Université de Neuchâtel

Handling auxiliary variables in survey sampling and nonresponse

by

Audrey-Anne Vallée

Institut de statistique
Faculté des sciences

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

Accepted by the examination committee:

Prof. Pascal Felber	Université de Neuchâtel	Jury president
Prof. Yves Tillé	Université de Neuchâtel	Thesis directeur
Prof. Patrice Bertail	Université Paris Nanterre	Jury member
Prof. Jean Opsomer	Colorado State University and Westat	Jury member
Prof. Maria Giovanna Ranalli	Università degli Studi di Perugia	Jury member

Thesis defended on April 5, 2019

UNIVERSITÉ DE NEUCHÂTEL

Faculté des Sciences
Secrétariat-décanat de Faculté
Rue Emile-Argand 11
2000 Neuchâtel – Suisse
Tél : + 41 (0)32 718 21 00
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 Audrey-Anne VALLÉE

Titre:

"Handling auxiliary variables in survey sampling and nonresponse"

sur le rapport des membres du jury composé comme suit:

- Prof. Yves Tillé, directeur de thèse, Université de Neuchâtel, Suisse
- Prof. Pascal Felber, Université de Neuchâtel, Suisse
- Prof. Patrice Bertail, Université Paris-Nanterre, France
- Prof. Giovanna Ranalli, Università degli studi di Perugia, Italie
- Prof. Jean Opsomer, Colorado State University, USA

Neuchâtel, le 11 avril 2019

Le Doyen, Prof. P. Felber

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, le professeur Yves Tillé. Dès mon arrivée en Suisse, il a su se montrer accueillant, disponible et surtout enthousiaste. Yves est un chercheur inspirant doté d'une intuition remarquable qui sait partager ses idées et ses passions. Son soutien dans ma recherche, mes conférences, mes projets, mes conférences, les activités d'équipe, mes conférences m'a beaucoup aidée. J'ai aussi beaucoup appris grâce à ses sages conseils, tel un simple et récurrent rappel qu'il faut faire sa thèse. Yves, j'ai eu beaucoup de plaisir à travailler avec toi et je suis très honorée de t'avoir eu comme directeur.

J'aimerais remercier les membres de mon jury de thèse, le professeur Pascal Felber, le professeur Patrice Bertail, le professeur Jean Opsomer et la professeure Maria Giovanna Ranalli. Je vous suis reconnaissante d'avoir pris le temps d'évaluer ma thèse avec minutie. Je remercie Bastien Ferland-Raymond et le professeur Louis-Paul Rivest avec qui j'ai réalisé une partie de cette thèse. Je remercie aussi David Haziza, Louis-Paul Rivest et Guillaume Chauvet pour le soutien qu'ils m'apportent.

Les années passées à l'Institut de Statistique ont été remplies de bons moments. J'ai beaucoup apprécié côtoyer tous les membres de cette belle équipe. Vous m'avez accompagnée tout au long de mon séjour et je suis très reconnaissante de votre accueil chaleureux. Mon séjour en Suisse m'a aussi permis de me lier d'amitié avec des gens merveilleux. Parmi ceux qui ont agrémenté mes journées, et mes soirées, je tiens à souligner mes sentiments pour Lionel, Matthieu, Denis, Raphaël, Marie-Hélène et mon ado Léo.

Les encouragements et le soutien de mes parents me suivent partout. Je les remercie de m'accompagner aussi chaleureusement dans toutes mes aventures. Merci pour tout. Je salue aussi Alexandre, Ali, Arthur et Cédric avec qui j'ai pu profiter de quelques voyages de ski et de nombreux voyages en train. Je ne peux évidemment pas oublier Marielle, Michel, Michel et les Mérons. Finalement, Julien, je te remercie d'égayer mon quotidien. Merci aussi pour ton continuel soutien.

Summary

This manuscript is dedicated to the use of auxiliary information in survey sampling and nonresponse. We are interested in the integration of auxiliary variables in sampling methods and in the treatment of nonresponse to improve the efficiency and the precision of surveys. We also deal with the calculation of the precision of estimators. Indeed, variances rapidly become difficult to calculate when the estimation methods are sophisticated. The thesis is organized as follows. The first chapter consists in an introduction to some concepts of survey sampling and nonresponse. In the second chapter, we develop a sampling design for a forest inventory in order to satisfy a number of requirements. The sample needs to optimize the work of the ground teams while ensuring the selection of every type of trees. To meet the objectives, stratified balanced sampling is used in a two-stage sample. In the third chapter, we discuss the calculation of the variance when two independent samples intersect. The variance and its estimator can be decomposed conditionally to one sample or conditionally to the other one. In specific situations, as in the nonresponse case, it results in convenient simplifications. The fourth chapter presents a linearization method for the estimation of the variance in the presence of nonresponse. In the fifth chapter, an imputation method for Swiss cheese nonresponse is developed. This imputation method uses stratified balanced sampling.

Keywords: Balanced sampling, calibration techniques, forest inventory, imputation method, Swiss cheese nonresponse, variance estimation.

Résumé

Ce manuscrit est consacré à l'utilisation d'informations auxiliaires en échantillonnage et en non-réponse. Nous nous intéressons à l'intégration de variables auxiliaires dans les méthodes d'échantillonnage et au traitement de la non-réponse afin d'améliorer l'efficacité et la précision des enquêtes. Nous traitons également du calcul de la précision d'estimateurs. En effet, les variances deviennent rapidement difficiles à calculer lorsque les méthodes d'estimation sont sophistiquées. La thèse est organisée comme suit. Le premier chapitre consiste en une introduction à quelques concepts d'échantillonnage et de non-réponse. Dans le deuxième chapitre, nous développons un plan d'échantillonnage pour un inventaire forestier afin de satisfaire un certain nombre d'exigences. L'échantillon doit optimiser le travail des équipes au sol tout en assurant la sélection de tous les types d'arbres. Pour atteindre les objectifs, un plan d'échantillonnage équilibré et stratifié est utilisé dans un échantillon à deux degrés. Dans le troisième chapitre, nous discutons du calcul de la variance dans le cas d'une intersection entre deux échantillons indépendants. La variance et son estimateur peuvent être décomposés conditionnellement à un échantillon ou conditionnellement à l'autre. Dans des situations spécifiques, comme dans le cas de la non-réponse, il en résulte des simplifications bien pratiques. Le quatrième chapitre présente une méthode de linéarisation pour l'estimation de la variance en présence de non-réponse. Dans le cinquième chapitre, une méthode d'imputation pour une non-réponse en fromage suisse est développée. Cette méthode d'imputation utilise un plan d'échantillonnage équilibré et stratifié.

Mots clés: Échantillonnage équilibré, estimation de la variance, inventaire forestier, non-réponse en fromage suisse, techniques de calage.

Contents

List of figures

List of tables

Introduction

Surveys collect important information on all aspects of society. Survey sampling provides a framework for conducting well structured and high quality studies. Statisticians participate in the different stages of those surveys. They develop refined techniques to address various challenges and to correct errors. They use all kinds of available information to improve the elaboration of the studies, the treatment of the data and the quality of the results. A first step of a survey is to create a data frame, which is the list of known units of the population used to select a sample. Auxiliary information in the data frame allows, for instance, to assign unequal probabilities of selection in the sample to the units and to oversample rare and special units. Afterwards, a sampling design can be elaborated to randomly select a sample of units. Sophisticated sampling designs are developed to improve the data collection and minimize errors due to the sampling. Next, sampled units are surveyed and the information is recorded. Typically, the data sets suffer from measurement errors and missing values. A variety of treatments using auxiliary information are elaborated to correct those kinds of problems. Finally, the estimation of the parameters of interest and the calculation of their precision are produced. They must take into account the complexity of the parameters, the sampling design and the treatments of the data.

The first part of this thesis is dedicated to the use of auxiliary information to address multiple challenges in the different stages of survey sampling. In particular, we will see that the design of a survey sample can rapidly become complex because of constraints that are imposed by the data collection and the limitations of the study. However, auxiliary information that is available before conducting the survey can be used to remain accurate while satisfying all kinds of requirements. The rest of the thesis is focussing on survey nonresponse. Indeed, data collection remains a difficult task to accomplish without errors. Measurement errors and missing values are almost inevitable in the recorded data sets. The presence of nonresponse can introduce a bias and an increase in variability of the estimators. It is therefore of common practice to use auxiliary information to impute the missing values, to reweight the responding units and to adapt the estimators in order to produce reliable estimates. We discuss methods to calculate the precision of estimators in

the presence of corrected data sets. We also suggest a treatment whose objective is to improve this precision.

A brief introduction to the concepts and notions of survey sampling and nonresponse is covered in Chapter 1. Chapters 2-4 are self-contained published or accepted papers in peer-reviewed journals. Chapter 5 is also in the form of a self-contained paper. Those articles have all been written with different collaborators.

Chapter 2, a reprint of Vallée et al. (2015), was co-written with Bastien Ferland-Raymond, Professor Louis-Paul Rivest and Professor Yves Tillé. This paper was motivated by a regular forest intentory of the Minister of Forests, Wildlife and Parks of the province of Quebec. The sample of trees had to include various types of trees and to cover the entire region while optimizing the routes of the ground teams. Auxiliary information that was made available by photographies and spectral imagery is incorporated in sophisticated sampling methods to satisfy the objectives. Thus stratified balanced sampling is used to select a controlled number of each type of trees and to improve the accuracy of the produced estimators.

In survey sampling, the calculation of the precision of estimators rapidly become difficult. It must take into account the sampling design, the treatment of the data and the complexity of the estimators. In Chapter 3, a reprint of Tillé and Vallée (2017), we discuss the decomposition of the variance of an estimated total when independent samples intersect. In this case, the variance can be computed conditionally either to one sample or the other one. Depending on the sampling designs, the conditioning can affect the computation of the variance and its estimator. A specific application in this article is survey nonresponse, which can be seen as the intersection of two independent samples.

Furthermore on variance estimation, Chapter 4, a reprint of Vallée and Tillé (2019), presents a linearization method for variance estimation in the presence of nonresponse. When the estimators are complex, it is difficult to compute an explicit variance estimator. In the absence of nonresponse, different linearization methods were proposed to estimate the variance. One of these methods linearizes with respect to the sampling indicators. In this chapter, we extend this method to deal with nonresponse and to produce explicit variance estimators.

In Chapter 5, co-written with Professor Yves Tillé, we propose an imputation method for Swiss cheese nonresponse, when every variable of a survey contains missing values. It is an extension of balanced K-nearest neighbor imputation to treat multivariate nonresponse. The method randomly selects donors to impute the missing values. Those donors are neighboring units of the nonrespondents and they are selected so that balancing equations are satisfied. We will see that stratified balanced sampling is required to implement the imputation method. Chapter 2 can be seen as a preliminary work to this chapter. It prepares to a complex use of the sampling method.

Chapter 1

An Introduction to Survey Sampling and Nonresponse

Abstract: This chapter introduces some concepts and notations of survey sampling and nonresponse. The basics of the framework for this thesis are presented. Section 1.2 covers the complete case, when all the values of the survey sample are observed. The notation is introduced in Section 1.2.1 and some sampling designs are presented in Section 1.2.2. Section 1.3 presents the concept of nonresponse in survey samples. The types of nonresponse are detailed in Section 1.3.1. The notation is presented in Section 1.3.2 and the frameworks for inference are discussed in Section 1.3.3.

Keywords: Imputation; sampling design; reweighting; variance estimation.

1.1. Survey sampling and nonresponse

Surveys are frequently used to investigate different aspects of a population. Typically, a sample of the population is surveyed to collect data and produce estimates. Survey sampling is the framework grouping the techniques used to conduct those studies. Statisticians play an important role in the conception, the execution, the analysis and the interpretation of surveys. In the first place, they participate to the creation of the list of units in the population, namely the data frame. The data frame is the basis on which the inclusion probabilities of the units are determined. From the data frame, the statisticians can determine the sampling design, which is the way to randomly select the sample. Once the data are collected, the statisticians are involved in the treatment of the outputs. Afterwards, they produce the estimation of the parameters of interest and they calculate their precision.

Several aspects can complicate the elaboration of survey samples. In the first place, the conception of the sampling design may be driven by some challenges. For example, the creation of the list of the sampling units can be cumbersome. In this case, it may be reasonable to consider a two-stage sampling design, where groups of sampling units are preliminarily selected and investigated. In other cases, the

sampling units may be spread over a large territory. The sampling design must take that into account to limit the costs and the study duration. Furthermore, there are different sources of errors in survey samples. There are sampling and non-sampling errors. Sampling errors are due to the fact that the estimates are produced based on a subset of the population. Non-sampling errors are for example coverage, measurement and nonresponse errors. Coverage issues arise when the data frame does not match the target population of the study. The data frame may contain units that are not supposed to be surveyed or, on the contrary, some targeted units may be absent of the data frame. Measurement errors are the differences between the real value of an item and the recorded value in the data set. Nonresponse is the presence of missing values in a data set. To address the problems in survey samples, sophisticated methods using auxiliary information are developed. In this thesis, we focus on sampling methods and the treatment of nonresponse. Section 1.2 covers basic notions of survey sampling and Section 1.3 reviews the baselines of survey nonresponse.

1.2. Complete case

Survey sampling concepts, notation and basic estimation methods are introduced in this section. Those notions are presented in the complete case, which corresponds to the absence of nonresponse. For comprehensive references on survey sampling, see Särndal et al. (1992); Tillé (2001); Ardilly (2006).

1.2.1. Concepts and notation

Consider a finite population U of size N. We are interested in estimating a population parameter θ, which is a function of a survey variable y. A sampling design is the probability distribution assigning to each subset $s \subset U$ a probability $p(s)$ to be selected. Define S the random sample such that $\text{pr}(S = s) = p(s)$. The sampling design is such that $p(s) \geq 0$ and

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

Figure 1.1 represents a sample S in a population U. Consider the indicator variable

$$a_k = \begin{cases} 1 & \text{if the unit } k \text{ is selected in the sample } S, \\ 0 & \text{otherwise,} \end{cases}$$

for $k \in U$. The first order inclusion probability π_k is the probability that unit $k \in U$ is selected in the sample,

$$\pi_k = \Pr(k \in S) = \mathrm{E}_p(a_k) = \sum_{\substack{s \subset U \\ s \ni k}} p(s),$$

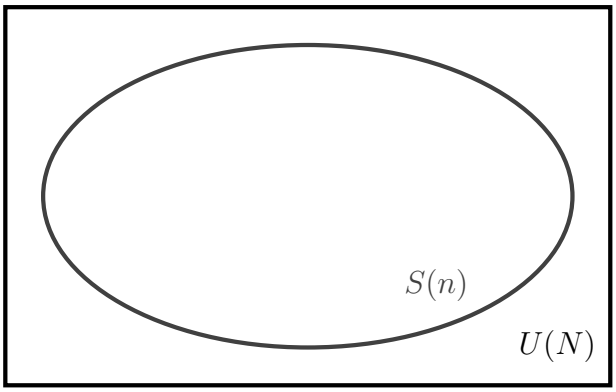

Fig. 1.1. Representation of a population U of size N and a sample S of size n.

where $\mathrm{E}_p(.)$ stands for the expectation with respect to the sampling design. The second order inclusion probability $\pi_{k\ell}$ is the probability that both $k \in U$ and $\ell \in U$ are selected in the sample,

$$\pi_{kl} = \Pr(k, \ell \in S) = \mathrm{E}_p(a_k a_\ell) = \sum_{\substack{s \subset U \\ s \ni \{k,\ell\}}} p(s)$$

if $k \neq \ell$ and $\pi_{kk} = \pi_k$ if $k = \ell$.

The value of the variable y is measured on the sampled units only. Consider y_k the value of variable y of unit $k \in U$. Suppose, for instance, that the parameter of interest is $\theta = Y$, the total of y, where

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

This population total can be estimated by the Hortvitz-Thompson total estimator

$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} a_k \frac{y_k}{\pi_k}.$$

This estimator is unbiased,

$$\mathrm{E}_p\left(\widehat{Y}\right) = \sum_{k \in U} \mathrm{E}_p\left(a_k\right) \frac{y_k}{\pi_k} = \sum_{k \in U} y_k = Y.$$

The variance of the total estimator is

$$\mathrm{V}_p\left(\widehat{Y}\right) = \sum_{k \in U} \sum_{\ell \in U} \Delta_{k\ell} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell}, \tag{1.2.1}$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$ if $k \neq \ell$ and $\Delta_{kk} = \pi_k(1 - \pi_k)$. An unbiased estimator of (1.2.1) is

$$\widehat{\mathrm{V}}\left(\widehat{Y}\right) = \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell}.$$

1.2.2. Some sampling designs

The conditions in which a sample is selected vary from one survey to another. Various sampling designs were developed in the literature to suit the different frameworks. Tillé (2006) covers many sampling designs along with algorithms to implement them. The following examples introduce sampling designs that are used in this manuscript.

Example 1.1. *Poisson sampling design:*
Consider a population U of N units. Consider also π_k the first order inclusion probability of unit $k \in U$. In the Poisson sampling design, each unit $k \in U$ is selected independently in the sample with probability π_k. In this case, the second order inclusion probability of units $k, \ell \in U$ such that $k \neq \ell$ is $\pi_{k\ell} = \pi_k \pi_\ell$. The sample size n_S is random and depends on S. The expectation of the sample size is

$$\mathrm{E}_p\left(n_S\right) = \mathrm{E}_p\left(\sum_{k \in U} a_k\right) = \sum_{k \in U} \pi_k.$$

Example 1.2. *Two-phase sampling design:*
Consider a population U of size N. A first phase sample S^A is selected in U according to a sampling design $p^A(.)$. The inclusion probability of unit k in S^A is π_k^A, for $k \in U$. At the second phase, a sample S^B is selected in S^A according to a sampling design $p^B(s^B|S^A) = \Pr(S^B = s^B|S^A)$. The unit $k \in S^A$ has a probability $\pi_k^{B|A} = \Pr(k \in S^B|S^A)$ to be in S^B conditionally to S^A.

Example 1.3. *Two-stage sampling design:*
Consider a population U of N sampling units, namely the secondary units. The population is partitioned into M subsets, or primary units, $U_1, \ldots, U_i, \ldots, U_M$. At the first stage, a sample S_1 of m primary units is selected with a sampling design $p^1(.)$. The first order inclusion probability of subset U_i is π_i^1. At the second stage, a sample S_i of n_i secondary units is selected in any $U_i \in S_1$. The final sample of $n = \sum_{i=1}^m n_i$ sampling units is $S = \bigcup_{i=1}^m S_i$.

1.3. Nonresponse

Missing values are difficult to avoid in survey samples. In this section, the notions, the impact and the treatments for nonresponse are introduced. For comprehensive references on survey nonresponse, see Särndal and Lundström (2005); Haziza (2009); Kim and Shao (2013).

1.3.1. Two types of nonresponse

Nonresponse can occur in many ways. More specifically, there are two types of nonresponse that are distinguished in the literature: total and item nonresponse. Total nonresponse corresponds to the cases where all items of a unit are missing.

This can happen, for instance, when a unit refuses to answer the survey. It can also be impossible to reach the unit because it is inaccessible. Item nonresponse occurs when some, but not all, items of a unit are missing. For example, a unit can voluntarily not answer a question because it is sensitive or not understandable.

Missing values generally have undesirable effects on survey estimates. It can introduce a bias and increase the variance of the estimators. Indeed, if the units with missing values have a similar profile, the nonresponse can affect the estimated parameters. For example, suppose that men and women are surveyed on their weight and that men are less susceptible to indicate their weight. The average weight of men is larger than the women's one. Then a higher nonresponse rate among men than the one among women will cause an underestimation of the average weight in the population. Furthermore, missing values can be seen as a decrease in the sample size. A smaller sample size usually leads to an increase in the variability of the estimators.

In order to reduce the negative effects of nonresponse, different treatments have been developed. Unit nonresponse is usually corrected by reweighting methods. The nonrespondents are removed from the data set and the weights of the respondent units are adjusted to account for the nonrespondents. Kalton and Flores-Cervantes (2003); Brick (2013) present overviews of reweighting methods for the correction of unit nonresponse. Regarding item nonresponse, it is generally treated by imputation methods. Imputation consists in replacing missing values with artificial values, which are determined by the data analysts. Recently, Chen and Haziza (2018) reviewed imputation methods for item nonresponse.

1.3.2. Notation and examples

Consider a sample S of n units from a population U. The set of n_r sampled respondents is noted $S_r \subset S$, see Figure 1.2. The set of sampled nonrespondents is S_m and it contains $n_m = n - n_r$ units. Define R_k the response indicator of unit k, such that

$$R_k = \begin{cases} 1 & \text{if the unit } k \text{ is a respondent,} \\ 0 & \text{otherwise.} \end{cases}$$

The distribution of R_k is the nonresponse mechanism and it is generally unknown. The assumptions made on the nonresponse mechanism are called the nonresponse model.

In the case of unit nonresponse, the sampling weights of the respondents are adjusted to account for the nonrespondents. The reweighted estimator of the total Y is

$$\widehat{Y}_R = \sum_{k \in S_r} w_k y_k = \sum_{k \in U} a_k R_k w_k y_k, \tag{1.3.1}$$

where w_k is the adjusted weight of unit k such that $a_k R_k = 1$.

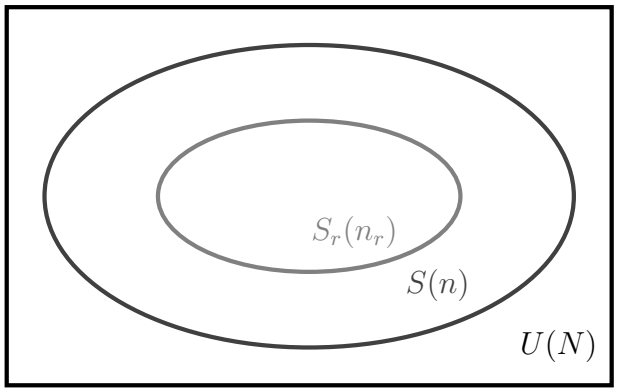

Fig. 1.2. Representation of a population U of size N, a sample S of size n and a respondent set in the sample, S_r of size n_r.

Example 1.4. *The adjusted weights can be found by calibration. Consider* $\mathbf{x}_k \in \mathbb{R}^p$*, the vector of* p *auxiliary variables of unit* k*. The vector* $\mathbf{x}_k$ *is known for any* $k \in S$ *and the vector of population totals* $\mathbf{X} = \sum_{k\in U} \mathbf{x}_k$ *is assumed to be known. The calibration weights are found by minimizing a distance between the sampling weight* $d_k = \pi_k^{-1}$ *and the calibration weight* w_k *for each* $k \in S_r$*, while respecting the calibration function*

$$\sum_{k\in S_r} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k.$$

Once the calibration weights are determined, they can be introduced in the calibrated estimator (1.3.1). Deville and Särndal (1992) present more precisely calibration techniques in the complete case. Särndal and Lundström (2005) discuss calibration reweighting for nonresponse. They also modify the calibration equations in function of the type of auxiliary variables.

In the case of item nonresponse, the missing values can be imputed. This means that a missing value is replaced by an artificial value. The imputed estimator of the total Y is

$$\widehat{Y}_I = \sum_{k\in S} d_k \widetilde{y}_k = \sum_{k\in S_r} d_k y_k + \sum_{k\in S_m} d_k y_k^*,$$

where

$$\widetilde{y}_k = R_k y_k + (1 - R_k) y_k^*$$

and y_k^* is the value imputed to the missing item of unit k. The imputation methods can either be deterministic or random. If repeated on the same data set, deterministic imputation will always lead to the same imputed values. In the case of random imputation, there is a random element making sure that the imputed values of a data set are different if the imputation is repeated under the same conditions. Random imputation increases the variability of the estimators, but it is relevant when the distribution of the imputed variable needs to be preserved. Indeed, deterministic imputation does not tend to preserve the distributions of the variables, while

random imputation does. So a deterministic method is sufficient to estimate a total or a mean. For quantile estimation, a random imputation method is favored.

Example 1.5. *Imputation methods are often motivated by a model on the variable of interest. Consider the following model for y, namely the imputation model,*

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

where $\mathbf{x}_k \in \mathbb{R}^p$ is the vector of p auxiliary variables of unit k and $\boldsymbol{\beta}$ is a vector of p regression coefficients. The error term ε_k is such that $\mathrm{E}_m(\varepsilon_k) = 0$, $\mathrm{E}_m(\varepsilon_k \varepsilon_\ell) = 0$ and $\mathrm{V}_m(\varepsilon_k) = \sigma^2$, for any $k \neq \ell \in U$ and for some fixed σ. The notations $\mathrm{E}_m(.)$ and $\mathrm{V}_m(.)$ stand for the expectation and the variance with respect to the imputation model. In the deterministic case, the missing item of unit k can be imputed using regression imputation,

$$y_k^* = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}},$$

where $\widehat{\boldsymbol{\beta}}$ is the vector of regression coefficients estimated in S_r. In the case of random imputation, a random term can be added and

$$y_k^* = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}} + e_k,$$

where e_k is a random residual. This residual can, for instance, be randomly selected among observed residuals in S_r.

