_r} \psi_{ij}^{[bk]} \mathbf{x}_i,$$

for each $j \in S_m$.

(A6): The approximation of the conditional imputation variance is exact, i.e.

$$\mathrm{Var}_I(\widehat{Y}_I) = \mathrm{Var}_I^{app}(\widehat{Y}_I) = \sum_{i \in S_r} \sum_{\substack{j \in S_m \\ \psi_{ij}^{[bk]} \neq 0}} c_{ij} d_j^2 \left(y_i - \mathbf{b}^\top \mathbf{x}_i\right)^2,$$

where c_{ij} and $\mathbf{b}$ are defined below Formula (3.24).

(A7): $\#\left\{\psi_{ij}^{[bk]} \psi_{i\ell}^{[bk]} > 0 \,|\, i \in S_r\right\} = O\left(\frac{k^2}{n_m}\right)$ for each $j, \ell \in S_m$ such that $j \neq \ell$.
This hypothesis states that the way the donors distribute from one nonrespondent to another is not very dependent. This constraint is incompatible with the fact that the same respondents are always used as donors.

Proposition 3.1 *Suppose assumptions (A1) to (A7) hold. Then*

$$\frac{\widehat{Y}_I - Y}{N}$$

converges in probability to 0.

The proof is given in Appendix C.

3.7 Simulation study

A brief simulation study is conducted to test the performance of the new imputation method and to test the accuracy of the proposed estimator for imputation variance.

3.7.1 The data

The MU284 population from Särndal et al. (1992) was considered here. This data set is available in the `R sampling` package (Tillé and Matei, 2007). The following variables were considered (the initial names of the variables are written in brackets):

- $\mathbf{y}$: revenues from 1985 municipal taxation, in millions of kronor (`RMT85`),
- $\mathbf{x}^1$: 1985 population, in thousands (`P85`),
- $\mathbf{x}^2$: 1975 population, in thousands (`P75`),
- $\mathbf{x}^3$: number of Conservative seats in municipal council (`CS82`).

The correlations between $\mathbf{y}$ and the variables $\mathbf{x}^1$, $\mathbf{x}^2$ and $\mathbf{x}^3$ are respectively 0.96, 0.97 and 0.52. The population size is $N = 284$. Two cases were considered, namely

Case 1: The three auxiliary variables ($\mathbf{x}^1$, $\mathbf{x}^2$, and $\mathbf{x}^3$) were considered,

Case 2: Only the auxiliary variable that is the less correlated to $\mathbf{y}$ (namely $\mathbf{x}^3$) was considered.

The model m defined in Section 3.6.1 induces a R^2 which is approximately 0.94 and 0.27 in Case 1 and in Case 2 respectively.

3.7.2 Simulation settings

A census was considered, which means that $\pi_i = d_i = 1$ for each unit i of the population $U = \{1, 2, \ldots, N\}$. The sample therefore matches the population, i.e. $S = U$. One hundred respondents sets were created by generating 100 response indicator vectors R. Each component $r_i, i \in U$ of R was generated from a Bernoulli distribution with parameter

$$\theta_i = \frac{1}{1 + \exp\left(1 - \beta x_{i\ell}\right)},$$

where β is a positive coefficient used to reach the mean response rate 70% (MAR), $x_{i\ell}$ is the value of the variable $\mathbf{x}^\ell$ for unit $i \in U$, and $\ell = 1, 3$ in Case 1 and in Case 2 respectively.

For each respondents set, 100 imputations were conducted with each of the following methods:

- NNI: nearest-neighbor,

- PMM: predictive mean matching proposed by Little (1988),
- SRS: random hot-deck, donors randomly selected with replacement in the respondents set,
- SRSWOR: same as SRS except that donors are selected without replacement as proposed in Kalton and Kish (1981, 1984),
- kNNI: k-nearest neighbor,
- bkNNI: proposed method, balanced k-nearest neighbor,

with $k = 20$. For each imputation, the total, the 10th percentile, the 90th percentile, and the variance of the imputed variable of interest were estimated. Note that, for bkNNI, the matrix of imputation probabilities $\boldsymbol{\psi}^{[bk]}$ was replaced by the matrix of imputation probabilities $\boldsymbol{\psi}^{[k]}$ defined in Expression (3.7) for the simulations in which Algorithm 3.1 failed to find a solution (see Section 3.4.4). Moreover, for each simulation, the imputation variance of the total obtained with the proposed method was estimated using Expression (3.24).

3.7.3 Measures of comparison

In order to measure the bias of the imputed estimator $\widehat{\theta}_I$ for a parameter θ, the Monte Carlo relative bias RB was considered. It is defined as

$$\text{RB}\left(\widehat{\theta}_I\right) = \frac{\widehat{\theta}_I^* - \theta}{\theta},$$

where

$$\widehat{\theta}_I^* = \frac{1}{M_R}\frac{1}{M_I}\sum_{r=1}^{M_R}\sum_{i=1}^{M_I}\widehat{\theta}_I^{r,i},$$

$M_R = 100$ is the number of respondents sets generated, $M_I = 100$ is the number of imputations conducted for each respondents set, and $\widehat{\theta}_I^{r,i}$ is the estimate obtained for the i-th imputation of the r-th respondents set generated. The quantity $\widehat{\theta}_I^*$ therefore represents the mean of the estimated value of the parameter θ over the $M_R M_I$ simulations. The variability of the imputed estimator $\widehat{\theta}_I$ was measured through the Monte Carlo relative root mean square error (RRMSE) defined as

$$\text{RRMSE}\left(\widehat{\theta}_I\right) = \frac{\sqrt{\text{MSE}\left(\widehat{\theta}_I\right)}}{\theta},$$

where

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

Finally, the Monte Carlo relative root imputation variance (RRIV), or relative imputation standard deviation, of the imputed estimator $\widehat{\theta}_I$ was computed in

order to measure the amount of variance due to imputation. It is defined as

$$\text{RRIV}\left(\widehat{\theta}_I\right) = \frac{\sqrt{\text{IV}\left(\widehat{\theta}_I\right)}}{\theta},$$

where

$$\text{IV}\left(\widehat{\theta}_I\right) = \frac{1}{M_R}\sum_{r=1}^{M_R}\frac{1}{M_I-1}\sum_{i=1}^{M_I}\left(\widehat{\theta}_I^{r,i} - \widehat{\theta}_I^{r}\right)^2,$$

and

$$\widehat{\theta}_I^{r} = \frac{1}{M_I}\sum_{i=1}^{M_I}\widehat{\theta}_I^{r,i}$$

represents the mean estimated value of θ for the r-th respondents set.

In order to test the accuracy of the variance formula of Expression (3.24), the average over the simulations of the approximated conditional imputation variance was computed, namely

$$\frac{1}{M_R}\sum_{r=1}^{M_R}\text{Var}_I^{app}(\widehat{Y}_I)^r,$$

where $\text{Var}_I^{app}(\widehat{Y}_I)^r$ is the imputation variance obtained with Expression (3.24) for the r-th respondents set generated. That one was then compared to the Monte Carlo imputation variance of the total IV $\left(\widehat{Y}_I\right)$ defined above.

3.7.4 Results of the simulations

Table 3.1 and Table 3.2 show measures of comparison for the six imputation methods considered in Case 1 and Case 2 respectively. Table 3.3 displays the average over the simulations of the estimated imputation variance of the total as well as the Monte Carlo imputation variance of the total.

The results confirm that the proposed method (bkNNI) performs particularly well when there is a strong linear relation between the variable of interest and the auxiliary variable, as in Case 1 ($R^2 \approx 0.94$). Indeed, results of Table 3.1 show that, in Case 1, bkNNI outperforms the other donor imputation methods considered. It provides the smallest RB and the smallest RRMSE for each parameter of interest considered.

Moreover, Table 3.1 and Table 3.2 show that the neighborhood principle and the balancing principle have an effect on the imputation variance of the total. This effect depends on the strength of the relation between the variable of interest and the auxiliary variables. Indeed, in Case 1 (strong linear relation) RRIV of the total is 0.094 for SRS, which reduces to 0.008 for kNNI (neighborhood

principle) and to 0.002 for bkNNI (neighborhood principle and balancing principle) whereas, in Case 2 (weak linear relation) these figures are 0.093, 0.019, and 0.016.

The results also show that selecting the donors without replacement (SRSWOR) among respondents induces a smaller imputation variance than selecting them with replacement (SRS), which is in agreement with Kalton and Kish (1981, 1984).

Finally, the results confirm that the performance of the proposed method relies on the strength of the linear relation that governs the data. Indeed, in Case 2 (Table 3.2) this linear relation is much weaker than in Case 1 and the proposed method shows diminished performance compared to that observed in Case 1. Note that, in Case 2, the proposed method nevertheless still performs better overall than the other methods considered.

Tab. 3.1.: Monte Carlo relative bias (RB), Monte Carlo relative root mean square error (RRMSE), and Monte Carlo relative root imputation variance (RRIV) for the total estimation, the 10-th percentile estimation, the 90-th percentile estimation, and the variance estimation of the variable of interest $\mathbf{y}$ in Case 1.

		Monte Carlo estimates		
Parameter of interest	Method	RB	RRMSE	RRIV
Total	NNI	0.008	0.010	0.000
	PMM	0.015	0.017	0.000
	SRS	0.281	0.297	0.094
	SRSWOR	0.278	0.288	0.070
	kNNI	0.030	0.032	0.008
	bkNNI	-0.001	0.003	0.002
10-th percentile	NNI	0.067	0.093	0.011
	PMM	0.039	0.093	0.010
	SRS	0.187	0.120	0.040
	SRSWOR	0.186	0.198	0.031
	kNNI	0.098	0.124	0.046
	bkNNI	0.006	0.083	0.053
90-th percentile	NNI	0.002	0.009	0.000
	PMM	0.005	0.013	0.000
	SRS	0.246	0.256	0.068
	SRSWOR	0.248	0.255	0.054
	kNNI	0.009	0.018	0.015
	bkNNI	0.000	0.006	0.005
Variance	NNI	-0.001	0.002	0.000
	PMM	-0.002	0.002	0.000
	SRS	0.389	0.533	0.361
	SRSWOR	0.379	0.466	0.267
	kNNI	-0.004	0.004	0.002
	bkNNI	0.000	0.001	0.000

Tab. 3.2.: Monte Carlo relative bias (RB), Monte Carlo relative root mean square error (RRMSE), and Monte Carlo relative root imputation variance (RRIV) for the total estimation, the 10-th percentile estimation, the 90-th percentile estimation, and the variance estimation of the variable of interest $\mathbf{y}$ in Case 2.

		Monte Carlo estimates		
Parameter of interest	Method	RB	RRMSE	RRIV
Total	NNI	-0.001	0.030	0.017
	PMM	0.000	0.030	0.017
	SRS	0.207	0.230	0.093
	SRSWOR	0.207	0.222	0.069
	kNNI	0.004	0.030	0.019
	bkNNI	-0.001	0.028	0.016
10-th percentile	NNI	-0.013	0.080	0.047
	PMM	-0.012	0.080	0.047
	SRS	0.092	0.108	0.036
	SRSWOR	0.092	0.105	0.027
	kNNI	0.023	0.076	0.046
	bkNNI	0.005	0.074	0.045
90-th percentile	NNI	0.004	0.051	0.034
	PMM	0.005	0.052	0.034
	SRS	0.193	0.211	0.072
	SRSWOR	0.193	0.207	0.056
	kNNI	0.004	0.053	0.036
	bkNNI	-0.001	0.052	0.034
Variance	NNI	-0.001	0.094	0.054
	PMM	-0.001	0.094	0.052
	SRS	0.372	0.525	0.356
	SRSWOR	0.376	0.473	0.268
	kNNI	-0.003	0.088	0.061
	bkNNI	-0.008	0.076	0.044

The results in Table 3.3 confirm that Formula (3.24) can underestimate the imputation variance. The magnitude of this underestimation goes along with the strength of the linear relation between the variable of interest and the auxiliary variables. Indeed, in Case 2 (weak linear relation), the average approximated conditional imputation variance represents more than the 90% of the Monte Carlo imputation variance of the total. This quantity drops to approximately 60% in Case 1 (strong linear relation).

3.8 Conclusion

In this paper, a new method of random hot-deck imputation, called balanced k-nearest neighbor, has been proposed. This method has the interesting property of being a donor imputation. It therefore produces observed and feasible values. The novelty of this method is that the selection of donors is viewed as a

Tab. 3.3.: Average over the simulations of the approximated conditional imputation variance of Expression (3.24), Monte Carlo imputation variance of the total and ratio of these two quantities in two different cases.

	Case	
	1	2
Average approx. imputation variance	8172.74	1244589.00
Monte Carlo imputation variance	13146.21	1327932.00
Ratio	0.62	0.94

sampling problem and uses calibration and balanced sampling. Also, selection of donors is achieved in a nonparametric manner as donors are selected in neighborhoods of recipients. As this method is random, it can be used for total estimation as well as for quantiles and variance estimation.

This method offers the nice advantage that it produces a total estimator with negligible imputation variance and a quasi-null bias in specified cases. Indeed, the method involves three underlying models or principles. They provide conditions for the imputed total estimator to be an unbiased estimator and for the imputation variance of that estimator to cancel. The method is resistant to model misspecification in terms of bias but a unique model results in a quasi-null imputation variance of the total.

A formula to approximate the conditional imputation variance of the total has been suggested. The procedure used is inspired by that applied to estimate the variance of the total for balanced sampling. The proposed approximation tends to underestimate the conditional imputation variance of the total.

Finally, a simulation study has been conducted to test the performance of the proposed method and that of the approximation formula of conditional imputation variance. It has been confirmed that the new method performs well when a strong linear relation governs the data and that this performance decreases as this linear relation weakens. Lastly, it was confirmed that the formula for imputation variance of the total tends to underestimate the conditional imputation variance of this one. Note that the estimation of the variance due to the rounding problem is still an unresolved problem. This variance can also be approximated by multiple imputations.

Acknowledgements

This research was supported by the Swiss National Science Foundation, project number P1NEP2_151904; and the Swiss Federal Statistical Office.

4 Nonparametric imputation for nonresponse in surveys

Abstract

Many imputation methods are based on statistical models that assume that the variable of interest is a noisy observation of a function of the auxiliary variables or covariates. Misspecification of this model may lead to severe errors in estimates and to misleading conclusions. A new imputation method for item nonresponse in surveys is proposed based on a nonparametric estimation of the functional dependence between the variable of interest and the auxiliary variables. We consider the use of smoothing spline estimation within an additive model framework to flexibly build an imputation model in the case of multiple auxiliary variables. The performance of our method is assessed via a simulation study, suggesting that our method performs better than competing imputation methods.[1]

Keywords: additive models, data imputation, sample survey, smoothing spline.

4.1 Introduction

Nonresponse in surveys is a commonly encountered problem that, when ignored, can affect the performance of the statistical estimators for the quantities of interest. Two general adjustment techniques that have been developed to alleviate the effects of nonresponse are *reweighting* and *imputation*. Reweighting procedures consist of increasing the initial weights of respondents in order to compensate for nonrespondents and are commonly used to treat unit nonresponse. Imputation procedures consist of filling in the missing values in the data with *imputed values* and are commonly used to treat item nonresponse. When dealing with nonresponse, both reweighting and imputation may rely on a statistical model. Imputation for the variable of interest can be more efficient if it is based on information contained in a number of auxiliary variables, specifically, through a model that estimates a functional link between the latter and the variable of interest. However, the validity of the model will have a direct effect on the accuracy of the estimated quantities. It is therefore crucial to be able to build flexible models that can capture a large spectrum of patterns and make only weak assumptions about the true underlying mechanism generating

[1] This chapter is a working paper co-written with Professor Radu V. Craiu.

the data. Given these constraints, it is not surprising that nonparametric models have been used to handle nonresponse in surveys.

Giommi (1987) focused on unit nonresponse and proposed two nonparametric reweighting procedures based on kernel density estimators to estimate response probabilities. Later, Niyonsenga (1994, 1997) used the nonparametric estimation of Giommi (1987) to handle nonresponse when unit nonresponse and item nonresponse occur together. Finally, Da Silva and Opsomer (2006) and Da Silva and Opsomer (2009) applied, respectively, kernel regression and local polynomial regression to estimate the response probabilities and derived asymptotic properties of the propensity score adjusted estimator for these approaches. These techniques are suitable when the number of auxiliary variables is relatively low. We propose here an imputation method for item nonresponse in surveys when the variable of interest is a noisy observation of a function of many auxiliary variables. We consider smoothing spline models within an additive regression framework which allows us to handle a large number of auxiliary variables. This improvement significantly expands the range of nonparametric methods for handling nonresponse. Moreover, the model considered is adaptable to a wide variety of functional patterns thus providing protection against model misspecification. Results of a simulation study confirm the performance of our method and highlight its capacity to adapt to many different situations.

The paper is organized as follows: Section 4.2 establishes the framework and introduces notation; Section 4.3 provides a motivation for the new imputation method; two nonparametric tools used in the new imputation method are reviewed in Section 4.4; Section 4.5 presents the new method as well as bootstrap procedures to estimate the variance of the total. The performance of the new method is compared to that of other imputation methods through a simulation study presented in Section 4.6. We close with concluding remarks and a discussion of future work.

4.2 Framework

Consider a finite population $U = \{1, 2, \ldots, N\}$ of possibly unknown size N. Suppose that the parameter of interest is the population total

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

for some unknown variable of interest y. A sample S of size n is selected from U according to a probabilistic sampling design $p(\cdot)$ with the aim of observing

y_i for $i \in S$. Consider

$$\pi_i = \Pr(i \in S) = \sum_{s \subset U; s \ni i} p(s),$$

the first-order inclusion probability of unit i and suppose that $\pi_i > 0$ for all $i \in U$. Let $d_i = 1/\pi_i$ represent the design weight of unit $i \in U$. In this paper we consider two widely used sampling designs, simple random sampling without replacement (SRSWOR) and stratified sampling. Under SRSWOR, each sample of (fixed) size n has the same probability of being selected and $\pi_i = n/N$ for all $i \in U$. Under stratified sampling, the population U is partitioned into H strata $U_1, \ldots, U_H$ of respective sizes $N_1, \ldots, N_H$ and SRSWOR is applied independently in each stratum h. A sample S_h of size n_h is hence selected in each stratum U_h, $h = 1, \ldots, H$ and $\pi_i = n_h/N_h$ for all $i \in U_h$.

Once a sample S is selected, each unit $i \in S$ is classified as either respondent or nonrespondent, depending on whether y_i is observed or missing. Consider the response indicator vector $(r_i | i \in S)^\top$ where r_i takes value 1 if y_i is observed and 0 if it is missing. This results in the set of respondents $S_r = \{i \in S | r_i = 1\}$ and in the set of nonrespondents $S_m = \{i \in S | r_i = 0\}$.