1.3.3. Frameworks for inference in the presence of nonresponse

In the presence of nonresponse, the framework to make inference on the estimators needs to be established. Two frameworks are distinguished to conduct inference: the imputation model approach and the nonresponse model approach. In the imputation model approach, the inference is made with respect to the sampling design, the imputation model and the nonresponse mechanism. The nonresponse mechanism is not specified, it is only assumed to be ignorable. This is, the nonresponse mechanism does not depend on y when accounting for the auxiliary variables. In the nonresponse model approach, the inference is made with respect to the sampling design and the nonresponse model. No assumptions are needed on the variable of interest.

Whichever framework is used, there are two approaches for variance estimation: the two-phase approach and the reverse approach. In the first case, the procedure on which the approach is based is the following:

(i) A sample S is selected in a population U with respect to a sampling design $p(s)$;

(ii) A set of respondents $S_r \subset S$ is determined with respect to a nonresponse mechanism.

This procedure is illustrated in Figure 1.2. The sample of respondents is thus seen as a two-phase sample. The first phase sample corresponds to the sample S and

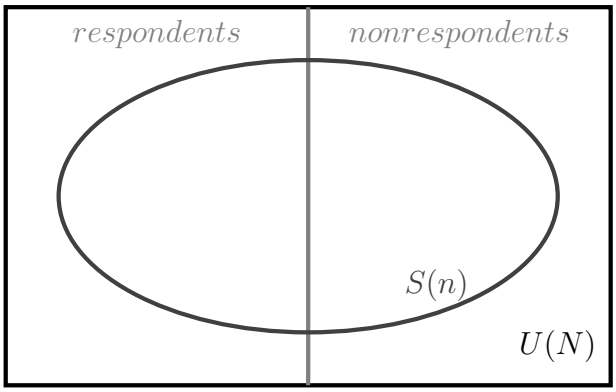

Fig. 1.3. Representation of the reverse approach. Consider a population U of size N. The population is divided in a set of respondents and a set of nonrespondents. A sample S of size n contains both respondents and nonrespondents.

the second phase sample corresponds to the set of respondents in S. In the reverse approach, the procedure is inverse. The procedure is seen as follow:

(i) The population U is divided in a set of respondents and a set of nonrespondents, with respect to a nonresponse mechanism;

(ii) A sample S is selected in U with respect to a sampling design $p(s)$. This sample contains respondents and nonrespondents.

This approach is represented in Figure 1.3.

To calculate the variance of an estimator, the framework needs to be well defined. We have to choose between the imputation model approach and the nonresponse model approach. Moreover, the variance is also decomposed in function of the two-phase approach or the reverse approach. Haziza (2009) gives a complete overview of the different combinations of frameworks. The author details the variances of an imputed estimator for each framework.

Chapter 2

Incorporating Spatial and Operational Constraints in the Sampling Designs for Forest Inventories

Abstract: Besides the evaluation of the volume of standing trees, goals of forest inventories include collecting geophysical information and monitoring fragile ecosystems. In the province of Quebec, Canada, their implementation faces challenging methodological problems. The survey area covers a large territory which is hardly accessible and has a diverse forest. The main operational goals are to spread the sampled plots throughout the survey area and to capture, in the sample, the forest heterogeneity while keeping the cost at a reasonable level. In many inventories, a two dimensional systematic sampling design is applied and the rich auxiliary information is only used at the estimation stage. We show how to use modern and advanced sampling techniques to improve the planning of forest inventories and meet complex operational goals. For the Quebec forest inventory, we build a two-stage sampling design that has clusters of plots to optimize field work and predetermined sample sizes for forest types. Constraints of spreading the sample in the whole territory and of balancing according to auxiliary variables are also implemented. To meet these requirements, we use unequal inclusion probabilities, balanced sampling, highly stratified balanced sampling, and sample spreading. The impact of these novel techniques on the implementation of requirements and on the precision of survey estimates is investigated using Quebec inventory data.[1]

Keywords: Balanced sampling; Spatially balanced sampling; Stratified sampling; Unequal probability sampling.

[1]This article is a reprint of Vallée, A.-A., Ferland-Raymond, B., Rivest, L.-P., and Tillé, Y. (2015). Incorporating spatial and operational constraints in the sampling designs for forest inventories. *Environmetrics*, 26(8):557–570

2.1. Introduction

Forest inventories are multipurpose surveys that collect extensive data on forest composition, such as tree species, tree height, tree diameter at breast height, dendrometric information, forest health and site quality. Besides planning forest operations, the goals of these inventories also include the monitoring of the forest ecology, see Lawrence et al. (2010) for a recent review. Many inventories use an areal frame with two dimensional systematic sampling to select the survey area. The main sampling methods of forest inventory are presented among others in McRoberts and Tomppo (2007); Gregoire and Valentine (2007); Mandallaz (2008). The rich auxiliary information available through various types of imaging techniques is generally used only at the estimation stage of the survey. For instance, Opsomer et al. (2007) use a generalized additive model to improve the efficiency of the estimates for the US survey inventory and Saarela et al. (2015) use model-assisted estimation while studying Western Finnish forest. It may be interesting to use this auxiliary information while elaborating the sampling design.

Recently, Fattorini et al. (2015) used auxiliary information to construct the inclusion probabilities and to spatially spread the sample of forest surveys. Fattorini et al. (2006) used auxiliary variables to stratify the different phases of its sampling design. Grafström et al. (2014) demonstrated that spreading the sample in the auxiliary variables space may be very efficient. Grafström and Ringvall (2013) show that using auxiliary information while planning the sampling design is very important; it is allowing to obtain better samples and to improve the estimates. This work uses the auxiliary information at the sampling design step to ensure that important operational constraints are met. It constructs a two stage sampling design with fixed sample sizes for both, secondary units and forest types (for two-stage sampling, see Särndal et al., 1992). Balanced sampling (Deville and Tillé, 2004; Tillé, 2006), sample spreading (Grafström et al., 2012; Grafström and Tillé, 2013; Grafström and Lundström, 2013; Grafström and Schelin, 2014) and balanced stratification (Chauvet, 2009; Hasler and Tillé, 2014) are used and their impact on the sampling properties of survey estimates are investigated.

These techniques are easily malleable and can fulfil various tasks as ensuring fixed sample sizes and consistency of the estimates. Moreover, from a user point of view, they do not change the way in which the survey data is exploited. Indeed, the inclusion probabilities are determined at the planning stage and a simple Horvitz-Thompson estimator can be used. Thus, implementing operational constraints using these advanced sampling methods does not change the way in which forest characteristics are estimated.

A description of the context of the Quebec inventory is given in Section 2.2. Section 2.3 presents the specific requirements of this inventory and the resulting inclusion probabilities of the sample units. The first stage of the sampling design

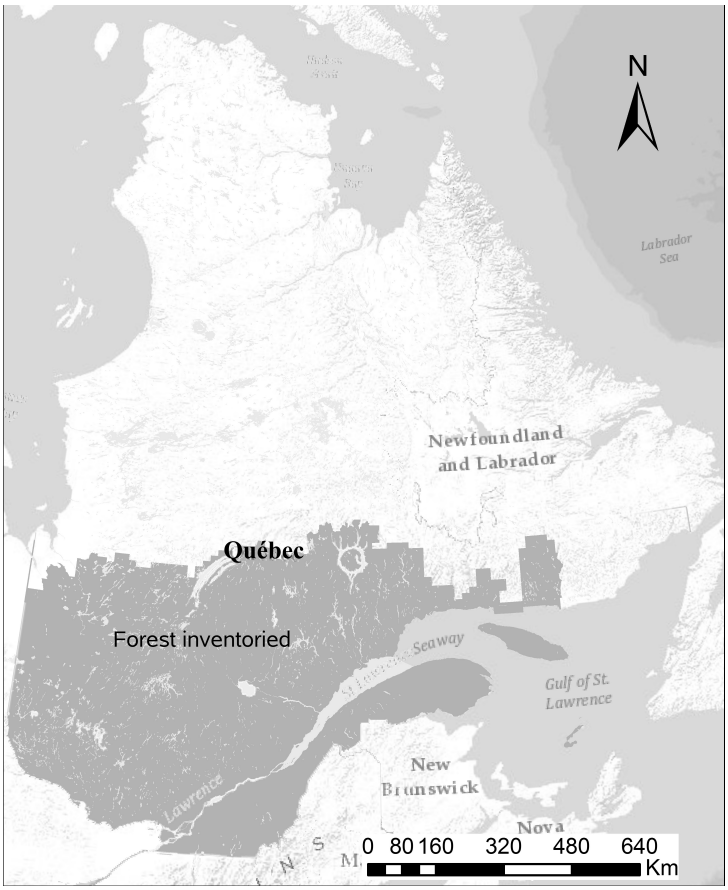

Fig. 2.1. Region of interest in the Quebec regular inventory, which is south of the 52nd parallel.

consists in balanced sampling and is elaborated in Section 2.4. In Section 2.5, a highly stratified sampling design is used to implement the second stage. The high points of the proposed design are detailed in Section 2.6. Additional improvements to the sampling design, through sample spreading, are investigated in Section 2.7. Finally, the impact of these novel techniques on the precision of survey estimates is investigated using field data from a Quebec inventory in Section 2.8.

2.2. Forest inventories in Quebec

In the province of Quebec, Canada, the forest south of the 52nd parallel is investigated by a regular inventory. This represents an area of about 600,000 km^2. As a representation, Figure 2.1 shows the region of interest in the Quebec inventory. Most of the forest is situated north of the populated area of Quebec and cannot be easily accessed. Indeed for the inventory discussed in Section 2.8 more than 60% of the plots sampled were reached from the air by helicopter and seaplane as ground transportation was often not possible for a lack of forest roads. The forest is varied as it covers several ecological zones. The most abundant species, black spruce (*Picea mariana*), Jack pine (*Pinus banksiana*) and balsam fir (*Abies balsamea*), representing 59% of the volume of growing stock, are all associated to the northern part of the

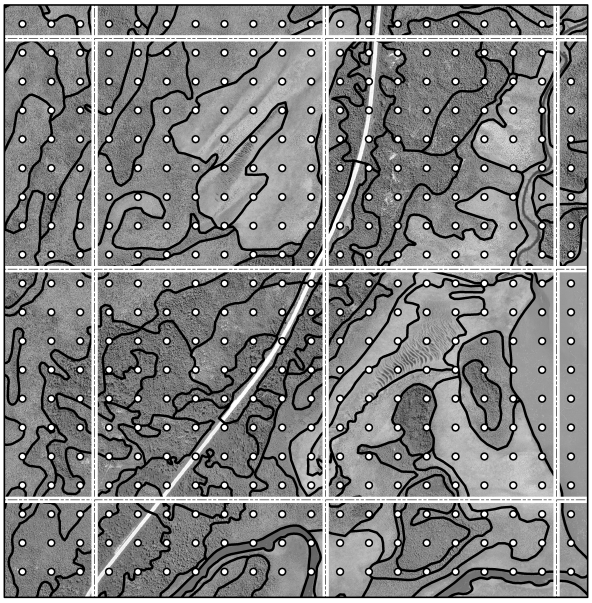

Fig. 2.2. Example of a set of polygons, tiles and plots. Plots are represented by the dots. Tiles are clusters of 64 plots and are represented by squares. Polygons are the rounded shapes.

inventory area. Indeed more than twenty different tree species are monitored by the inventory. The Quebec forest area can be compared to that of Finland which covers 300,000 km^2. Its forest is however more diverse as, in Finland, three species, Scots pine *(Pinus sylvestris L.)*, Norway spruce (*Picea abies*) and Birch (*Betula*) account for more than 95% of the growing stock, see (Metla, 2015).

The Quebec inventory area is divided into forest management units (FMU) that are typically surveyed once every ten years; they are the areal sampling frames used for selecting plots. The objective of the inventory discussed in this paper is to provide a snapshot of the forest in a FMU at a given time point; this is a cross-sectional survey. Monitoring temporal changes in the Quebec forest is not an objective of this survey as this is done through a network of permanent plots (Minister of Forests, Wildlife and Parks, 2003).

Two or three years before the survey, the FMU is photographed and subdivided into homogeneous polygons by photo-interpreters (see Figure 2.2). These polygons are the basic compilation units for all the inventory operations. Photo-interpreted variables are recorded for each polygon including tree height, basal area and information about the tree species found in the polygon. This information is then used to divide the survey area into H types of polygons and to set targeted sample sizes for each type.

Following chapter 7 of Gregoire and Valentine (2007), this inventory uses areal sampling frames. The sampled units are circular plots of 400 m^2 centered at points randomly selected in an areal frame. Thus the forest to be inventoried is regarded

more as a continuous landscape than as a finite set of individual trees. There are simply too many trees in the 600,000 km^2 study area to collect tree level information. The results are reported at the plot level, in terms of units per hectare. A sampled plot is visited by a team of foresters who collect ground information. Besides tree measurements such as volume and height, many other characteristics such as soil richness, land cover and wood quality are investigated. Note that the photo-interpreted auxiliary information has been collected at the polygon level. Thus all the plots belonging to a polygon have the same values for all photo-interpreted variables (see Figure 2.2). Besides the photo-interpretation, Landsat 5 multi-spectral imagery provides a second source of auxiliary information. These images have a 30 m resolution and the Landsat information is available at the plot level. It comes as a vector containing the 7 spectral band values ranging between 1 and 256, band number 6 is usually not used in forestry.

A key operational constraint is to optimize the number of sampled plots given a fixed number of forester working days and a predetermined transportation budget. In the example of Section 2.8, plots in the surveyed FMU can be as much as 100 km apart so the field work has to be organized carefully. Another operational constraint is to account for the various forest types in the sample selection. For instance, mixed forest is more diversified in species and should thus be oversampled compared with the resinous forest of black spruce. Indeed, this resinous forest is found in the northern part of the inventory area and is less diversified.

Selecting the plots according to a simple random sampling design stratified by type is not appropriate. Most plots are not easily accessible; going to and coming back from a sampled plot may take a whole working day. Visiting a single sample unit each day is far from being optimal and consistent with the first operational constraint. Once on location, a ground team should visit a cluster of plots.

In order to enable the selection of clusters of plots, a systematic sample of plots is first selected on a 125 m square grid. Then tiles, defined as a cluster of 8×8 adjacent plots on the grid, are created (see Figure 2.2). Each tile contains 64 plots and covers 1 km^2; the study area can be seen as a grid of tiles. Once on location, in a tile, a ground team is able to collect the information for 4 plots of the tile in a single working day.

To account for the two operational constraints highlighted in this section, the plots need to be selected according to a two-stage sampling design. First, a random sample of tiles is selected at the first stage. At the second stage, four plots are randomly selected in each sampled tile. A methodological challenge is to enforce the operational constraint of a fixed sample size by type of plots in this two-stage sampling design. Also, the Landsat band values are related to key forest attributes and it would be interesting to use them while planning the survey. Methods for doing so are proposed and investigated in this work.

2.3. An overview of the proposed sampling design

2.3.1. Notation

As mentioned in Section 2.2, the results of this forest inventory are reported at the plot level. Thus, let U denote the population of plots, and let N denote its size. Population U is divided into M tiles denoted by $U_{\cdot i}, i = 1, \ldots, M$, (see Figure 2.2). They are the primary sampling units. Considering Figure 2.2, the number of plots $N_{\cdot i}$ in $U_{\cdot i}$ satisfies $N_{\cdot i} \leq 8 \times 8 = 64$ as plots in a non-forest environment and unaccessible plots are excluded.

The plots are of H different types, which are defined in this paper in terms of species and age classes. Thus, population U can be divided into sub-populations U_{hi}, which contain every plots of type h in tile i, where $h = 1, \ldots, H$ and $i = 1, \ldots, M$. Let N_{hi} be the number of plots in sub-population U_{hi}. The set of plots of type h is denoted by $U_{h\cdot}$ and its size is $N_{h\cdot}$.

The goal is to select $n = m \times r$ plots in population U using two stages of sampling. First a sample of m tiles is selected. Next a sample of r plots is selected within the tiles chosen at the first step. Let n_{hi} denotes the number of sampled plots of type h in tile i. The number of plots selected in sampled tile i is $n_{\cdot i} = r$. The notation $n_{h\cdot}$ stands for the number of plots of type h within the sample. Observe that $\sum_h n_{h\cdot} = n$. As mentioned in Section 2.2, a suitable value for $n_{\cdot i} = r$ is 4 as a forester is able to visit 4 plots per day if they are in the same tile.

2.3.2. Calculation of the inclusion probabilities

From a statistical point of view, stratifying the plots by forest types and using stratified random sampling, possibly balanced according to key auxiliary variables, would be satisfactory. Such a design would however be very costly and sampling tiles of plots is needed to control expenses. Still we would like to retain the first order selection probabilities of a stratified design and have selection probabilities given by $\pi_k = n_{h\cdot}/N_{h\cdot}$, for all $k \in U_{h\cdot}$, where $n_{h\cdot}$ is the predetermined sample size for plots of type h. A consequence of this constraint would be that the expected number of plots of type h is equal to $n_{h\cdot}$. Moreover, we would like the number of plots sampled in a tile selected at stage 1 to be equal to r. The next results give the selection probabilities necessary for these two constraints to be met.

Proposition 2.1. *Let $\pi_k = \pi_{1i}\pi_{k|i}$, where π_{1i} is the first stage probability for selecting tile i and $\pi_{k|i}$ is the probability of selecting plot k given that tile i is selected. In order to meet the two constraints:*

(i) the number of plots sampled in each tile is r, and

(ii) $\pi_k = n_{h\cdot}/N_{h\cdot}$ if plot k is of type h,

the first and second stage selection probabilities π_{1i} and $\pi_{k|i}$ must be given by

$$\pi_{1i} = \frac{1}{r}\sum_{h=1}^{H}\frac{n_{h\cdot}}{N_{h\cdot}}N_{hi}, \tag{2.3.1}$$

$$\pi_{k|i} = \frac{r\frac{n_{h\cdot}}{N_{h\cdot}}}{\sum_{h=1}^{H}\frac{n_{h\cdot}}{N_{h\cdot}}N_{hi}}, \tag{2.3.2}$$

for k in tile i and $i = 1, \ldots, M$.

Proof. From constraint *(i)*, the average number of plots sampled in a tile is r. Thus, we must have

$$\sum_{k\in U_{\cdot i}} \pi_{k|i} = r, \tag{2.3.3}$$

for $i = 1, \ldots, M$. On the other hand, from constraint *(ii)*, $\pi_k = n_{h\cdot}/N_{h\cdot}$ for $k \in U_{h\cdot}$. We thus have

$$\begin{aligned}\sum_{k\in U_{\cdot i}} \pi_{k|i} &= \sum_{k\in U_{\cdot i}} \frac{\pi_k}{\pi_{1i}} = \frac{1}{\pi_{1i}}\sum_{h=1}^{H}\sum_{k\in U_{hi}} \pi_k \\ &= \frac{1}{\pi_{1i}}\sum_{h=1}^{H}\sum_{k\in U_{hi}} \frac{n_{h\cdot}}{N_{h\cdot}} = \frac{1}{\pi_{1i}}\sum_{h=1}^{H}\frac{n_{h\cdot}}{N_{h\cdot}}N_{hi}.\end{aligned} \tag{2.3.4}$$

Expressions (2.3.3) and (2.3.4) lead to

$$\pi_{1i} = \frac{1}{r}\sum_{h=1}^{H}\frac{n_{h\cdot}}{N_{h\cdot}}N_{hi}.$$

The value of $\pi_{k|i}$ is derived from the fact that $\pi_k = \pi_{1i}\pi_{k|i}$. □

Constraint *(ii)* implies that the expectation of the number of plots of type h is $n_{h\cdot}$. Indeed, the expectation of the sample size in a subpopulation is equal to the sum of the inclusion probabilities in this subpopulation, that is

$$\sum_{k\in U_{h\cdot}} \pi_k = \sum_{k\in U_{h\cdot}} \frac{n_{h\cdot}}{N_{h\cdot}} = n_{h\cdot}.$$

In some rare cases, Expressions (2.3.1) and (2.3.2) could lead to values larger than 1, which is unacceptable for inclusion probabilities. In our applications, this problem does not appear for the first-stage inclusion probabilities π_{1i}. However, in very rare cases, the problem occurs within some tiles of the population. In this case, instead of Expression (2.3.2), the inclusion probabilities are computed as

$$\pi_{k|i} = \min\left(C_i\frac{n_{h\cdot}}{N_{h\cdot}}, 1\right),\ k \in U_{hi},$$

where C_i is defined in such a way that

$$\sum_{h=1}^{H} N_{hi}\min\left(C_i\frac{n_{h\cdot}}{N_{h\cdot}}, 1\right) = n_{\cdot i},\ i = 1, \ldots, M.$$

An algorithm to compute these inclusion probabilities is presented in Tillé (2006, p. 19).

Selection probabilities alone cannot constrain the sample size to be exactly r in each tile and to be exactly $n_{h\cdot}$ for plots of type h. Balanced sampling, at the two stages of the design, is needed to have such predetermined sample sizes. The next sections show how this can be obtained.

2.4. First stage of the sampling design

This section reviews balanced sampling, as introduced by Deville and Tillé (2004). Then it shows how this technique is applied to select a sample of m tiles for the forest inventory presented in Section 2.2.

2.4.1. Balanced sampling

Let $\mathbf{x}_k \in \mathbb{R}^p$ be a column vector of p auxiliary variables available for every units k in the population U. A sampling design assigns a probability $p(s)$ to each sample or subset s of U. Let S denotes the random sample, i.e. $\Pr(S = s) = p(s)$. A sampling design $p(.)$ is said to be balanced on the vector of variables $\mathbf{x}_k$ if

$$\sum_{k \in s} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k, \tag{2.4.1}$$

for every sample $s \subset U$, such that $p(s) > 0$. Note that it is not always possible to satisfy exactly Equation (2.4.1), because the selection of a sample is an integer problem. Indeed, each unit is selected or not in the sample. Most of the time, it is impossible to find a sample $S \subset U$ which exactly satisfies the balancing equations. Several algorithms have been proposed to select a balanced sample or an approximately balanced sample if an exact solution does not exists.

Deville and Tillé (2004) have proposed a general solution called the 'cube method'. This method satisfies exactly Equation (2.4.1), if it is possible, and otherwise, finds a good approximation for the rounding problem. A rapid algorithm has been proposed by Chauvet and Tillé (2006). This procedure has two phases named the flight phase and the landing phase.

The flight phase starts with the vector $\boldsymbol{\pi}$ of selection probabilities and produces a random vector $\boldsymbol{\pi}^*$ containing mostly zeros and ones. When all the components of $\boldsymbol{\pi}^*$ are either 0 or 1, the sample is selected. At most p values of $\boldsymbol{\pi}^*$ can fail to be 0 or 1, see (Deville and Tillé, 2004). Then an exact solution cannot be reached for the balancing equations and a landing phase is necessary to find a sample that is approximately balanced.

The landing phase starts with $\boldsymbol{\pi}^*$ and gives a vector containing only 0s and 1s corresponding to a sample for which the balancing equations (2.4.1) are approximately satisfied. There are several ways of implementing the landing phase. In the application discussed in this paper, the method by suppression of variables is used. The balancing constraints are relaxed by removing variables one at a time.

2.4.2. First stage of the inventory sample

The inclusion probabilities for the selection of the tiles, at the first stage of the sampling design, are derived in Proposition 2.1. Here we propose to balance on the variables in the following vector,

$$\mathbf{x}_{1i} = \left(\pi_{1i},\ \ N_{1i}\frac{n_{1\cdot}}{N_{1\cdot}}, \ldots, N_{hi}\frac{n_{h\cdot}}{N_{h\cdot}}, \ldots, N_{Hi}\frac{n_{H\cdot}}{N_{H\cdot}},\ \ \mathbf{z}_{1i}^{\top}\right)^{\top},\ i = 1, \ldots, M. \tag{2.4.2}$$

Vector $\mathbf{z}_{1i}$ contains the totals of the landstat band colors in tile i:

$$\mathbf{z}_{1i} = \sum_{k \in U_{\cdot i}} \mathbf{z}_k,$$

where $\mathbf{z}_k$ is the vector of the landstat band colors in plot k.

The balancing equation system is

$$\sum_{i \in S_T} \frac{\mathbf{x}_{1i}}{\pi_{1i}} = \sum_{i=1}^{M} \mathbf{x}_{1i},$$

where S_T denotes the sample of tiles.

The population total of the first balancing variable, π_{1i}, is m. The first balancing equation becomes

$$\sum_{i \in S_T} \frac{\pi_{1i}}{\pi_{1i}} = \sum_{i \in S_T} 1 = \sum_{i=1}^{M} \pi_{1i} = m, \tag{2.4.3}$$

which means that the tile sample size is exactly m.

Balancing equations 2 to $H + 1$ determine the distribution of plot types in the sample of tiles. The one for type h is

$$\sum_{i \in S_T} \frac{N_{hi}}{\pi_{1i}} \frac{n_{h\cdot}}{N_{h\cdot}} \approx \sum_{i=1}^{M} N_{hi} \frac{n_{h\cdot}}{N_{h\cdot}} = n_{h\cdot}. \tag{2.4.4}$$

Finally, balancing on the Landsat variables makes the sample representative for these variables.

2.5. Second stage of the sampling design

Given the first stage sample of tiles, the second stage sampling design is stratified by tiles. The second stage has three types of balancing equations, for the tile sample sizes, $n_{\cdot i} = r$, for sample sizes by forest types, $n_{\cdot h}$, and for the Landsat color variables. Conditionally to the first stage, the second stage can be seen as a stratified design where the stratification is on the tiles. The number of strata is the number of tiles. The sampling design is then highly stratified. Special sampling algorithms are needed for its implementation. These algorithms are briefly reviewed next.

2.5.1. Highly stratified balanced sampling

Suppose that population U is divided into L strata U_ℓ, $\ell = 1, \ldots, L$. A stratified sampling design is said to be balanced on $\mathbf{x}_k$ if the sampling design is balanced in

each stratum, that is

$$\sum_{k \in S \cap U_\ell} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U_\ell} \mathbf{x}_k, \tag{2.5.1}$$

where S is the sample of plots and $\ell = 1, \ldots, L$.

Chauvet (2009) has proposed a method to select a sample satisfying Equation (2.5.1). This method is decomposed in three steps. First, flight phases are run within each stratum independently to balance the sample within the strata as much as possible. Second, the set of all the plots that were not selected or rejected in the first step are merged together. A general flight phase is run in this set of plots to balance the sample, as much as possible, over all the plots in the population. Thirdly, to manage the plots that are not selected or rejected at the end of the second step, Chauvet (2009) uses an unequal probability sampling design with fixed sample size.

The method of Chauvet can rapidly become untractable when the number of strata is large. The second step of Chauvet's method may be impossible because, in the second step, stratum indicator variables must be added in the matrix of balancing constraint to conserve a fixed sample size in the strata. The size of the matrix is then too large to run the flight phase of the cube method.

Hasler and Tillé (2014) have proposed a new selection algorithm which is fast and leads to a solution even if the sample is highly stratified. The first step is to run independent flight phases in each stratum as in Chauvet (2009); the second step is changed however. The flight phase is not run over all the plots with conditional selection probabilities in $(0, 1)$ at the end of step 1. It is carried out sequentially. First a flight phase is run over the first two strata combined, $U_1 \cup U_2$. Then the third stratum U_3 is added to the set and another flight phase is run; this action is repeated until the last stratum is added and a last flight phase is run. Before adding a new stratum, the current vector of selection probabilities, $\pi(t)$, is checked to determine whether the sample for one of the strata considered in the current flight phase is complete, that is whether $\pi_k(t) = 0$ or 1 for all $k \in U_\ell$. When this happens, all the balancing variables for this stratum are taken out; they are not used in the subsequent flight phases ran when additional strata are added. This sequential method involves only a relatively small matrix of balancing variables at each step, as variables for stratum balancing are brought in and taken out throughout the process. Finally, the third step is a landing phase by suppression of variables over units for which the decision about selection or rejection is not yet taken. The modification at the second step allows balancing the sampling design even if it is highly stratified; indeed there is no limit to the number of strata that can be considered. Note that if the sum of the inclusion probabilities in each stratum is an integer, the method of Hasler and Tillé (2014) allows exactly the integer number of units to be selected in each stratum.

2.5.2. Second stage of the inventory sample

The first step of the design gave a sample S_T containing m tiles. The second stage of the sampling design selects a stratified sample of $r = n_{\cdot i}$ plots per tiles for a total of $n = r \times m$ plots. Let $S_{\cdot i}$ denote the sample of r plots selected in tile $i \in S_T$. The final sample of plots is denoted by S, where

$$S = \bigcup_{i \in S_T} S_{\cdot i}.$$

The inclusion probabilities for the selection of plots, conditionally to the first stage, are set in Proposition 2.1. Here we propose to balance on the variables in the following vector,

$$\mathbf{x}_{2k} = \left(\mathbb{1}_{k \in U_1 \cdot} \pi_{k|i}, \ldots, \mathbb{1}_{k \in U_H \cdot} \pi_{k|i}, \ \ \mathbf{z}_k^\top \right)^\top, \tag{2.5.2}$$

for $k \in U_{\cdot i}$, $i \in S_T$, where $\mathbf{z}_k$ is the vector of the Landsat band colors of plot k. The sampling design is stratified on the tiles: balancing equations (2.5.1), with the vector (2.5.2), must be satisfied for every tiles $U_{\cdot i}$, $i \in S_T$. That is, the balancing equation for tile $U_{\cdot i}$ is

$$\sum_{k \in S_{\cdot i}} \frac{\mathbf{x}_{2k}}{\pi_{k|i}} = \sum_{k \in U_{\cdot i}} \mathbf{x}_{2k},$$

for $i \in S_T$. Since the number of tiles in S_T is large, the method of Hasler and Tillé (2014) for highly stratified balanced sampling is used, with vector (2.5.2). As a result, the sample of plots has total estimators of variables included in vector (2.5.2) balanced on the totals of every plots in S_T:

$$\sum_{i \in S_T} \sum_{k \in S_{\cdot i}} \frac{\mathbf{x}_{2k}}{\pi_{k|i}} = \sum_{i \in S_T} \sum_{k \in U_{\cdot i}} \mathbf{x}_{2k}.$$

Moreover, as the sum of the inclusion probabilities of plots is an integer in each tile:

$$\sum_{k \in U_{\cdot i}} \pi_{k|i} = n_{\cdot i} = r, \tag{2.5.3}$$

for $i \in S_T$, there are exactly r selected plots in every tiles.

Balancing equations 1 to H determine the distribution of plot types in the sample. The one for type h is

$$\sum_{i \in S_T} \sum_{k \in S_{\cdot i}} \frac{\mathbb{1}_{k \in U_{h \cdot}} \times \pi_{k|i}}{\pi_{k|i}} \approx \sum_{i \in S_T} \sum_{k \in U_{\cdot i}} \mathbb{1}_{k \in U_{h \cdot}} \times \pi_{k|i} \approx n_{h \cdot}. \tag{2.5.4}$$

The second equality is a result of Equation (2.4.4) and

$$\sum_{k \in U_{\cdot i}} \mathbb{1}_{k \in U_{h \cdot}} \times \pi_{k|i} = \frac{N_{hi} n_{h \cdot}}{\pi_{1i} N_{h \cdot}}.$$

It ensures the selection, as exactly as possible, of $n_{h \cdot}$ plots of type h, for $h = 1, \cdots, H$. Balancing on the Landsat variables makes the sample balanced for these variables, as mentioned is Section 2.4.

The algorithm is highly stratified balanced sampling. Thus exactly $r = 4$ plots are selected in each tile. The constraint on the fixed sample size for each forest type can only be approximately satisfied because it may not be possible to find an exactly balanced design for the types in a two-stage sampling design.

2.6. High points of the design

The proposed sampling design has balancing constraints at the two stages. The totals on which the samples are balanced are the same in both stages, but the balancing variables must be defined differently for each one. At the first stage, they are totals at the tile level. At the second stage, plots are studied individually; the auxiliary variables are then defined at the plot level. Even though the units are not the same, at the end, the samples are balanced on the same totals. The techniques used for balancing have to be adapted and the balancing vectors are different at the two stages. An important fact is that the totals on which the sample is balanced at the second stage are the estimated totals in the first stage. Then, in order to exactly balance on the same totals at both stages, the first one needs to be perfectly balanced.

Equation (2.4.4) and Equation (2.5.4) are both balancing on the sample sizes of each type, $n_{h\cdot}$. The last components of vectors (2.4.2) and (2.5.2) are both for balancing on Landsat band colors. The sample sizes are managed differently however. At the first stage, balancing on the tile inclusion probabilities ensures a fixed tile sample size, see Equation (2.4.3). At the second stage, the method of Hasler and Tillé, with the inclusion probabilities of plots, is ensuring a fixed sample size per tile, see equation (2.5.3).

Since at the first stage, the distribution of types needs to be handled at the tile level, and at the second stage, the plots are directly managed, the stages of the design have distinct roles. At the first stage, the balancing equations prevent the selection of a bad sample where one type would be under represented, jeopardizing the operational constraint of having a predetermined number of plots for each forest type. At the second stage, the highly stratified balanced design proposed allows capturing the local forest variability as it favors selecting plots of several types within each tile, when they are available. To illustrate this point consider two tiles, sampled at stage 1, containing plots of types 1 and 2, where the first and second tiles contain respectively 85% and 15% of type 1 plots. Overall, one expects 4 plots of type 1 and 4 plots of type 2 to be selected from these two tiles if the two types have the same sampling fraction. Highly stratified balanced sampling favors the selection of plots of types 1 and 2 in both tiles. This is not so with other sampling methods, such as systematic sampling within tiles or simultaneous balanced sampling, for tile and forest type sample sizes, for all tiles together. Thus highly stratified balanced

sampling could reduce the intra cluster correlation typical of data collected in multi stage designs.

2.7. Spreading the sample

The study area may be characterized by large, homogeneous regions with similar forest coverage. It can be attractive to avoid selecting neighboring tiles. Indeed, it would be interesting to maximize the spreading of the sample in a systematic way as this may improve the estimations. To that end, Stevens Jr and Olsen (2004) developed Generalized Random Tessellation Stratified (GRTS) sampling to select a spread sample. According to their method, the two-dimensional map should be reduced to a one-dimensional line. The length of one sampling unit on this line is proportional to the inclusion probability of this unit. Then a systematic sample is selected and the map is reconstructed. The problem is that balancing constraints on auxiliary variables cannot be taken into account with the GRTS method. In order to respect balancing constraints while spatially spreading a sample, Grafström and Tillé (2013) proposed spatially balanced sampling, also known as doubly balanced sampling. The method allows to simultaneously balance the sample on auxiliary variables and to spread it using location variables. More, Grafström and Tillé (2013) compare their method, among others, with GRTS sampling and show that their method is more precise because of the balancing constraints. Considering the constraints of the inventory, doubly balanced sampling is way more adequate.

Let p be the number of auxiliary variables. The method of Grafström and Tillé uses an algorithm to select a cluster of $p+1$ neighboring units. This algorithm randomly selects, with equal probabilities, one unit in the population. The p closest units of the chosen one are also selected. In the next step, the mean position of this cluster of $p+1$ units is computed. The sum of squares of the distances between units of the cluster and their mean position is calculated. Then a new cluster of the $p+1$ nearest units to this mean position is selected. The mean position of this new cluster is computed. This resampling process is repeated until the sum of squares of the distances between units of the cluster and their mean position is minimal. The $p+1$ units in the final cluster are in the same neighborhood.

The first step of the method of Grafström and Tillé is to select a cluster of $p+1$ units in the population with the above algorithm. Then a flight phase of the cube method is run into this cluster. Because the number of balancing equations is p, a minimum of one unit is either selected or rejected in this cluster. Moreover, the balancing Equation (2.4.1) is still respected. Next, another cluster of $p+1$ units is selected with the proposed algorithm, but this time, only among the set of units which are not yet selected or rejected. A flight phase is run in this new cluster and at least one unit has a decided sampling outcome. The experience is repeated until less than $p+1$ units are neither selected nor rejected. That is, flight phases are

run into different clusters of $p+1$ neighboring units (chosen with the algorithm) until the number of units with non integer inclusion probability is less than $p+1$. Then a landing phase is run into this group to finalize the sample. The resulting sample is as balanced as possible and well spread. Nearby units have small joint inclusion probabilities of selection because these probabilities are updated locally, see Grafström and Tillé (2013).

In order to spatially spread the sample, the method of Grafström and Tillé can be integrated in the sampling design proposed in this paper. The spreading techniques are incorporated to both stages. At the first stage of this design, a doubly balanced sample of tiles is selected to ensure a dispersion of the sampled tiles all over the inventoried territory. From this sample, a highly stratified doubly balanced sample of plots is selected. This allows, if possible, the selection of scattered plots in sampled tiles. Then the requirements presented in Section 2.3.2 are satisfied. The estimations of the auxiliary variable totals are approximately equal to their true population values. Finally, the selection of nearby tiles and nearby plots in every tile is avoided.

2.8. Numerical comparisons and illustrations

This section is an application of the sampling design presented in this paper using data of the Quebec forest inventory, obtained from a FMU located north of the 50th parallel, with an area of 2663 km^2. This FMU is represented in Figure 2.3.

2.8.1. Data of the Quebec inventory

The inventory area is about 50% of the total area, because the young forest, which is characterized by a height lower than 7m, is excluded from the survey. The photo-interpreted variables were based on pictures taken in 2010. As this is a northern forest, the dominant tree species is the black spruce (*Picea mariana*). Nine other species are found in this area including Jack pine (*Pinus banksiana*), trembling Aspen (*Populus tremuloides*), and white birch (*Betula papyrifera*).

Types of forests are constructed in terms of the forest *Cover type*, which is 89% resinous (R), 3% deciduous (F), and 8% mixed (M), and *Age*. The *Age* clusters labeled 10, 30, 50, 70 and 90 have an amplitude of 20 years. For instance, the cluster 30 has trees aged between 21 and 40 years. Trees older than 101 years old are gathered in the cluster 120. In some polygons the ages are not homogeneous. In that case, *Age* has 4 levels : JIR, JIN, VIR, VIN. Label JIR gathers trees of varying age but younger than 80 years old with various heights. Label JIN groups trees of varying age but younger than 80 years old with a homogeneous height. Labels VIR and VIN gather trees of varying age but older than 80 years old with various and homogeneous heights respectively. The forest types are defined by cross-classifying the plots according to the factor *Age* of the tree and the factor

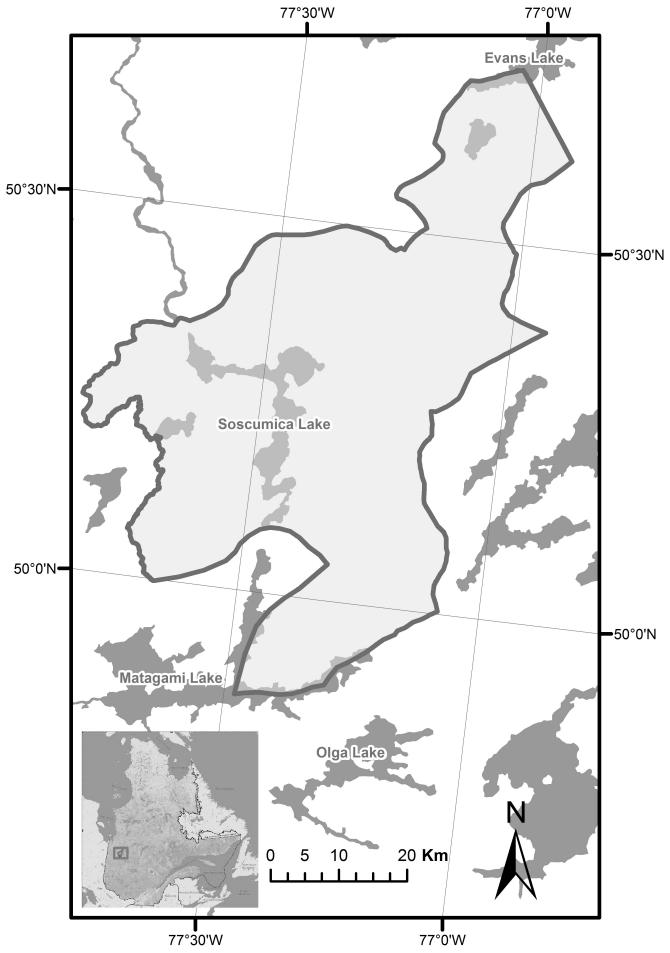

Fig. 2.3. Region of Quebec forest inventory investigated in this simulation study. This FMU is located north of the 50th parallel and represents only a fraction of the global Quebec inventory area.

Cover type. Rare types were combined with adjacent ones; this led to a total of $H = 15$ forest types, see Table 2.1.

In order to investigate the sampling properties of estimators produced by the sampling design considered in this paper, the photo-interpreted $Height$ is used as the variable of interest y. It is strongly related to the type distribution. So considering the type distribution in the sampling design should improve the precision of the estimated mean height. Landsat 5 TM satellite images acquired in June 2010 (path 18, row 25) provide additional auxiliary information. Two color band values (*band* 2 and *band* 4) are more correlated with $Height$ and will be used in the sampling design investigations. Information on the population under study is provided in Table 2.1.

2.8.2. Sampling designs in the simulation study

In order to evaluate the precision associated to the various components of the sampling design described previously, several sampling designs are implemented. The impact of using tiles as primary sampling units, of spreading and of balancing the

Tab. 2.1. Description of the population by type. The values of $n_{h\cdot}$ is a requirement of the sampling design. The number of tiles is $M = 2522$ and the number of sampled tiles is $m = 75$. Band stands for the color band provided by satellite images.

					Height		Band 2		Band 4	
Type	Age	Type	$N_{h\cdot}$	$n_{h\cdot}$	mean	sd	mean	sd	mean	sd
F-30	10, 30	F	151	11	8.5	1.1	25.4	2.2	95.6	13.3
F-70	50, 70	F	248	10	20.8	2.2	24.1	1.3	70.4	11.3
F-90	90	F	3739	28	21.6	1.7	24.0	1.2	77.6	12.9
JIN	JIN	F, M, R	71	7	9.2	1.9	25.6	1.4	64.6	7.8
JIR	JIR	F, M, R	122	9	12.6	3.5	25.0	1.5	76.8	12.4
M-30	30	M	378	15	7.8	0.9	26.3	1.9	86.2	11.9
M-70	50, 70	M	504	12	15.8	2.4	24.5	1.3	67.2	10.9
M-90	90, 120	M	1858	20	18.3	2.9	24.1	1.0	70.3	10.9
R-30	10, 30	R	265	13	7.8	1.2	27.3	1.7	70.9	8.8
R-50	50	R	318	19	8.1	1.0	27.1	1.5	65.6	7.2
R-70	70	R	24995	39	11.6	2.7	25.7	1.7	59.1	5.9
R-90	90	R	15263	36	13.7	2.2	24.7	1.5	58.8	7.4
R-120	120	R	9679	34	12.2	2.5	27.0	1.6	62.2	5.3
VIN	VIN	M, R	1379	22	11.7	2.2	26.4	1.8	61.4	5.7
VIR	VIR	F, M, R	1943	25	12.7	2.8	26.2	1.5	65.1	8.3
All			60913	300	13.1	3.7	25.5	1.8	61.7	9.2

sample on types and on color band values is evaluated. The precision of a given sample design is measured using the estimated mean of the variable *Height* and the observed sample size $n_{h\cdot}$ for forest type h.

Four sampling designs with only one stage of selection are simulated. They are obtained by selecting plots directly without considering tiles. Inclusion probabilities of plots depend on types only. Then only probabilities $\pi_k = n_{h\cdot}/N_{h\cdot}$ are necessary in these designs. In order to label the designs, the notation 1S is used for one stage sampling designs. The notation B concerns Balanced sampling, N denotes Not Balanced, T is used when there are balancing constraints on Types and C is used for balancing constraints on Color band values.

1. The first sampling design (1S-NB) is an unequal probability sampling design. Plots are selected according to a random systematic sampling design, which is presented in Tillé (2006, pp. 127-128). This design is used as a benchmark because tiles, types and color band values are not taken into account.
2. The second sampling design (1S-BT) is stratified by forest types.
3. The third sampling design (1S-BC) is balanced on color band values only. This design is used to evaluate the gain in precision caused by the color band values.

4. The fourth sampling design (1S-BCT) is balanced on types and on color band values.

Four other sampling designs with two stages of selection are implemented. At first, a sample of tile is selected. Then a sample of plots is selected through tiles. The notation 2S means two-stage sampling.

1. The first sampling design (2S-NB) is a regular two-stage sampling design. That is, a fixed number of tiles m is selected at the first stage. The only balancing vector at this step is the inclusion probability vector. At the second stage, a stratified sampling design is used to select exactly $n_{.i} = 4$ plots in the sampled tiles. Types are not considered in this sampling process.
2. The second sampling design (2S-BT) is a two-stage sampling design with balancing constraints on the types. At the first stage, balancing constraints on the sample size and on types are considered. At the second stage, balancing constraints are used to stratify and to obtain the targeted sample sizes by types.
3. The third sampling design (2S-BC) is a two-stage sampling design balanced on color band values.
4. The fourth sampling design (2S-BCT) is the one described in sections 2.4 and 2.5. That is, the balancing vector at the first stage is Equation (2.4.2). The balancing vector at the second stage is Equation (2.5.2).

To evaluate the gain of spreading the sample, the eight sampling designs are realized twice: once without spreading and once with spatially balanced sampling in every stages. It leads to a total of sixteen different simulated designs presented in Table 2.2.

Tab. 2.2. Studied designs. The labels are the following: one stage (1S), two stages (2S), not balanced (NB), balanced on type (BT), balanced on colors (BC), balanced on types and colors (BCT), spread (S)

		Auxiliary information used			
Sampling design	Label	Tiles	Types	Colors	Spreading
Unequal probabilities not balanced	1S-NB				
Unequal probabilities not balanced and spread	1S-NB-S				✓
Stratified on types	1S-BT		✓		
Stratified on types and spread	1S-BT-S		✓		✓
Balanced on colors	1S-BC			✓	
Balanced on colors and spread	1S-BC-S			✓	✓
Balanced on colors and types	1S-BCT		✓	✓	
Balanced on colors and types, spread	1S-BCT-S		✓	✓	✓
Two-stage	2S-NB	✓			
Two-stage and spread	2S-NB-S	✓			✓
Two-stage, balanced on types	2S-BT	✓	✓		
Two-stage, balanced on types and spread	2S-BT-S	✓	✓		✓
Two-stage, balanced on colors	2S-BC	✓		✓	
Two-stage, balanced on colors and spread	2S-BC-S	✓		✓	✓
Two-stage, balanced on colors and types	2S-BCT	✓	✓	✓	
Two-stage, balanced on colors and types, spread	2S-BCT-S	✓	✓	✓	✓

2.8.3. Comparison of the sampling designs

Two aspects of the sixteen sampling designs are studied: the ability to obtain exactly $n_{h\cdot}$ plots of type h, for $h = 1, \ldots, H$, and the precision of the estimator of the *Height* mean. To that end, for each of the sixteen sampling designs presented in Section 2.8.2, a total of $R = 10,000$ samples were selected. For each sample, the observed number of plots of each type is noted. In order to satisfy the requirements, the number of plots of each type must be near the values $n_{h\cdot}$, for $h = 1, \ldots, H$, given in Table 2.1. In a sample, the estimator of the mean height is also computed,

$$\widehat{\overline{Y}} = \frac{\sum_{k \in S} y_k/\pi_k}{\sum_{k \in S} 1/\pi_k},$$

where S is the sample of plots, y_k is the height value for plot k and π_k is defined in Section 2.3.

The Monte Carlo relative root mean square error (RRMSE) is used to compare sampling designs. For estimator $\widehat{\theta}$, it is calculated as

$$\text{RRMSE}_{MC}\left(\widehat{\theta}\right) = \frac{1}{\theta}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}^{(r)} - \theta\right)^2},$$

where $\widehat{\theta}^{(r)}$ is the value of $\widehat{\theta}$ for the r-th sample. To evaluated the extent to which the constraint of having $n_{h\cdot}$ plots of type h, $h = 1, \ldots, H$ is met, the Monte Carlo RRMSEs of the observed number of plots of type h is computed for the sixteen sampling designs under study. The precision of the mean *Height* estimator $\widehat{\overline{Y}}$, is also evaluated using a Monte Carlo RRMSE for each sampling design.

2.8.4. Results

Table 2.3 presents the Monte Carlo RRMSE of the number of plots of each type for every sampling designs. In this table, each row corresponds to a different type. There are two rows for each type. The first row contains the RRMSE for the sampling designs without spreading. The second row contains the RRMSE for the sampling designs with spreading. The columns correspond to the eight sampling designs presented in Table 2.2.

Clearly, the RRMSE values for the stratified sampling designs (1S-BT and 1S-BCT) are null. Indeed, these designs select exactly $n_{h\cdot}$ plots of each type, where $h = 1, \ldots, H$. For the unequal probability sampling, the RRMSE is going from 0.13 to 0.36. By balancing on color band values, the variability decreases a little with the RRMSE going from 0.06 to 0.35.

By using two-stage sampling designs, the $n_{h\cdot}$ are more scattered. The two-stage design is always less stable than the other designs, with RRMSE values going from 0.18 to 0.55. The use of balancing constraints on the types reduces the variation of the number of plots of each type, with RRMSE values ranging between 0.04 to 0.23,

Tab. 2.3. RRMSE of number of plots of each type for sixteen sampling designs. For each type, the first row corresponds to the non-spread design and the second row to the spread design.