Under complete response, the Horvitz-Thompson estimator

$$\widehat{Y} = \sum_{i \in S} \frac{1}{\pi_i} y_i, \tag{4.1}$$

is a design unbiased estimator for Y, i.e. $\mathrm{E}_p(\widehat{Y}) = Y$. In the case of a survey with nonresponse, however, the estimator (4.1) cannot be computed since some of the y_i's, $i \in S$ are missing. One remedy is to impute each missing value y_i, $i \in S_m$ with an imputed value y_i^*. The population total Y can then be estimated through the *imputed estimator*

$$\widehat{Y}_I = \sum_{i \in S} \frac{1}{\pi_i} \left[y_i r_i + y_i^* (1 - r_i) \right] = \sum_{i \in S_r} \frac{1}{\pi_i} y_i + \sum_{i \in S_m} \frac{1}{\pi_i} y_i^* = \sum_{i \in S} \frac{1}{\pi_i} \widetilde{y}_i, \tag{4.2}$$

where

$$\widetilde{y}_i = \begin{cases} y_i & \text{if } i \in S_r; \\ y_i^* & \text{if } i \in S_m. \end{cases}$$

If the imputation process exactly reconstructs the missing values, that is if $y_i^* = y_i$ for $i \in S_m$, then $\widehat{Y}_I$ is a design unbiased estimator for the population total Y. Hence, an imputation method that reconstructs the missing data well can provide protection against nonresponse bias. Design weights can optionally be taken into account when constructing the imputed values, the resulting method being referred to as *survey weighted imputation*.

Consider a vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})^\top$ of values taken by q auxiliary variables x_1, x_2, …, x_q and known for all $i \in U$ or at least for all $i \in S$. Auxiliary information can be used at different stages of the survey, namely in establishing the sampling design, for estimation, and handling of nonresponse. Reliable auxiliary information can explain the variation in the variable of interest and/or in the response probabilities and helps reduce error due to sampling and nonresponse.

4.3 Motivation

We consider a variable of interest, y, that is measured along with q auxiliary variables, $x_1, \ldots, x_q$. In situations in which the variable of interest is not recorded for some sampled units, one may rely on the auxiliary variables to impute the missing values if there is a way to connect these variables via an *imputation model* (Särndal, 1992). For instance, consider a general model of the type

$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{iq}) + \varepsilon_i, \tag{4.3}$$

where f is a function from $\mathbb{R}^q$ to $\mathbb{R}$, and ε_i are zero-mean independent errors with variance σ^2. A deterministic imputation method estimates first the function f based on those individuals/items $i \in S_r$ for which $(y_i, \mathbf{x}_i) = (y_i, x_{i1}, \ldots, x_{iq})$ are fully observed, and then imputes values for $i \in S_m$ using the estimated function and the observed $\mathbf{x}_i$. The challenging issue of estimating f naturally arises because the choice of the imputation model crucially impacts the accuracy of the imputed values. A misspecified model may result in highly biased estimates for the parameters of interest.

Without prior knowledge on the form of f in (4.3), it is natural to use a nonparametric regression model since the resulting estimate $\hat{f}$ is known to adapt to the shape of f based on the information provided by the data. When handling survey data, however, several auxiliary variables are often available and one needs to include most of them in the model. Unfortunately, a few nonparametric smoothers such as kernel-based ones tend to break down in high dimension, unless the sample size is very large. This phenomenon is known as the *curse of dimensionality* (Bellman, 1961; Stones, 1985) and can be alleviated if an Additive Model (AM, Hastie and Tibshirani, 1986) is used. Such a model is additive in the predictor variables and takes the form

$$y_i = a_0 + \sum_{j=1}^{q} a_j(x_{ij}) + \varepsilon_i, \tag{4.4}$$

where $(y_i, \mathbf{x}_i) = (y_i, x_{i1}, \ldots, x_{iq})$, $i = 1, \ldots, N$, are observations, a_0 is a constant, a_j, $j = 1, \ldots, q$, are univariate smooth functions, and ε_i are zero-mean

independent errors with common variance σ^2. The functions a_j, $j = 1, \ldots, q$, are each individually estimated by univariate smoothers so the curse of dimensionality is avoided because the original problem of nonparametric estimation in $\mathbb{R}^q$ has been replaced by q estimation problems in $\mathbb{R}$. Without loss of generality, henceforth we suppose that the $\mathbf{x}_i$, $i = 1, \ldots, N$, lie in the interval $[0, 1]^q$.

We propose an imputation method for nonresponse in survey based on AM. The new method is based on imputation model (4.4). The nonparametric tools used to estimate the regression function are presented in Section 4.4 and the new method is presented in Section 4.5.

4.4 Nonparametric tools

This section introduces two nonparametric tools used in the new imputation method, smoothing spline regression and additive models. The main idea of smoothing spline regression is to fit a data set with a curve that maximizes a measure of goodness-of-fit while achieving a fixed degree of smoothness. There is an extensive literature devoted to spline regression and we refer the reader to Green and Silverman (1994), Eubank (1999), and Wang (2011). Smoothing spline regression (SSR) assumes model (4.4) with a unique predictor variable, that is

$$y_i = a(x_i) + \varepsilon_i, \quad 1 \leq i \leq N,$$

where ε_i are zero-mean independent errors with common variance σ^2, and a is a smooth function in the sense that $a \in W_2^m[0, 1]$ where $W_2^m[0, 1]$ is the Sobolev space

$$W_2^m[0,1] = \Big\{ g : g, g', \ldots, g^{(m-1)} \text{ are absolutely continuous,} \\ \int_0^1 g^{(m)}(t)^2 < +\infty \Big\}.$$

We consider a basis of functions b_k, $k \in 1, \ldots K$, called *spline basis functions*, for $W_2^m[0, 1]$. The SSR yields the best approximation of function a in $W_2^m[0, 1]$ while controlling the degree of smoothness. The resulting *smoothing spline estimator* $\widehat{a}$ is the minimizer of the following penalized least square (PLS) criterion

$$\frac{1}{N} \sum_{i=1}^{N} (y_i - g(x_i))^2 + \lambda \int_0^1 g^{(m)}(t)^2 dt, \tag{4.5}$$

over all functions in $W_2^m[0, 1]$. The parameter λ is the *smoothing parameter* and its size decides the balance between goodness-of-fit, as measured by

the mean squared residual, and smoothness, as measured by the integral. There exist different basis of functions, each of which can produce a different smoothing spline estimator. In what follows, we will consider the thin plate spline basis (see Wood, 2003) and the smoothing parameter λ will be selected by generalized cross validation.

With survey data, it is often desirable to consider design weights when estimating parameters of interest. Indeed, a design weight $d_i = 1/\pi_i$ can be interpreted as the number of population units that sampled unit i represents. Hence, when units are selected with unequal inclusion probabilities it might be unreasonable to assume that each sampled unit has the same influence on the parameters of interest. A weighted version of the smoothing spline estimator was proposed by Zhang et al. (2013) who suggested adding design weights in the general PLS criterion in equation (4.5). Hence, they consider the smoothing spline estimator adapted for survey data which is the minimizer over g of

$$\frac{1}{\widehat{N}} \sum_{i \in S} d_i \left(y_i - g(x_i)\right)^2 + \lambda \int_0^1 g^{(m)}(t)^2 dt, \tag{4.6}$$

where $\widehat{N} = \sum_{i \in S} d_i$ is the estimated population size. Note that Zhang et al. (2013) restrict themselves to the case $m = 2$.

A flexible way to combine the contributions of each auxiliary variable to the variable of interest is provided by the additive model paradigm. A class of generalized additive models was proposed by Hastie and Tibshirani (1986) and was discussed in depth in the book Hastie and Tibshirani (1990). We focus here on the additive regression model (AM), which assumes

$$y_i = a_0 + \sum_{j=1}^{q} a_j(x_{ij}) + \varepsilon_i,$$

where a_0 is a constant, a_j, $j = 1, \ldots, q$, are smooth functions, and ε_i are zero-mean independent errors with common variance σ^2. SSR is used to estimate each function a_j, $j = 1, \ldots, q$. A backfitting algorithm (Hastie and Tibshirani, 1986) or a direct fitting approach (Wood, 2008) can be considered.

When appropriate, an additive model allows us to handle multiple predictor variables in a reasonable computation time and avoids the curse of dimensionality problem as it breaks a high-dimensional nonparametric estimation problem into a number of one-dimensional ones.

4.5 The method

In this section, we propose a nonparametric model-based imputation method for nonresponse in surveys and discuss bootstrap procedures to estimate the resulting variance of the total estimator for the population U.

4.5.1 Estimation and imputation

Assume that the sample S contains respondents S_r for which the values of the variable of interest $\{y_i : i \in S_r\}$ are observed and nonrespondents for which these values $\{y_i : i \in S_m\}$ are missing. For each unit $i \in S$ we have available auxiliary variables values $\mathbf{x}_i = \{x_{i1}, \ldots, x_{iq}\}$. We consider the following additive imputation model

$$y_i = a_0 + \sum_{j=1}^{q} a_j(x_{ij}) + \varepsilon_i, \tag{4.7}$$

where a_0 is a constant, a_j, $j = 1, \ldots, q$, are univariate functions in the functional space defined in Section 4.4, and ε_i are zero-mean independent errors with common variance σ^2. Smoothing spline estimates $\widehat{a}_j$, $j = 1, \ldots, q$, of functions a_j, $j = 1, \ldots, q$, and an estimate $\widehat{a}_0$ of a_0 are obtained using the complete data $(y_i, \mathbf{x}_i)$, $i \in S_r$. Two different smoothing splines estimators can be obtained based on expression (4.5) (unweighted imputation) or expression (4.6) (survey weighted imputation), respectively. Finally, missing values y_i, $i \in S_m$, are imputed with predictions based on imputation model (4.7) as follows

$$y_i^* = \widehat{a}_0 + \sum_{j=1}^{q} \widehat{a}_j(x_{ij}). \tag{4.8}$$

It is sometimes desirable to apply hot-deck imputation methods, i.e. to impute values that are possible (e.g. integers) or have been already observed. The new method based on AM mostly imputes values that do not match observed values. The simple extension presented here leads to a hot-deck imputation method and is built on the idea of predictive mean matching (Little, 1988). The idea is to replace each imputed value with the closest observed value. Consider the observed y_i, $i \in S_r$, and the imputed values y_i^*, $i \in S_m$, obtained via expression (4.8). To obtain a hot-deck imputation method, imputed observed values y_i^{**}, $i \in S_m$, are obtained as follows

$$y_i^{**} = y_{j(i)} \quad \text{where} \quad \left|y_i^* - y_{j(i)}\right| = \min_{j \in S | r_j = 1} \left|y_i^* - y_j\right|.$$

Alternative implementations are possible. For instance, if each value of the variable of interest is restricted to be an integer, one can choose $y_i^{**} = \text{int}(y_i^*)$

where int$(\cdot)$ is the integer part function. With this choice, the imputed values are possible but may not have been already observed in the sample.

4.5.2 Variance estimation for the imputed total

A valid method for estimating the variance of the estimator of the population total must account for the extra variability due to imputing the missing values. In turn, this variability is due to the variance of predicted values y_i^* produced via the additive model. Since an analytical expression for the asymptotic error of AM predictive value is not available, we pursue a bootstrap-based approach. Bootstrap procedures to estimate the variance of parameters of interest are available for different imputation methods and sampling designs. In this Section, we follow Shao and Sitter (1996) to devise bootstrap procedures to estimate the variance of the total under AM imputation for simple random sampling without replacement (SRSWOR) and stratified sampling. The bootstrap proposed in Shao and Sitter (1996) is asymptotically valid irrespective of the sampling design, or the imputation method.

We follow Shao and Sitter (1996) and apply the without-replacement bootstrap (BWO) proposed by Gross (1980) to estimate the variance of the total under AM imputation for SRSWOR. Procedure 4.1 presents the applied procedure which proceeds as follows. Given a sample of size n from a population of size N, we set $k = N/n$ and assume k is an integer (otherwise we round it off). In step 1 we construct a pseudopopulation of size N by replicating the sample k times. In step 2, a simple random sample of size n is selected from the pseudopopulation. Because the pseudopopulation consists of sampled units, the bootstrap sample is very likely to contain both units with missing y_i and units with observed y_i. In step 3, AM imputation is applied to the bootstrap sample. Steps 2 and 3 are repeated to obtain B analogs of the imputed total estimator. In step 5, the bootstrap variance of the imputed total is obtained using the standard bootstrap formulae.

For stratified sampling, we also follow Shao and Sitter (1996) and apply the mirror-match bootstrap (MMB) proposed by Sitter (1992) to estimate the variance of the total under AM imputation. Procedure 4.2 presents the applied procedure. In steps 1 and 2, the procedure mimics the stratified sampling by selecting several times SRSWOR of size n'_h in stratum h. If n'_h is such that $n'_h = f_h n_h$, then the size of the bootstrap sample S_h^* is the same as that of S_h, i.e. $n_h^* = n_h$. This procedure is repeated independently in each stratum h times to obtain a bootstrap sample $\mathbf{S}^*$. Because the bootstrap sample consists of sampled units, it is very likely to contain both units with missing y_i and units with observed y_i. Hence, in step 4, AM imputation is applied to the bootstrap sample $\mathbf{S}^*$ and the bootstrap analog $\widehat{Y}_I^{(b)}$ of the imputed total estimator $\widehat{Y}_I$ is obtained. Depending on the choice of n'_h and on whether randomization is

Procedure 4.1 Variance of the imputed total estimator under SRSWOR.

Step 1: Suppose $N = kn$ for an integer k.
Construct a pseudopopulation by replicating the sample k times.

Step 2: Draw a SRSWOR of size n from the pseudopopulation of step 1.

Step 3: Apply AM imputation to impute the missing y_i's of the sample selected in step 2.

Step 4: Repeat steps 2 and 3 a large number of times B to obtain $\widehat{Y}_I^{(1)}, \ldots, \widehat{Y}_I^{(B)}$ where $\widehat{Y}_I^{(b)}$ is the analog of $\widehat{Y}_I$ for the b-th bootstrap sample.

Step 5: Obtain the bootstrap variance of $\widehat{Y}_I$ by

$$\mathrm{V}_{boot}(\widehat{Y}_I) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{Y}_I^{(b)} - \widehat{Y}_I^{(.)}\right)^2,$$

where $\widehat{Y}_I^{(.)}$ is the mean bootstrap analog of $\widehat{Y}_I$

$$\widehat{Y}_I^{(.)} = \frac{1}{B}\sum_{b=1}^{B}\widehat{Y}_I^{(b)}.$$

applied to round n'_h and/or k_h, the bootstrap procedure might mimic a stratified sampling in a population whose size differs from N. Fraction N/n^* appears in the computation of the bootstrap analog of the imputed total estimator $\widehat{Y}_I$ to take this into account. Steps 1 to 4 are repeated to obtain B analogs of the imputed total estimator. In step 6, the bootstrap variance of the imputed total is obtained using the standard bootstrap formulae.

The computational time involved in the bootstrap evaluation of variance can be shortened if multiple processors are available. The embarrassing parallel structure of the procedure implies that the sample-specific calculation can be performed on a separate processor and the merging of simulated values is needed only in Step 5 (for Procedure 1) and in Step 6 (for Procedure 2).

4.6 Simulations

A simulation study was conducted to test the performance of the proposed imputation method. Simulated data and real data were considered. In Sections 4.6.1 and 4.6.2, the simulation settings for the simulated data and for the real data are respectively presented. Measures used to compare the new imputation method with existing imputation methods and to test the accuracy of the bootstrap procedures for the variance estimation are described in Section 4.6.3. Finally, the results of the simulations in each setting are displayed and commented in Sections 4.6.4 and 4.6.5 respectively.

Procedure 4.2 Variance of the imputed total estimator under stratified sampling.

Step 1: Choose $1 \leq n'_h < n_h$ and select a SRSWOR of size n'_h without replacement from S_h.
If n'_h is not integer, apply a randomization (see Sitter, 1992).

Step 2: Repeat step 1 $k_h = n_h(1-f^*_h)/(n'_h(1-f_h))$ times independently to obtain a sample $S^*_h = \{hi : i = 1, \ldots, n^*_h\}$ of size $n^*_h = n'_h k_h$, where $f_h = n_h/N_h$ and $f^*_h = n'_h/n_h$.
If k_h is not integer, apply a randomization (see Sitter, 1992)

Step 3: Repeat steps 1 and 2 independently for each stratum h to obtain a bootstrap sample $\mathbf{S}^* = \{S^*_1, \ldots, S^*_H\} = \{hi : h = 1, \ldots, H; i = 1, \ldots, n^*_h\}$ of size $n^* = \sum_{h=1}^{H} n^*_h$.

Step 4: Apply AM imputation to impute the bootstrap sample $\mathbf{S}^*$ and obtain the bootstrap analog of the imputed total estimator $\widehat{Y}_I$ by

$$\widehat{Y}_I^{(b)} = \frac{N}{n^*} \sum_{hi \in \mathbf{S}^*} \frac{\widetilde{y}^{(*)}_{hi}}{f^*_h} = \frac{N}{n^*} \sum_{h=1}^{H} \frac{n_h}{n'_h} \sum_{hi \in S^*_h} \widetilde{y}^{(*)}_{hi},$$

where $\widetilde{y}^{(*)}_{hi}$ is the value of the variable of interest of unit hi if this one is observed and the imputed value otherwise.

Step 5: Repeat steps 1 to 4 a large number of times B to obtain $\widehat{Y}_I^{(1)}, \ldots, \widehat{Y}_I^{(B)}$ where $\widehat{Y}_I^{(b)}$ is the analog of $\widehat{Y}_I$ for the b-th bootstrap sample.

Step 6: Obtain the bootstrap variance of $\widehat{Y}_I$ by

$$\mathrm{V}_{boot}(\widehat{Y}_I) = \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{Y}_I^{(b)} - \widehat{Y}_I^{(.)} \right)^2,$$

where $\widehat{Y}_I^{(.)}$ is the mean bootstrap analog of $\widehat{Y}_I$

$$\widehat{Y}_I^{(.)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{Y}_I^{(b)}.$$

4.6.1 Setting 1: simulated data

Populations of size $N = 10000$ were considered. Four auxiliary variables x_1, x_2, x_3, and x_4 were generated. The values x_{i1}, x_{i2}, and x_{i3}, $i = 1, \ldots, N$, are independent draws from a Uniform$[0, 1]$ random variable and x_{i4}, $i = 1, \ldots, N$, are independent draws of a gamma density with shape and scale parameters, respectively, 3 and 1/6 that were mapped into the $[0, 1]$ interval via the transformation $x_{i4} \to (x_{i4} - \min(x_4)) \,/\, (\max(x_4) - \min(x4))$.

Five populations were then generated as follows:

$$
\begin{aligned}
y_i^{(1)} &= 1 + 5x_{i1} + x_{i2} + x_{i3} + x_{i4} + \varepsilon_i, \\
y_i^{(2)} &= 2 + \cos(\pi x_{i1} + \pi) + \sin(4\pi x_{i2}) + \exp(-(x_{i3} - 0.5)^2) \\
&\quad + (x_{i4} - 0.5)^2 + \varepsilon_i, \\
y_i^{(3)} &= 1 + \cos(2\pi x_{i1}) + x_{i1}x_{i2} + x_{i3}^2 x_{i4} + \varepsilon_i, \\
y_i^{(4)} &= 2 + \cos(\pi(x_{i1} + x_{i2}))\sin(\pi(x_{i3} + x_{i4})) + \varepsilon_i, \\
y_i^{(5)} &= 1 + \varepsilon_i,
\end{aligned}
$$

where $i = 1, \ldots, N$, and where ε_i are N independent draws of a normal random variable with mean 0 and standard deviation 0.1. In the first four populations, the variable of interest is linked to the auxiliary variables. In the first two populations the link is correctly specified by an AM, even a linear model in population 1. In populations 3 and 4 the AM is not a valid representation of the truth, while in the last population there is no link between the variable of interest and the auxiliary variables.

Two different sampling designs were used for the selection of samples: simple random sampling without replacement (SRSWOR) and stratified sampling. For simple random sampling, a sampling rate of $f = 0.2$ was considered. For stratified sampling, strata were created as follows. First, units were classified into two groups, depending whether their value x_{i1} is larger than the median of x_1 or not. In each group created, units were then subdivided into two other groups, depending on whether their value x_{i2} is larger than the median of x_2 in each group or not. The procedure was repeated for variables x_3 and x_4. This resulted in creating 16 strata of size 625 that are somewhat homogeneous with respect to the auxiliary variables. Then, SRSWOR was applied within strata with a sampling rate of $f = 0.2$ in each stratum.

The response probabilities were obtained from

$$
p_i = \frac{\exp(b_0 + b_1 x_{i1})}{1 + \exp(b_0 + b_1 x_{i1})},
$$

where b_0 and b_1 were set to obtain an overall mean response rate which is approximately 75%.

Ten thousand simulations were then conducted as follow. For each simulation, a sample S was selected according to either SRSWOR or stratified sampling. For each sample S selected, a respondents set S_r and a nonrespondents set S_m were then created by generating a response indicator vector $(r_i | i \in S)^\top$, where r_i, $i \in S$, was generated from a Bernoulli distribution with parameter p_i. Then, for each set of respondents and of nonrespondents obtained, the missing y_i,

$i \in S_m$, were replaced with imputed y_i^* using the four following imputation methods:

- **Regression imputation:** Imputed values y_i^*, $i \in S_m$, are obtained by

$$y_i^* = \widehat{\beta}_0 + \sum_{j=1}^{q} \widehat{\beta}_j x_{ij},$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_q)^\top$ is defined by

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{j \in S_r} d_j (1, \mathbf{x}_j)^\top (1, \mathbf{x}_j) \right)^{-1} \sum_{i \in S_r} d_i (1, \mathbf{x}_i)^\top y_i.$$

Regression imputation is based on the following imputation model

$$y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + \varepsilon_i.$$

- **Mean imputation:** The missing y_i, $i \in S_m$, are replaced by the respondents' mean value, that is the imputed values y_i^*, $i \in S_m$, are obtained by

$$y_i^* = \frac{1}{\sum_{j \in S_r} d_j} \sum_{i \in S_r} d_i y_i.$$

Mean imputation is a particular case of regression imputation where only a constant covariate is considered and is based on the following imputation model

$$y_i = \beta_0 + \varepsilon_i.$$

- **Nearest neighbor imputation:** The missing y_i, $i \in S_m$, are replaced by their respective nearest neighbor in the complete data. The proximity is quantified through the auxiliary variables. Imputed values y_i^*, $i \in S_m$, are obtained by

$$y_i^* = y_{j(i)} \quad \text{where} \quad d(\mathbf{x}_i, \mathbf{x}_{j(i)}) = \min_{j \in S | r_j = 1} d(\mathbf{x}_i, \mathbf{x}_j),$$

where $d(\cdot, \cdot)$ is the Euclidean distance.

- **AM imputation:** An AM was fitted using the complete data $(y_i, \mathbf{x}_i)$, $i \in S_r$, and imputed values y_i^*, $i \in S_m$, were obtained through predictions with this model, as explained in Section 4.5. Survey weights were considered in the smoothing spline estimator computation of each term, as in the PLS equation of expression (4.6). The model was fitted using

function `gam` of `R` package `mgcv` (Wood, 2014). Function `gam` uses $m = 2$ and thin plate splines basis by default. The model is fitted by penalized likelihood maximization and the smoothing parameter is selected by generalized cross validation.

The imputed total estimator $\widehat{Y}_I$ was computed for each method and each simulation. Note that all the considered imputation methods use auxiliary information when computing imputed values, except mean imputation.

Moreover, ten thousand simulations were conducted to test the accuracy of the bootstrap procedures presented in Section 4.5.2 to estimate the variance of the total. SRSWOR and stratified sampling were considered. For each simulation, a sample S, a set of respondents S_r and of nonrespondents S_m were created as described above. The missing values were replaced with imputed values using AM imputation. The imputed total estimator $\widehat{Y}_I$ and its bootstrap variance $\mathrm{V}_{boot}(\widehat{Y}_I)$ were computed for each simulation. For the bootstrap variance under SRSWOR, procedure 4.1 was applied where, in step 1, the sample was replicated $k = 1/f = 5$ times to create a pseudopopulation of size 10000 and $B = 100$ bootstrap replicates were generated. For the bootstrap variance under stratified sampling, procedure 4.2 was applied where, in step 1, a sample of size 125 was selected in each stratum, that is $n'_h = f \cdot n_h = 125$ for each stratum h. This results in integer n'_h and k_h for each stratum h.

4.6.2 Setting 2: real data

We consider the data from the 1992 family expenditure survey (FES), see Central Statistical Office (1993). The data is made available by the UK data archive at the University of Essex. To test our method, we considered that the households having a non-missing and larger than zero disposable income (disposable income and self-supply and in kind) of the 1992 FES form the population of interest. The size of this population is $N = 7409$. The variable disposable income was modified as follows. First, it was divided by its mean value. Because income distributions are often right skewed, the natural logarithm of the obtained value plus one was computed. One was added before computing the logarithm to avoid negative values. We suppose that the aim of the survey is to estimate the population total of the modified disposable income. The population was stratified into 12 regions and simple random sampling with a sampling rate of $f = 0.2$ was applied within each region (stratum). The sample size was randomly rounded for 8 strata for which this sampling rate led to a non-integer sample size. For each sampled household, we supposed that the following characteristics were observed:

x_{i1}: number of adults in household i,
x_{i2}: number of children in household i,
x_{i3}: number of persons economically active in household i,

x_{i4}: age of the head of household i,
x_{i5}: age of the chief economic supporter of household i.
Such variables could for instance come from a register. It was supposed that the willingness of a household to respond depends on the number of adults in this household and that the households respond independently from each other. Hence, the response probabilities were obtained from

$$p_i = \frac{\exp(b_0 + b_1 x_{i1})}{1 + \exp(b_0 + b_1 x_{i1})},$$

where b_0 and b_1 were set to obtain an overall mean response rate which is approximately 70%. Then, for each sampled household, a response indicator was generated from a Bernoulli distribution with parameter p_i. The modified disposable income was then recorded for respondents and erased for nonrespondents. Ten thousand simulations were conducted. The same imputation methods as in Section 4.6.1 were considered.

Moreover, ten thousand simulations were conducted to test the accuracy of the bootstrap procedures presented in section 4.5.2 to estimate the variance of the total. For each simulation, a sample and a set of respondents and of nonrespondents were created as described above. The missing values were replaced with imputed values using AM imputation. The imputed total estimator $\widehat{Y}_I$ and its bootstrap variance $\mathrm{V}_{boot}(\widehat{Y}_I)$ were computed for each simulation. For the bootstrap variance, procedure 4.2 was applied with $B = 100$ bootstrap replicates. We set $n'_h = f \cdot n_h$ and a randomization was applied to round the non-integer n'_h and the non-integer k_h (see Sitter, 1992).

4.6.3 Measures of comparison

For each simulation and each imputation method of both settings, the population total for the variable of interest was estimated through the imputed estimator of expression (4.2). To compare the performance of the methods, four comparison measures were recorded. First, to quantify the accuracy of imputed values, the Monte Carlo mean relative prediction error was computed, which is defined as

$$\mathrm{MRPE} = \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{n_m^{(\ell)}}\sum_{i\in S_m^{(\ell)}}\left|\frac{y_i^{*(\ell)} - y_i}{y_i}\right|,$$

where $S_m^{(\ell)}$ is the nonrespondents set obtained at the ℓ-th simulation, $n_m^{(\ell)}$ is the size of $S_m^{(\ell)}$, $y_i^{*(\ell)}$ is the imputed value obtained for $i \in S_m^{(\ell)}$ at the ℓ-th simulation, and L represents the number of simulations. Then, for each imputation method, the performance of the imputed estimator of expression (4.2) was studied through three comparison measures, namely

- the Monte Carlo relative bias (RB) defined as

$$\text{RB} = \frac{B}{Y},$$

where $\text{B} = \widehat{Y}_I^{(\cdot)} - Y$, $\widehat{Y}_I^{(\cdot)}$ represents the mean imputed estimator over the L simulations

$$\widehat{Y}_I^{(\cdot)} = \frac{1}{L}\sum_{\ell=1}^{L} \widehat{Y}_I^{(\ell)},$$

and $\widehat{Y}_I^{(\ell)}$ is the imputed estimator $\widehat{Y}_I$ obtained at the ℓ-th simulation,

- the Monte Carlo relative root variance (or relative standard deviation) defined as

$$\text{RRVAR} = \frac{(\text{VAR})^{1/2}}{Y},$$

where

$$\text{VAR} = \frac{1}{L-1}\sum_{\ell=1}^{L} \left(\widehat{Y}_I^{(\ell)} - \widehat{Y}_I^{(\cdot)}\right)^2,$$

- the Monte Carlo relative root mean square error defined as

$$\text{RRMSE} = \frac{\left(\text{B}^2 + \text{VAR}\right)^{1/2}}{Y}.$$

For AM imputation, the following measures were computed to test the accuracy of the bootstrap variance estimator:

- The Monte Carlo variance of the total estimator:

$$\text{VAR} = \frac{1}{L-1}\sum_{\ell=1}^{L} \left(\widehat{Y}_I^{(\ell)} - \widehat{Y}_I^{(\cdot)}\right)^2,$$

- The Monte Carlo expectation of the bootstrap variance estimator:

$$\text{VAR}_{boot} = \frac{1}{L}\sum_{\ell=1}^{L} \text{V}_{boot}^{(\ell)}(\widehat{Y}_I),$$

where $\text{V}_{boot}^{(\ell)}(\widehat{Y}_I)$ is the bootstrap variance $\text{V}_{boot}(\widehat{Y}_I)$ obtained at the ℓ-th simulation,

- The coverage rate CR: the proportion of times the true total Y falls into the 95% confidence interval

$$\widehat{Y}_I \pm 1.96\sqrt{\text{V}_{boot}(\widehat{Y}_I)}.$$

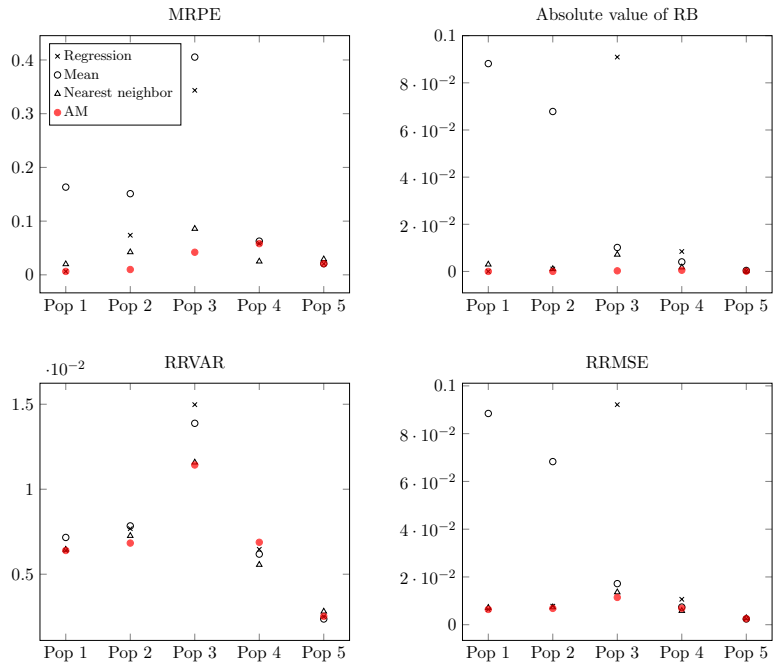


Fig. 4.1.: Comparison measures of four imputation methods in five populations under SRSWOR.

4.6.4 Results of setting 1

Figure 4.1, Figure 4.2, and Table 4.2 display the results of Setting 1. Table 4.1 reports the average ranks over the populations of each imputation method for each measure of comparison. The absolute value of RB was considered.

We first comment the results shown in Figures 4.1 and 4.2. When functional dependence between the variable of interest and the auxiliary variables is additive (populations 1 and 2), AM imputation provides the best results. If, moreover, this dependence functional is linear (population 1), regression imputation performs as well as AM imputation. When there is no dependence between the variable of interest and the auxiliary variables (population 5), all four methods perform fairly similarly. Because the functional dependence between the variable of interest and the auxiliary variables is not additive in populations 3 and 4, the results for these two populations allow us to study the performance of AM imputation under model misspecification. We can see that AM imputation still performs the best in population 3. The reason for this is that, even though the functional dependence is not additive, it can be well approximated by an additive function. In population 4, the situation is less obvious and it is difficult to rank the imputation methods. Indeed, in this population, nearest neighbor performs better than the other methods in terms of MRPE, RRVAR, and RMSE, and AM performs better in terms of RB. In order to produce a global index of performance we ranked the imputing methods

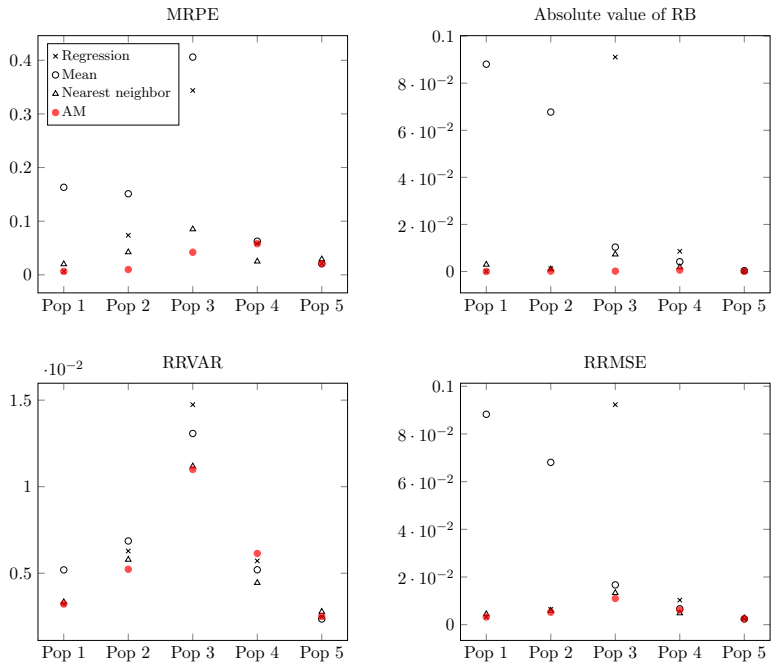


Fig. 4.2.: Comparison measures of four imputation methods in five populations under stratified sampling.

for each population and each performance criterion. The results, reported in Table 4.1 show that, globally, AM imputation performs better than the other imputation methods considered here.

The performance of the bootstrap-based estimators of variance is assessed in Table 4.2. Whether the functional dependence between the variable of interest and the auxiliary variables is additive (populations 1 and 2) or not (populations 3, 4, 5), the bootstrap variance is very close to the variance obtained by simulation. Also, it leads to very good coverage rates (between 93% and 95%) across all five populations considered.

4.6.5 Results of setting 2

Table 4.3 and Table 4.4 display the results of our analysis performed under setting 2. The numbers in brackets in Table 4.3 report the ranks of each imputation method for each measure of comparison.

We can see that AM imputation clearly outperforms the computing imputation methods in terms of MRPE and in terms of RB and performs slightly better than the other methods in terms of RRVAR. With this data, the bootstrap variance yields a coverage rate of 94% that is close to the theoretically stated value of 95%.

Tab. 4.1.: Average ranks over five populations of each imputation method for each measure of comparison (in absolute value).

Imputation method	MRPE	RB	RRVAR	RRMSE
	Simple random sampling (SRSWOR)			
Regression	2.4	2.6	2.8	3.0
Mean	3.4	3.6	2.8	3.0
Nearest Neighbor	2.4	2.6	2.4	2.4
AM	1.8	1.2	2.0	1.6
	Stratified sampling			
Regression	2.4	3.0	2.8	3.0
Mean	3.4	3.6	2.8	3.0
Nearest Neighbor	2.4	2.4	2.4	2.4
AM	1.8	1.0	2.0	1.6

4.7 Conclusion

A new imputation method for nonresponse in surveys based on spline smoothing within the additive model paradigm was proposed. The simulations indicate that the new method is very flexible and can capture a large spectrum of functional dependencies between the variable of interest and the auxiliary variables. Since the model requires only weak assumptions, it is less susceptible to model misspecification than other models such as parametric ones. Most importantly, the AM formulation makes it possible to consider several auxiliary variables in the imputation process without running into the curse of dimensionality phenomenon. A bootstrap procedure to estimate the variance of the total under SRSWOR and stratified sampling was suggested.

Through a simulation study, the new imputation method was confirmed to perform well in many different situations. The main conclusions of the simulation study are the following. AM imputation performs better than the other imputation methods considered when the functional dependence between the variable of interest and the auxiliary variables is additive or when this dependence can be well approximated by an additive function. When this dependence is not well approximated by an additive function or when there is no dependence between the variable of interest and the auxiliary variables, AM imputation shows a performance similar to that of the other imputation methods considered. In all the cases studied, the proposed bootstrap-based variance estimates were close to the true Monte Carlo variance and produced very good coverage rates.

Future work include extending the current method to situations in which the

Tab. 4.2.: Monte Carlo variance of the total, Monte carlo expectation of the bootstrap variance and coverage rate associated with AM imputation for two different sampling designs and five populations.

	VAR	VAR_{boot}	CR
	Simple random sampling (SRSWOR)		
Population 1	91424.08	90745.57	0.94
Population 2	39926.89	40811.53	0.95
Population 3	23726.96	22933.32	0.94
Population 4	16351.94	14464.70	0.93
Population 5	625.62	590.12	0.94
	Stratified sampling		
Population 1	22958.24	22836.67	0.95
Population 2	24436.15	24189.18	0.95
Population 3	21844.84	21267.10	0.94
Population 4	12414.24	10894.36	0.93
Population 5	641.01	589.73	0.94

Tab. 4.3.: Comparison measures for four imputation methods for FES data.