	Sampling design							
	One stage (1S)				Two stages (2S)			
Type	1S-NB	1S-BT	1S-BC	1S-BCT	2S-NB	2S-BT	2S-BC	2S-BCT
F_30	0.29	0.00	0.28	0.00	0.42	0.15	0.40	0.15
	0.22	0.00	0.20	0.00	0.36	0.15	0.33	0.15
F_70	0.31	0.00	0.30	0.00	0.48	0.15	0.47	0.15
	0.24	0.00	0.21	0.00	0.40	0.14	0.36	0.14
F_90	0.18	0.00	0.17	0.00	0.25	0.04	0.18	0.04
	0.16	0.00	0.14	0.00	0.22	0.04	0.16	0.04
JIN	0.36	0.00	0.35	0.00	0.55	0.23	0.54	0.23
	0.26	0.00	0.24	0.00	0.47	0.23	0.42	0.23
JIR	0.32	0.00	0.32	0.00	0.47	0.17	0.46	0.17
	0.26	0.00	0.24	0.00	0.44	0.17	0.41	0.17
M_30	0.25	0.00	0.24	0.00	0.37	0.11	0.36	0.11
	0.19	0.00	0.18	0.00	0.32	0.11	0.30	0.11
M_70	0.28	0.00	0.27	0.00	0.42	0.11	0.42	0.11
	0.25	0.00	0.24	0.00	0.39	0.12	0.38	0.12
M_90	0.22	0.00	0.21	0.00	0.31	0.06	0.28	0.06
	0.19	0.00	0.18	0.00	0.27	0.06	0.24	0.06
R_30	0.26	0.00	0.26	0.00	0.40	0.12	0.38	0.12
	0.22	0.00	0.20	0.00	0.35	0.13	0.32	0.13
R_50	0.21	0.00	0.21	0.00	0.37	0.11	0.35	0.11
	0.14	0.00	0.12	0.00	0.27	0.10	0.24	0.10
R_70	0.15	0.00	0.09	0.00	0.22	0.07	0.15	0.07
	0.14	0.00	0.06	0.00	0.19	0.07	0.12	0.07
R_90	0.15	0.00	0.15	0.00	0.23	0.05	0.21	0.05
	0.13	0.00	0.10	0.00	0.18	0.05	0.14	0.05
R_120	0.16	0.00	0.16	0.00	0.26	0.09	0.24	0.09
	0.14	0.00	0.12	0.00	0.22	0.09	0.20	0.09
VIN	0.20	0.00	0.20	0.00	0.29	0.07	0.28	0.07
	0.19	0.00	0.17	0.00	0.27	0.07	0.25	0.07
VIR	0.19	0.00	0.19	0.00	0.27	0.05	0.26	0.05
	0.17	0.00	0.16	0.00	0.25	0.06	0.23	0.06
Minimum	0.15	0.00	0.09	0.00	0.22	0.04	0.15	0.04
	0.13	0.00	0.06	0.00	0.18	0.04	0.12	0.04
Median	0.22	0.00	0.21	0.00	0.37	0.11	0.35	0.11
	0.19	0.00	0.18	0.00	0.27	0.10	0.25	0.10
Maximum	0.36	0.00	0.35	0.00	0.55	0.23	0.54	0.23
	0.26	0.00	0.24	0.00	0.47	0.23	0.42	0.23

Tab. 2.4. RRMSE of mean estimator $\widehat{\overline{Y}}$

	RRMSE	
Design	Not spread	spread
1S-NB	0.020	0.018
1S-BT	0.016	0.015
1S-BC	0.018	0.016
1S-BCT	0.015	0.013
2S-NB	0.027	0.024
2S-BT	0.020	0.019
2S-BC	0.022	0.020
2S-BCT	0.018	0.017

for designs 2S-BT and 2S-BCT. On the other hand, balancing on the color band values and spreading do not seem helpful in terms of the stability of the observed $n_{h\cdot}$.

Table 2.4 presents the Monte Carlo RRMSEs of the estimator $\widehat{\overline{Y}}$. In this table, rows present the eight sampling designs. The first column contains RRMSE values of designs without spreading. The second column contains values for spread designs. The stratification variable *Type* is very linked to variable *Height*. Consequently, the gain in accuracy is very important for the estimator $\widehat{\overline{Y}}$ when designs are stratified. These designs have a RRMSE going from 0.013 to 0.016. The unequal probability sampling has a RRMSE value of 0.020.

There is always a loss of accuracy when using a two-stage design. This 'cluster effect' is associated to a within tile correlation of the studied variable *Height*. By using a two-stage sampling design without balancing variables, the RRMSE of $\widehat{\overline{Y}}$ is very large (0.027). Balancing on the types reduces the RRMSE value to 0.020. Balancing on color band values and spreading does seem to improve a little the precision of the estimator $\widehat{\overline{Y}}$.

2.9. Conclusion

The simulation study presented has shown that complex method can be used to elaborate a sampling design even though there are several requirements. By using balanced sampling, the requirements of the Quebec inventory can be satisfied very well. In a two-stage sampling design, it is possible to select exactly 4 plots in every sampled tiles and to satisfy a large set of balancing constraints.

The 'cluster effect', i.e. the loss of accuracy when using two-stage sampling is not inevitable. Our simulations show that balanced sampling enables us to maintain the advantage of stratification through the two stages. Considering constraints on the types provides an estimation of the mean height almost as good as with a usual stratified sample. Moreover the integration of other balancing variables as colors or

of spreading can be integrated in the sampling design in order to plan a very efficient data collection.

Chapter 3

Revisiting Variance Decomposition when Independent Samples Intersect

Abstract: The variance and the estimated variance of the expanded estimator in the intersection of two independent samples can be decomposed into two ways. Due to the inclusion probabilities, it is generally more practical to compute the variance with one decomposition. With the other one, it is more convenient to estimate the variance. [1]

Keywords: Reverse approach; Two-phase sampling; Two-stage sampling; Variance estimation.

3.1. Introduction

In survey sampling, when the sampling designs include several stages or phases, variance estimation suddenly becomes much more intricate. In two-stage sampling, when secondary units are selected in a sample of primary units, one already obtains a curious result. The variance of the expanded estimator and the estimator of the variance can both be decomposed into two terms. However each term of the estimator does not estimate the corresponding term of the variance (see among others Särndal et al., 1992, pp. 137-139). Beaumont et al. (2015) linked the variance estimator to the reverse approach of the decomposition of the variance. However, these authors did not explain the fact that the variance is obtained from a different conditioning than the variance estimator. Variance decomposition is also crucial in nonresponse because questionnaire nonresponse can be modeled as a second phase of sampling. Several options exist to estimate the variance with nonresponse as the two-phase approach (Särndal, 1992) and the reverse approach (Fay, 1991; Shao and Steel, 1999).

In this paper, we discuss the variance and its estimation in samples that are the intersection of two independent samples. We show that two different decompositions

[1]This article is a reprint of Tillé, Y. and Vallée, A.-A. (2017). Revisiting variance decomposition when independent samples intersect. *Statistics & Probability Letters*, 130:71–75

of the variance can be obtained. One of them is more interesting for variance estimation. This is explained by possible simplifications of the joint inclusion probabilities of one of the two samples.

3.2. General case

We consider the case where two independent samples intersect. Define $U = \{1, \ldots, N\}$ a population of size N and let s^A and s^B be samples of U. Two sampling designs are defined on U, say $p^A(s^A)$ and $p^B(s^B)$ such that $p^A(s^A) \geq 0$, $p^B(s^B) \geq 0$,

$$\sum_{s^A \subset U} p^A(s^A) = 1 \text{ and } \sum_{s^B \subset U} p^B(s^B) = 1.$$

Define the two random samples S^A and S^B such that $\text{pr}(S^A = s^A) = p^A(s^A)$ and $\text{pr}(S^B = s^B) = p^B(s^B)$. The two random samples are assumed to be independent in the sense that $\text{pr}(S^A = s^A, S^B = s^b) = p^A(s^A)p^B(s^B)$.

Let I_k^A and I_k^B be respectively the indicator variables of the presence of unit k in sample S^A and S^B. The first-order inclusion probabilities are $\pi_k^A = \text{E}(I_k^A) = \text{pr}(k \in S^A)$ and $\pi_k^B = \text{E}(I_k^B) = \text{pr}(k \in S^B)$. The joint inclusion probabilities are $\pi_{k\ell}^A = \text{E}(I_k^A I_\ell^A) = \text{pr}(k, \ell \in S^A)$ and $\pi_{k\ell}^B = \text{E}(I_k^B I_\ell^B) = \text{pr}(k, \ell \in S^B)$, with $\pi_{kk}^A = \pi_k^A$ and $\pi_{kk}^B = \pi_k^B$, for $k, \ell \in U$. Moreover define $\Delta_{k\ell}^A = \pi_{k\ell}^A - \pi_k^A \pi_\ell^A$ and $\Delta_{k\ell}^B = \pi_{k\ell}^B - \pi_k^B \pi_\ell^B$. Consider the sampling design obtained by intersecting two independent samples $S = S^A \cap S^B$. Due to the independence, we have $I_k = I_k^A I_k^B$, $\pi_k = \text{pr}(k \in S) = \pi_k^A \pi_k^B$ and $\pi_{k\ell} = \text{pr}(k, \ell \in S) = \pi_{k\ell}^A \pi_{k\ell}^B$. Next define $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell, k, \ell \in U$.

Beaumont and Haziza (2016) define a two-phase sampling design as strongly invariant if $\text{pr}(S^B|S^A) = \text{pr}(S^B)$, where S^A is the first phase sample and S^B is the second phase sample. In other words, the selection of the second phase sample does not depend on the selection of the first phase sample, which means that the two samples are independent. This definition does not contain the two-phase sampling design as defined for instance by Särndal and Swensson (1987). Indeed, these authors admit that the second phase of the design can depend on the first phase, that is π_k^B and $\pi_{k\ell}^B$ are functions of S^A. In this case, the designs are not independent and the theory below does not apply.

The two-stage design can be seen as a specific case of the two-phase design that is strongly invariant. The first stage corresponds to the selection of primary units, e.g. municipalities regrouping households, and the second stage consists in selecting the secondary units, e.g. households. Särndal et al. (1992, pp. 137-139) explain that a two-stage sampling design must satisfy the principles of invariance and independence. For these authors, invariance means that the selection of the secondary units of the second stage does not depend on the first stage. Independence means that the secondary units are selected independently from one primary unit to another one. The definition of strongly invariant samples of Beaumont and Haziza

(2016) corresponds to the invariance of Särndal et al. (1992, pp. 137-139). Two-stage sampling can be viewed as the intersection of two independent samples selected by a cluster design and a stratified design. In the cluster sample, a set of clusters, which correspond to the primary units, is selected. All the secondary units in this set are therefore selected. In the stratified sample, one set of secondary units is selected per stratum, where a stratum corresponds to a primary unit. The intersection of this stratified sample and the chosen secondary units of the cluster sample is a two-stage sample. Samples of secondary units are selected in sampled primary units.

Another specific case of independent samples is questionnaire nonresponse. A sample is selected in the population and some units in this sample are respondents, the others are nonrespondents. The sample of units in the population is seen as a first sample, selected according to a sampling design $p^A(.)$. The set of respondents is seen as a second sample which is independent from the first one. Moreover the second sample $p^B(.)$ is in general assumed to be a Poisson design, which means that $\Delta^B_{k\ell} = 0$ when $k \neq \ell$.

3.3. Estimation and variance estimation

Suppose that the variable of interest y takes value y_k on unit k of the population. The variable is observed on units selected in a strongly invariant two-phase sample $S = S^A \cap S^B$. In order to estimate the total

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

one can use the expanded estimator

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

(Narain, 1951; Horvitz and Thompson, 1952).

The variance of $\widehat{Y}$ is

$$\operatorname{var}(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}$$

and can be unbiasedly estimated by

$$\widehat{\mathrm{v}}(\widehat{Y}) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}}.$$

The delicate elements are the decompositions of $\Delta_{k\ell}$ and $\Delta_{k\ell}/\pi_{k\ell}$. They need to be decomposed in function of $p^A(.)$ and $p^B(.)$ by means of the law of total variance. With this law, the variance of a random variable can be decomposed conditionally to another random variable. Consider for instance two random variables x_1 and x_2, the variance of x_1 is decomposed as

$$\operatorname{var}(x_1) = \mathrm{E} \operatorname{var}(x_1|x_2) + \operatorname{var} \mathrm{E}(x_1|x_2).$$

This can be extended to a covariance. Consider a third random variable x_3. The covariance between x_1 and x_2 can be decomposed as

$$\operatorname{cov}(x_1, x_2) = \mathrm{E}\operatorname{cov}(x_1, x_2|x_3) + \operatorname{cov}\left[\mathrm{E}(x_1|x_3), \mathrm{E}(x_2|x_3)\right].$$

For $\Delta_{k\ell}$, there are two possible decompositions. The usual one consists in using the law of total variance by conditioning with respect to S^A:

$$\begin{aligned}\Delta_{k\ell} &= \operatorname{cov}(I_k, I_\ell) = \mathrm{E}\operatorname{cov}(I_k, I_\ell \mid S^A) + \operatorname{cov}\left[\mathrm{E}(I_k \mid S^A), \mathrm{E}(I_\ell \mid S^A)\right] \\ &= \Delta^B_{k\ell}\pi^A_{k\ell} + \pi^B_k\pi^B_\ell\Delta^A_{k\ell}. \end{aligned} \tag{3.3.1}$$

The reverse decomposition consists in conditioning with respect to S^B:

$$\begin{aligned}\Delta_{k\ell} &= \operatorname{cov}(I_k, I_\ell) = \mathrm{E}\operatorname{cov}(I_k, I_\ell \mid S^B) + \operatorname{cov}\left[\mathrm{E}(I_k \mid S^B), \mathrm{E}(I_\ell \mid S^B)\right] \\ &= \Delta^A_{k\ell}\pi^B_{k\ell} + \pi^A_k\pi^A_\ell\Delta^B_{k\ell}.\end{aligned}$$

This reverse approach is usable because the sampling design is strongly invariant, we can inverse the order of the two phases.

We can also obtain two decompositions for the ratio $\Delta_{k\ell}/\pi_{k\ell}$. The usual one is

$$\frac{\Delta_{k\ell}}{\pi_{k\ell}} = \frac{\Delta^B_{k\ell}}{\pi^B_{k\ell}} + \frac{\pi^B_k\pi^B_\ell\Delta^A_{k\ell}}{\pi^A_{k\ell}\pi^B_{k\ell}},$$

and the reverse one is

$$\frac{\Delta_{k\ell}}{\pi_{k\ell}} = \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + \frac{\pi^A_k\pi^A_\ell\Delta^B_{k\ell}}{\pi^A_{k\ell}\pi^B_{k\ell}}. \tag{3.3.2}$$

The decompositions obtained by conditioning with respect to S^A and with respect to S^B are obviously equal.

3.4. Poisson sampling

A sample with nonresponse is often seen as a two-phase sampling design. The first phase is the sampling design of the survey. The second phase is a Poisson sampling design corresponding to the nonresponse phase. Two-phase sampling with a Poisson sample at the second phase is strongly invariant and there are two decompositions of $\Delta_{k\ell}$. If S^B is selected according to a Poisson design, $\pi^B_{k\ell} = \pi^B_k\pi^B_\ell + I\{k = \ell\}\pi^B_k(1-\pi^B_k)$ and thus $\Delta^B_{k\ell} = I\{k = \ell\}\pi^B_k(1-\pi^B_k)$, for all $k, \ell \in U$, where $I\{A\}$ equals 1 if A is true and 0 otherwise. The quantity $\Delta^B_{k\ell}$ vanishes when $k \neq \ell$ and only depends on the first-order inclusion probabilities.

The usual decomposition, called the two-phase approach in nonresponse theory, for $\Delta_{k\ell}$ is

$$\Delta_{k\ell} = I\{k = \ell\}\pi^B_k(1-\pi^B_k)\pi^A_{k\ell} + \Delta^A_{k\ell}\pi^B_k\pi^B_\ell,$$

and the reverse approach is

$$\begin{aligned}\Delta_{k\ell} &= \Delta^A_{k\ell}\left[\pi^B_k\pi^B_\ell + I\{k=\ell\}\pi^B_k(1-\pi^B_k)\right] + \pi^A_k\pi^A_\ell I\{k=\ell\}\pi^B_k(1-\pi^B_k)\\ &= \Delta^A_{k\ell}\pi^B_k\pi^B_\ell + \pi^A_k(1-\pi^A_k)I\{k=\ell\}\pi^B_k(1-\pi^B_k) + \pi^A_k\pi^A_\ell I\{k=\ell\}\pi^B_k(1-\pi^B_k)\\ &= \Delta^A_{k\ell}\pi^B_k\pi^B_\ell + I\{k=\ell\}\pi^B_k(1-\pi^B_k)\pi^A_k.\end{aligned}$$

The results are equal but the computation of the usual decomposition is simpler and faster than the computation with the reverse one.

The usual decomposition for the ratio $\Delta_{k\ell}/\pi_{k\ell}$ is:

$$\begin{aligned}\frac{\Delta_{k\ell}}{\pi_{k\ell}} &= I\{k=\ell\}(1-\pi^B_k) + \frac{\pi^B_k\pi^B_\ell\Delta^A_{k\ell}}{\pi^A_{k\ell}\left[\pi^B_k\pi^B_\ell + I\{k=\ell\}\pi^B_k(1-\pi^B_k)\right]}\\ &= I\{k=\ell\}(1-\pi^B_k) + \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + I\{k=\ell\}\pi^B_k(1-\pi^A_k) - I\{k=\ell\}(1-\pi^A_{k\ell})\\ &= I\{k=\ell\}\pi^A_k(1-\pi^B_k) + \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}}.\end{aligned}$$

The reverse decomposition, is straightforward:

$$\frac{\Delta_{k\ell}}{\pi_{k\ell}} = \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + I\{k=\ell\}\pi^A_k(1-\pi^B_k).$$

This result is interesting. While the usual decomposition of $\Delta_{k\ell}$ is more direct than the reverse decomposition, the reverse decomposition of $\Delta_{k\ell}/\pi_{k\ell}$ is more straightforward than the usual one.

The reverse decomposition is used in the reverse approach to estimate the variance in the presence of nonresponse proposed by Fay (1991) and Shao and Steel (1999). This approach often simplifies the estimation of variance under nonresponse. The estimator of variance becomes

$$\widehat{\mathrm{v}}(\widehat{Y}) = \sum_{k\in S}\sum_{\ell\in S}\frac{y_k y_\ell}{\pi_k\pi_\ell}\frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + \sum_{k\in S}\frac{y_k^2}{(\pi_k)^2}\pi^A_k(1-\pi^B_k).$$

If π^B_k is the probability of response, it must obviously be estimated and plugged into the variance estimator.

3.5. Two-stage sampling

Consider the strongly invariant two-stage sampling design as defined in Beaumont and Haziza (2016). The population is partitioned into M subsets $U_1,\ldots,U_i,\ldots,U_M$. A sample of primary units S_1 is a list of m randomly selected primary units. The sample S^A is the union of the secondary units in the m primary units. All the units of the selected subsets are in sample $S^A = \bigcup_{i\in S_1} U_i$. Define π^1_i the probability of selecting the primary unit i, π^1_{ij} the probability of selecting primary units i and j together with $\pi^1_{ii} = \pi^1_i$ and $\Delta^1_{ij} = \pi^1_{ij} - \pi^1_i\pi^1_j$ $(i = 1,\ldots,M;\ j = 1,\ldots,M)$. Then $\pi^A_k = \pi^1_i$ if unit k is in subset U_i, $\pi^A_{k\ell} = I\{k,\ell\in U_i\}\pi^1_i + I\{k\in U_i\}I\{\ell\in U_j\}\pi^1_{ij}$,

where $i \neq j$ and $\Delta^A_{k\ell} = I\{k,\ell \in U_i\}\pi^1_i(1-\pi^1_i) + I\{k \in U_i\}I\{\ell \in U_j\}(\pi^1_{ij} - \pi^1_i\pi^1_j)$. Sample S^B is a stratified sample where a stratum is a subset U_i $(i = 1,\ldots,M)$. Subsamples are selected in each subset U_i independently from one subset to another one. Then $\pi^B_{k\ell} = I\{k,\ell \in U_i\}\pi^B_{k\ell} + I\{k \in U_i\}I\{\ell \in U_j\}\pi^B_k\pi^B_\ell$ and $\Delta^B_{k\ell} = I\{k,\ell \in U_i\}(\pi^B_{k\ell} - \pi^B_k\pi^B_\ell)$. The two-stage sample is the intersection $S = S^A \cap S^B$.

The expanded estimator can be written

$$\widehat{Y} = \sum_{i\in S}\frac{y_k}{\pi_k} = \sum_{i\in S_1}\frac{\widehat{Y}_i}{\pi^1_i},$$

where

$$\widehat{Y}_i = \sum_{k\in U_i\cap S^B}\frac{y_k}{\pi^B_k}$$

is an unbiased estimator of

$$Y_i = \sum_{k\in U_i} y_k$$

when subset U_i is selected at the first stage.

The usual decomposition is:

$$\Delta_{k\ell} \;=\; I\{k,\ell \in U_i\}\pi^1_i\Delta^B_{k\ell} + \Delta^A_{k\ell}\pi^B_k\pi^B_\ell.$$

The reverse decomposition is:

$$\begin{aligned}
\Delta_{k\ell} &= I\{k,\ell \in U_i\}(\pi^1_i)^2\Delta^B_{k\ell} + I\{k,\ell \in U_i\}\pi^1_i(1-\pi^1_i)\pi^B_{k\ell}\\
&\quad + I\{k \in U_i\}I\{\ell \in U_j\}\Delta^1_{ij}\pi^B_k\pi^B_\ell\\
&= I\{k,\ell \in U_i\}\left[\pi^B_{k\ell}\pi^1_i - \pi^B_k\pi^B_\ell(\pi^1_i)^2 + \pi^B_k\pi^B_\ell\pi^1_i - \pi^B_k\pi^B_\ell\pi^1_i\right]\\
&\quad + I\{k \in U_i\}I\{\ell \in U_j\}\Delta^1_{ij}\pi^B_k\pi^B_\ell\\
&= I\{k,\ell \in U_i\}\Delta^B_{k\ell}\pi^1_i + I\{k,\ell \in U_i\}\pi^B_k\pi^B_\ell\pi^1_i(1-\pi^1_i)\\
&\quad + I\{k \in U_i\}I\{\ell \in U_j\}\Delta^1_{ij}\pi^B_k\pi^B_\ell\\
&= I\{k,\ell \in U_i\}\Delta^B_{k\ell}\pi^1_i + \Delta^A_{k\ell}\pi^B_k\pi^B_\ell.
\end{aligned}$$

The usual decomposition of $\Delta_{k\ell}/\pi_{k\ell}$ is

$$\begin{aligned}
\frac{\Delta_{k\ell}}{\pi_{k\ell}} &= I\{k,\ell \in U_i\}\frac{\Delta^B_{k\ell}}{\pi^B_{k\ell}} + I\{k,\ell \in U_i\}\frac{\pi^1_i(1-\pi^1_i)\pi^B_k\pi^B_\ell}{\pi^1_i\pi^B_{k\ell}}\\
&\quad + I\{k \in U_i\}I\{\ell \in U_j\}\frac{\Delta^1_{ij}\pi^B_k\pi^B_\ell}{\pi^1_{ij}\pi^B_k\pi^B_\ell}\\
&= I\{k,\ell \in U_i\}\left[\frac{\Delta^B_{k\ell}}{\pi^B_{k\ell}} - (1-\pi^1_i)\frac{\Delta^B_{k\ell}}{\pi^B_{k\ell}}\right] + \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}}\\
&= \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + I\{k,\ell \in U_i\}\frac{\pi^1_i\Delta^B_{k\ell}}{\pi^B_{k\ell}}
\end{aligned}$$

and the reverse one is

$$\frac{\Delta_{k\ell}}{\pi_{k\ell}} \;=\; \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + I\{k,\ell \in U_i\}\frac{(\pi^1_i)^2\Delta^B_{k\ell}}{\pi^1_i\pi^B_{k\ell}} = \frac{\Delta^A_{k\ell}}{\pi^A_{k\ell}} + I\{k,\ell \in U_i\}\frac{\pi^1_i\Delta^B_{k\ell}}{\pi^B_{k\ell}}.$$

The usual decomposition leads to the following variance:

$$\begin{aligned}\operatorname{var}\left(\sum_{i\in S_1}\frac{\widehat{Y}_i}{\pi_i}\right) &= \operatorname{E}\operatorname{var}\left(\sum_{i\in S_1}\frac{\widehat{Y}_i}{\pi_i^1}\;\middle|\; S_1\right)+\operatorname{var}\operatorname{E}\left(\sum_{i\in S_1}\frac{\widehat{Y}_i}{\pi_i^1}\;\middle|\; S_1\right)\\ &= \sum_{i=1}^{M}\frac{\operatorname{var}\left(\widehat{Y}_i\right)}{\pi_i^1}+\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{Y_iY_j}{\pi_i^1\pi_j^1}\Delta_{ij}^1. \qquad (3.5.1)\end{aligned}$$

The two terms are in general interpreted as the variance due to the first and the second stage respectively, but this interpretation is probably misleading. Indeed, if we use the reverse approach, we obtain a completely different variance decomposition:

$$\begin{aligned}\operatorname{var}\left(\sum_{i\in S_1}\frac{\widehat{Y}_i}{\pi_i^1}\right) &= \operatorname{E}\operatorname{var}\left(\sum_{i\in S_1}\frac{\widehat{Y}_i}{\pi_i^1}\;\middle|\; S^B\right)+\operatorname{var}\operatorname{E}\left(\sum_{i\in S_1}\frac{\widehat{Y}_i}{\pi_i^1}\;\middle|\; S^B\right)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{\operatorname{E}\left(\widehat{Y}_i\widehat{Y}_j\right)}{\pi_i^1\pi_j^1}\Delta_{ij}^1+\sum_{i=1}^{M}\operatorname{var}\left(\widehat{Y}_i\right) \qquad (3.5.2)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{Y_iY_j}{\pi_i^1\pi_j^1}\Delta_{ij}^1+\sum_{i=1}^{M}\frac{\operatorname{var}\left(\widehat{Y}_i\right)}{\pi_i^1}(1-\pi_i^1)+\sum_{i=1}^{M}\operatorname{var}\left(\widehat{Y}_i\right)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{Y_iY_j}{\pi_i^1\pi_j^1}\Delta_{ij}^1+\sum_{i=1}^{M}\frac{\operatorname{var}\left(\widehat{Y}_i\right)}{\pi_i^1}.\end{aligned}$$

Expression (3.5.2) can be expanded because $\operatorname{E}\left(\widehat{Y}_i\widehat{Y}_j\right)=Y_iY_j+I\{i=j\}\operatorname{var}(\widehat{Y}_i)$.