Imputation method	MRPE $\times 10^{1}$	RB $\times 10^{-2}$	RRVAR $\times 10^{-2}$	RRMSE $\times 10^{-2}$
Regression	3.24(3)	0.75(2)	1.41(2)	1.60(2)
Mean	4.49(4)	5.52(4)	1.52(3)	5.73(4)
Nearest Neighbor	3.21(2)	0.79(3)	1.52(3)	1.72(3)
AM	2.88(1)	0.03(1)	1.39(1)	1.39(1)

samples are dependent and improving the computational speed of the variance via parallel processing.

Acknowledgements

The authors thank Yves Tillé for his constructive suggestions. This research was supported by the Swiss National Science Foundation, project number P1NEP2_151904 (CH) and the Natural Science and Engineering Research Council of Canada (RVC).

Tab. 4.4.: Monte Carlo variance of the total, Monte carlo expectation of the bootstrap variance and coverage rate associated with AM imputation for FES data.

VAR	VAR_{boot}	CR
4198.03	4031.72	0.94

5 Weighting adjustment for nonignorable nonresponse with a heterogeneous structure of the variable of interest

Abstract

We consider a setup in which nonignorable nonresponse is present in the survey. In such a case, unit response probabilities depend on the variable of interest. It is assumed that the values of the variable of interest are sampled from a superpopulation which can be described as a mixture of some hidden components or subpopulations; a typical example of such a variable is the income. We consider that auxiliary information is available for all the sampled units. Three solutions that underline the hidden structure of the variable of interest in a logistic regression model for the response probabilities are proposed. Maximum likelihood and generalized calibration are applied to estimate the response model parameters. The estimated response probabilities are then used in a two-phase estimator for the population mean. We hypothesize that incorporating information about the heterogeneous structure of the variable of interest in the model for the response probabilities makes it possible to better control the nonresponse bias and the variance of the two-phase estimator. A variance estimator of the two-phase estimator is discussed, while the performance of the proposed procedures is studied through simulations. An application to real data is presented.[1]

Keywords: mixture distribution, survey sampling, unit response probability, two-phase estimation.

5.1 Introduction

Reweighting procedures are commonly used to compensate for unit nonresponse in surveys. The main idea is to increase the sampling weight of each respondent in order to compensate for the nonrespondents. Such procedures are referred to as nonresponse weighting adjustment (NWA) methods. Nonresponse can be viewed as a second phase of the survey. Theory of two-phase sampling hence suggests a two-phase estimator which extends the usual

[1]This chapter is a working paper co-written with Dr. Alina Matei.

Horvitz-Thompson estimator by multiplying the sampling weights of the respondents by the inverse of their response probabilities. As the response probabilities are unknown, a preliminary step consists of estimating them. The sampling weights of the respondents are then multiplied by the inverse of their estimated response probabilities and a two-phase estimator adjusted for nonresponse is obtained. In the literature, several approaches have been used to estimate the response probabilities, as for example response homogeneity groups, calibration, or parametric modeling as in Cassel et al. (1983) and Kim and Kim (2007). Auxiliary information available at the sample or population level plays a central role in the estimation process. It can simultaneously decrease variance and nonresponse bias of estimators if it is adequately used in the response probabilities estimation. The reader may refer to Särndal and Lundström (2005) for an overview of the NWA methods.

Nonignorable nonresponse refers to a nonresponse mechanism which depends on the variable of interest itself (see Little, 1982, for a formal definition). It is particularly difficult to handle, as the process that leads to nonresponse is defined through characteristics of interest which are partially or completely missing. Sophisticated techniques must therefore be used to control for nonresponse bias and variance in this framework. The problem of nonignorable nonresponse in surveys has already been addressed as for instance in Greenlees et al. (1982), Little and Rubin (1987), Beaumont (2000), and Fang et al. (2010).

We consider survey data along with nonignorable nonresponse. We assume that the values of the variable of interest are independent and identically distributed (i.i.d.) draws of a random variable with a mixture distribution, such as a mixture of normal distributions. A typical example of application in which the proposed methods can be used is a survey whose main variable of interest is the income. Indeed, it makes sense to suppose that the willingness to answer questions related to income depends on the income itself. On the other hand, the population can be broken down into several subpopulations and the income can be modeled using a mixture distribution (see e.g. Flachaire and Nuñez, 2007). In this framework we propose three NWA procedures for handling nonignorable nonresponse. Since the structure of the variable of interest is a priori unknown, the components (or subpopulations) are considered latent and we propose to reconstruct them in two steps. In the first step, a mixture model is fitted to the values of the variable of interest of the respondents via an EM algorithm. The respondents are assigned to components based on the fitted model. In the second step, the missing values of the components of the nonrespondents are imputed using auxiliary information available for both respondents and nonrespondents. In the three presented NWA procedures, the response probabilities are modeled through logistic regression based on

these reconstructed components. The estimated response probabilities are then used in two-phase estimator for the population mean. We hypothesize that incorporating the components in the model for the response probabilities makes it possible to better control the nonresponse bias and the variance of the two-phase estimator.

The paper is organized as follows: Section 5.2 introduces the framework and notation. Section 5.3 discusses the estimation of response probabilities for nonignorable nonresponse using logistic regression. The proposed procedures are presented in Section 5.4; an estimator of the variance is introduced in Section 5.5. Section 5.6 assesses the performance of the proposed procedures through three simulation studies. An application to real data is presented in Section 5.7. Section 5.8 closes the paper with concluding remarks.

5.2 Framework

Consider a finite population U of size N, indexed by i from 1 to N. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})^\top$ be a vector of values of q auxiliary variables $\mathbf{x}$ attached to unit i and suppose that the parameter of interest is the population mean

$$\overline{Y} = \frac{1}{N} \sum_{i \in U} y_i,$$

for a continuous or categorical variable of interest y, where y_i is the value of y for unit $i \in U$.

In a first phase, a sample s of size n is selected from population U using a sampling design $p(s)$. Let $\pi_i = \sum_{s; s \ni i} p(s)$ denote the first-order inclusion probability of unit i and suppose thereafter that $\pi_i > 0$ for all $i \in U$. The vector $\mathbf{x}_i$ of auxiliary variables is assumed to be available for each population unit $i \in U$ or at least for each sampled unit $i \in s$. In the presence of unit nonresponse, some selected units do not respond to the survey. This results in two subsets which form a partition of s: the survey *respondents* (the set r) and the survey *nonrespondents* (the set $\overline{r}$). The value y_i of the variable of interest is observed for each respondent $i \in r$ but is missing for each nonrespondent $i \in \overline{r}$. For $i \in s$, let R_i be the response indicator of y_i which takes value 1 if unit i is a respondent (i.e. if $i \in r$) and 0 if unit i is a nonrespondent (i.e. if $i \in \overline{r}$). Let p_i be the response probability of unit i, that is $p_i = \Pr(i \in r | s; i \in s)$. It is supposed that the units respond independently from each other. The response indicator R_i is therefore generated from a Bernoulli random variable with parameter p_i. Moreover, it is thereafter assumed that $p_i > 0$ for all $i \in U$. In the ideal case of complete response, the Horvitz-Thompson estimator

$$\widehat{\overline{Y}}_\pi = \frac{1}{N} \sum_{i \in s} \frac{1}{\pi_i} y_i, \tag{5.1}$$

is a design unbiased estimator of $\overline{Y}$. In the presence of nonresponse, however, this latter is unavailable, as the values y_i of the variable of interest are missing for nonrespondents $i \in \overline{r}$. Nonresponse can be viewed as a second phase of the survey; a subsample r of s is selected according to a Poisson sampling design $q(r|s) = \prod_{i \in r} p_i \prod_{i \in \overline{r}} (1 - p_i)$. Theory of two-phase sampling suggests, in this case, the double expansion estimator

$$\widehat{\overline{Y}}_{true} = \frac{1}{N} \sum_{i \in r} \frac{1}{\pi_i} \frac{1}{p_i} y_i,$$

which extends the Horvitz-Thompson estimator in Expression (5.1). This estimator would be unbiased for $\overline{Y}$ if the response probabilities p_i were known. Unfortunately, this is never the case. A preliminary step therefore consists of estimating the response probabilities. Those are then replaced by the estimated response probabilities $\widehat{p}_i$ in the previous estimator and the two-phase estimator adjusted for nonresponse

$$\widehat{\overline{Y}}_{e0} = \frac{1}{N} \sum_{i \in r} \frac{1}{\pi_i} \frac{1}{\widehat{p}_i} y_i, \tag{5.2}$$

is obtained. If the response probabilities are fitted using logistic regression with maximum likelihood estimating its parameters, it was shown by Kim and Kim (2007) that Estimator (5.2) has a lower variance than the estimator using the true response probabilities. As pointed out by Kim and Kim (2007), the weights in Estimator (5.2) do not sum to one and this estimator can be very unstable when the estimated response probabilities are close to 0. A more stable estimator for $\overline{Y}$ is a Hájek type estimator, which is considered further

$$\widehat{\overline{Y}}_{e} = \left(\sum_{j \in r} \frac{1}{\pi_j} \frac{1}{\widehat{p}_j} \right)^{-1} \sum_{i \in r} \frac{1}{\pi_i} \frac{1}{\widehat{p}_i} y_i. \tag{5.3}$$

We assume a superpopulation model where the values of the variable of interest are generated via i.i.d. draws of a random variable that follows a mixture of normal distributions with t components, i.e. a random variable that has density

$$f(y|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_t) = \sum_{\ell=1}^{t} \lambda_\ell f_\ell(y|\boldsymbol{\theta}_\ell), \tag{5.4}$$

$\lambda_\ell \geq 0$, $\sum_{\ell=1}^{t} \lambda_\ell = 1$, where λ_ℓ is the probability of component ℓ (y_i is drawn from a mixture of densities of underlying components or subpopulations in unknown mixing proportions $\lambda_1, \ldots, \lambda_t$), $\boldsymbol{\theta}_\ell = (\mu_\ell, \sigma_\ell^2)$ is the specific parameter vector for the density function f_ℓ of the normal distribution in the ℓ-th component. Inference on the number of components in a mixture model can be

obtained using a likelihood ratio test, where its p-value is assessed using bootstrap (see McLachaln and Peel, 2000, chapter 6). The number of components can also be derived using a penalized likelihood approach, as the Bayesian information criterion, or by cross-validation.

We assume that the sample inherits of the structure of y within the population and that the respondents set inherits of the structure of y within the sample. In particular, the components of the mixture distribution of y within the respondents set are the same as the components of the mixture distribution of y within the sample. They also correspond to the components of the mixture distribution within the population.

5.3 Estimating response probabilities

Under nonignorable nonresponse, a solution to estimate the response probabilities consists of modeling them through a logistic regression in which the variable of interest plays the role of a covariate. Hence, the following model can be considered:

$$p_i = \mathrm{E}\left(R_i|\mathbf{z}_i\right) = \frac{1}{1+\exp\left(-\mathbf{z}_i^\top \boldsymbol{\beta}\right)}, \tag{5.5}$$

where $\mathbf{z}_i = (1, y_i)^\top$ or $\mathbf{z}_i = (1, \mathbf{x}_i^\top, y_i)^\top$ and $\boldsymbol{\beta}$ is a vector of parameters. Two available estimation methods are maximum likelihood and generalized calibration (Deville, 2000, 2002; Kott, 2006). In what follows, we present three existing solutions to estimate the response probabilities via Model (5.5).

Solution 1: estimation via maximum likelihood with the auxiliary variables only
In the presence of nonresponse, the vector of parameters $\boldsymbol{\beta}$ in Model (5.5) can not be estimated directly via maximum likelihood since the values y_i of the variable of interest are missing for the nonrespondents. The first solution follows Cassel et al. (1983) who suggest, when y and $\mathbf{x}$ are well correlated, to only consider the auxiliary variables in the logistic regression. They propose to use Model (5.5) with $\mathbf{z}_i = (1, \mathbf{x}_i^\top)^\top$, that is

$$p_i = \mathrm{E}\left(R_i|\mathbf{x}_i\right) = \frac{1}{1+\exp\left[-\left(1, \mathbf{x}_i^\top\right)\boldsymbol{\beta}\right]}, \tag{5.6}$$

where $\boldsymbol{\beta}$ is a vector of parameters. The values $\mathbf{x}_i$ of the auxiliary variables being known for each sampled unit $i \in s$, the parameters can be estimated via maximum likelihood by considering $(R_i, \mathbf{x}_i)$ for $i \in s$. Let $\widehat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$. The estimated response probabilities $\widehat{p}_i$ are obtained by replacing the vector of parameters $\boldsymbol{\beta}$ with its estimate $\widehat{\boldsymbol{\beta}}$ in Expression (5.6). If the auxiliary variables are good linear predictors of the variable of interest or of the response

probabilities, then using $\widehat{p}_i$ in Estimator (5.3) provides protection against nonresponse bias (see Cassel et al., 1983).

Solution 2: estimation via maximum likelihood with imputed variable of interest A second solution based on maximum likelihood estimation is proposed in Laaksonen and Chambers (2006). Their solution consists of first imputing the variable of interest using an appropriate imputation method for units $i \in \bar{r}$. Then, the imputed variable y^* (with $y_i^* = y_i$, if $i \in r$) plays the role of a covariate in the logistic regression, which consists of considering $\mathbf{z}_i = (1, y_i^*)^\top$ in Model (5.5), that is

$$p_i = \mathrm{E}\left(R_i | y_i^*\right) = \frac{1}{1 + \exp\left[-\left(1, y_i^*\right)\boldsymbol{\beta}\right]}, \tag{5.7}$$

where $\boldsymbol{\beta}$ is a vector of parameters. Note that we could additionally include the auxiliary variables in the model. In this case, we would consider $\mathbf{z}_i = (1, \mathbf{x}_i^\top, y_i^*)^\top$ instead of $\mathbf{z}_i = (1, y_i^*)^\top$. The values y_i^* of the imputed variable being known for each sampled unit $i \in s$, the vector of parameters in (5.7) can be estimated via maximum likelihood by considering (R_i, y_i^*) for $i \in s$. Let $\widehat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$. We obtain the estimated response probabilities $\widehat{p}_i$ by replacing the vector of parameters $\boldsymbol{\beta}$ with its estimate $\widehat{\boldsymbol{\beta}}$ in Expression (5.7). This procedure provides protection against nonresponse bias if the method applied to impute the variable of interest reconstructs accurately the missing values.

Solution 3: estimation via generalized calibration The third solution follows Deville (2000, 2002) and Kott (2006) and uses generalized calibration to estimate the response probabilities. With Model (5.5) in mind, we postulate the following model for the inverse of the response probabilities: $p_i^{-1} = 1 + \exp(-\mathbf{z}_i^\top \boldsymbol{\beta})$, where $\mathbf{z}_i = (1, y_i)^\top$ or $\mathbf{z}_i = (1, \mathbf{x}_i^\top, y_i)^\top$. Let $\mathbf{x}_i^g$ be a vector of calibration variables attached to unit i known for each respondent and with total known at the population level, where vectors $\mathbf{x}^g$ and $\mathbf{z}$ have the same dimension. The idea is to modify the initial design weights π_i^{-1} and to find final weights $w_i = \pi_i^{-1} p_i^{-1} = \pi_i^{-1}\left[1 + \exp(-\mathbf{z}_i^\top \boldsymbol{\beta})\right]$ satisfying the calibration equation

$$\sum_{i \in r} w_i \mathbf{x}_i^g = \sum_{i \in U} \mathbf{x}_i^g. \tag{5.8}$$

To this end, raking method is applied and we calibrate on the set $U \setminus r$. We obtain the estimate $\widehat{\boldsymbol{\beta}}$ of β and the estimated response probabilities $\widehat{p}_i = \left[1 + \exp(-\mathbf{z}_i^\top \widehat{\boldsymbol{\beta}})\right]^{-1}$. Note that it is possible to consider $\sum_{i \in s} \frac{1}{\pi_i} \mathbf{x}_i^g$ in the right hand side of the calibration equation if the total is unknown at the population level.

Generalized calibration only requires the values of the variables of interest at the respondents level in order to estimate the response probabilities. By contrast, the solutions based on maximum likelihood estimation require variables with values available at the sample level, reason why the variable of interest can not be directly included in the model for the response probabilities when maximum likelihood is applied. However, generalized calibration sometimes fails to estimate the response probabilities since the existence of a solution to the calibration equation is not guaranteed. On the other hand, Lesage and Haziza (2015) showed that the generalized calibration estimator may be highly biased and unstable. In particular, they showed that: 1) the generalized calibration estimator has negligible bias when the calibration variables and the nonresponse mechanism are uncorrelated conditional to the response model variables but that it may be very unstable when the calibration variables are weakly related to the response model variables; 2) the calibration estimator is biased when the calibration variables and the nonresponse mechanism are correlated conditional to the response model variables and both the bias and the variance are amplified as the relationship between the calibration variables and response model variables weakens.

5.4 Proposed procedures

Our proposal builds on the solutions presented in Section 5.3 and adds on latent information on the variables of interest to the models for the response probabilities. As specified in Section 5.1, we adopt a superpopulation model for y (the values of the variable of interest are i.i.d. draws of a random variable with a mixture of normal distributions with t components) and we are working in the model-assisted approach (Särndal et al., 1992). Our idea is that a gain in terms of reduction of nonresponse bias and variance can be obtained if these components are included in the model for the response probabilities. In the presence of nonresponse, however, these components are latent since the values y_i of the variable of interest are observed only for the respondents. In the current section, a procedure to reconstruct the components is presented. Then, three solutions to include them in the response probabilities estimation process are proposed.

5.4.1 Reconstruction of latent components

A popular method to fit a mixture distribution is to use an EM algorithm (see McLachaln and Peel, 2000, chapter 2). Because the values of the variable of interest are observed for the respondents only, it is not possible here to fit a mixture distribution at the sample level in order to highlight the components of the sampled units. We propose to reconstruct the components with the following two steps.

In a first step, the components of the variable of interest y of the respondents are estimated via an EM algorithm. This paragraph briefly presents the applied algorithm (Benaglia et al., 2009). One considers that (y_i, k_i) are realizations of i.i.d. random variables (Y_i, K_i) where $K_i = (K_{i1}, \ldots, K_{it})$ is the vector of latent components membership indicator variables. The density (5.4) yields

$$\begin{aligned} Y_i|K_{i\ell} = 1 &\sim \mathcal{N}(\boldsymbol{\theta}_\ell), \quad \boldsymbol{\theta}_\ell = (\mu_\ell, \sigma_\ell^2), \\ K_i|\boldsymbol{\lambda} &\sim \text{Multinomial}(1, \boldsymbol{\lambda}), \quad \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_t), \end{aligned}$$

see Ishwaran and Zarepour (2002). The aim of the EM algorithm is to estimate the parameters $(\boldsymbol{\lambda}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t)$. The algorithm initializes with mixing proportions random from a uniform Dirichlet distribution and with starting values of the vectors of component means and variances. See (McLachaln and Peel, 2000, chapter 2) for the specification of the starting values for the EM algorithm. E-steps and M-steps are alternated. At each E-step, the conditional distribution of $K_{i\ell}$ is obtained via Bayes' rule given the current estimate of the parameters. At each M-step, the conditional expected log-likelihood is maximized; the parameters (the mixing proportions and the vectors of components means and variances) are updated. These two steps are alternated until convergence. When applied to the y_i values of respondents, the algorithm returns, for each respondent, posterior probabilities of components. We assign each respondent to the component with the largest posterior probability. Let $c_{i\ell}$, $i \in r, \ell = 1, \ldots, t$ be the components membership indicators where $c_{i\ell}$ takes value 1 if respondent i belongs to component ℓ and 0 otherwise.