The reverse decomposition can seem more intricate. Nevertheless, the unbiased estimator of the variance is

$$\widehat{\mathrm{v}}(\widehat{Y})=\sum_{i\in S_1}\sum_{j\in S_1}\frac{\widehat{Y}_i\widehat{Y}_j}{\pi_i^1\pi_j^1}\frac{\Delta_{ij}^1}{\pi_{ij}^1}+\sum_{i\in S_1}\frac{\widehat{\mathrm{v}}\left(\widehat{Y}_i\right)}{\pi_i^1} \qquad (3.5.3)$$

(see for instance Särndal et al., 1992, pp. 137-139). The first term of the estimator of variance is not an unbiased estimator of the second term of the usually used variance (3.5.1). The same happens with the second term of the estimated variance and the first term of variance (3.5.1). The reverse decomposition is not usually used to calculate the variance of the two-stage estimator. As noticed by Beaumont et al. (2015), the interesting thing about this decomposition is that clearly, estimator (3.5.3) is an unbiased estimator of Expression (3.5.2). Each term of estimator (3.5.3) estimates each term of variance (3.5.2) without bias. The reverse approach is then more appropriate to rapidly find an estimator of the variance.

The two terms of (3.5.1) are often interpreted as variances corresponding to the first and second stages. This interpretation is in fact misleading. The two terms of (3.5.1) are just the ones obtained by one of the two decompositions given by the law of total variance.

3.6. Detecting the briefest decomposition

For the two-phase sampling with Poisson sampling at the second phase and for the two-stage sampling, the usual variance decomposition is more straightforward than the reverse one. Oppositely, for the variance estimation, the reverse approach is simpler. A common aspect of the two sampling designs is that the element $\Delta^B_{k\ell}$ vanishes in many situations. For the two-phase sampling with Poisson sampling at the second phase, $\Delta^B_{k\ell} = I\{k = \ell\}\pi^B_k(1 - \pi^B_k)$ is non-null only when $k = \ell$. For the two-stage sampling, $\Delta^B_{k\ell} = I\{k, \ell \in U_i\}(\pi^B_{k\ell} - \pi^B_k\pi^B_\ell)$ is non-null only when units k and ℓ are in the same primary unit.

When most of the $\Delta^B_{k\ell}$ are null, the simplest decomposition occurs when $\Delta^B_{k\ell}$ is multiplied or divided by $\pi^A_{k\ell}$ and not by $\pi^A_k\pi^A_\ell$. For the decomposition of the variance, this is the case for the usual approach given in Expression (3.3.1). For the estimation of variance, this is the case for the reverse approach given in Expression (3.3.2). A convenient way of computing the variance can thus be misleading when estimating the variance.

Chapter 4

Linearization for Variance Estimation by Means of Sampling Indicators: Application to Nonresponse

Abstract: In presence of nonresponse, it is usual to impute missing values, to reweight responding units and to adapt the estimators accordingly. The computation of the precision of the estimators becomes rapidly complex, it must take into account the sampling design, the treatment and refinement of the estimators. In the absence of nonresponse, it is possible to linearize estimators with respect to the sampling indicators to compute explicit variance estimators. In this paper, we extend this linearization method to deal with nonresponse. It becomes particularly straightforward to compute explicit variance estimators. Some known results are revisited in a simpler way than the usual developments and new results for complex estimators are proposed. A simulation study evaluates the proposed methodology. [1]

Keywords: calibration; imputation; response indicator; reverse approach; reweighting.

4.1. Introduction

In survey sampling, variance estimation is a delicate matter. It depends on the sampling design, the complexity of the parameter, the nonresponse, the imputation and the reweighting processes. An important part of the literature on survey sampling is devoted to this problem. Resampling methods like jackknife (Quenouille, 1949) or bootstrap (Efron, 1979) were adapted to complexe estimators (see among others Rao and Shao, 1992; Booth et al., 1994; Shao and Sitter, 1996; Berger and Skinner, 2005; Antal and Tillé, 2011; Beaumont and Patak, 2012). These methods may be demanding computationnally and explicit variance estimators may be

[1]This article is a reprint of Vallée, A.-A. and Tillé, Y. (2019). Linearisation for variance estimation by means of sampling indicators: application to non-response. *International Statistical Review*. https://doi.org/10.1111/insr.12313

preferred. Taylor linearization have been used to approximate variance estimators explicitly. In the complete case, when all the values of a survey are observed, Woodruff (1971); Binder (1996) proposed to linearize estimators with respect to estimated totals, while Demnati and Rao (2004, 2010) linearized with respect to the sampling weights. Binder (1983); Binder and Patak (1994); Deville (1999) proposed a method using estimating equations and influence functions. Graf (2011) returned to the classical definition of Taylor linearization, presented for instance in Stuart and Ord (1994, Chapter 10), to ensure that a variance estimator can be calculated. She linearized the estimators with respect to the sampling indicators, which are dummy variables indicating if each unit is selected in the sample. All these methods are of common practice, they usually lead to similar results. For instance, Särndal (1982); Deville and Särndal (1992) proposed variance estimators for calibrated total estimators. Berger (2008); Langel and Tillé (2013) compared variance estimators for the Gini index.

In the presence of nonresponse, Kott (2006) linearizes with respect to the vector of parameters $\boldsymbol{\lambda}$ in the calibration weights. Kim and Kim (2007); Kim and Rao (2009) used estimating equations and linearized with respect to parameters in the nonresponse model and the imputation model. When the nonresponse treatment and the estimator are complex, for example when the estimator is not linear in the variable of interest, it might still be challenging to find a variance estimator with those linearization methods.

In this paper, the linearization method of Graf (2011) is extended to deal with nonresponse. The estimators are linearized with respect to the response indicators and the variable of interest. It is then possible to compute variance estimators even if the parameter and the nonresponse treatment are complex. In section 4.2, the method of Graf (2011) in the complete case is reviewed. Her results on calibrated totals are reproduced and extended to any calibrated estimator. Two frameworks used for inference with nonresponse are introduced in Section 4.3. The linearization method is extended to deal with variance estimation in the case of inference based on a nonresponse model in Section 4.4. The method is applied to revisit the results of Kott (2006) in a simpler manner and to a general class of calibrated estimators for which there was no solution with Kott's method. In Section 4.5, the linearization method dealing with inference based on an imputation model is presented and applied to any imputed estimator. Some variance estimators are evaluated through simulations in Section 4.6.

4.2. Linearization in the complete case

In order to compute variance estimators, Graf (2011) proposed to linearize estimates with respect to the sampling indicators. The linearization method is detailed in this section.

4.2.1. Graf's method

Consider a finite population U of size N. We are interested in estimating a parameter $\theta = \theta(\mathbf{y})$, where $\mathbf{y} = (y_1 \cdots y_k \cdots y_N)^\top$ and y_k is the value of the variable of interest y of unit k. A sample s of size n is randomly selected in U by means of a sampling design $p(s)$ such that $p(s) \geq 0$ for all $s \subset U$ and $\sum_{s \subset U} p(s) = 1$. Define $\mathbf{a} = (a_1 \cdots a_k \cdots a_N)^\top$, the vector of sampling indicators, where a_k is 1 if unit k is selected in the sample and 0 otherwise. The first order inclusion probability of unit k is π_k, and $\mathrm{E}_p(\mathbf{a}) = \boldsymbol{\pi} = (\pi_1 \cdots \pi_k \cdots \pi_N)^\top$, where $\mathrm{E}_p(.)$ denotes the expectation with respect to the sampling design.

Consider $\widehat{\theta} = \widehat{\theta}(\mathbf{y}, \mathbf{a})$, an estimator of $\theta = \theta(\mathbf{y})$, such that $\widehat{\theta}(\mathbf{y}, \mathbf{a})$ is twice differentiable with respect to a_ℓ, $\ell = 1, \ldots, N$. Graf (2011) proposed two ways to linearize $\widehat{\theta}$ with respect to the sampling indicators: in the neighborhood of the population parameter and in the neighborhood of the estimator. In the first case,

$$\widehat{\theta} = \widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) + \sum_{\ell \in U} \widetilde{z}_\ell (a_\ell - \pi_\ell) + R(\tau_1), \tag{4.2.1}$$

where the linearization variable is

$$\widetilde{z}_\ell = \left. \frac{\partial \widehat{\theta}}{\partial a_\ell} \right|_{\mathbf{a} = \boldsymbol{\pi}},$$

and the remainder is

$$R(\tau_1) = \frac{1}{2} \sum_{k \in U} \sum_{\ell \in U} \left. \frac{\partial^2 \widehat{\theta}}{\partial a_k \partial a_\ell} \right|_{\mathbf{a} = \tau_1 \mathbf{a} + (1 - \tau_1)\boldsymbol{\pi}} (a_k - \pi_k)(a_\ell - \pi_\ell),$$

for some $\tau_1 \in (0, 1)$ (see for instance Edwards, 1994, p. 133). In the second case,

$$\widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) = \widehat{\theta} + \sum_{\ell \in U} z_\ell (\pi_\ell - a_\ell) + R(\tau_2), \tag{4.2.2}$$

where $\tau_2 \in (0, 1)$, the linearization variable is

$$z_\ell = \left. \frac{\partial \widehat{\theta}(\mathbf{y}, \boldsymbol{\pi})}{\partial \pi_\ell} \right|_{\boldsymbol{\pi} = \mathbf{a}}.$$

The approximations are used to estimate the variance of $\widehat{\theta}$. Suppose that

$$\theta^{-1} R(\tau_1) = O_p(h_n), \tag{4.2.3}$$

$$\theta^{-1} R(\tau_2) = O_p(\eta_n), \tag{4.2.4}$$

where h_n and η_n are two sequences of real numbers such that $\lim_{n \to \infty} h_n = \lim_{n \to \infty} \eta_n = 0$, when the sample size n and the population size N are large. If (4.2.3) holds, expression (4.2.1) can be used to approximate the variance of $\widehat{\theta}$ by using $\widetilde{z}_\ell$ in the Horvitz-Thompson variance,

$$\mathrm{V}(\widehat{\theta}) \approx \sum_{k \in U} \sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) \widetilde{z}_k \widetilde{z}_\ell,$$

where $\pi_{k\ell}$ is the probability that units k and ℓ are both selected in the sample (Horvitz and Thompson, 1952). The value $\widetilde{z}_\ell$ is not necessarily known for $\ell \in U$; it is estimated by some $\widehat{z}_\ell$ and the variance is estimated by

$$\widehat{\mathrm{V}}(\widehat{\theta}) = \sum_{k\in U}\sum_{\ell\in U} a_k a_\ell \frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_{k\ell}} \widehat{z}_k \widehat{z}_\ell. \tag{4.2.5}$$

Considering equations (4.2.1)-(4.2.4), a natural estimator is $\widehat{z}_\ell = z_\ell$ (Graf, 2011).

Remark 4.1. *In common cases $h_n = \eta_n = n^{-1}$, but this is not a general rule; for instance in quantile estimation, $h_n \geq n^{-1}$.*

Remark 4.2. *Define*

$$\widehat{\theta}_{approx,1} = \widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) + \sum_{\ell\in U} \widetilde{z}_\ell (a_\ell - \pi_\ell).$$

If $\widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) = \theta$, then $\mathrm{E}_p(\widehat{\theta} - \widehat{\theta}_{approx,1}) = \mathrm{E}_p(\widehat{\theta} - \theta)$ is the design-bias of the estimator.

Remark 4.3. *Define*

$$\widehat{\theta}_{approx,2} = \widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) + \sum_{\ell\in U} z_\ell (a_\ell - \pi_\ell).$$

If equations (4.2.3) and (4.2.4) hold, $\widehat{\theta}_{approx,1}$ and $\widehat{\theta}_{approx,2}$ are consistent estimators of $\widehat{\theta}$. Equations (4.2.3) and (4.2.4) must be evaluated on a case-by-case basis.

Example 4.1. *Consider the population total of the variable of interest, $Y = \sum_{k\in U} y_k$, and its estimator $\widehat{Y} = N\widehat{Y}/\widehat{N}$, where*

$$\widehat{Y} = \sum_{k\in U} a_k d_k y_k, \quad \widehat{N} = \sum_{k\in U} a_k d_k$$

and $d_k = \pi_k^{-1}$. Then $\widetilde{z}_\ell = d_\ell(y_\ell - Y/N)$. A natural estimator of $\widetilde{z}_\ell$ would be $\widehat{z}_\ell = d_\ell(y_\ell - \widehat{Y}/\widehat{N})$, but using (4.2.2), we find $z_\ell = d_\ell N(y_\ell - \widehat{Y}/\widehat{N})/\widehat{N}$.

Example 4.2. *Consider the ratio $R_{yx} = Y/X$, where $X = \sum_{k\in U} x_k$, which is estimated by $\widehat{R}_{yx} = \widehat{Y}/\widehat{X}$, where $\widehat{X} = \sum_{k\in U} a_k d_k x_k$. Then*

$$z_\ell = d_\ell \frac{y_\ell - x_\ell \widehat{R}_{yx}}{\widehat{X}}, \quad \widetilde{z}_\ell = d_\ell \frac{y_\ell - x_\ell R_{yx}}{X}$$

and

$$\frac{\partial^2 \widehat{R}_{yx}}{\partial a_k \partial a_\ell} = d_k d_\ell \frac{2\widehat{R}_{yx} x_k x_\ell - x_\ell y_k - x_k y_\ell}{\widehat{X}^2}.$$

Suppose that $X^{-1}(\widehat{Y} - Y) = O_p(n^{-1/2})$ and $X^{-1}(\widehat{X} - X) = O_p(n^{-1/2})$, then

$$R(\tau) = \frac{1}{\widehat{X}_\tau^2}\left[\widehat{R}_{yx\tau}(\widehat{X} - X)^2 - (\widehat{X} - X)(\widehat{Y} - Y)\right] = O_p\left(\frac{1}{n}\right),$$

where $\widehat{R}_{yx\tau} = \widehat{X}_\tau^{-1}\widehat{Y}_\tau$, $\widehat{X}_\tau = \tau\widehat{X} + (1-\tau)X$ for $\tau \in (0,1)$. Expressions (4.2.3) and (4.2.4) are $O_p(n^{-1})$.

Example 4.3. *Let the population geometric mean be*

$$g = \left(\prod_{k \in U} y_k \right)^{1/N}, \tag{4.2.6}$$

with $y_k > 0$. *It is estimated by*

$$\widehat{g} = \prod_{k \in s} y_k^{1/(\widehat{N}\pi_k)} = \exp\left(\frac{1}{\widehat{N}} \sum_{k \in U} \frac{a_k}{\pi_k} \log y_k \right).$$

The linearization variables are $z_\ell = \widehat{N}^{-1} \widehat{g} d_\ell \log(y_\ell / \widehat{g})$ *and* $\widetilde{z}_\ell = N^{-1} g d_\ell \log(y_\ell / g)$. *Then,*

$$\frac{\partial^2 \widehat{g}}{\partial a_k a_\ell} = d_k d_\ell \frac{\widehat{g}}{\widehat{N}^2} \left[-\log \frac{y_k}{\widehat{g}} + \log \frac{y_k}{\widehat{g}} \log \frac{y_\ell}{\widehat{g}} - \log \frac{y_\ell}{\widehat{g}} \right]$$

and

$$R = \frac{\widehat{g}_\tau}{2\widehat{N}_\tau^2} \left[\left(\widehat{L}_\tau - L_\tau \right)^2 - 2 \left(\widehat{L}_\tau - L_\tau \right) \left(\widehat{N} - N \right) \right],$$
$$\widehat{L}_\tau = \sum_{k \in U} a_k d_k \log \frac{y_k}{\widehat{g}_\tau}, \quad L_\tau = \sum_{k \in U} \log \frac{y_k}{\widehat{g}_\tau}$$

and $\widehat{g}_\tau$ *is* $\widehat{g}$ *where* $\mathbf{a}$ *is replaced by* $\tau \mathbf{a} + (1 - \tau)\boldsymbol{\pi}$.

4.2.2. Calibrated total estimator

In calibrated estimation, the original sampling weight $d_k = \pi_k^{-1}$ is modified for $k \in s$ (Deville and Särndal, 1992; Deville et al., 1993). Consider the vector $\mathbf{x}_k$ containing the values taken by p auxiliary variables on unit k. The new weights are such that

$$\sum_{k \in U} a_k w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k, \tag{4.2.7}$$

where w_k is the calibrated weight of unit k. The weights are defined by

$$w_k = d_k F_k(\mathbf{x}_k^\top \boldsymbol{\lambda}). \tag{4.2.8}$$

The calibration function $F_k(.)$ is strictly increasing, $F_k(0) = 1$ and $F_k'(0) = q_k$, where $F_k'(.)$ is the derivative of $F_k(.)$ and q_k is a tuning parameter. The vector $\boldsymbol{\lambda}$ contains the Lagrange multiplier (see Deville and Särndal, 1992). The calibrated total estimator of a variable y is

$$\widehat{Y}_c = \sum_{k \in U} a_k d_k F_k(\mathbf{x}_k^\top \boldsymbol{\lambda}) y_k = \sum_{k \in U} a_k w_k y_k.$$

Graf (2015) linearized the calibrated estimator to estimate its variance (see also Appendix 4.8.1).

Proposition 4.1. *The linearization variable of $\widehat{Y}_c$ obtained when linearizing with respect to a_ℓ is $z_\ell = w_\ell e_\ell$, where $e_\ell = y_\ell - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{y|x}$, $\widehat{\mathbf{B}}_{y|x} = \widehat{\mathbf{T}}_{xx}^{-1}\widehat{\mathbf{t}}_{xy}$,*

$$\widehat{\mathbf{T}}_{xx} = \sum_{k\in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda})\mathbf{x}_k\mathbf{x}_k^\top, \quad \widehat{\mathbf{t}}_{xy} = \sum_{k\in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda})\mathbf{x}_k y_k. \tag{4.2.9}$$

The linearization variables are weighted residuals of the regression of y by the auxiliary variables, where the regression coefficients are weighted by $d_\ell F_\ell'(\mathbf{x}_\ell^\top \boldsymbol{\lambda})$. Deville and Särndal (1992) stated that the calibrated estimator is asymptotically equivalent to the Generalized Regression estimator (GREG) and their regression coefficients are weighted by $w_k q_k$ instead of $d_\ell F_\ell'(\mathbf{x}_\ell^\top \boldsymbol{\lambda})$. Demnati and Rao (2004) obtained the linearization variable of Proposition 4.1. The difference is that the proposed methodology targets the sampling indicators, without the sampling weights. This difference is important in the proposed extension to nonresponse.

Example 4.4. *In the GREG-estimator, the calibration function is $F_k(u) = 1 + q_k u$ and the regression coefficients in Proposition 4.1 are weighted by $d_\ell F_\ell'(\mathbf{x}_\ell^\top \boldsymbol{\lambda}) = d_\ell q_\ell$. Then $z_\ell = d_\ell g_\ell e_\ell$, where $g_\ell = w_\ell/d_\ell$ is called g-weight. Särndal (1982) advocated for the use of the g-weights when estimating the variance of the GREG-estimator.*

Example 4.5. *Consider the linear truncated function $F_k(u) = \max(L, \min(1 + q_k u, H))$, where $L < 1 < H$ are respectively a lower and an upper bound for the g-weights (Deville and Särndal, 1992). In this case, the derivative of $F_k(u)$ is null when u is outside the interval $[(L-1)/q_k, (H-1)/q_k]$. This means that the regression coefficients of Proposition 4.1 are computed in the set of units that have weights strictly between the bounds.*

4.2.3. Calibration in a complex estimator

Consider $\mathbf{d} = (d_1 \cdots d_k \cdots d_N)^\top$, where $d_k = \pi_k^{-1}$ is the sampling weight of unit k, and suppose that an estimator of θ is $\widehat{\theta}_d = \widehat{\theta}(\mathbf{y}, \mathbf{a}, \mathbf{d})$. A vector of calibrated weights $\mathbf{w} = (w_1 \cdots w_k \cdots w_N)^\top$, where w_k is defined in (4.2.8), can be used to estimate θ in the calibrated estimator $\widehat{\theta}_w = \widehat{\theta}(\mathbf{y}, \mathbf{a}, \mathbf{w})$. The linearization method is used to compute the variance estimator of any calibrated estimator in Proposition 4.2, which is proved in Appendix 4.8.2.

Condition 4.1. *Assume that, in $\widehat{\theta}_d$, a_k is always multiplied by its associated weight d_k. Then the linearization variable of $\widehat{\theta}_d$ is $z_\ell^d = d_\ell h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{d})$, where $h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{d}) = d_\ell \partial\widehat{\theta}_d/\partial a_\ell$. This implies that $h_{\ell 2}(\mathbf{y}, \mathbf{a}, \mathbf{d}) = \partial\widehat{\theta}_d/\partial d_\ell = a_\ell h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{d})$.*

Proposition 4.2. *If Condition 4.1 holds, the linearization variable of $\widehat{\theta}_w$ obtained when linearizing with respect to a_ℓ is $z_\ell^w = w_\ell v_\ell^w$, where*

$$\begin{aligned} v_\ell^w &= h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{h1|x}, \\ \widehat{\mathbf{B}}_{h1|x} &= \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k\in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda})\mathbf{x}_k h_{k1}(\mathbf{y}, \mathbf{a}, \mathbf{w}), \end{aligned} \tag{4.2.10}$$

and $\widehat{\mathbf{T}}_{xx}$ is defined in Equation (4.2.9).

Condition 4.1 assumes that w_k and a_k appear in $\widehat{\theta}_w$ only as a product $w_k a_k$, which is common in practice. In this case, $h_{k2}(.) = a_k h_{k1}(.)$ and expression (4.2.10) is the residual of the regression of h_{k1} by $\mathbf{x}_k$. The computation of the linearization variable can then be done in two steps. First the linearization variable $h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w})$ is computed naïvely, as if the vector $\mathbf{w}$ was not depending on $\mathbf{a}$. Next, the residual v_ℓ^w is calculated with the regression coefficients in $\widehat{\mathbf{B}}_{h1|x}$. The linearization method is also applicable if Condition 4.1 does not hold, but the calculation of z_ℓ^w is more complicated and it does not result in Proposition 4.2.

Example 4.6. *Consider the calibrated estimator of the Gini index*

$$\widehat{G} = \frac{1}{2\widehat{N}_c \widehat{Y}_c} \sum_{i \in U} \sum_{k \in U} a_i a_k w_i w_k |y_i - y_k|, \quad \textit{where} \quad \widehat{N}_c = \sum_{k \in U} a_k w_k.$$

If the weights w_k are considered as fixed, Langel and Tillé (2013) showed that $z_\ell = w_\ell h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w})$, where

$$h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) = \frac{1}{\widehat{N}_c \widehat{Y}_c} \left[\widehat{Y}_c - \widehat{N}_c y_\ell + 2\widehat{N}_\ell (y_\ell - \widehat{\overline{Y}}_\ell) - \widehat{G}(\widehat{Y}_c + \widehat{N}_c y_\ell) \right],$$

$$\widehat{N}_\ell = \sum_{i \in U} a_i w_i \mathrm{I}_{y_i \leq y_\ell}, \widehat{\overline{Y}}_\ell = \frac{1}{\widehat{N}_\ell} \sum_{i \in U} a_i w_i y_i \mathrm{I}_{\widehat{N}_i \leq \widehat{N}_\ell}.$$

The linearization variable considering that the weights are calibrated on random totals is $z_\ell^w = w_\ell \left[h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{h1|x} \right]$.

4.3. Dealing with nonresponse

Two types of nonresponse are distinguished in survey sampling. The first one is unit nonresponse, in which all variables are missing for some units. Usually, it is addressed by reweighting, each sampled respondent k receives a new weight w_k. The second type is item nonresponse, in which some (but not all) items are missing. In this case, the missing values can be imputed and y_k is replaced by $\widetilde{y}_k = R_k y_k + (1 - R_k) y_k^*$, where y_k^* is the imputed value of nonrespondent k. The response indicator R_k is 1 if unit k is a respondent and 0 otherwise.

For the purpose of inference, three sources of randomness are now considered. As in the complete case, the sampling indicator is random and generated by the sampling design. The response indicator is a random variable motivated by the nonresponse model. The variable of interest is modeled by a random superpopulation model, also called an imputation model. Those sources of randomness lead to two approaches to inference: the nonresponse model approach and the imputation model approach.

The variance can be decomposed in two ways. The first one is the two-phase approach, in which a sample is selected in the population and then respondents are chosen in the sample. Rao (1990); Rao and Sitter (1995) discussed this approach under the imputation model context and Särndal (1992), under the nonresponse

model context. The second decomposition is the reverse approach, in which the respondents are identified in the population and then a sample containing respondents and nonrespondents is chosen in the population (Fay, 1991). The respondent sample is seen as a strongly invariant two-phase design (Beaumont and Haziza, 2016). This means that the selection of the respondents in the second phase is independent of the first phase. Shao and Steel (1999) discussed the reverse approach under the two inference contexts. This approach is often favored when estimating variances on account of simplifications due to the nonresponse mechanism. The linearization method is extended to compute variances with the reverse approach and inference based on a nonresponse model in Section 4.4, and with inference based on an imputation model in Section 4.5. In the interest of brevity, the two-phase approach is omitted, but it could be treated similarly.

4.4. Inference based on a nonresponse model

In the nonresponse model approach, the vector of sampling indicators $\mathbf{a}$ and the vector of response indicators $\mathbf{R} = (R_1 \cdots R_k \cdots R_N)^\top$ are random, while the variable of interest is seen as fixed. A nonresponse model is assumed; components of $\mathbf{R}$ have independent bernoulli distributions with parameter p_k, for $k = 1, \ldots, N$, where p_k is the probability that unit k responds. This probability is usually unknown and estimated by $\widehat{p}_k$.

Suppose that θ has an estimator $\widehat{\theta}_{NR}$ which is treated for nonresponse. The reverse variance of estimator $\widehat{\theta}_{NR}$ is

$$\mathrm{V}(\widehat{\theta}_{NR}) = \mathrm{E}_q \mathrm{V}_p(\widehat{\theta}_{NR}) + \mathrm{V}_q \mathrm{E}_p(\widehat{\theta}_{NR}), \tag{4.4.1}$$

where $\mathrm{E}_q(.)$ and $\mathrm{V}_q(.)$ are respectively the expectation and the variance with respect to the nonresponse model.

4.4.1. Linearization for inference based on a nonresponse model

The linearization method in Section 4.2 is extended, in Methodology 4.1, to approximate variances in the case of inference based on a nonresponse model. Estimator $\widehat{\theta}_{NR}$ is written $\widehat{\theta}_{NR} = \widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R})$ and it is linearized with respect to a_ℓ and R_ℓ.

Methodology 4.1 Linearization method to estimate Equation (4.4.1), the reverse variance with inference based on a nonresponse model

1: The approximation of the parameter with respect to the sampling indicators is,

$$\widehat{\theta}_{NR} \approx \widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) + \sum_{\ell \in U} \tilde{z}^a_\ell (a_\ell - \pi_\ell), \tag{4.4.2}$$

where

$$\tilde{z}^a_k = z^a_k \Big|_{\mathbf{a}=\boldsymbol{\pi}}, \quad z^a_k = \frac{\partial \widehat{\theta}_{NR}}{\partial a_\ell}.$$

2: The first component of the term on the right hand side of (4.4.1) becomes

$$\mathrm{V}_1 = \mathrm{E}_q \mathrm{V}_p(\widehat{\theta}_{NR}) \approx \mathrm{E}_q \left[\sum_{k \in U} \sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) \tilde{z}_k^a \tilde{z}_\ell^a \right]$$

and is estimated by

$$\widehat{\mathrm{V}}_1 = \sum_{k \in U} a_k \sum_{\ell \in U} a_\ell \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}} z_k^a z_\ell^a. \tag{4.4.3}$$

3: The second component of the expression on the right hand side of (4.4.1) is approximated using (4.4.2), $\mathrm{V}_2 = \mathrm{V}_q \mathrm{E}_p(\widehat{\theta}_{NR}) \approx \mathrm{V}_q \left[\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) \right]$.

4: V_2 is approximated by linearizing $\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R})$ with respect to the response indicators,

$$\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) \approx \widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{p}) + \sum_{\ell \in U} \tilde{z}_\ell^{aR} (R_\ell - p_\ell),$$

where $\mathbf{p} = (p_1 \cdots p_k \cdots p_N)^\top$ and

$$\tilde{z}_\ell^{aR} = z_\ell^{aR} \Big|_{\mathbf{R}=\mathbf{p}}, \quad z_\ell^{aR} = \frac{\partial \widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R})}{\partial R_\ell}.$$

5: Expression V_2 becomes

$$\mathrm{V}_2 \approx \mathrm{V}_q \left[\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) \right] \approx \sum_{k \in U} \sum_{\ell \in U} (p_{k\ell} - p_k p_\ell) \tilde{z}_k^{aR} \tilde{z}_\ell^{aR} = \sum_{k \in U} p_k (1 - p_k) (\tilde{z}_k^{aR})^2,$$

where $p_{k\ell}$ is the joint response probability of units k and ℓ. The respondent set is seen as a Poisson sample, $p_{k\ell} = p_k p_\ell$ if $k \neq \ell$, $p_{kk} = p_k$, and the variance reduces to a single sum. Expression V_2 is estimated by

$$\widehat{\mathrm{V}}_2 = \sum_{k \in U} a_k R_k \frac{(1 - \widehat{p}_k)}{\pi_k} (\hat{z}_k^{aR})^2, \tag{4.4.4}$$

where $\widehat{p}_k$ and $\hat{z}_k^{aR}$ are estimators of p_k and $\tilde{z}_k^{aR}$ respectively.

Remark 4.4. *The estimation of* V_2 *is based on the approximation of* $\mathrm{E}_p(\widehat{\theta}_{NR})$*, which may introduce a bias. As seen in Section 4.2, this biais depends on the remainder.*

Remark 4.5. *Beaumont et al. (2015) discuss situations in which it is possible to ignore some terms of the variance. The simplifications can greatly shorten the proposed methodology.*

With this methodology, variances can be estimated even in complicated cases. Linearizing with respect to the sampling and response indicators guaranties the ability to solve the expectations and the variances. This extension to nonresponse was not necessarily possible with other linearization methods. Kott (2006) proposed a one-step variance approximation method which is applicable only if a_k and R_k are present in the estimated parameter as a product. The quantity $a_k R_k$ becomes a single indicator $\delta_k = a_k R_k$, which is 1 if unit k is selected in the sample and is a

respondent, 0 otherwise. Similarly to Methodology 4.1, the estimated parameter may be linearized with respect to the variable δ_k to obtain a variance estimator in one single step. In Section 4.4.2, the proposed methodology is applied to revisit the results of Kott (2006). In Section 4.4.3, the linearization method is used to find variance estimators of calibrated estimators that are not accessible with Kott's method.

4.4.2. Revisiting Kott's method for calibration on totals

Consider $\mathbf{x}_k$, a vector of auxiliary variables which is completely known for unit k such that $a_k R_k = 1$, and consider $\sum_{k\in U} \mathbf{x}_k$, known population totals. The total Y is estimated in the presence of unit nonresponse by

$$\widehat{Y}_R = \sum_{k\in U} a_k R_k w_k y_k,$$

where the calibration weights are such that $w_k = d_k F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})$ and

$$\sum_{k\in U} a_k R_k w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k. \tag{4.4.5}$$

Indicators a_k and R_k appear only as a product $\delta_k = a_k R_k$. Kott (2006) supposes that the set of respondents is selected according to a Poisson sampling design and that the estimated response probability of unit k is $\hat{p}_k = 1/F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})$. Reweighting thus consists in assuming an underlying nonresponse model. For instance, the choice of the calibration function $F(u) = 1 + \exp u$ ensures that $0 < \hat{p}_k < 1$ and then the underlying model is logistic.

The variance estimator is obtained by considering the indicator δ_k associated to probability $\pi_k^* = \mathrm{E}(\delta_k) = \pi_k p_k$. The joint inclusion probability of unit k and ℓ in the respondent sample is $\pi_{k\ell}^* = \pi_{k\ell} p_k p_\ell$ if $k \neq \ell$ and $\pi_{kk}^* = \pi_k p_k$. Kott (2006) proposed to expand $F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})$ around $\boldsymbol{\lambda}$, which is an extension of Binder (1983); Demnati and Rao (2004). This leads to

$$\mathrm{V}(\widehat{Y}_R) = \sum_{k\in U} \left(\frac{1}{\pi_k^*} - \frac{1}{\pi_k}\right) e_k^2 + \sum_{k\in U}\sum_{\ell\in U} (\pi_{k\ell} - \pi_k \pi_\ell) \frac{e_k}{\pi_k} \frac{e_\ell}{\pi_\ell},$$

where

$$e_k = y_k - \mathbf{x}_k^\top \left[\sum_{k\in U} F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) p_k \mathbf{x}_k \mathbf{x}_k^\top\right]^{-1} \sum_{k\in U} F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) p_k \mathbf{x}_k y_k.$$

The proposed variance estimator is

$$\widehat{\mathrm{V}}(\widehat{Y}_R) = \sum_{k\in U} a_k R_k (w_k^2 \pi_k - w_k) v_k^2 + \sum_{k\in U} a_k R_k \sum_{\ell\in U} a_\ell R_\ell \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}} w_k v_k w_\ell v_\ell, \tag{4.4.6}$$

where $v_k = y_k - \mathbf{x}_k^\top \widehat{\mathbf{B}}_{y|x}$, $\widehat{\mathbf{B}}_{y|x} = \widehat{\mathbf{T}}_{xx}^{-1} \widehat{\mathbf{t}}_{xy}$,

$$\widehat{\mathbf{T}}_{xx} = \sum_{k\in U} a_k R_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k \mathbf{x}_k^\top, \quad \widehat{\mathbf{t}}_{xy} = \sum_{k\in U} a_k R_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k y_k.$$

The derivation of this result is relatively convoluted. The proposed linearization method can be used to find kott's result in a straightforward manner. In Proposition 4.3, the estimator is linearized with respect to the indicator δ_k in one single step (see Appendix 4.8.3).

Proposition 4.3. *The linearization variable of $\widehat{Y}_R$ obtained when linearizing with respect to $\delta_k = a_k R_k$ is $w_\ell e_\ell$, where $e_\ell = y_\ell - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{y|x}$.*

4.4.3. Simultaneous calibration and reverse approach

In practical applications, auxiliary information can be available at different levels, for instance at the population level and at the sample level (see among others Dupont, 1995; Hidiroglou and Särndal, 1998; Estevao and Särndal, 2002). The set of respondents may be simultaneously calibrated on these two levels. Consider a vector of auxiliary variables, which is decomposed into two parts, i.e. $\mathbf{x}_k = (\mathbf{x}_k^{\bullet\top} \quad \mathbf{x}_k^{\circ\top})^\top$. Vector $\mathbf{x}_k$ is known for all units in the sample. The vector of totals $\mathbf{X}^\bullet = \sum_{k\in U} \mathbf{x}_k^\bullet$ is assumed to be known while vector $\mathbf{X}^\circ = \sum_{k\in U} \mathbf{x}_k^\circ$ is unknown.

The reweighted total estimator of variable y is

$$\widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R}) = \sum_{k\in U} a_k R_k w_{k2} y_k,$$

where the weights are obtained by solving the following two systems of equations,

$$\sum_{k\in U} a_k R_k w_{k2} \mathbf{x}_k = \begin{pmatrix} \mathbf{X}^\bullet \\ \sum_{k\in U} a_k w_{k1} \mathbf{x}_k^\circ \end{pmatrix} \tag{4.4.7}$$

and

$$\sum_{k\in U} a_k w_{k1} \mathbf{x}_k^\bullet = \mathbf{X}^\bullet.$$

The weights are defined by $w_{k1} = d_k F_{k1}(\mathbf{x}_k^{\bullet\top}\boldsymbol{\lambda}_1)$ and $w_{k2} = d_k F_{k2}(\mathbf{x}_k^\top \boldsymbol{\lambda}_2)$, where $F_{k1}(.)$ and $F_{k2}(.)$ are two calibration functions. The response probabilities can be estimated by $\widehat{p}_k = 1/F_{k2}(\mathbf{x}_k^\top \boldsymbol{\lambda}_2)$. It is interesting to consider $F_{2k}(.) = 1 + \exp(.)$ to ensure $0 < \widehat{p}_k < 1$.

The proposed methodology leads to a variance estimator even if the calibration is on population or sample totals. That means it is possible to find a variance estimator if the indicator a_k is not always accompanied by the indicator R_k, $k = 1, \ldots, N$, which is not the case of Kott's method. Following Methodology 4.1, two linearization variables are needed: in proposition 4.4, the linearization variable z_ℓ^a is computed and in Proposition 4.5, the linearization variable z_ℓ^{aR} is computed (see Appendix 4.8.4).

Proposition 4.4. *The linearization variable of $\widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R})$ when linearizing with respect to a_ℓ is*

$$z_\ell^a = R_\ell w_{\ell 2} e_\ell + \begin{pmatrix} \mathbf{0} \\ w_{\ell 1} \mathbf{e}_\ell^\circ \end{pmatrix}^\top \widehat{\mathbf{B}}_{y|x},$$

where $e_\ell = y_\ell - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{y|x}$, $\mathbf{e}_\ell^\circ = \mathbf{x}_\ell^\circ - \mathbf{x}_\ell^{\bullet\top} \widehat{\mathbf{B}}_{x^\circ|x^\bullet}$, $\mathbf{0}$ *is a null vector of the same length as* $\mathbf{x}_k^\bullet$,

$$\widehat{\mathbf{B}}_{x^\circ|x^\bullet} = \widehat{\mathbf{T}}_{x^\bullet x^\bullet}^{-1} \widehat{\mathbf{T}}_{x^\bullet x^\circ}, \qquad \widehat{\mathbf{B}}_{y|x} = \widehat{\mathbf{T}}_{xx}^{-1} \widehat{\mathbf{t}}_{xy},$$

$$\widehat{\mathbf{T}}_{x^\bullet x^\bullet} = \sum_{k\in U} a_k d_k F'_{k1}(\mathbf{x}_k^{\bullet\top}\boldsymbol{\lambda}_1)\mathbf{x}_k^\bullet \mathbf{x}_k^{\bullet\top}, \qquad \widehat{\mathbf{T}}_{x^\bullet x^\circ} = \sum_{k\in U} a_k d_k F'_{k1}(\mathbf{x}_k^{\bullet\top}\boldsymbol{\lambda}_1)\mathbf{x}_k^\bullet \mathbf{x}_k^{\circ\top},$$

$$\widehat{\mathbf{T}}_{xx} = \sum_{k\in U} a_k R_k d_k F'_{k2}(\mathbf{x}_k^\top\boldsymbol{\lambda}_2)\mathbf{x}_k \mathbf{x}_k^\top, \quad \widehat{\mathbf{t}}_{xy} = \sum_{k\in U} a_k R_k d_k F'_{k2}(\mathbf{x}_k^\top\boldsymbol{\lambda}_2)\mathbf{x}_k y_k.$$

Proposition 4.5. *The linearization variable of* $\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R})$ *when linearizing with respect to* R_ℓ *is* $z_\ell^{aR} = F_{\ell 2}(\mathbf{x}_\ell^\top \boldsymbol{\lambda}_2^\pi) e_\ell^\pi$, *where* $\boldsymbol{\lambda}_2^\pi$ *and* e_ℓ^π *are* $\boldsymbol{\lambda}_2$ *and* e_ℓ, *respectively, when* a_k *is replaced by* π_k.

The linearization variable z_ℓ^a in Proposition 4.4 can be integrated to estimator $\widehat{\mathrm{V}}_1$ in (4.4.3). The linearization variable z_ℓ^{aR} can be estimated by $\hat{z}_\ell^{aR} = F_{\ell 2}(\mathbf{x}^\top \boldsymbol{\lambda}_2) e_\ell$ and it can be plugged in estimator $\widehat{\mathrm{V}}_2$ given in (4.4.4). When the vector of auxiliary variables of unit k is $\mathbf{x}_k = \mathbf{x}_k^\bullet$, this variance estimator reduces to the one obtained by Kott (2006).

4.5. Inference based on an imputation model

In the imputation model approach, vectors $\mathbf{a}$ and $\mathbf{R}$ and a model for the y-variable are seen as random. The nonresponse mechanism is only assumed to be ignorable; it does not depend on the variable of interest when auxiliary information is taken into account. Some assumptions on the imputation model are made. For instance, assume that the model of y_k is $m:\ y_k = \mathbf{x}_k\boldsymbol{\beta} + \varepsilon_k$, where $\mathbf{x}_k$ is a vector of auxiliary variables of length p and $\boldsymbol{\beta}$ is a vector of regression coefficients. The random variable is ε_k, an error term such that $\mathrm{E}_m(\varepsilon_k) = 0$, $\mathrm{E}_m(\varepsilon_k^2) = \sigma^2$ and $\mathrm{E}_m(\varepsilon_k \varepsilon_j) = 0$ if $k \neq j$. Notations $\mathrm{E}_m(.)$ and $\mathrm{Var}_m(.)$ hold respectively for expectation and variance with respect to the imputation model. Considering $\widetilde{\theta}_{NR} = \mathrm{E}_p(\widehat{\theta}_{NR})$, the reverse variance is

$$\begin{aligned}
\mathrm{V}(\widehat{\theta}_{NR} - \theta) &= \mathrm{E}_q\mathrm{E}_m\mathrm{E}_p\left(\widehat{\theta}_{NR} - \widetilde{\theta}_{NR} + \widetilde{\theta}_{NR} - \theta\right)^2 \\
&= \mathrm{E}_q\mathrm{E}_m\mathrm{V}_p(\widehat{\theta}_{NR} - \widetilde{\theta}_{NR}) + \mathrm{E}_q\mathrm{V}_m\mathrm{E}_p(\widetilde{\theta}_{NR} - \theta) + 2\mathrm{E}_q\mathrm{E}_m\mathrm{E}_p\left[(\widehat{\theta}_{NR} - \widetilde{\theta}_{NR})(\widetilde{\theta}_{NR} - \theta)\right] \\
&= \mathrm{E}_q\mathrm{E}_m\mathrm{V}_p(\widehat{\theta}_{NR}) + \mathrm{E}_q\mathrm{V}_m\mathrm{E}_p(\widehat{\theta}_{NR} - \theta). \qquad (4.5.1)
\end{aligned}$$

4.5.1. Linearization for inference based on an imputation model

The linearization method in Section 4.2 is extended, in Methodology 4.2, to approximate variances in the case of inference based on an imputation model. Estimator $\widehat{\theta}_{NR} = \widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R})$ is linearized with respect to a_ℓ and y_ℓ.

Methodology 4.2 Linearization method to estimate Equation (4.5.1), the reverse variance with inference based on an imputation model

1: The approximation of the parameter with respect to the sampling indicators is calculated in (4.4.2).

2: The first component of the term on the right hand side of (4.5.1) becomes

$$\mathrm{V}_1 = \mathrm{E}_q \mathrm{E}_m \mathrm{V}_p(\widehat{\theta}_{NR}) \approx \mathrm{E}_q \mathrm{E}_m \left[\sum_{k \in U} \sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) \tilde{z}_k^a \tilde{z}_\ell^a \right]$$

and is estimated by Expression (4.4.3).

3: The second component of the expression on the right hand side of (4.5.1) is approximated using (4.4.2),

$$\mathrm{V}_2 = \mathrm{E}_q \mathrm{V}_m \mathrm{E}_p \left[\widehat{\theta}_{NR} - \theta(\mathbf{y}) \right] \approx \mathrm{E}_q \mathrm{V}_m \left[\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) - \theta(\mathbf{y}) \right].$$

4: Variance V_2 is approximated by linearizing $\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) - \theta(\mathbf{y})$ with respect to the variable of interest,

$$\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) - \theta(\mathbf{y}) \approx \widehat{\theta}_{NR} \left[\mathrm{E}_m(\mathbf{y}), \boldsymbol{\pi}, \mathbf{R} \right] - \theta \left[\mathrm{E}_m(\mathbf{y}) \right] + \sum_{\ell \in U} \tilde{z}_\ell^{ay} \left[y_\ell - \mathrm{E}_m(y_\ell) \right],$$

where

$$\tilde{z}_\ell^{ay} = z_\ell^{ay} \Big|_{\mathbf{y} = \mathrm{E}_m(\mathbf{y})} - z_\ell^{y} \Big|_{\mathbf{y} = \mathrm{E}_m(\mathbf{y})}, \quad z_\ell^{ay} = \frac{\partial \widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R})}{\partial y_\ell}, \quad z_\ell^{y} = \frac{\partial \theta(\mathbf{y})}{\partial y_\ell}.$$

5: Expression V_2 becomes

$$\mathrm{V}_2 \approx \mathrm{E}_q \mathrm{V}_m \left[\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) - \theta(\mathbf{y}) \right] \approx \sigma^2 \sum_{k \in U} (\tilde{z}_k^{ay})^2,$$

where σ^2 is the variance of y_ℓ. Expression V_2 is estimated by

$$\widehat{\mathrm{V}}_2 = \widehat{\sigma}^2 \sum_{k \in U} a_k \frac{(\hat{z}_k^{ay})^2}{\pi_k}, \tag{4.5.2}$$

where $\widehat{\sigma}^2$ and $\hat{z}_k^{ay}$ are estimators of σ^2 and $z_k^{ay} - z_\ell^y$ respectively.

This methodology is used to find an explicit variance estimator for imputed complex estimators in Section 4.5.2.

4.5.2. Imputation in a complex estimator

Item nonresponse can be treated by imputation. Consider a variable $\mathbf{y}$ with missing values imputed by a regression imputation method. Parameter $\theta = \theta(\mathbf{y})$ is estimated by $\widehat{\theta}_{NR} = \widehat{\theta}_{NR}(\tilde{\mathbf{y}}, \mathbf{a}, \mathbf{R})$, where $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_N)^\top$, $\tilde{y}_k = R_k y_k + (1 - R_k) y_k^*$, $y_k^* = \mathbf{x}_k^\top \widehat{\mathbf{B}}_{y|x}$, $\widehat{\mathbf{B}}_{y|x} = \widehat{\mathbf{T}}_{xx}^{-1} \widehat{\mathbf{t}}_{xy}$,

$$\widehat{\mathbf{T}}_{xx} = \sum_{k \in U} a_k R_k \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\pi_k}, \quad \widehat{\mathbf{t}}_{xy} = \sum_{k \in U} a_k R_k \frac{\mathbf{x}_k y_k}{\pi_k}$$

and $\mathbf{x}_k$ is a vector of auxiliary variables available for all of the units in the sample.

The linearization method is used to estimate the variance (4.5.1) of this imputed estimator. Following Methodology 4.2, the linearization variable z_ℓ^a is computed in

Proposition 4.6 and the linearization variable z_ℓ^{ay} is computed in Proposition 4.7 (see Appendix 4.8.5 for more details).