In a second step, we propose to impute the missing components membership indicators of nonrespondents by nearest neighbor using auxiliary information. That is, we set, for $i \in s, \ell = 1, \ldots, t$

$$c_{i\ell}^* = \begin{cases} c_{i\ell}, & \text{if } i \in r, \\ c_{j(i)\ell}, & \text{if } i \in \overline{r}, \end{cases}$$

where

$$j(i) = \underset{j \in r}{\operatorname{argmin}}\, d(\mathbf{x}_i, \mathbf{x}_j),$$

for some distance measure $d(\cdot, \cdot)$.

Because our procedure fits a mixture distribution for the respondents only, it is based on the idea that the components of the mixture distribution of y within the respondents set are the same as the components of the mixture distribution of y within the sample. Note that this represents a strong assumption because the structure of y for the respondents may be different from that for the nonrespondents, e.g. individuals with a low income are more likely to be

nonrespondents to a survey related to income than individuals with a high income, or vice versa.

5.4.2 The proposed solutions

The assumption that the values of the variable of interest are generated via i.i.d. draws of a random variable that follows a mixture distribution with t components suggests a possibly different model for the variable of interest in each component, and consequently, a possibly different model for the response probabilities in each component. We propose to update the solutions presented in Section 5.3 by considering the components when estimating the response probabilities. The resulting solutions are presented in what follows.

Solution 1a: estimation via maximum likelihood with the auxiliary variables, different model in each component Based on the idea that the model for the variable of interest may vary from one component to the other, we propose to fit Model (5.6) independently in each reconstructed component. Hence, we set $\mathbf{z}_i = (c_{i1}^*, \ldots, c_{it}^*, c_{i1}^*\mathbf{x}_i^\top, \ldots, c_{it}^*\mathbf{x}_i^\top)^\top$ in Model (5.5), that is we propose the following model:

$$\begin{aligned} p_i &= \mathrm{E}\left(R_i|\mathbf{x}_i, c_{i1}^*, c_{i2}^* \ldots, c_{it}^*\right) \\ &= \frac{1}{1+\exp\left[-\left(\sum_{\ell=1}^t c_{i\ell}^*\beta_{0\ell} + \sum_{\ell=1}^t c_{i\ell}^*\mathbf{x}_i^\top\boldsymbol{\beta}_{1\ell}\right)\right]}, \end{aligned} \tag{5.9}$$

where $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{0t}, \boldsymbol{\beta}_{11}^\top, \ldots, \boldsymbol{\beta}_{1t}^\top)^\top$ is a vector of parameters. Maximum likelihood estimation is then applied to fit this model by considering $(R_i, \mathbf{x}_i, c_{i1}^*, c_{i2}^*, \ldots, c_{it}^*)$ for $i \in s$. This leads to estimate $\widehat{\boldsymbol{\beta}}$ which we plug in (5.9) to obtain $\widehat{p}_i$.

Solution 2a: estimation via maximum likelihood with the variable of interest imputed independently in each component Because a mixture of distribution underlines a possibly different model for the variable of interest within each component, we propose to impute the missing values of the variable of interest independently in each reconstructed component when the response probabilities are estimated via Model (5.7). Thus, the reconstructed components become imputation classes in the spirit of Haziza and Beaumont (2007). Note, however, that we do not use an imputed estimator of the population mean, but we use the imputed variable in the model for the response probabilities.

We impute the variable of interest via regression imputation, independently in each component. That is, we obtain values y_i^* of the imputed variable y^* as

follows

$$y_i^* = \begin{cases} y_i, & \text{if } i \in r, \\ \sum_{\ell=1}^{t} \mathbf{u}_i^\top \widehat{\mathbf{b}}_\ell c_{i\ell}^*, & \text{if } i \in \overline{r}, \end{cases}$$

where

$$\widehat{\mathbf{b}}_\ell = \left(\sum_{i \in r} w_i \mathbf{u}_i \mathbf{u}_i^\top c_{i\ell}^* \right)^{-1} \left(\sum_{j \in r} w_j \mathbf{u}_j y_j c_{j\ell}^* \right),$$

where $\mathbf{u}_i$ is a vector of auxiliary variables available for all units $i \in r$ that can be different from $\mathbf{x}_i$ and that contains a constant variable. The regression imputation model is specific for each component and may be different from one component to the other. Using a separate regression model in each component may lead to a better imputation of missing values of y compared to an overall regression imputation model.

The values y_i^* of the imputed variable are available for all sampled units $i \in s$ and for all components. The response probabilities are estimated using logistic regression with the imputed variable of interest as a covariate as in Laaksonen and Chambers (2006). Hence, we set $\mathbf{z}_i = (1, y_i^*)^\top$ in Model (5.5), which leads to Model (5.7). The model parameters are estimated through maximum likelihood by considering (R_i, y_i^*) for $i \in s$. A first option is to fit Model (5.7) independently in each component $\ell = 1, \ldots, t$; a second option is to fit a single logistic model over all the components. To better preserve the variability of the original variable of interest y, random regression imputation may be used instead of the deterministic one. However, that may lead to a larger variability of the estimated response probabilities and to a variance inflation.

Solution 3a: estimation via generalized calibration within the components The third solution we propose follows Deville (2000, 2002) and Kott (2006) and applies generalized calibration within each reconstructed component to estimate the response probabilities. We set $\mathbf{z}_i = (c_{i1}^*, \ldots, c_{it}^*, c_{i1}^* y_i, \ldots, c_{it}^* y_i)^\top$ and $\mathbf{x}_i^g = (c_{i1}^*, \ldots, c_{it}^*, c_{i1}^* x_{i(1)}, \ldots, c_{it}^* x_{i(t)})^\top$, where $x_{i(\ell)}$ is the value of the auxiliary variable that is the most strongly correlated to the variable of interest within component ℓ (Deville, 2000, 2002). We apply generalized calibration as described in Solution 3, page 92.

The proposed procedures underlie two models, a superpopulation model and a nonresponse model. The superpopulation model assumes that the values of the variable of interest are i.i.d. draws from a mixture distribution and that a linear regression model may be fitted within each component. The nonresponse model assumes that the response probabilities are written as a logistic function of the values of the variable of interest, see Model (5.5). If both models are

correct, our hypothesis is that the proposed procedures provide additional protection against nonresponse bias and variance compared to the procedures in Section 5.3. However, the components of the mixture distribution might not be perfectly reconstructed; the reconstruction precision may impact on the bias and variance of the mean estimators obtained with our procedures.

Finally, even though y_i^* is essentially a linear combination of the outer product of the auxiliary variables and the reconstructed component membership indicators, Model (5.7) is different from Model (5.9) because it uses the original y_i for the respondents and thus performs closer to the assumed response model. This remark is also sustained by the examples shown further.

5.5 Variance estimation

We follow Kim and Kim (2007) to estimate the variance of the two-phase estimator adjusted for nonresponse of Expression (5.3) when the proposed procedures are applied to estimate the response probabilities. Suppose a logistic regression model for the response probabilities p_i with covariates $\mathbf{z}_i$ and parameters $\boldsymbol{\beta}$ and assume that the response probabilities are estimated through maximum likelihood. In addition, we assume the same conditions as those in Kim and Kim (2007) on the data and on the nonresponse mechanism. Hence, we assume that there is a sequence of samples and finite populations such that the sequence of $\mathbf{v}_i = (1, \mathbf{z}_i^\top, y_i)^\top$ has bounded fourth moment, such that the sample moments of $\mathbf{v}$ converge to their population moments, and such that no extreme weights dominate the others (see Kim and Kim, 2007, for details). The following conditions on the nonresponse mechanism are also set

- (R1) The responses are independent from each other,
- (R2) The response probabilities are parametrically modeled,
- (R3) The response probability is bounded below,
- (R4) The response probability of a unit does not depend on the characteristics of the other elements in the sample.

Kim and Kim (2007) consider the reverse approach (Fay, 1991; Shao and Steel, 1999), which is the reason why condition (R4) is set. Finally, we assume that the covariates $\mathbf{z}_i$ have finite second moment. In the ideal case of complete response, the estimator

$$\widehat{V}\left(\widehat{\overline{Y}}_\pi\right) = \sum_{i \in s} \sum_{j \in s} \Omega_{ij} y_i y_j,$$

can be used to estimate the variance of $\widehat{\overline{Y}}_\pi$. Suppose the coefficients Ω_{ij} are chosen such that estimator $\widehat{V}(\widehat{\overline{Y}}_\pi)$ is unbiased for the variance of $\widehat{\overline{Y}}_\pi$. Considering the estimator proposed in Kim and Kim (2007) (see Expression

(33) in the paper of Kim and Kim) to estimate the variance of $\widehat{\overline{Y}}_e$ leads

$$\widehat{V}_{rev}\left(\widehat{\overline{Y}}_e\right) = \widehat{V}_{e1} + \widehat{V}_{e2}, \tag{5.10}$$

where

$$\begin{aligned}
\widehat{V}_{e1} &= \sum_{i\in r}\sum_{j\in r}\Omega_{ij}\widehat{\eta}_i\widehat{\eta}_j,\\
\widehat{\eta}_i &= y_i - \widehat{\overline{Y}}_e,\\
\widehat{V}_{e2} &= N^{-2}\sum_{i\in r}\pi_i^{-1}\widehat{p}_i^{-2}\left(1-\widehat{p}_i\right)\left(y_i - \widehat{\overline{Y}}_e - \pi_i\widehat{p}_i\mathbf{z}_i^\top\widehat{\boldsymbol{\alpha}}_n\right)^2,\\
\widehat{\boldsymbol{\alpha}}_n &= \left\{\sum_{i\in r}(1-\widehat{p}_i)\mathbf{z}_i\mathbf{z}_i^\top\right\}^{-1}\sum_{i\in r}\pi_i^{-1}(\widehat{p}_i^{-1}-1)\mathbf{z}_i\left(y_i - \widehat{\overline{Y}}_e\right).
\end{aligned}$$

Under the assumptions stated above, estimator $\widehat{V}_{rev}(\widehat{\overline{Y}}_e)$ is approximately unbiased for the variance of the mean.

In the construction of the variance estimator proposed in Kim and Kim (2007), it is supposed that the estimate of the vector of parameters is the solution to an estimating equation (Equation (7) in Kim and Kim, 2007). When the vector of parameters of the nonresponse model is estimated via maximum likelihood, the estimate satisfies this estimating equation. It is not obviously the case when the vector of parameters is estimated via generalized calibration. However, we hypothesize that, for a given model, the two estimation methods, maximum likelihood and generalized calibration, lead to variances that are of the same order, and we also use Formula (5.10) to estimate the variance of $\widehat{\overline{Y}}_e$ when the response probabilities are estimated via generalized calibration (solutions 3 and 3a). We will see that this assumption is sometimes inadequate.

5.6 Simulations

A simulation study was conducted to evaluate the performance of the procedures proposed in Section 5.4. Three different settings were considered and are presented below. A census was considered in all cases, which implies that we set $U = s$ and $\pi_i = 1$ for each $i \in s$, in order to focus only on the nonresponse error. Ten thousand simulations were conducted. Functions of the R package `mixtools` (Benaglia et al., 2009) were used in order to randomly generate mixture distributions and to perform the EM algorithm on the resulting y.

For each setting, the simulations were conducted as follows. First, for each unit i, the response probabilities were obtained from the logistic function

$$p_i = \frac{1}{1+\exp\left[-\left(d_0 + d_1 y_i\right)\right]},$$

where d_0 and d_1 were fixed to obtain a mean response rate close to 70%. Next, 10,000 response sets were created by generating 10,000 response indicator vectors R. Each component $R_i, i \in U$ of R was generated from a Bernoulli distribution with parameter p_i. For each response set generated, the population mean of the variable of interest was estimated through the two-phase estimator adjusted for nonresponse given in Expression (5.3) by considering different solutions to estimate the response probabilities. The different estimators considered are

1. $\widehat{\overline{Y}}_1$: estimator with solution 1 to estimate the response probabilities (Cassel et al., 1983), i.e. the response probabilities were estimated via maximum likelihood estimation of the parameters of the logistic regression model with the auxiliary variables as covariates (Model (5.6)),
2. $\widehat{\overline{Y}}_{1a}$: estimator with solution 1a to estimate the response probabilities, i.e. the response probabilities were estimated via maximum likelihood estimation of the parameters of the logistic regression model with the auxiliary variables as covariates. A different model was considered in each component (Model (5.9)),
3. $\widehat{\overline{Y}}_2$: estimator with solution 2 to estimate the response probabilities; the variable of interest was imputed through regression (the components of the distribution of the variable of interest were not considered) and the response probabilities were estimated via maximum likelihood estimation of the parameters of the logistic regression model with the imputed variable as covariate (Model (5.7)),
4. $\widehat{\overline{Y}}_{2a}$: estimator with solution 2a, first option, to estimate the response probabilities. The variable of interest was imputed through regression within each reconstructed component; the response probabilities were estimated via maximum likelihood estimation of the parameters of the logistic regression model within each reconstructed component with the imputed variable as covariate (Model (5.7)),
5. $\widehat{\overline{Y}}_{2aa}$: estimator with solution 2a, second option, to estimate the response probabilities. The variable of interest was imputed through regression within each reconstructed component; the response probabilities were estimated via maximum likelihood estimation of the parameters of the global logistic regression model with the imputed variable as covariate (Model (5.7)),
6. $\widehat{\overline{Y}}_3$: estimator with solution 3 to estimate the response probabilities, i.e. the response probabilities were estimated with generalized calibration (the components were not considered); we set $\mathbf{z}_i = (1, y_i)^\top$ and $\mathbf{x}_i^g = (1, x_i)^\top$, where x is the auxiliary variable that is the most strongly correlated to the variable of interest (Deville, 2000, 2002; Kott, 2006),

7. $\widehat{\overline{Y}}_{3a}$: estimator with solution 3a to estimate the response probabilities, i.e. the response probabilities were estimated via generalized calibration within each component; we set $\mathbf{z}_i = (c^*_{i1}, \ldots, c^*_{it}, c^*_{i1}y_i, \ldots, c^*_{it}y_i)^\top$ and $\mathbf{x}^g_i = (c^*_{i1}, \ldots, c^*_{it}, c^*_{i1}x_{i(1)}, \ldots, c^*_{it}x_{i(t)})^\top$, where $x_{i(\ell)}$ is the value of the auxiliary variable that is the most strongly correlated to the variable of interest within component ℓ,
8. $\widehat{\overline{Y}}_4$: estimator with constant estimated response probabilities (equal to the response rate); the response probabilities were estimated with solution 1 with $\mathbf{z}_i = 1$,
9. $\widehat{\overline{Y}}_{4a}$: estimator with constant estimated response probabilities within the components (equal to the response rate within the components); the response probabilities were estimated with solution 1a with $\mathbf{z}_i = (c^*_{i1}, \ldots, c_{it^*})^\top$,
10. $\widehat{\overline{Y}}_{true}$: the true response probabilities were considered in the two-phase estimator. The variance of this estimator could not be estimated with Formula (5.10) because the response probabilities were not parametrically modeled.

For each simulation run, the variance associated with each mean estimator was estimated with Formula (5.10). When solutions 2 and 2a were applied, the variance was estimated with Formula (5.10) with $\mathbf{z}_i = (1, \mathbf{u}_i^\top \widehat{\mathbf{b}})^\top$ and $\mathbf{z}_i = (1, \sum_{\ell=1}^t \mathbf{u}_i^\top \widehat{\mathbf{b}}_\ell c^*_{i\ell})^\top$, respectively, where

$$\widehat{\mathbf{b}} = \left(\sum_{i\in r} w_i \mathbf{u}_i \mathbf{u}_i^\top\right)^{-1} \left(\sum_{j\in r} w_j \mathbf{u}_j y_j\right),$$
$$\widehat{\mathbf{b}}_\ell = \left(\sum_{i\in r} w_i \mathbf{u}_i \mathbf{u}_i^\top c^*_{i\ell}\right)^{-1} \left(\sum_{j\in r} w_j \mathbf{u}_j y_j c^*_{j\ell}\right).$$

The following measures were considered for these estimators and their associated variance estimators, here generically denoted by $\widehat{\overline{Y}}$ and $\widehat{V}(\widehat{\overline{Y}})$:

- The Monte Carlo relative bias:

$$\text{RB} = \frac{B}{\overline{Y}},$$

where $B = \text{E}_{sim}(\widehat{\overline{Y}}) - \overline{Y}$,

$$\text{E}_{sim}(\widehat{\overline{Y}}) = \frac{1}{M}\sum_{i=1}^{M} \widehat{\overline{Y}}_i,$$

$\widehat{\overline{Y}}_i$ is the estimate of $\widehat{\overline{Y}}$ obtained at the i-th simulation run, and M is the number of simulation runs,

- The Monte Carlo variance:

$$\text{VAR} = \frac{1}{M-1}\sum_{i=1}^{M}\left[\widehat{\overline{Y}}_i - \text{E}_{sim}\left(\widehat{\overline{Y}}\right)\right]^2,$$

- The Monte Carlo mean square error: $\text{MSE} = \text{B}^2 + \text{VAR}$,
- The simulation expected value of the variance estimator:

$$\text{E}_{sim}\left[\widehat{V}(\widehat{\overline{Y}})\right] = \frac{1}{M}\sum_{i=1}^{M}\widehat{V}(\widehat{\overline{Y}})_i,$$

  where $\widehat{V}(\widehat{\overline{Y}})_i$ is the variance estimator of $\widehat{\overline{Y}}$ obtained at the i-th simulation run using Expression (5.10),
- The coverage rate CR: the proportion of times the true mean $\overline{Y}$ falls into a 95% confidence interval based on the normal approximation

$$\widehat{\overline{Y}} \pm 1.96\sqrt{\widehat{V}(\widehat{\overline{Y}})}.$$

Setting 1

A population of size 500 was generated. The values y_i of the variable of interest y were generated from a normal mixture distribution. We considered two components with prior probability $\lambda_1 = \lambda_2 = 0.5$, with mean $\mu_1 = 0$, $\mu_2 = 5$, and standard deviation $\sigma_1 = \sigma_2 = 1$, respectively. Next, the EM algorithm was applied on y and posterior component membership probabilities were obtained. We assigned to each unit the component with the largest posterior probability, which resulted in partitioning the population into two components. Next, the values x_i of one auxiliary variable x were generated as follows:

$$\begin{aligned} x_i &= 4 + 3y_i + \varepsilon_i \qquad \text{if unit } i \text{ belongs to component 1,} \\ x_i &= 1 + 10y_i + \varepsilon_i \qquad \text{if unit } i \text{ belongs to component 2,} \end{aligned}$$

where the values ε_i of the random noises ε were generated from independent draws of a $N(0,1)$ distribution. Simulations were then conducted according to the scheme described above. The results are presented in Table 5.1.

In this setting, incorporating the components in the model for the response probabilities decreases both the bias and the variance of the mean estimator (compare $\widehat{\overline{Y}}_1$ with $\widehat{\overline{Y}}_{1a}$, $\widehat{\overline{Y}}_2$ with $\widehat{\overline{Y}}_{2a}$, etc). The generalized calibration estimators $\widehat{\overline{Y}}_3$ and $\widehat{\overline{Y}}_{3a}$ perform the best in terms of bias and the former performs extremely well even though the model considered to estimate the response probabilities does not include the components. The reason is that the calibration variables and the nonresponse mechanism are uncorrelated conditional to the response model variables (see last paragraph of Section 5.3). This is also the case in the next two settings. The relative bias of estimators $\widehat{\overline{Y}}_{2a}$ to

Tab. 5.1.: Comparison measures of estimators in Setting 1.

	RB ($\times 10^{-3}$)	Var ($\times 10^{-4}$)	MSE ($\times 10^{-3}$)	$\mathrm{E}_{sim}(\widehat{V})$ ($\times 10^{-4}$)	CR
$\widehat{\overline{Y}}_1$	16.04	7.40	2.54	7.06	0.63
$\widehat{\overline{Y}}_{1a}$	3.02	2.16	0.28	2.08	0.88
$\widehat{\overline{Y}}_2$	−3.14	9.92	1.06	14.87	0.98
$\widehat{\overline{Y}}_{2a}$	0.88	2.43	0.25	6.76	0.99
$\widehat{\overline{Y}}_{2aa}$	−0.09	5.35	0.54	6.39	0.96
$\widehat{\overline{Y}}_3$	−0.17	4.48	0.45	4.54	0.93
$\widehat{\overline{Y}}_{3a}$	−0.07	1.09	0.11	1.29	0.81
$\widehat{\overline{Y}}_4$	239.01	68.75	405.88	57.89	0.00
$\widehat{\overline{Y}}_{4a}$	35.26	11.05	9.79	10.27	0.17
$\widehat{\overline{Y}}_{true}$	0.71	100.78	10.08		

$\widehat{\overline{Y}}_{3a}$ is of the same order than the bias of estimator $\widehat{\overline{Y}}_{true}$, which uses the true response probabilities and is therefore unbiased (the residual bias is due to the simulations). These four estimators have, in this setting, an excellent performance in terms of bias. The estimators using equal response probabilities ($\widehat{\overline{Y}}_4$) and equal probabilities within the components ($\widehat{\overline{Y}}_{4a}$) yield poor results which is not a surprise since they do not take advantage of the auxiliary information available.

The variance estimator globally performs well: the simulation expected value of the variance estimator in (5.10) is generally close to the variance obtained by simulations for all the estimators. For estimator $\widehat{\overline{Y}}_{2a}$, the former is however more than twice the latter, which explains the large coverage rate associated. The inverse phenomenon appears for estimators $\widehat{\overline{Y}}_4$ and $\widehat{\overline{Y}}_{4a}$. The associated coverage rates are around the expected value of 0.95 for estimators $\widehat{\overline{Y}}_{2aa}$ and $\widehat{\overline{Y}}_3$, slightly higher for estimators $\widehat{\overline{Y}}_2$ and $\widehat{\overline{Y}}_{2a}$, and lower for the other estimators. This questions the normality assumption of the mean estimator upon which is built the confidence interval and the assumption that, for a given model, the two estimation methods (maximum likelihood and generalized calibration) lead to variances that are of the same order.

Finally, the estimators using the response probabilities estimated through maximum likelihood show a decrease in variance compared to the estimator with the true probabilities ($\widehat{\overline{Y}}_{true}$). This phenomenon also appears in the next two settings. This confirms the result of Kim and Kim (2007) stating that if the response probabilities are parametrically modeled, then the estimator using the estimated response probabilities is more efficient than the estimator using the

true response probabilities when maximum likelihood is used to estimate the model parameters. Note that estimator $\widehat{\overline{Y}}_{true}$ is in practice unavailable because the true response probabilities are unknown.

Setting 2

As in the first setting, a population of size 500 was generated and the values y_i of the variable of interest y were generated from a normal mixture distribution. Two components with prior probability $\lambda_1 = \lambda_2 = 0.5$, with mean $\mu_1 = 0$, $\mu_2 = 5$ and standard deviation $\sigma_1 = 1$, $\sigma_2 = 2$, respectively, were considered. Two auxiliary variables x_1 and x_2 were used, with values x_{i1} and x_{i2} generated as follows:

$$
\begin{aligned}
x_{i1} &= 4 + 3y_i + \varepsilon_{i1} && \text{if unit } i \text{ belongs to component 1,} \\
x_{i2} &= 1 - 6y_i + \varepsilon_{i11} && \text{if unit } i \text{ belongs to component 1,} \\
x_{i1} &= 1 + 10y_i^2 + \varepsilon_{i2} && \text{if unit } i \text{ belongs to component 2,} \\
x_{i2} &= 4 + \varepsilon_{i22} && \text{if unit } i \text{ belongs to component 2,}
\end{aligned}
$$

where the values ε_{ij}, $j = 1, 11, 2, 22$ of the random noises were generated from independent draws of a normal random variable with mean 0 and standard deviation 1, 3, 3, 10, respectively. In this setting, different standard deviations were considered in the mixture distribution used to generate y, random noises have standard deviations that differ from one component to the other, and two auxiliary variables were used. These characteristics represent the main differences with respect to Setting 1. Simulations were then conducted according to the scheme described above. The results are presented in Table 5.2.

Tab. 5.2.: Comparison measures of estimators in Setting 2.

	RB ($\times 10^{-3}$)	Var ($\times 10^{-4}$)	MSE ($\times 10^{-3}$)	$\mathrm{E}_{sim}(\widehat{V})$ ($\times 10^{-4}$)	CR
$\widehat{\overline{Y}}_1$	26.84	11.36	6.71	10.10	0.35
$\widehat{\overline{Y}}_{1a}$	2.71	2.39	0.30	1.70	0.82
$\widehat{\overline{Y}}_2$	3.65	14.19	1.52	18.41	0.96
$\widehat{\overline{Y}}_{2a}$	1.78	2.54	0.28	4.92	0.95
$\widehat{\overline{Y}}_{2aa}$	1.05	4.43	0.45	4.58	0.94
$\widehat{\overline{Y}}_3$	−0.19	9.66	0.97	3.45	0.72
$\widehat{\overline{Y}}_{3a}$	0.43	2.49	0.25	1.07	0.65
$\widehat{\overline{Y}}_4$	205.71	59.78	333.40	50.03	0.00
$\widehat{\overline{Y}}_{4a}$	30.53	9.92	8.20	8.80	0.20
$\widehat{\overline{Y}}_{true}$	0.91	87.51	8.76		

In this setting, the results follow a similar pattern that those in Setting 1, except that estimator $\widehat{\overline{Y}}_3$ performs better than estimator $\widehat{\overline{Y}}_{3a}$ in terms of bias even though the model considered to estimate the response probabilities does not include the components for the former whereas it does for the latter. However, both these estimators yield a relative bias smaller than that of estimator $\widehat{\overline{Y}}_{true}$, which indicate that they both have an excellent performance in terms of bias.

The variance estimator globally performs well with a simulation expected value generally close to the variance obtained by simulations. It yields excellent coverage rates for estimators $\widehat{\overline{Y}}_2$, $\widehat{\overline{Y}}_{2a}$, and $\widehat{\overline{Y}}_{2aa}$. For the other estimators, the simulation expected value of the variance estimator (5.10) is smaller than the variance obtained by simulation, which indicates that the variance is underestimated. Again, the low coverage rates question the normality assumption of the mean and the assumption that maximum likelihood and generalized calibration lead to variances that are of the same order.

Setting 3: MU284 population

We considered the MU284 population data available in Appendix B of Särndal et al. (1992). The revenues from 1985 municipal taxation (in millions of kronor) and the 1985 population (in thousands) served as variable of interest and as auxiliary variable, respectively. These two variables were modified as follows. First, six outliers were removed. Hence, only the observations with a revenue from 1985 municipal taxation smaller than 1000 were considered. Second, the variable of interest was divided by its standard deviation. Finally, the natural logarithm was applied to both variables. Figure 5.1 shows the scatter plot of the final auxiliary variable x against the final variable of interest y (left) and the density estimate of the final variable of interest y (right). Note that the y does not indicate the presence of a mixture distribution as such. Rather a spurious clustering was presumed as shown in Figure 5.1. Simulations were then conducted according to the scheme described above to check if the proposed methods are robust to the departure from the mixture structure assumption for y. Two normal components of the mixture distribution of y were considered. In this setting, the calibration failed (non-convergence or impossible calibration) for 82 simulations out of 10,000. We decided to leave out these 82 simulations but checked that this did not modify the performance of the others estimators. The results are presented in Table 5.3.

In this setting, $\widehat{\overline{Y}}_3$ performs the best, closely followed by estimator $\widehat{\overline{Y}}_{3a}$. These two estimators show here not only a very small bias but also a very small variance compared to the other estimators. As it was also the case in the first two settings, the calibration variables and the nonresponse mechanism are uncorrelated conditional to the response model variables, which explains the small bias associated with these two estimators. In this setting, the calibration

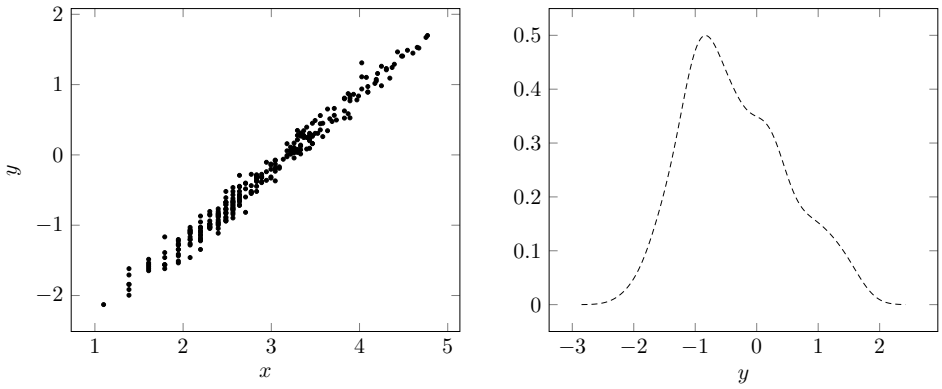


Fig. 5.1.: Left panel: scatter plot of the auxiliary variable x against the variable of interest y. Right panel: density estimate of y.

Tab. 5.3.: Comparison measures of estimators in Setting 3.

	RB ($\times 10^{-3}$)	Var ($\times 10^{-4}$)	MSE ($\times 10^{-3}$)	$\mathrm{E}_{sim}(\widehat{V})$ ($\times 10^{-4}$)	CR
$\widehat{\overline{Y}}_1$	7.33	6.83	6.89	6.35	0.89
$\widehat{\overline{Y}}_{1a}$	14.77	4.12	4.38	3.71	0.85
$\widehat{\overline{Y}}_2$	1.07	6.23	6.24	6.26	0.92
$\widehat{\overline{Y}}_{2a}$	3.84	3.89	3.91	7.12	0.97
$\widehat{\overline{Y}}_{2aa}$	1.13	6.27	6.27	6.46	0.92
$\widehat{\overline{Y}}_3$	−0.36	0.22	0.22	5.70	1.00
$\widehat{\overline{Y}}_{3a}$	−0.72	0.24	0.24	3.45	0.99
$\widehat{\overline{Y}}_4$	610.23	7.52	454.00	5.56	0.00
$\widehat{\overline{Y}}_{4a}$	285.18	13.44	110.95	3.25	0.00
$\widehat{\overline{Y}}_{true}$	3.77	28.60	28.62		

variables are moreover strongly correlated the the response model variables, which explains the small variance (see last paragraph of Section 5.3). The hypothesis of a mixture distribution for y does not hold and it seems that incorporating the components in the model for the response probabilities tends to increase the bias and to decrease the variance of the mean estimator (compare for example $\widehat{\overline{Y}}_1$ with $\widehat{\overline{Y}}_{1a}$ or $\widehat{\overline{Y}}_2$ with $\widehat{\overline{Y}}_{2a}$) but this is not a rule (compare $\widehat{\overline{Y}}_4$ with $\widehat{\overline{Y}}_{4a}$).

In this setting, the variance estimator globally performs well for the first five estimators: the simulation expected value of the variance estimator in (5.10) is relatively close to the variance obtained by simulations. Note that, for estimators $\widehat{\overline{Y}}_3$ and $\widehat{\overline{Y}}_{3a}$, the former is much larger that the latter. The reason for the overestimation of the variance of these two estimators is that the two

estimation methods available to estimate the response probabilities, maximum likelihood and generalized calibration, do not yield similar variance when there is a strong correlation between the variable of interest and the auxiliary variables. Indeed, when the response probabilities are estimated via generalized calibration and when the auxiliary variables are included in the calibration variables $\mathbf{x}^g$, the estimated response probabilities satisfy Equation (5.8) and the variance of the total estimator of the auxiliary variables vanishes. Because the variable of interest and the auxiliary variables are highly correlated, it implies that the variance of the total estimator of the variable of interest nearly vanishes. Hence, in this particular case, estimation of the model parameters via generalized calibration yield to a mean estimator with a possibly much smaller variance than if the model parameters were estimated via maximum likelihood and Formula (5.10) overestimates the variance.

5.7 Application to real data

We consider the Swiss Statistics on Income and Living Conditions (SILC) survey data of 2009. In this survey, units are households made up of permanent residents in Switzerland in which, whenever possible, all individuals aged 16 or over were interviewed. For individuals in the sample, the survey variable income measured by Computer Assisted Telephone Interview (CATI) was coupled with income data from an available register (Central Compensation Office register). We considered as sampled units only the individuals with available income data from the register and older than 16. This implied a sample of size $n = 8762$. The variable recorded from the register is more accurate than that measured by interview. The first one was used as the variable of interest y. On the other hand, the CATI variable is affected by an unknown nonresponse mechanism. The 2009 SILC survey data was previously analyzed by Graf (2014). This author applied the observed nonresponse mechanism affecting the CATI variable on the corresponding variable from the register y. Graf (2014) employed a segmentation technique (Kass, 1980) on y using 41 auxiliary variables that produced 73 response homogeneous groups and noted that the correlation between the response rate of y and the median of y in each response homogeneous group was about 0.39, indicating that the nonresponse on y (and affecting the CATI variable since both are highly correlated) is not ignorable.

Using the same idea as in Graf (2014), the set of nonrespondents to the CATI variable was considered as $\overline{r}$. Choosing y to be the variable from the register allows us to compute the bias due to nonresponse (the second phase) because the sample mean estimator is known for the first phase. In set r, the correlation between the variable from the register y and the variable measured by CATI was about 0.84. The difference between the two variables is due to a measurement

error in collecting the CATI variable. The following auxiliary variables x_1, x_2, and x_3 were used: age, sex, and number of years of education, respectively. Finally, the survey weights w_i were considered. These weights are issued from a complex survey sampling and include a calibration nonresponse adjustment factor.

Among the 8762 sampled units, 6884 were therefore classified as respondents, which corresponds to a response rate of approximately 78%. A mixture of normal distributions was used for y. The EM algorithm applied on y highlighted three components of the variable of interest. Figure 5.2 shows the density estimates (dashed curves) of the income from the register (y) on s and r, respectively. The solid curves represent gaussian densities of three individual components, each of which being scaled by its posterior probability. For graphical reasons, the abscissa axes were cut at $1 \cdot 10^6$.

The plots highlight two interesting features of the data. First, both graphs show three bumps (the last one being less visible because of the long right tail of the y distribution), which indicates the possible presence of three components in the population. Second, the height of the bumps differs from one graph to the other. Indeed, the first bump is higher than the second when all sampled units are considered, whereas they are of same height when the respondents only are considered. This means that units with smaller income tend to be less likely to respond to the question of income than units with higher income, which supports the hypothesis of nonignorable nonresponse.

The response rates computed in each component on the sample level were 0.63, 0.82 and 0.83, respectively, showing different response behaviors between the groups and sustaining the hypothesis of nonignorable nonresponse. When the EM algorithm was applied on r, the first component of y underlined small incomes with a range between 100 and 24,505 Swiss francs, the second component underlined incomes with a range between 24,578 and 177,151 Swiss francs, while the third component underlined incomes with a range between 177,560 and 1,758,845 Swiss francs. The number of respondents in each group was 1466, 5160 and 258, respectively. When computed within the reconstructed components obtained with the procedure presented in Section 5.4.1, we obtained the following response rates: 0.80, 0.78 and 0.81, respectively, overestimating the response rate of units with smaller income (see the true response rates above).

To find the nearest neighbors to impute the components, we applied the function `nn2` included in the R package `RANN` (Arya et al., 2014). Hence, for each nonrespondent, 300 nearest neighbors (among both respondents and nonrespondents) were found and one donor was randomly selected among the respondents that appeared in these 300 nearest neighbors. In each group, a

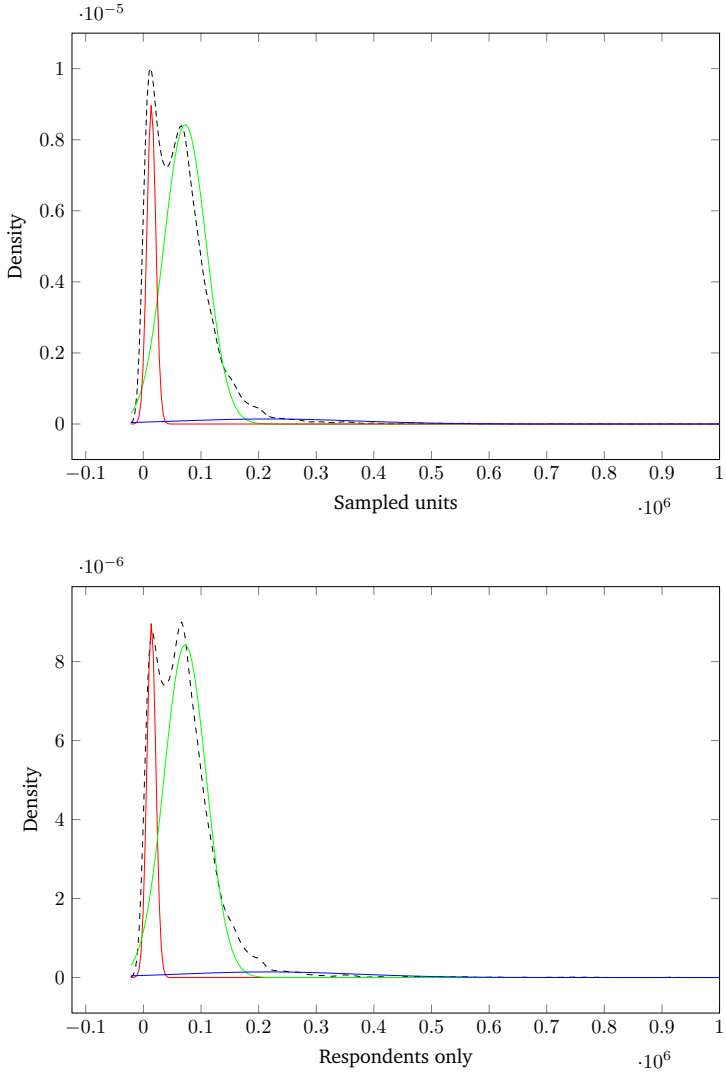


Fig. 5.2.: Density estimates of the income from the register (dashed curves) and gaussian densities of three components (solid curves) for sampled units (top panel) and for respondents only (bottom panel).

different regression model was used to impute the missing values of y. Thus, in the first and second group, the regression models used age and education as covariates, while in the third one, all three covariates (age, sex, and education) were used. In the first two components, the square of the imputed variable of interest was considered as a covariate in the logistic regression used to model the response probabilities.