Proposition 4.6. *The linearization variable of* $\widehat{\theta}_{NR} = \widehat{\theta}_{NR}(\widetilde{\mathbf{y}}, \mathbf{a}, \mathbf{R})$ *when linearizing with respect to* a_ℓ *is*

$$z_\ell^a = \frac{h_{\ell 1}(\widetilde{\mathbf{y}}, \mathbf{a}, \mathbf{R})}{\pi_\ell} + \frac{R_\ell e_\ell}{\pi_\ell} \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k \in U} a_k (1 - R_k) \mathbf{x}_k h_{k2}(\widetilde{\mathbf{y}}, \mathbf{a}, \mathbf{R}),$$

where $e_\ell = y_\ell - \mathbf{x}_\ell \widehat{\mathbf{B}}_{y|x}$,

$$\frac{h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{R})}{\pi_\ell} = \frac{\partial \widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R})}{\partial a_\ell}, \quad h_{\ell 2}(\mathbf{y}, \mathbf{a}, \mathbf{R}) = \frac{\partial \widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R})}{\partial y_\ell}.$$

Proposition 4.7. *The linearization variable of* $\widehat{\theta}_{NR}(\widetilde{\mathbf{y}}^\pi, \boldsymbol{\pi}, \mathbf{R})$*, where* $\widetilde{\mathbf{y}}^\pi$ *is the vector* $\widetilde{\mathbf{y}}$ *with* $\mathbf{a} = \boldsymbol{\pi}$*, when linearizing with respect to* y_ℓ *is*

$$z_\ell^{ay} = R_\ell h_{\ell 2}(\widetilde{\mathbf{y}}^\pi, \boldsymbol{\pi}, \mathbf{R}) + R_\ell \mathbf{x}_\ell^\top \left(\sum_{k \in U} R_k \mathbf{x}_k^\top \mathbf{x}_k \right)^{-1} \sum_{k \in U} (1 - R_k) \mathbf{x}_k h_{k2}(\widetilde{\mathbf{y}}^\pi, \boldsymbol{\pi}, \mathbf{R}).$$

The linearization variable z_ℓ^a in Proposition 4.6 is integrated to estimator $\widehat{\mathrm{V}}_1$ in (4.4.3). The linearization variable $z_\ell^{ay} - z_\ell^y$ is estimated and plugged in estimator $\widehat{\mathrm{V}}_2$ given in (4.5.2).

Remark 4.6. *When the parameter of interest is the population total, the variance estimator in Kim and Rao (2009) coincides with the results in this Section. However, as the authors point out, their linearization method needs to be adjusted if the parameter of interest is not linear in the* y_k*.*

Example 4.7. *The geometric mean in (4.2.6) can be estimated with the imputed estimator*

$$\widehat{g}_I = \prod_{k \in s} \widetilde{y}_k^{d_k / \widehat{N}} = \exp \frac{1}{\widehat{N}} \sum_{k \in U} a_k d_k \log \widetilde{y}_k.$$

Using $d_\ell h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{R}) = \widehat{N}^{-1} \widehat{g} d_\ell \log(y_\ell / \widehat{g})$ *and* $h_{\ell 2}(\mathbf{y}, \mathbf{a}, \mathbf{R}) = \widehat{g} a_\ell d_\ell / (\widehat{N} y_\ell)$*, the linearization variable with respect to* a_ℓ *is*

$$z_\ell^a = \frac{\widehat{g}_I d_\ell}{\widehat{N}} \left[\log \left(\frac{\widetilde{y}_\ell}{\widehat{g}_I} \right) + R_\ell e_\ell \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k \in U} a_k d_k (1 - R_k) \frac{\mathbf{x}_k}{y_k^*} \right].$$

The estimator of $z_\ell^{ay} - z_\ell^y$*, the linearization variable with respect to* y_ℓ *is*

$$\begin{aligned} \widehat{z}_\ell^{ay} &= \frac{\widehat{g}_I R_\ell}{\widehat{N}} \left[\frac{1}{y_\ell} + \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k \in U} a_k d_k (1 - R_k) \frac{\mathbf{x}_k}{y_k^*} \right] - \frac{\widehat{g}_I}{\widehat{N} \widetilde{y}_\ell} \\ &= \frac{\widehat{g}_I}{\widehat{N}} \left[R_\ell \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k \in U} a_k d_k (1 - R_k) \frac{\mathbf{x}_k}{y_k^*} - (1 - R_\ell) \frac{1}{y_\ell^*} \right]. \end{aligned}$$

Example 4.8. *Consider the estimator of the Gini index with values imputed by regression imputation, such that* $\widetilde{y}_k \neq \widetilde{y}_i$*, for* $k \neq i$*,*

$$\widehat{G}_I = \frac{1}{2\widehat{N}\widehat{Y}_I} \sum_{k\in U} a_k d_k \sum_{i\in U} a_i d_i |\widetilde{y}_i - \widetilde{y}_k|,$$

where $\widehat{Y}_I = \sum_{k\in U} a_k d_k \widetilde{y}_k$*. Using*

$$h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{R}) = \frac{1}{\widehat{N}\widehat{Y}} \left[\widehat{Y} - \widehat{N} y_\ell + 2\widehat{N}_\ell (y_\ell - \widehat{\overline{Y}}_\ell) - \widehat{G}(\widehat{Y} + \widehat{N} y_\ell)\right],$$

$$h_{\ell 2}(\mathbf{y}, \mathbf{a}, \mathbf{R}) = \frac{a_\ell d_\ell}{\widehat{N}\widehat{Y}} \left(\sum_{k\in U} a_k d_k \frac{y_\ell - y_k}{|y_\ell - y_k|} - \widehat{G}\widehat{N}\right)$$

and $\widehat{N}_\ell$ *and* $\widehat{\overline{Y}}_\ell$ *defined in Example 4.6, the linearization variable with respect to* a_ℓ *is*

$$z_\ell^a = \frac{d_\ell}{\widehat{N}\widehat{Y}_I} \left[\widehat{Y}_I - \widehat{N}\widetilde{y}_\ell + 2\widehat{N}_\ell (\widetilde{y}_\ell - \widehat{\overline{Y}}_{I\ell}) - \widehat{G}_I(\widehat{Y}_I + \widehat{N}\widetilde{y}_\ell)\right]$$
$$+ R_\ell d_\ell e_\ell \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k\in U} a_k (1 - R_k) \mathbf{x}_k \frac{d_k}{2\widehat{N}\widehat{Y}_I} (2\widehat{N}_k - \widehat{N} - 2\widehat{N}\widehat{G}_I + d_k)$$

and the estimator of $z_\ell^{ay} - z_\ell^y$ *is*

$$\widehat{z}_\ell^{ay} = R_\ell \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k\in U} a_k (1 - R_k) d_k \frac{\mathbf{x}_k}{\widehat{N}\widehat{Y}_I} \left(\sum_{k\in U} a_k d_k \frac{\widetilde{y}_\ell - \widetilde{y}_k}{|\widetilde{y}_\ell - \widetilde{y}_k|} - \widehat{G}_I \widehat{N}\right)$$
$$- \frac{(1 - R_\ell)}{\widehat{N}\widehat{Y}_I} \left(\sum_{k\in U} a_k d_k \frac{\widetilde{y}_\ell - \widetilde{y}_k}{|\widetilde{y}_\ell - \widetilde{y}_k|} - \widehat{G}_I \widehat{N}\right).$$

4.6. Simulation study

Some variance estimators obtained in this paper were evaluated in two simulated datasets and in a real dataset. The accuracy of the linearization method was compared to the one of a resampling method.

4.6.1. Simulated data

Two populations of $N = 500$ and $N = 1000$ units were generated. A vector of auxiliary variables was generated for each unit, $\mathbf{x}_k = (x_{1k} \quad x_{2k} \quad x_{3k})^\top$, where $x_{1k} = 1$, x_{2k} and x_{3k} have a Gamma distribution with expectation 100 and 200 and variance 2000 and 4000 respectively. The variable of interest was generated according to the model $m:\ y_k = 20 + 3x_{2k} + 2x_{3k} + \varepsilon_k$, where ε_k was normally distributed with mean 0 and variance 9. This model is adequate to estimate a geometric mean and to use regression imputation in this study. The correlation between the y-variable and x_2 was 0.73 and the one between y and x_3 was 0.68.

4.6.2. Real data

The dataset '`Ilocos`' is available in the `R` package '`ineq`' (Zeileis, 2014). The data originally come from the Philippines' National Statistics Office. More especially, they are from the 1997 Family and Income and Expenditure Survey and the 1998 Annual Poverty Indicators Survey. The dataset contains the income and seven auxiliary variables of $N = 632$ households. In this simulation study, the natural logarithm of the total income of the household was used as the variable of interest. The vector of auxiliary variables for unit k was $\mathbf{x}_k = (x_{1k} \quad x_{2k} \quad x_{3k})^\top$, where $x_{1k} = 1$, x_{2k} was the family size and x_{3k} was a dummy variable indicating if the household is rural or urban. The correlation between the y-variable and x_2 is 0.25 and the one between y and x_3 is 0.28.

4.6.3. Simulation framework

For each population, four samples were selected with a simple random sampling design without replacement so that the sampling rates were $n/N = 0.1, 0.2, 0.3, 0.4$. In each sample, a set of respondents was selected with Poisson sampling, where the unit probability of response was a logistic function of x_2 and the expected response rate was approximately 0.70. The following estimators were computed:

- in the complete sample:
  - the Gini index $\widehat{G}$, with calibrated weights and its variance, using the linearization variable in Example 4.6 and the calibration vector $\mathbf{x}_k = (x_{1k} \quad x_{2k} \quad x_{3k})^\top$,
  - the geometric mean $\widehat{g}$ and its variance, using Example 4.3,
- in the sample with nonresponse:
  - the total $\widehat{Y}_I$, calibrated on population and sample totals, and its variance, using $\mathbf{x}_k^\bullet = (x_{1k} \quad x_{2k})^\top$, $\mathbf{x}_k^\circ = x_{3k}$ and results of section 4.4.3,
  - the Gini index $\widehat{G}_I$, with regression imputation and its variance, as in Example 4.8,
  - the geometric mean $\widehat{g}_I$, with regression imputation and its variance, as in Example 4.7.

Each variance estimator was compared to a bootstrap variance estimator (Efron, 1979; Shao and Sitter, 1996), with the following procedure to estimate the variance of an estimator $\widehat{\theta}$ (Haziza, 2009):

1. A bootstrap sample s^* of size $n^* = n - 1$ of the original sample s is selected with simple random sampling with replacement.
2. Consider R_k^*, the response indicator of unit k in s^*. The missing values are imputed with regression imputation, where the regression coefficients are computed using s^*.
3. The estimator $\widehat{\theta}^*$ is computed using s^* and the new imputed data.

4. Steps 1-3 are repeated $B = 1000$ times.
5. The bootstrap variance estimator of $\widehat{\theta}$ is

$$\widehat{\mathrm{V}}_B(\widehat{\theta}) = \frac{N-n}{N}\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}^*_{(b)} - \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^*_{(b)}\right)^2,$$

where $\widehat{\theta}^*_{(b)}$ is the value of $\widehat{\theta}^*$ at the replication b, $b = 1, \ldots, B$.

The sample selection was repeated $R = 10,000$ times. For each population and sampling rate, the Monte Carlo relative bias and relative root mean squared error of each variance estimator were calculated. Consider $\widehat{\theta}$, an estimator of the parameter θ, and $\widehat{\mathrm{V}}$, an estimator of $\mathrm{V}(\widehat{\theta})$. The Monte Carlo relative bias of $\widehat{V}$ is

$$\mathrm{RB}_{MC}(\widehat{\mathrm{V}}) = \frac{\mathrm{E}_{MC}(\widehat{\mathrm{V}}) - \mathrm{V}_{MC}(\widehat{\theta})}{\mathrm{V}_{MC}(\widehat{\theta})} \times 100,$$

where

$$\mathrm{E}_{MC}(\widehat{\mathrm{V}}) = \frac{1}{R}\sum_{r=1}^{R}\widehat{\mathrm{V}}^{(r)}, \quad \mathrm{V}_{MC}(\widehat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\left[\widehat{\theta}^{(r)} - \mathrm{E}_{MC}(\widehat{\theta})\right]^2$$

and $\widehat{\mathrm{V}}^{(r)}$ and $\widehat{\theta}^{(r)}$ are the values of estimators $\widehat{\mathrm{V}}$ and $\widehat{\theta}$ at simulation r, $r = 1, \ldots, R$. The Monte Carlo relative root mean squared error of $\widehat{V}$ is,

$$\mathrm{RRMSE}_{MC}(\widehat{\mathrm{V}}) = \frac{\sqrt{\mathrm{E}_{MC}\left\{\left[\widehat{\mathrm{V}} - \mathrm{V}_{MC}(\widehat{\theta})\right]^2\right\}}}{\mathrm{V}_{MC}(\widehat{\theta})} \times 100.$$

4.6.4. Simulation results

Tables 4.1-4.3 present the results of the simulated data with $N = 500$, with $N = 1000$ and of the dataset '`Ilocos`' respectively. The first five rows of the tables correspond to the variance estimators obtained with the linearization method (Lin) and the last five rows correspond to the bootstrap variance estimators (Boot). The relative biases and relative root mean squared errors (RRMSE), between brackets, are presented for each sampling rate in columns.

Consider the results of the linearization method (Lin) in Tables 4.1-4.3. Except for the Gini index variance estimator, all the relative biases are smaller than 5.00% in absolute terms. Globally, the sample size and the population size do not seem to have a large impact on the relative biases, but as the sample size and the population size increase, the relative root mean squared error seems to decrease. The variance of the Gini index have relative biases smaller than 12.00% in absolute terms and it tends to decrease as n and N increase. This difference in terms of relative biases could be explained by the instability of the variance estimators of dispersion parameters. As a matter of fact, the relative root mean squared errors of the Gini variance estimators are greater than the others.

In the simulated datasets, Tables 4.1 and 4.2, the relative biases of the bootstrap variance estimators (Boot) are smaller than 5.00% in absolute terms. Except for

the Gini index with smaller population and sample sizes, the bootstrap method is comparable to the linearization one, in terms of relative biases and relative root mean squared errors. In the real dataset, Table 4.3, the relative biases of the bootstrap estimators in the presence of nonresponse tend to increase, up to 21.19% in absolute terms, when the sample size increases. Indeed, in the presence of imputed data, this bootstrap variance estimator is known to be reliable when the sampling rate is small (Shao and Sitter, 1996).

More generally, the objectives of the linearization method are to obtain explicit variance estimators and to have a small computation cost. In other words, the variance estimators should have a comparable precision, but a lower computational cost than a resampling method. In this simulation study, the precision of both methods is broadly comparable. The main exceptions concern the Gini index in function of the sampling rate and the nonresponse.

Tab. 4.1. Relative biases (RB) and relative root mean squared errors (RRMSE) of the variance estimators in the simulated population of $N = 500$ units.

		$n = 50$		$n = 100$		$n = 150$		$n = 200$	
$\widehat{\theta}$	Method	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
$\widehat{G}$	Lin	-11.35	(34.31)	-6.10	(23.79)	-3.49	(17.66)	-5.13	(14.35)
$\widehat{g}$	Lin	0.76	(20.97)	0.71	(14.16)	-1.03	(10.63)	-2.18	(8.69)
$\widehat{Y}_I$	Lin	-3.77	(21.55)	-1.42	(14.33)	-1.14	(10.82)	-0.27	(8.67)
$\widehat{G}_I$	Lin	1.23	(32.19)	-0.62	(21.19)	0.63	(16.39)	-1.50	(13.00)
$\widehat{g}_I$	Lin	0.73	(20.94)	0.72	(14.15)	-0.99	(10.62)	-2.17	(8.68)
$\widehat{G}$	Boot	2.85	(39.89)	-0.03	(25.52)	0.36	(18.79)	-2.35	(14.84)
$\widehat{g}$	Boot	0.77	(21.34)	0.69	(14.88)	-1.08	(11.59)	-2.21	(9.76)
$\widehat{Y}_I$	Boot	-1.26	(22.24)	-0.28	(15.18)	-0.48	(11.74)	0.31	(9.79)
$\widehat{G}_I$	Boot	-0.07	(31.41)	-1.39	(21.31)	0.08	(16.82)	-1.88	(13.67)
$\widehat{g}_I$	Boot	0.74	(21.31)	0.70	(14.87)	-1.05	(11.58)	-2.21	(9.75)

4.7. Discussion

An important aspect of the proposed linearization method is that it is usable in any circumstance, if the derivatives of the parameter exist and the remainder is small. This is not necessarily the case of other linearization methods. As seen in Sections 4.4 and 4.5, our linearization method gives similar results as in Kott (2006) or in Kim and Rao (2009). Nevertheless, the derivation of the linearized variable is simpler and faster.

Tab. 4.2. Relative biases (RB) and relative root mean squared errors (RRMSE) of the variance estimators in the simulated population of $N = 1000$ units.

		$n = 100$		$n = 200$		$n = 300$		$n = 400$	
$\widehat{\theta}$	Method	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
$\widehat{G}$	Lin	-6.11	(24.13)	-2.49	(15.48)	-2.49	(11.75)	-1.15	(9.15)
$\widehat{g}$	Lin	0.55	(13.70)	3.67	(10.00)	-0.62	(6.87)	1.27	(5.72)
$\widehat{Y}_I$	Lin	-2.42	(15.45)	0.46	(10.21)	-2.74	(8.01)	0.10	(6.19)
$\widehat{G}_I$	Lin	-1.48	(21.67)	0.20	(14.66)	-0.16	(11.25)	0.50	(9.02)
$\widehat{g}_I$	Lin	0.57	(13.70)	3.69	(10.01)	-0.61	(6.87)	1.27	(5.72)
$\widehat{G}$	Boot	0.50	(26.33)	0.68	(16.79)	-0.53	(12.6)	0.41	(10.37)
$\widehat{g}$	Boot	0.67	(14.46)	3.70	(10.99)	-0.53	(8.27)	1.25	(7.31)
$\widehat{Y}_I$	Boot	-1.32	(16.1)	1.00	(11.28)	-2.35	(9.10)	0.38	(7.69)
$\widehat{G}_I$	Boot	-2.37	(21.63)	-0.21	(15.16)	-0.54	(11.94)	0.30	(10.01)
$\widehat{g}_I$	Boot	0.69	(14.46)	3.71	(10.99)	-0.53	(8.27)	1.25	(7.30)

Tab. 4.3. Relative biases (RB) and relative root mean squared errors (RRMSE) of the variance estimators in the dataset Ilocos of $N = 632$ units.

		$n = 63$		$n = 126$		$n = 190$		$n = 253$	
$\widehat{\theta}$	Method	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
$\widehat{G}$	Lin	-5.31	(30.92)	-2.32	(20.46)	-1.81	(15.54)	-2.43	(12.37)
$\widehat{g}$	Lin	0.80	(17.24)	-0.50	(11.29)	1.30	(8.89)	1.12	(7.11)
$\widehat{Y}_I$	Lin	-4.98	(23.90)	-3.28	(16.02)	-1.21	(12.65)	-1.27	(10.83)
$\widehat{G}_I$	Lin	1.07	(37.53)	-1.78	(24.54)	-6.34	(19.21)	-9.25	(17.54)
$\widehat{g}_I$	Lin	-3.43	(22.48)	-2.71	15.43)	0.35	(12.45)	-0.22	(10.51)
$\widehat{G}$	Boot	-2.10	(31.10)	-1.09	(20.79)	-1.03	(16.07)	-1.82	(13.02)
$\widehat{g}$	Boot	0.76	(17.78)	-0.51	(12.15)	1.23	(9.94)	1.11	(8.45)
$\widehat{Y}_I$	Boot	-3.66	(24.39)	-8.21	(17.16)	-11.32	(16.13)	-16.88	(19.21)
$\widehat{G}_I$	Boot	-2.03	(36.15)	-7.42	(24.51)	-14.79	(22.29)	-21.19	(24.89)
$\widehat{g}_I$	Boot	-3.98	(23.00)	-8.16	(16.81)	-10.00	(15.24)	-15.75	(18.20)

The interest of our method is that it enables to linearize the variables in such a way that all the sources of variability and the statistical treatments are taken into account: the random sample, the nonresponse, the calibration, the reweighting for non-response and the imputation.

We considered the case of deterministic imputation in Section 4.5. The extension of the linearization method to deal with random imputation should be relatively straightforward. However, the application to the case of nearest neighbor imputation does not seem direct.

4.8. Proofs of the Propositions

4.8.1. Proof of Proposition 4.1

PROOF. First, let derive constraint (4.2.7) with respect to indicator a_ℓ and obtain

$$d_\ell F_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda})\mathbf{x}_\ell + \sum_{k\in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda})\mathbf{x}_k\mathbf{x}_k^\top \frac{\partial \boldsymbol{\lambda}}{\partial a_\ell} = 0.$$

This leads to

$$\frac{\partial \boldsymbol{\lambda}}{\partial a_\ell} = -\widehat{\mathbf{T}}_{xx}^{-1} d_\ell F_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda})\mathbf{x}_\ell = -\widehat{\mathbf{T}}_{xx}^{-1} w_\ell \mathbf{x}_\ell. \tag{4.8.1}$$

The partial derivative of the calibration estimator is then given by

$$\frac{\partial \widehat{Y}}{\partial a_\ell} = w_\ell y_\ell - \sum_{k\in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) y_k \mathbf{x}_k^\top \widehat{\mathbf{T}}_{xx}^{-1} w_\ell \mathbf{x}_\ell = w_\ell e_\ell. \tag{4.8.2}$$

□

4.8.2. Proof of Proposition 4.2

PROOF. Noting that vector $\mathbf{w}$ depends on $\mathbf{a}$, the linearization variable is $z_\ell^w = w_\ell v_\ell^w$, with

$$v_\ell^w = h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) + \sum_{k\in U} a_k h_{k2}(\mathbf{y}, \mathbf{a}, \mathbf{w}) \frac{\partial w_k}{\partial a_\ell}.$$

The derivative of w_k is obtained by differentiating Equation (4.2.7) with respect to a_ℓ:

$$w_\ell \mathbf{x}_\ell + \sum_{k\in U} a_k d_k \mathbf{x}_k \mathbf{x}_k^\top F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \frac{\partial \boldsymbol{\lambda}}{\partial a_\ell} = \mathbf{0}.$$

Thus

$$\begin{aligned}
\frac{\partial \boldsymbol{\lambda}}{\partial a_\ell} &= -w_\ell \widehat{\mathbf{T}}_{xx}^{-1} \mathbf{x}_\ell, \\
\frac{\partial w_k}{\partial a_\ell} &= d_k \frac{\partial F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})}{\partial a_\ell} = -d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k^\top \widehat{\mathbf{T}}_{xx}^{-1} w_\ell \mathbf{x}_\ell
\end{aligned}$$

and

$$\begin{aligned}
v_\ell^w &= h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) - \mathbf{x}_\ell^\top \widehat{\mathbf{T}}_{xx}^{-1} \sum_{k\in U} a_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda}) \mathbf{x}_k h_{k2}(\mathbf{y}, \mathbf{a}, \mathbf{w}) \\
&= h_{\ell 1}(\mathbf{y}, \mathbf{a}, \mathbf{w}) - \mathbf{x}_\ell^\top \widehat{\mathbf{B}}_{h1|x}.
\end{aligned}$$

□

4.8.3. Proof of Proposition 4.3

PROOF. The partial derivative of equation (4.4.5) with respect to δ_ℓ is

$$d_\ell F_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda})\mathbf{x}_\ell + \sum_{k\in U} \delta_k d_k F_k'(\mathbf{x}_k^\top \boldsymbol{\lambda})\mathbf{x}_k\mathbf{x}_k^\top \frac{\partial \boldsymbol{\lambda}}{\partial \delta_\ell} = 0.$$

This implies that $\partial\boldsymbol{\lambda}/\partial\delta_\ell = -\widehat{\mathbf{T}}_{xx}^{-1} w_\ell \mathbf{x}_\ell$. The linearization variable of $\widehat{Y}_R$ is

$$z_\ell = \frac{\partial \widehat{Y}_R}{\partial \delta_\ell} = d_\ell F_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda}) y_\ell - \sum_{k\in U} \delta_k d_k F_k(\mathbf{x}_k^\top \boldsymbol{\lambda}) y_k \mathbf{x}_k^\top \widehat{\mathbf{T}}_{xx}^{-1} w_\ell \mathbf{x}_\ell = w_\ell e_\ell.$$

□

4.8.4. Proofs of propositions 4.4 and 4.5

Proof. By considering Proposition 4.1, the linearization variable with respect to a_ℓ of $\sum_{k\in U} a_k w_{k1} \mathbf{x}_k^\circ$ is $\mathbf{z}_\ell^\circ = w_{\ell 1}\mathbf{e}_\ell^\circ$. The partial derivative of the calibration constraint (4.4.7) with respect to a_ℓ is

$$R_\ell w_{\ell 2} \mathbf{x}_\ell + \widehat{\mathbf{T}}_{xx} \frac{\partial \boldsymbol{\lambda}_2}{\partial a_\ell} = \begin{pmatrix} \mathbf{0} \\ w_{\ell 1}\mathbf{e}_\ell^\circ \end{pmatrix},$$

which leads to

$$\frac{\partial \boldsymbol{\lambda}_2}{\partial a_\ell} = \widehat{\mathbf{T}}_{xx}^{-1} \left[\begin{pmatrix} \mathbf{0} \\ w_{\ell 1}\mathbf{e}_\ell^\circ \end{pmatrix} - R_\ell w_{\ell 2} \mathbf{x}_\ell \right].$$