Because we had no information about the design plan or about the inclusion probabilities, we supposed that the sample was selected from the population via Poisson sampling and we considered the inverse of the survey weights as response probabilities. The two-phase estimator adjusted for nonresponse was therefore computed as follows:

$$\widehat{\overline{Y}}_e = \left(\sum_{j\in r} w_j \frac{1}{\widehat{p}_j}\right)^{-1} \sum_{i\in r} w_i \frac{1}{\widehat{p}_i} y_i.$$

To quantify the nonresponse error, we considered as target mean the estimator under complete response computed on s

$$\widehat{\overline{Y}}_w = \left(\sum_{j\in s} w_j\right)^{-1} \sum_{i\in s} w_i y_i. \tag{5.11}$$

We then estimated the total with the estimators presented in Section 5.6 along with their associated variance estimates, here generically denoted by $\widehat{\overline{Y}}$ and $\widehat{V}(\widehat{\overline{Y}})$. For each estimator, we computed

- the relative error (in%): RE $= 100 \cdot (\widehat{\overline{Y}} - \widehat{\overline{Y}}_w)/\widehat{\overline{Y}}_w$,
- the variance estimator given in Expression (5.10): $\widehat{V}(\widehat{\overline{Y}})$,
- the resulting 95% confidence interval: CI $= \widehat{\overline{Y}} \pm 1.96\sqrt{\widehat{V}(\widehat{\overline{Y}})}$, and
- its length: L $= 2 \cdot 1.96\sqrt{\widehat{V}(\widehat{\overline{Y}})}$.

The results are presented in Table 5.4, where $\widehat{\overline{Y}}_{ptrue}$ is the estimator using the true value of y (known here) instead of the imputed one in Model (5.7). The model parameters were estimated based on the entire data, without using the components. Note that the mean square error of the estimators was not displayed in Table 5.4, because the bias was estimated only with respect to the second phase, while the variance estimator accounts for both phases.

In the SILC framework, all the considered estimators yield similar results in terms of variance. Estimators $\widehat{\overline{Y}}_{3a}$ and $\widehat{\overline{Y}}_{2a}$ which consider the hidden structure of the variable of interest perform the best in terms of bias. All the other estimators (except for $\widehat{\overline{Y}}_{ptrue}$) overestimate or underestimate by at least 4.5% the mean income with respect to the estimate under complete response in (5.11). Finally, the mean income estimate under complete response (see (5.11)) equals

Tab. 5.4.: Results on SILC 2009 data.

Estimator	RE	$\widehat{V}(\widehat{\overline{Y}})$ $(\times 10^4)$	95% CI $(\times 10^3)$	L $(\times 10^3)$
$\widehat{\overline{Y}}_1$	5.32	30.81	[66.97;69.15]	2.18
$\widehat{\overline{Y}}_{1a}$	5.42	30.81	[67.04;69.22]	2.18
$\widehat{\overline{Y}}_2$	4.79	30.80	[66.63;68.81]	2.18
$\widehat{\overline{Y}}_{2a}$	3.53	30.79	[65.82;67.99]	2.17
$\widehat{\overline{Y}}_{2aa}$	4.61	30.79	[66.52;68.69]	2.18
$\widehat{\overline{Y}}_3$	−4.50	31.06	[60.62;62.81]	2.18
$\widehat{\overline{Y}}_{3a}$	−3.27	31.01	[61.42;63.60]	2.18
$\widehat{\overline{Y}}_4$	6.60	30.85	[67.80;69.98]	2.18
$\widehat{\overline{Y}}_{4a}$	6.68	30.84	[67.85;70.03]	2.18
$\widehat{\overline{Y}}_{ptrue}$	1.27	30.80	[64.36;66.53]	2.18

$64.63 \cdot 10^3$, which lies outside each of the confidence intervals constructed, except for the confidence interval associated with $\widehat{\overline{Y}}_{ptrue}$ (in practice unavailable since some y values are missing). Because the true population mean is unknown, it is however not possible to check if it falls within the different confidence intervals.

5.8 Conclusion

We have proposed three NWA procedures for handling nonignorable nonresponse when the values of the variable of interest are i.i.d. draws from a mixture distribution. Our procedures take into account the latent structure of y in the estimation of the response probabilities. While not always true, this tends to reduce the nonresponse bias and the variance of the two-phase estimator when the superpopulation model is correctly specified. A variance estimator for the mean associated with our procedures was also described. The proposed estimators together with their associated variance estimators were studied through simulations and applied to real data.

The proposed procedures take into account a superpopulation model and a nonresponse model. The superpopulation model corresponds to the mixture distribution of y. The goal of using such a model is to create model-clusters or groups of respondents homogeneous with respect to the y distribution. When faced with nonignorable nonresponse, the response probabilities are intrinsically related to the structure of y. Accordingly, the proposed procedures estimate p_i as an outcome of the latent structure of y.

Our proposal builds on existing solutions and adds on latent information on

the variables of interest to the models for the response probabilities. Based on the idea that the model for the variable of interest may vary from one component to the other, the first proposed solution consists of fitting Model (5.6) independently in each reconstructed component and to estimate the response probabilities via maximum likelihood. One expects to reduce the bias and the variance of the resulting estimator compared to the case in which Model (5.6) is fitted globally. This does not seem to happen in the case where x and y are highly correlated as in Setting 3. In such a case, incorporating the components membership indicators in the response model may result in overfitting and in a larger nonresponse bias compared to the case in which only x is used. However, in general, a very large correlation between x and y is infrequent.

When Model (5.7) is considered to estimate the response probabilities, we propose to impute the missing values of the variable of interest independently in each reconstructed component. Two options were used to estimate the response probabilities with the imputed variable of interest y^*: an estimation within the reconstructed components and an estimation on the entire data. In the three simulation settings, the first option produces a smaller variance, while the second one a smaller bias. When the superpopulation model holds (Settings 1 and 2), the two resulting estimators perform better that the estimator with the response probabilities estimated via Model (5.7) where the variable of interest is imputed globally (when the components are not considered).

The third solution consists of estimating the response probabilities via generalized calibration within each component. While the resulting estimator performed the best on real data, we have seen that incorporating the component when the response probabilities are estimated via generalized calibration does not necessarily decrease the bias and the variance of the two-phase estimator. Also, we have seen that the estimator using the response probabilities estimated via generalized calibration (without the components) performs very well in most cases, even though it does not include the component structure of the variable of interest. This simple estimator challenges our estimators which incorporate the components in the model for the response probabilities. However, generalized calibration sometimes fails to estimate the response probabilities since the existence of a solution to the calibration equation is not guaranteed. When this is the case, the response probabilities have to be estimated via maximum likelihood and our estimators (solutions 1a and 2a) perform globally better than the other estimators (solutions 1 and 2). Moreover, in the three simulation settings, the calibration variables and the nonresponse mechanism were uncorrelated conditional to the response model variables, which explains the very small bias associated with the generalized calibration estimator (see last paragraph of Section 5.3). This estimator would yield a weaker performance in a setting where the calibration variables and

the nonresponse mechanism are correlated conditional to the response model variables.

Finally, the proposed procedures model the response probabilities as an outcome of the latent component structure of y. Because they are latent, the components need to be reconstructed prior to estimating the response probabilities. The procedure applied for this aim is based on the strong assumption that the sample inherits of the structure of y within the population and that the respondents set inherits of the structure of y within the sample. A departure from this assumption reduces the performance of the proposed estimators. Compared to the results shown in this paper, our estimators would perform better if we applied a procedure to reconstruct the components that requires a weaker assumption. This is still an unsolved problem.

Acknowledgements

This research was partially supported by the Swiss National Science Foundation, project number P1NEP2_151904. We would like to thank Eric Graf for his help provided in understanding the 2009 SILC data structure and Pr. Yves Tillé for his constructive comments.

General conclusion

In this document, the problem of nonresponse was addressed from a data processing and estimation point of view. We did not discuss either methods to reduce nonresponse or follow-up procedures. Rather, we put ourselves in the skin of a statistician who receives a sample survey data file containing missing values with the task of producing point and variance estimation. Two imputation methods, three reweighting procedures, and one sampling algorithm were presented. In what follows, the limitations of these procedures are briefly commented, further improvements are proposed, and directions for future research are outlined.

Chapter 2 is a reprint of Hasler and Tillé (2014) and presents a fast algorithm for stratified balanced sampling. This algorithm makes it possible to sample from highly stratified populations. Developed for the purpose of the imputation method presented in Chapter 3, this algorithm turns out to be valuable for many other applications, such as some large-scale surveys. In this chapter, we propose a variance estimator for the total and we illustrate one of the possible applications of the proposed algorithm to handle nonresponse.

Chapter 3 is devoted to the first imputation method proposed in this document. This random donor imputation method is based on three imputation models and takes advantage of each of these to provide protection against model misspecification. Limitations and possible improvements of this method as well as avenues for further research are mentioned in what follows. First, because the proposed method is computationally very intensive, its scope of application is limited to standard data of official statistics. It is, in the current form, inapplicable to high-dimensional data. Second, this chapter proposes a variance estimation of the total estimator which is restricted to imputation variance. The other two terms that appear in the variance, sampling variance and nonresponse variance, are not estimated and inference for the total is unavailable. A procedure to estimate the total variance needs to be developed for the new method to appeal to survey practitioners. Third, the proposed method is suitable for handling nonresponse when a single variable of interest is present in the survey, which represents a restrictive framework. Most of the

surveys include several variables of interest, each of which containing a possibly different nonresponse pattern. An adaptation of the proposed method to such a framework will increase its attractiveness and represents an interesting avenue to explore. Finally, the formula proposed in Section 3.5 underestimates the imputation variance of the total because it does not account for the variance due to the rounding problem (see also Section 2.2). Estimation of this part of the variance when balanced sampling is applied is an unsolved problem and represents an important line for further research.

Chapter 4 presents the second imputation method proposed in this document. The main interest of this deterministic predicted value imputation method lies in the flexible building of the imputation model, which provides protection against model misspecification. The absence of an analytical expression for the imputation error represents Achilles heel of this method. A consequence is that the theoretical properties of the total estimator, such as bias and variance, are unavailable whether the assumptions embodied by the imputation model are verified or not. So is also the case for the asymptotical properties. In particular, we could not propose a variance estimator for the total along with its theoretical properties, reason why we pursued a bootstrap-based approach. Finally, as it is the case for the imputation method presented in Chapter 3, this second imputation method is suitable for handling nonresponse when a single variable of interest is present in the survey. An adaptation of the method to allow for handling nonresponse in a survey with several variables of interest represents an interesting avenue for further research.

Finally, Chapter 5 proposes three reweighting procedures to handle the more complicated type of nonresponse mechanism: nonignorable nonresponse. The response probabilities modeling approach is applied. The response probabilities are viewed as an outcome of the latent component structure of the variable of interest. Three general features of the procedures are noted here. First, because the structure of the variable of interest is latent, the components need to be reconstructed prior to estimating the response probabilities. The procedure that we suggest for this aim is based on the strong assumption that the sample inherits of the structure of the variable of interest within the population and that the respondents set inherits of this structure within the sample. A departure from this assumption reduces the performance of the estimators. Second, the proposed procedure is based on a fully parametric model for the response probabilities. If the user does not have a prior knowledge of the appropriateness of this model, he will prefer a more flexible procedure. Last, even though a single variable of interest is considered in this chapter, the proposed reweighting procedures allow to handle unit nonresponse when several variables of interest are present in the survey.

Despite the constantly growing material on nonresponse in sample surveys, many practical problems remain unsolved and are sparsely studied. Without claiming to have provided solutions to some of these unsolved problems in any way, I hope that our work will at least contribute to the progresses in sample surveys theory by bringing new perspectives and by raising new questions.

Proof of Property 2.1

A

Proof.

(i) Proof by induction.

(a) For $j = 2$ in step 2 of Algorithm 2.2, two strata are considered. Therefore $q + 2$ balancing variables are used in the flight phase. Thus

$$\#\left\{k \in U_1 \cup U_2 \,\middle|\, 0 < \phi_k^{(2)} < 1\right\} \leq q + 2 \leq 2q + 2.$$

The result is valid for $j = 2$.

(b) Assume that the result is valid for $j = \ell$, i.e. assume that

$$\#\left\{k \in \bigcup_{i=1}^{\ell} U_i \,\middle|\, 0 < \phi_k^{(\ell)} < 1\right\} \leq 2q + 2.$$

As it is impossible to have a rounding problem for a single unit of a stratum (the sum of the inclusion probabilities is an integer in each stratum), the number of strata containing units such that $0 < \phi_k^{(\ell)} < 1$ is at most $q + 1$. Then a strata is added for the flight phase for step $j = \ell + 1$. Therefore, at most $q + 2$ strata are considered, which means that at most $q + 2$ balancing variables of the type of $\phi_k^{(j-1)} \mathbb{1}(k \in U_i)$ are required. Moreover, the q balancing variables

$$\frac{\phi_k^{(j-1)} \mathbf{x}_k^\top}{\pi_k}$$

are considered. In total, at most $2q + 2$ balancing variables are used in the flight phase for $j = \ell + 1$. This implies that

$$\#\left\{k \in \bigcup_{i=1}^{\ell+1} U_i \,\middle|\, 0 < \phi_k^{(\ell+1)} < 1\right\} \leq 2q + 2,$$

which means that the result is true for $j = \ell + 1$.

(ii) $\mathbf{Z}^{(j)}$ represents the matrix whose columns are the balancing variables used in the j-th flight phase of step 2 of Algorithm 2.2. In the previous point, it is shown that at most $2q + 2$ units are considered in each flight phase of step 2. It has also been explained that, in total, at most $2q + 2$ balancing variables are used in each of these flight phases. It results that the number of non-null columns of matrix $\mathbf{Z}^{(j)}$ is less than or equal to $2q + 2$. □

B Proofs of the properties of Chapter 3

Proof of Property 3.1.

1.

$$\begin{aligned}\mathrm{E}_m\mathrm{E}_I\left(\widehat{Y}_I-\widehat{Y}\right)&=\mathrm{E}_m\left(\sum_{i\in S_r}d_iy_i+\sum_{j\in S_m}d_j\sum_{i\in S_r}\psi_{ij}^{[bk]}y_i-\sum_{i\in S}d_iy_i\right)\\&=\left(\sum_{i\in S_r}d_i\boldsymbol{\beta}^\top\mathbf{x}_i+\sum_{j\in S_m}d_j\sum_{i\in S_r}\psi_{ij}^{[bk]}\boldsymbol{\beta}^\top\mathbf{x}_i-\sum_{i\in S}d_i\boldsymbol{\beta}^\top\mathbf{x}_i\right)\\&=\boldsymbol{\beta}^\top\left[\mathrm{E}_I\left(\widehat{\mathbf{X}}_I\right)-\widehat{\mathbf{X}}\right]=0,\end{aligned}$$

where the last equality comes from Equation (3.11). As it is supposed that the data is MAR, the expectation with respect to m, the one with respect to p, and the one with respect to q can be reversed. It therefore produces

$$\begin{aligned}\mathrm{Bias}\left(\widehat{Y}_I\right)&=\mathrm{E}_m\mathrm{E}_p\mathrm{E}_q\mathrm{E}_I\left(\widehat{Y}_I-Y\right)=\mathrm{E}_m\mathrm{E}_p\mathrm{E}_q\mathrm{E}_I\left(\widehat{Y}_I-\widehat{Y}+\widehat{Y}-Y\right)\\&=\mathrm{E}_m\mathrm{E}_p\mathrm{E}_q\mathrm{E}_I\left(\widehat{Y}_I-\widehat{Y}\right)=\mathrm{E}_p\mathrm{E}_q\mathrm{E}_m\mathrm{E}_I\left(\widehat{Y}_I-\widehat{Y}\right)=0.\end{aligned}$$

2. If $y_i=\boldsymbol{\beta}^\top\mathbf{x}_i$,

$$\widehat{Y}_I-\mathrm{E}_I\left(\widehat{Y}_I\right)=\boldsymbol{\beta}^\top\widehat{\mathbf{X}}_I-\mathrm{E}_I\left(\boldsymbol{\beta}^\top\widehat{\mathbf{X}}_I\right)=\boldsymbol{\beta}^\top\left[\widehat{\mathbf{X}}_I-\mathrm{E}_I\left(\widehat{\mathbf{X}}_I\right)\right]\approx 0,$$

where the last approximation comes from Remark 3.3 (page 56). Therefore

$$\mathrm{Var}_I\left(\widehat{Y}_I\right)=\mathrm{E}_I\left[\widehat{Y}_I-\mathrm{E}_I\left(\widehat{Y}_I\right)\right]^2\approx 0,$$

and

$$\mathrm{Var}_{imp}=\mathrm{E}_p\mathrm{E}_q\mathrm{Var}_I\left(\widehat{Y}_I\right)\approx 0.$$

□

Proof of Property 3.2.

$$\begin{aligned}\mathrm{E}_I\left(\widehat{Y}_I\right) &= \sum_{i\in S_r} d_i y_i + \sum_{j\in S_m} d_j \sum_{i\in S_r} \psi_{ij}^{[bk]} y_i = \sum_{i\in S_r} d_i \left(1 + \sum_{j\in S_m} \frac{d_j}{d_i}\psi_{ij}^{[bk]}\right) y_i \\ &= \sum_{i\in S_r} d_i \frac{1}{\theta_i} y_i.\end{aligned}$$

If θ_i is the true response probability, this last expression represents the propensity score adjusted estimator. Therefore

$$\mathrm{E}_p\mathrm{E}_q\mathrm{E}_I\left(\widehat{Y}_I\right) = \mathrm{E}_p\mathrm{E}_q\left(\sum_{i\in S_r} d_i \frac{1}{\theta_i} y_i\right) = Y,$$

and consequently

$$\mathrm{Bias}\left(\widehat{Y}_I\right) = \mathrm{E}_p\mathrm{E}_q\mathrm{E}_I\left(\widehat{Y}_I - Y\right) = 0.$$

□

Proof of Property 3.3.
We have

$$\begin{aligned}\mathrm{E}_I\left(\widehat{Y}_I - \widehat{Y}\right) &= \sum_{j\in S_m} d_j \sum_{i\in S_r} \psi_{ij}^{[bk]} y_i - \sum_{j\in S_m} d_j y_j \\ &= \sum_{j\in S_m} d_j \sum_{i\in S_r} \psi_{ij}^{[bk]} y_i - \sum_{j\in S_m} d_j \sum_{i\in S_r} \psi_{ij}^{[bk]} y_j \\ &= \sum_{j\in S_m} d_j \sum_{i\in S_r} \psi_{ij}^{[bk]} (y_i - y_j) \\ &= 0,\end{aligned}$$

where the last equality comes from the fact that $\psi_{ij}^{[bk]}$ is nonzero only if $i \in knn(j)$ and from the hypothesis of the property. Therefore, it produces

$$\begin{aligned}\mathrm{Bias}\left(\widehat{Y}_I\right) &= \mathrm{E}_p\mathrm{E}_q\mathrm{E}_I\left(\widehat{Y}_I - Y\right) = \mathrm{E}_p\mathrm{E}_q\mathrm{E}_I\left(\widehat{Y}_I - \widehat{Y} + \widehat{Y} - Y\right) \\ &= \mathrm{E}_p\mathrm{E}_q\mathrm{E}_I\left(\widehat{Y}_I - \widehat{Y}\right) = 0.\end{aligned}$$

□

Proof of Proposition 3.1

C

The proof of Proposition 3.1 requires the following four lemmas.

Lemma 1 *Suppose that assumptions (A1) and (A2) hold. Then* $\frac{\widehat{Y}-Y}{N}$ *converges in probability to 0.*

Proof. As $\widehat{Y}$ is a design unbiased estimator of Y, we have

$$\mathrm{E}_p\left(\frac{\widehat{Y}-Y}{N}\right)=0.$$

Moreover, we have

$$\begin{aligned}\mathrm{Var}_p\left(\frac{\widehat{Y}-Y}{N}\right)&=\frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}y_iy_j\\&\leq\frac{1}{N^2}\sum_{i\in U}\sum_{\substack{j\in U\\j\neq i}}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}y_iy_j+\frac{1}{N^2}\sum_{i\in U}\frac{1}{\pi_i}y_i^2\\&=O\left(\frac{1}{n}\right),\end{aligned}$$

where the last equality follows from assumptions (A1) and (A2). By Bienayme-Chebychev inequality, we conclude that $\frac{\widehat{Y}-Y}{N}$ converges in probability to 0. □

Lemma 2 *Suppose that assumptions (A2) and (A6) hold. Then*

$$\mathrm{Var}_I\left(\frac{\widehat{Y}_I-\widehat{Y}}{N}\right)=O\left(\frac{1}{n}\right).$$

Proof. Assumption (A6) implies

$$\begin{aligned}\mathrm{Var}_I\left(\frac{\widehat{Y}_I-\widehat{Y}}{N}\right)&=\frac{1}{N^2}\mathrm{Var}_I^{app}\left(\widehat{Y}_I\right)\\&=\frac{1}{N^2}\sum_{i\in S_r}\sum_{\substack{j\in S_m\\\psi_{ij}^{[bk]}\neq 0}}c_{ij}d_j^2\left(y_i-\mathbf{b}^\top\mathbf{x}_i\right)^2\\&\leq\frac{1}{N^2}\sum_{j\in S_m}\sum_{\substack{i\in S_r\\\psi_{ij}^{[bk]}\neq 0}}\psi_{ij}^{[bk]}\frac{n_mk}{n_mk-q}d_j^2\left(y_i-\mathbf{b}^\top\mathbf{x}_i\right)^2\\&=O\left(\frac{1}{n}\right)\end{aligned}$$

where the last inequality comes from assumption (A2), from $y_i-\mathbf{b}^\top\mathbf{x}_i=O(1)$, and from $\psi_{ij}^{[bk]}=O\left(\frac{1}{k}\right)$. □

Lemma 3 *Suppose that assumptions (A2), (A4), and (A7) hold. Then*

$$\mathrm{Var}_m \mathrm{E}_I \left(\frac{\widehat{Y}_I - \widehat{Y}}{N} \right) = O\left(\frac{1}{n}\right).$$

Proof. Using $\mathrm{E}_I\left(\psi_{ij}^{[bk]}\right) = \phi_{ij}^{[bk]}$ and assumption (A4), we get

$$\begin{aligned}
\mathrm{E}_I \left(\frac{\widehat{Y}_I - \widehat{Y}}{N} \right) &= \frac{1}{N} \left(\sum_{j \in S_m} d_j \sum_{i \in S_r} \psi_{ij}^{[bk]} y_i - \sum_{j \in S_m} d_j y_j \right) \\
&= \frac{1}{N} \left[\sum_{j \in S_m} d_j \sum_{i \in S_r} \psi_{ij}^{[bk]} (\boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i) - \sum_{j \in S_m} d_j (\boldsymbol{\beta}^\top \mathbf{x}_j + \varepsilon_j) \right].
\end{aligned}$$

Therefore, from assumption (A4) again, we obtain

$$\begin{aligned}
&\mathrm{Var}_m \mathrm{E}_I \left(\frac{\widehat{Y}_I - \widehat{Y}}{N} \right) \\
&\quad = \frac{1}{N^2} \left[\sum_{i \in S_r} \left(\sum_{j \in S_m} d_j \psi_{ij}^{[bk]} \right)^2 \mathrm{Var}_m(\varepsilon_i) + \sum_{j \in S_m} d_j^2 \mathrm{Var}_m(\varepsilon_j) \right] \\
&\quad = \frac{1}{N^2} \sigma^2 \left[\sum_{i \in S_r} \left(\sum_{j \in S_m} d_j \psi_{ij}^{[bk]} \right)^2 + \sum_{j \in S_m} d_j^2 \right] \\
&\quad = \frac{1}{N^2} \sigma^2 \left[\sum_{i \in S_r} \sum_{j \in S_m} d_j^2 {\psi_{ij}^{[bk]}}^2 + \sum_{i \in S_r} \sum_{j \in S_m} \sum_{\substack{\ell \in S_m \\ \ell \neq j}} d_j d_\ell \psi_{ij}^{[bk]} \psi_{i\ell}^{[bk]} + \sum_{j \in S_m} d_j^2 \right] \\
&\quad = \frac{1}{N^2} \sigma^2 \left[\sum_{j \in S_m} d_j^2 \sum_{i \in S_r} {\psi_{ij}^{[bk]}}^2 + \sum_{j \in S_m} \sum_{\substack{\ell \in S_m \\ \ell \neq j}} d_j d_\ell \sum_{i \in S_r} \psi_{ij}^{[bk]} \psi_{i\ell}^{[bk]} + \sum_{j \in S_m} d_j^2 \right] \\
&\quad = O\left(\frac{1}{n}\right)
\end{aligned}$$

where the last equality follows from assumption (A2), from assumption (A7), and from $\psi_{ij}^{[bk]} = O\left(\frac{1}{k}\right)$. □

Lemma 4 *Suppose that assumptions (A2) to (A7) hold. Then* $\frac{\widehat{Y}_I - \widehat{Y}}{N}$ *converges in probability to 0.*

Proof. Assumption (A4) implies that (see proof of Property 3.1)

$$\mathrm{E}_m \mathrm{E}_I \left(\frac{\widehat{Y}_I - \widehat{Y}}{N} \right) = 0. \tag{C.1}$$

Then, by assumption (A3), we get

$$\mathrm{E}_{mpqI}\left(\frac{\widehat{Y}_I-\widehat{Y}}{N}\right)=\mathrm{E}_p\mathrm{E}_q\mathrm{E}_m\mathrm{E}_I\left(\frac{\widehat{Y}_I-\widehat{Y}}{N}\right)=0.$$

Moreover, from Lemma 2 and Lemma 3, we have

$$\mathrm{Var}_m\mathrm{E}_I\left(\frac{\widehat{Y}_I-\widehat{Y}}{N}\right)+\mathrm{E}_m\mathrm{Var}_I\left(\frac{\widehat{Y}_I-\widehat{Y}}{N}\right)=O\left(\frac{1}{n}\right). \tag{C.2}$$

Assumption (A3), Equation (C.1) and Equation (C.2) together imply

$$\mathrm{Var}_{mpqI}\left(\frac{\widehat{Y}_I-\widehat{Y}}{N}\right)=\mathrm{Var}_{pqmI}\left(\frac{\widehat{Y}_I-\widehat{Y}}{N}\right)=O\left(\frac{1}{n}\right).$$

By Bienayme-Chebychev inequality, we conclude that $\frac{\widehat{Y}_I-\widehat{Y}}{N}$ converges in probability to 0. □

Proof of Proposition 3.1.
The conclusion follows directly from equality

$$\frac{\widehat{Y}_I-Y}{N}=\frac{\widehat{Y}_I-\widehat{Y}}{N}+\frac{\widehat{Y}-Y}{N},$$

Lemma 1, and Lemma 4. □

Bibliography

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

Arya, S., Mount, D., Kemp, S. E., and Jefferis, G. (2014). *RANN: Fast Nearest Neighbour Search (wraps Arya and Mount's ANN library)*. R package version 2.4.1.

Baillargeon, S. and Rivest, L.-P. (2011). *stratification: Univariate Stratification of Survey Populations*. R package version 2.0-3.

Beaumont, J.-F. (2000). An estimation method for nonignorable nonresponse. *Survey Methodology*, 26(2):131–136.

Beaumont, J.-F., Haziza, D., and Bocci, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21(2):515–537.

Bellman, R. (1961). *Adaptive control processes: A guided tour*. Princeton University Press.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Brewer, K. R. W. (1999). Design-based or prediction-based inference? Stratified random vs stratified balanced sampling. *International Statistical Review*, 67:35–47.

Brick, J. M., Kalton, G., and Kim, J. K. (2004). Variance estimation with hot-deck imputation using a model. *Survey Methodology*, 30(1):57–66.

Bülher, W. and Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. *Metrika*, 22:161–175.

Cassel, C. M., Särndal, C.-E., and Wretman, J. H. (1983). Some uses of statistical models in connexion with the nonresponse problem. In Madow, W. G. and Olkin, I., editors, *Incomplete Data in Sample Surveys*, volume 3, pages 143–160. Academic Press, New York.

Central Statistical Office (1993). Family expenditure survey, 1992 [computer file]. Technical report, Colchester, Essex: UK Data Archive [distributor]. SN: 3064, http://dx.doi.org/10.5255/UKDA-SN-3064-1.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95:555–571.

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, 35:115–119.

Chauvet, G., Bonnéry, D., and Deville, J. C. (2011a). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141(2):984–994.

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

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

Chen, H. L., Rao, J. N. K., and Sitter, R. R. (2000). Efficient random imputation for missing survey data in complex survey. *Statistica Sinica*, 10:1153–1169.

Chen, H. L. and Shao, J. (1997). Biases and variances of survey estimators based on nearest neighbor imputation. In *Proceedings of the Section on Survey Research*, pages 365–369, American Statistical Association.

Chen, H. L. and Shao, J. (2000). Nearest-neighbour imputation for survey data. *Journal of Official Statistics*, 16:113–131.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96:260–269.

Da Silva, D. N. and Opsomer, J. D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *The Canadian Journal of Statistics*, 34(4):563–579.

Da Silva, D. N. and Opsomer, J. D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology*, 35(2):165–176.

Dahl, F. A. (2007). Convergence of random k-nearest-neighbour imputation. *Computational Statistics & Data Analysis*, 51:5913–5917.

Dalenius, T. and Hodges, J. L. J. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54:88–101.

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

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. In *Compstat - Proceedings in Computational Statistics: 14th Symposium Held in Utrecht, The Netherlands*, pages 65–76, New York. Springer.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. In *Actes des Journées de Méthodologie Statistique*, Paris. Insee-Méthodes.

Deville, J.-C., Grosbras, J.-M., and Roth, N. (1988). Efficient sampling algorithms and balanced sample. In *COMPSTAT, Proceedings in Computational Statistics*, pages 255–266, Heidelberg. Physica Verlag.

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

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

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

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

Díaz-García, J. A. and Garay-Tápia, M. M. (2007). Optimum allocation in stratified surveys: Stochastic programming. *Computational Statistics & Data Analysis*, 51(6):3016–3026.

Eltinge, J. L. and Yansaneh, I. S. (1997). Diagnostics for formation of nonresponse adjustment aells, with an application to ancome nonresponse in the U. S. consumer expenditure survey. *Survey Methodology*, 23(1):33–40.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing (Second Edition)*. Marcel Dekker, New York.

Fang, F., Hong, Q., and Shao, J. (2010). Empirical likelihood estimation for samples with nonignorable nonresponse. *Statistica Sinica*, 20:263–280.

Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81(394):354–365.

Fay, R. E. (1991). A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference*, pages 429–440. U.S. Census Bureau.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91:490–498.

Fellegi, P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71(353):17–35.

Flachaire, E. and Nuñez, O. (2007). Estimation of the income distribution and detection of subpopulations: An explanatory model. *Computational Statistics & Data Analysis*, 51:3368–3380.

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

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

Giommi, A. (1987). Nonparametric methods for estimating individual response probabilities. *Survey Methodology*, 13(2):127–134.

Graf, E. (2014). *Imputation of income variables in a survey context and estimation of variance for indicators of poverty and social exclusion*. PhD thesis, University of Neuchâtel, Switzerland.

Grafström, A., Matei, A., Qualité, L., and Tillé, Y. (2012). Size constrained unequal probability sampling with a non-integer sum of inclusion probabilities. *Electronic Journal of Statistics*, 6:1477–1489.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. Chapman & Hall/CRC, Boca Raton.

Greenlees, J. S., Reece, W. S., and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77(378):251 – 261.

Gross, S. T. (1980). Mean estimation in sample surveys. In *Proceedings of the Survey Research Methods Section*, pages 181–184. American Statistical Association.

Hasler, C. and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics & Data Analysis*, 74:81–94.

Hastie, T. J. and Tibshirani, R. J. (1986). Generalized additive models. *Statistical Science*, 1(3):297–318.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive Models*. Chapman & Hall/CRC, Boca Raton.

Haziza, D. and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75(1):25–43.

Haziza, D. and Lesage, E. (2015). A discussion of weighting procedures for unit nonresponse. To appear in *Journal of Official Statistics*.

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

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

Ishwaran, H. and Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12:941–963.

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

Kalton, G. and Kish, L. (1981). Two efficient random imputation procedures. In *Proceedings of the Section on Survey Research Methods*, pages 146–153. American Statistical Association.

Kalton, G. and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, A13:1919–1939.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29:119–127.

Kim, J. K. (2001). Variance estimation after imputation. *Survey Methodology*, 27(1):75–83.

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

Kim, J. K. and Kim, J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 35(4):501–514.

Kim, J. K. and Park, H. (2006). Imputation using response probability. *Canadian Journal of Statistics*, 34(1):171–182.

Kott, P. (2012). Why one should incorporate the design weights when adjusting for unit nonresponse using response homogeneity groups. *Survey Methodology*, 38(1):95–99.

Kott, P. S. (1994). A note on handling nonresponse in surveys. *Journal of the American Statistical Association*, 89(426):693–696.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133–142.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491):1265–1275.

Laaksonen, S. and Chambers, R. (2006). Survey estimation under informative nonresponse with follow-up. *Journal of Official Statistics*, 22(1):81–95.

Lavallée, P. and Hidiroglou, M. A. (1988). On the stratification of skewed populations, sur la stratification de populations asymétriques. *Survey Methodology, Techniques d'enquête*, 14:35–45.

Lee, H., Rancourt, E., and Särndal, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal Official Statistics*, 10(3):231–243.

Lesage, E. and Haziza, D. (2015). On the problem of bias and variance amplification of the instrumental calibration estimator in the presence of unit nonresponse. Working paper.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77(378):237–250.

Little, R. J. A. (1983). The nonignorable case. In Madow, W. G., Olkin, I., and Rubin, D. B., editors, *Incomplete Data in Sample Surveys*, volume 2, pages 383–413.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimated of means. *International Statistical Review*, 54(2):139–157.

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

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

McLachaln, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

Nedyalkova, D. and Tillé, Y. (2008). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95:521–537.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.

Niyonsenga, T. (1994). Nonparametric estimation of response probabilities in sampling theory. *Survey Methodology*, 20(2):177–184.

Niyonsenga, T. (1997). Response probability estimation. *Journal of Statistical Planning and Inference*, 59:111–126.

Rancourt, E., Särndal, C.-E., and Lee, H. (1994). Estimation in the presence of nearest neighbour imputation. In *Proceedings of the Section on Survey Research Methods*, pages 888–893. American Statistical Association.

Rao, J. N. K. (1990). Variance estimation under imputation for missing data. Technical report, Statistics Canada, Ottawa.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434):499–506.

Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, 79:811–822.

Rao, J. N. K. and Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82(2):453–460.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonreponse in Surveys*. Wiley, New York.

Sande, I. G. (1981a). Imputation in survey: coping with reality. *Survey Methodology*, 7(1):21–43.

Sande, I. G. (1981b). Imputation in surveys: coping with reality. *The American Statistician*, 36(3):145–152.

Särndal, C.-E. (1990). Methods for estimating the precision of survey estimates when imputation has been used. In *Proceedings of Statistics Canada's Symposium '90: measurement and Improvement of data quality*.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18(2):241–252.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2):99–119.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in surveys with nonresponse*. Wiley, New York.

Särndal, C.-E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55(3):279–294.

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

Shao, J. (2000). Cold deck and ratio imputation. *Survey Methodology*, 26(1):79–85.

Shao, J. (2009). Nonparametric variance estimation for nearest neighbor imputation. *Journal of Official Statistics*, 25(1):55–62.

Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91:1278–1288.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonneglible sampling fractions. *Journal of the American Statistical Association*, 94:254–265.

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

Sitter, R. R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87(416):755–765.

Stones, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705.

Tillé, Y. and Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74:31–37.

Tillé, Y. and Matei, A. (2007). *The R Package Sampling*. The Comprehensive R Archive Network, http://cran. r-project. org/, Manual of the Contributed Packages.

Vartivarian, S. and Little, R. J. (2002). On the formation of weighting adjustment cells for unit nonresponse. In *Proceedings of the Joint Statistical Meetings - Section on Survey Research Methods*, pages 3553–3558. American Statistical Association.

Wang, Y. (2011). *Smoothing splines: methods and applications*. Chapman & Hall/CRC, Boca Raton.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer, New York.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Serie B (Statistical Methodology)*, 65(1):95–114.

Wood, S. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society. Serie B (Statistical Methodology)*, 70(3):495–518.

Wood, S. (2014). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation*. R package version 1.7-28, http://CRAN.R-project.org/package=mgcv.

Wright, T. (2012). The equivalence of neyman optimum allocation for sampling and equal proportions for apportioning the U.S. house of representatives. *The American Statistician*, 66, No. 4:217–224.

Zeileis, A. (2013). *ineq: Measuring Inequality, Concentration, and Poverty*. R package version 0.2-11.

Zhang, G., Christensen, F., and Zheng, W. (2013). Nonparametric regression estimators in complex surveys. *Journal of Statistical Computation and Simulation*, 85(5):1026–1034.