The partial derivative of the estimator with respect to a_ℓ is

$$\frac{\partial \widehat{\theta}_{NR}(\mathbf{y}, \mathbf{a}, \mathbf{R})}{\partial a_\ell} = R_\ell w_{\ell 2} y_\ell + \widehat{\mathbf{t}}_{xy}^\top \frac{\partial \boldsymbol{\lambda}_2}{\partial a_\ell} = R_\ell w_{\ell 2} y_\ell + \left[\begin{pmatrix} \mathbf{0} \\ w_{\ell 1}\mathbf{e}_\ell^\circ \end{pmatrix} - R_\ell w_{\ell 2} \mathbf{x}_\ell \right]^\top \widehat{\mathbf{B}}_{y|x}.$$

□

Proof. When $\mathbf{a} = \boldsymbol{\pi}$, the estimator becomes

$$\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) = \sum_{k\in U} \pi_k R_k w_{k2}^\pi y_k,$$

where the weights are obtained by solving

$$\sum_{k\in U} \pi_k R_k w_{k2}^\pi \mathbf{x}_k = \begin{pmatrix} \mathbf{X}^\bullet \\ \sum_{k\in U} \pi_k w_{k1}^\pi \mathbf{x}_k^\circ \end{pmatrix},$$

and

$$\sum_{k\in U} \pi_k w_{k1}^\pi \mathbf{x}_k^\bullet = \mathbf{X}^\bullet.$$

The partial derivative of $\widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R})$ with respect to R_ℓ is

$$\frac{\partial \widehat{\theta}_{NR}(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R})}{\partial R_\ell} = \pi_\ell w_{2\ell}^\pi y_\ell + \sum_{k\in U} R_k F_{k2}'(\mathbf{x}_k^\top \boldsymbol{\lambda}_2^\pi) \mathbf{x}_k^\top \frac{\partial \boldsymbol{\lambda}_2^\pi}{\partial R_\ell} y_k = F_{\ell 2}(\mathbf{x}_\ell \boldsymbol{\lambda}_2^\pi) e_\ell^\pi,$$

where $\partial\boldsymbol{\lambda}_2^\pi/\partial R_\ell$ is the solution to

$$\frac{\partial \sum_{k\in U} \pi_k R_k w_{k2}^\pi \mathbf{x}_k}{\partial R_\ell} = \pi_\ell w_{2\ell}^\pi \mathbf{x}_\ell + \sum_{k\in U} R_k F_{k2}'(\mathbf{x}_k^\top \boldsymbol{\lambda}_2^\pi) \mathbf{x}_k \mathbf{x}_k^\top \frac{\partial \boldsymbol{\lambda}_2^\pi}{\partial R_\ell} = \mathbf{0}.$$

□

4.8.5. Proof of Proposition 4.6

PROOF. The partial derivative of $\widehat{\theta}_{NR}$ with respect to a_ℓ is

$$z_\ell^a = \frac{\partial \widehat{\theta}_{NR}}{\partial a_\ell} = \frac{h_{\ell 1}(\widetilde{\mathbf{y}}, \mathbf{a}, \mathbf{R})}{\pi_\ell} + \sum_{k \in U} a_k (1 - R_k) h_{k2}(\widetilde{\mathbf{y}}, \mathbf{a}, \mathbf{R}) \frac{\partial y_k^*}{\partial a_\ell}.$$

In the case of regression imputation,

$$\frac{\partial y_k^*}{\partial a_\ell} = \frac{R_\ell e_\ell \mathbf{x}_\ell^\top}{\pi_\ell} \widehat{\mathbf{T}}_{xx}^{-1} \mathbf{x}_k.$$

□

Chapter 5

Balanced Imputation for Swiss Cheese Nonresponse

Abstract: Swiss cheese nonresponse, also known as non-monotone nonresponse, occurs when every variables of a survey contains missing values without a particular pattern. The estimators of the parameters of interest can be considerably affected by the missing values which introduce a bias and an increase variability. To reduce the effects of nonresponse, the missing values are usually imputed. When several variables of a dataset need to be imputed, it may be difficult to preserve the distributions and the relations between the variables. In this work, balanced K-nearest neighbor imputation (Hasler and Tillé, 2016) is extended to treat Swiss cheese nonresponse. The method uses random imputations by donors. It is constructed to meet the following requirements. First, a nonrespondent should be imputed by neighboring donors. The distances are calculated with the observed values. Next, all missing values of a nonrespondent should be imputed by the same donor. Last, the donors are selected in order to respect balancing constraints allowing to decrease the variance of the estimators. To meet all the requirements, a matrix of imputation probabilities is constructed using calibration techniques. The donors are then selected with these imputation probabilities and balanced sampling methods.[1]

Keywords: calibration; non-monotone nonresponse; random imputation; variance.

5.1. Introduction

In large scale surveys, nonresponse is often inevitable. Two types of nonresponse are distinguished: unit nonresponse and item nonresponse. Unit nonresponse occurs when the values of every variables are missing for a group of units. Item nonresponse occurs when some, but not all, variables are missing for a set of units. The estimators of the parameters of interest may be considerably affected by the missing values, which can introduce a bias and cause an additional variability. Reweighting the

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

respondent units and imputing the missing values allow to reduce the undesirable effects of nonresponse.

In surveys with multiple variables of interest, different types of item nonresponse can occur. In a first case, the nonresponse can be univariate and only one variable contains missing values. The other variables of the survey are completely observed and they can be used to impute the missing values. Several imputation methods have been developed in this context. Haziza (2009) presents an overview of deterministic and random imputation methods, including multiple and fractional imputation. Andridge and Little (2010) present an overview of donor imputation methods. In a second case, the nonresponse can be multivariate and affect multiple variables. Multivariate nonresponse can have a monotone or a non-monotone pattern. The monotonous case is suitable to longitudinal studies, when units quit their study over time. Recently, Hasler et al. (2018) used vine copulas for imputation of monotone nonresponse. The authors also presented an overview of other imputation methods for this pattern. The non-monotone nonresponse, also known as Swiss cheese nonresponse, is of interest in this paper. This type of nonresponse occurs when all the variables of a survey contain missing values without a particular pattern. There are few treatments for such a multivariate nonresponse. It is difficult to preserve the distributions of the variables and the relationships between the variables with such missing values in the dataset. Judkins (1997) proposes donor imputation methods and Andridge and Little (2010) present an overview of existing methods. Some methods allow to impute iteratively; for instance Raghunathan et al. (2001) use a sequence of regression models between the variables.

Hasler and Tillé (2016) develop balanced K-nearest neighbor imputation to treat the univariate case of item nonresponse. The imputation method has advantages that would be interesting in the case of Swiss cheese nonresponse. In this paper, balanced K-nearest neighbor imputation is extended to the multivariate case. This extension is not trivial since it needs to manage missing values simultaneously for several variables. The imputation methods is elaborated in order to respect four requirements. First, it should be a donor imputation method. A random imputation method tends to preserve the distributions of the variables and a donor imputation method allows to impute both continuous and categorical variables. Second, the same donor is selected to replace all missing values of a nonrespondent. This should preserve the relations between the variables. Third, a nonrespondent can be imputed by donors that are close to it in order to ensure coherence between imputed and observed values. Last, balancing constraints are used to reduce the additional variability of the estimated parameters. The context and the requirements of the method are presented in Sections 5.2.1 and 5.2.2. The construction of the matrix of imputation probabilities is detailed in Section 5.3. The selection of the donors is

presented in Section 5.4.1 and the imputation in Section 5.4.2. A simulation study is detailed in Section 5.5.

5.2. Motivation

5.2.1. Swiss Cheese nonresponse

Consider a finite population U of size N with J variables of interest. A random sample S of size n is selected in U. The first order inclusion probability of unit k is π_k, the second order inclusion probability of units k and ℓ is $\pi_{k\ell}$ and $\pi_{kk} = \pi_k$, for any $k, \ell \in U$. The vector of J variables of interest, $\mathbf{x}_k = (x_{k1} \dots x_{kj} \dots x_{kJ})^\top$, is not necessarily fully observed for all $k \in S$. The vector of response indicators of unit $k \in S$ is $\mathbf{r}_k = (r_{k1} \dots r_{kj} \dots r_{kJ})^\top$, where r_{kj} is 1 if the variable j of unit k is observed and 0 otherwise. Consider $S_r \subset S$ the set of $n_r > 0$ units for which the J variables are completely observed. That is, $r_{ij} = 1$ for all $j = 1, \dots J$ and any $i \in S_r$. Consider $S_m = S - S_r$, a set of $n_m = n - n_r$ units such that some values, but not all, are missing. The nonresponse is non-monotone, it has no particular pattern. Figure 5.1 illustrates a dataset with Swiss cheese nonresponse.

Suppose that the missing values are treated by imputation and that the imputed value of unit k for a variable j is x^*_{kj}. Then the population total of the variable j, $X_j = \sum_{k \in U} x_{kj}$, can be estimated by

$$\widehat{X}_j = \sum_{k \in S} r_{kj} d_k x_{kj} + \sum_{k \in S} (1 - r_{kj}) d_k x^*_{kj},$$

where $d_k = \pi_k^{-1}$ is the sampling weight of unit k.

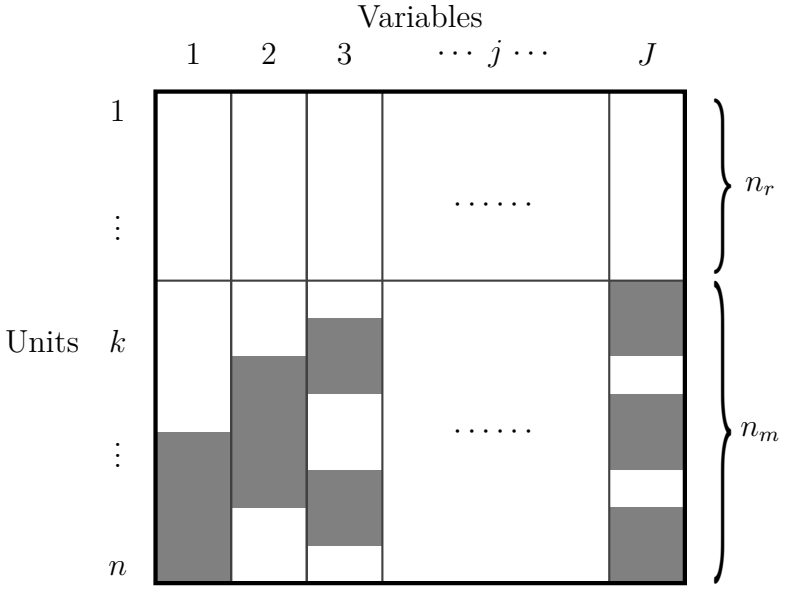

Fig. 5.1. Representation of Swiss cheese nonresponse in a dataset of n units and J variables. The first n_r rows correspond to the respondents and the next n_m rows correspond to the nonrespondents. The gray rectangles cover the missing values.

5.2.2. Requirements

The proposed method is elaborated to ensure coherence and accuracy in the imputed dataset. It is based on the following four requirements:

(i) The imputed values should be selected among the values of the n_r completely observed units: a donor imputation method should be used.

(ii) Only one donor should be selected per nonrespondent: all the missing values of a unit should be imputed by the same donor.

(iii) The donors should be selected among the K nearest neighbors of the unit with missing values.

(iv) If the observed values of the nonrespondents were imputed, the total estimator of each variable should remain unchanged.

Requirement (i) ensures that the imputed values are observed, therefore realistic, for both categorical and continuous variables. Also, a random imputation method tends to preserve the distributions of the variables. To illustrate requirement (ii), consider $J = 3$ variables and a nonrespondent $k \in S_m$ such that $\mathbf{r}_k = (1, 0, 0)^\top$. The missing values of unit k, x_{k2} and x_{k3}, are imputed by observed values selected on the same donor. This means that x_{k2} and x_{k3} are imputed by x_{i2} and x_{i3} of a selected donor $i \in S_r$. Requirement (ii) aims to preserve the relationships between the variables. Requirement (iii) allows the imputation of a nonrespondent by a similar unit. This ensures a coherence between the imputed values and the observed values of a nonrespondent. For instance, if the gender and the height of people are measured, a missing height of a man should be imputed by the height of another man. The idea behind requirement (iv) is that the observed information is unchanged if the units with missing values were completely imputed. This requirement results in a decrease in the variance of the estimators.

To implement a donor imputation method, each fully observed unit receives a probability to donate its values to each nonrespondent. Next, one donor per nonrespondent can be selected with respect to those imputation probabilities. The imputation probabilities respecting requirements (i)-(iv) are detailed in Section 5.3. The selection of the donors is detailed in Section 5.4.1.

5.3. Matrix of imputation probabilities

5.3.1. Determination of the imputation probabilities

The first step of a donor imputation method is to assign imputation probabilities to complete respondents. Consider $\boldsymbol{\psi} = (\psi_{ik})$, where $(i, k) \in S_r \times S_m$, the matrix of imputation probabilities. The element ψ_{ik} is the probability that respondent i gives its values to nonrespondent k and

$$\psi_{ik} \geq 0. \tag{5.3.1}$$

Only one donor should be randomly selected for each unit $k \in S_m$, see requirement (ii). If the donors are chosen using balanced sampling as suggested in Section 5.4.1, it is sufficient to ensure that the sum of the imputation probabilities associated with a nonrespondent is 1. This means

$$\sum_{i \in S_r} \psi_{ik} = 1, \tag{5.3.2}$$

for $k \in S_m$. When the donors are selected, the missing values of unit k are imputed by the corresponding values of the chosen donor and requirement (ii) will be fulfilled. Requirement (iii) limits the set of possible donors of a unit to its K nearest neighbors. The probability that unit $i \in S_r$ gives its values to the nonrespondent $k \in S_m$ should be non-zero only if i is one of the K nearest neighbors of k. Thus,

$$\psi_{ik} = 0 \text{ if } i \notin \text{knn}(k), \tag{5.3.3}$$

where $\text{knn}(\ell) = \{j \in S_r | \, \text{rank}\,(d(j,\ell)) \leq K\}$ and $d(\cdot,\cdot)$ is a distance function. The distance between units $\ell \in S_m$ and $j \in S_r$, $d(j,\ell)$, could for instance be the Mahalanobis distance calculated with the variables that are not missing for unit ℓ.

Requirement (iv) suggests that if the observed values of any $k \in S_m$ were imputed by the corresponding values of the donor, the total estimator of each variable would remain the same as the total estimator calculated with the observed values only. The imputation probabilities are therefore chosen so that if the known values of the units in S_m were imputed by the expectation of their imputed value, the total estimators would correspond to the estimators based on the observed values. So the imputation probabilities must satisfy

$$\sum_{k \in S_m} d_k r_{kj} \sum_{i \in S_r} \psi_{ik} x_{ij} = \sum_{k \in S_m} d_k r_{kj} x_{kj}, \tag{5.3.4}$$

for $j = 1, \ldots, J$, see also Figure 5.2. The right hand side of Equation (5.3.4) is the estimated total of the j^{th} variable calculated on the observed values in S_m, see Figure 5.2a. The left hand side of Equation (5.3.4) is the same estimated total, but calculated with the values that would be imputed if the observed values were imputed. Thus suppose that each observed x_{kj} such that $k \in S_m$ and $r_{kj} = 1$ is imputed by

$$\sum_{i \in S_r} \psi_{ik} x_{ij}.$$

The hatched region in Figure 5.2e represents those imputed values. Then the total of those imputed values corresponds to the right hand side of Equation (5.3.4), see Figure 5.2f.

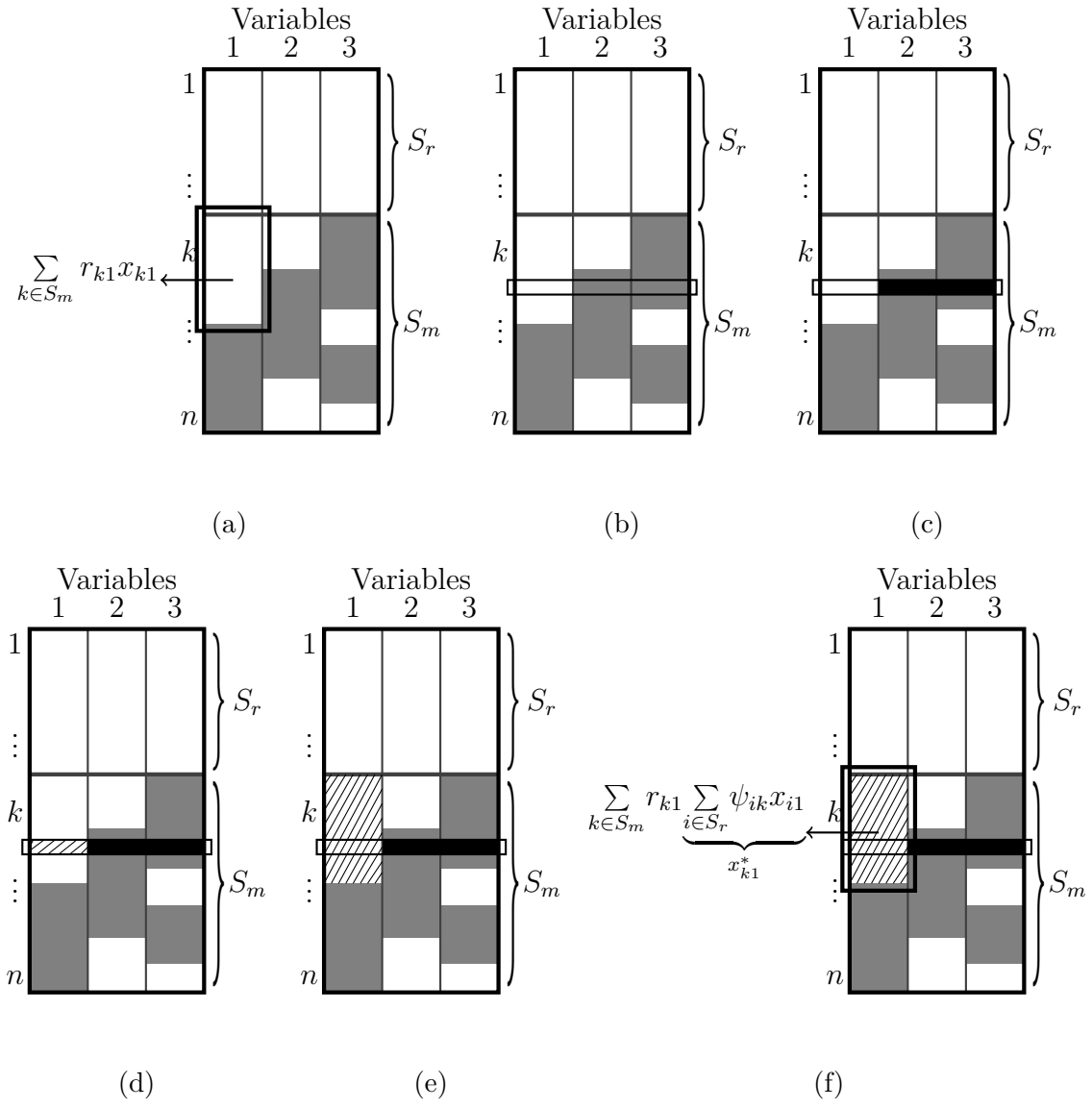

Fig. 5.2. Representation of Swiss cheese nonresponse in a dataset of n units and $J = 3$ variables. The gray rectangles cover the missing values. Figure 5.2a represents the right hand side of Equation (5.3.4) for variable x_1. For some nonrespondent (Figure 5.2b), the missing values are imputed in black (Figure 5.2c). Suppose that the observed value of the nonrespondent is also imputed in hatched (Figure 5.2d), as well as for each nonrespondent (Figure 5.2e). The left hand side of Equation (5.3.4) is then represented in Figure 5.2f.

Equation (5.3.4) can be rewritten as

$$\sum_{i \in S_r} \left(\sum_{k \in S_m} d_k r_{kj} \psi_{ik} \right) r_{ij} x_{ij} = \sum_{k \in S_m} d_k r_{kj} x_{kj}, \tag{5.3.5}$$

for $j = 1, \cdots, J$. The imputation probabilities respecting (5.3.5) can be found by calibration (Deville and Särndal, 1992). Consider the initial imputation probabilities

$$\psi_{ik}^0 = \begin{cases} \dfrac{1}{K} & \text{if } i \in \text{knn}(k), \\ 0 & \text{otherwise.} \end{cases} \tag{5.3.6}$$

Final imputation probabilities ψ_{ik} close to ψ_{ik}^0 respecting (5.3.5) are sought. Consider the pseudo-distance function

$$G\left(\psi_{ik}, \psi_{ik}^0\right) = \psi_{ik} \log\left(\frac{\psi_{ik}}{\psi_{ik}^0}\right) + \psi_{ik}^0 - \psi_{ik},$$

then the elements of $\boldsymbol{\psi}$ are obtained by minimizing

$$\begin{aligned} &\mathcal{L}(\psi_{11}, \ldots, \psi_{n_r n_m}, \lambda_1, \ldots, \lambda_J) \\ &\quad = \sum_{k \in S_m} \sum_{i \in S_r} G\left(\psi_{ik}, \psi_{ik}^0\right) - \sum_{j=i}^{J} \lambda_j \left(\sum_{i \in S_r} \sum_{k \in S_m} d_k r_{kj} \psi_{ik} r_{ij} x_{ij} - \sum_{k \in S_m} d_k r_{kj} x_{kj} \right). \end{aligned}$$

Using

$$\frac{\partial \mathcal{L}}{\partial \psi_{ik}} = \log \frac{\psi_{ik}}{\psi_{ik}^0} - \sum_{j=1}^{J} \lambda_j d_k r_{kj} r_{ij} x_{ij} = 0,$$

the imputation probabilities are

$$\psi_{ik} = \psi_{ik}^0 \exp\left(\sum_{j=1}^{J} \lambda_j d_k r_{kj} r_{ij} x_{ij} \right). \tag{5.3.7}$$

Calibration techniques are used to find $\boldsymbol{\lambda} = (\lambda_1 \ldots \lambda_j \ldots \lambda_J)^\top$ respecting Equation (5.3.4). Because of the initial probabilities (5.3.6) and expression (5.3.7), the resulting matrix $\boldsymbol{\psi}$ respects Equation (5.3.3). In contrast, nothing ensures that Equation (5.3.1) holds and that one donor per nonrespondent will be selected. We thus propose an algorithm to find the matrix $\boldsymbol{\psi}$ respecting Equations (5.3.1)-(5.3.4) simulatneously.

5.3.2. Algorithms

An algorithm is needed to find imputation probabilities that sum to 1 for each $k \in S_m$, as in Equation (5.3.2), while respecting the calibration equations (5.3.5). Algorithm 5.1 produces imputation probabilities satisfying the four requirements. First, imputation probabilities respecting the calibration constraints for all variables are found simultaneously. Then, they are normalized to sum to one for each nonrespondent. Those two steps are repeated until Equation (5.3.2) is perfectly respected and the calibration equations (5.3.5) are approximately respected. A major

problem with Algorithm 5.1 is that it is computationally demanding. A total of $n_r \times n_m$ weights (imputation probabilities) need to be found by calibration. This may be intensive for a calibration program. Alternatively, Algorithm 5.2 reduces the computer effort by considering a smaller calibration problem. Imputation probabilities are calibrated for one variable at the time iteratively and next, they are normalized. The number of calibration constraints is dramatically reduced, but the number of iteration is increased. Indeed, each calibration problem is adjusting n_r weights respecting a calibration equation for one variable only. Altogether, the decrease in computer effort is non negligible and thanks to the exponential form of the imputation probabilities in (5.3.7), it does not affect the accuracy.

It is worth noting that the initialization of the algorithms is a delicate matter. The initial imputation probabilities ψ_{ik}^0 given in (5.3.6) are directly linked to K, the number of neighbors that are possible donors. The number K should not be too large in order to be as accurate as possible. On the other hand, if K is too small, the algorithms may not converge, there might be no solutions to the requirements. Thus we suggest to begin with some not too large K. If the algorithms do not converge, increase by one the value of K until a solution is found.

Algorithm 5.1 Imputation probabilities for J variables simultaneously

---

Initialize

1: Vector of initial imputation probabilities

$$\boldsymbol{\psi}_v^{(1)} = \left(\psi_{11}^0, \cdots, \psi_{1n_m}^0, \psi_{21}^0, \cdots, \psi_{n_r n_m}^0\right)^\top.$$

2: Matrix of calibration variables

$$\widehat{\mathbf{A}} = \begin{pmatrix} d_1 r_{11} r_{11} x_{11} & d_1 r_{12} r_{12} x_{12} & \cdots & d_1 r_{1J} r_{1J} x_{1J} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n_m} r_{n_m 1} r_{11} x_{11} & d_{n_m} r_{n_m 2} r_{12} x_{12} & \cdots & d_{n_m} r_{n_m J} r_{1J} x_{1J} \\ d_1 r_{11} r_{21} x_{21} & d_1 r_{12} r_{22} x_{22} & \cdots & d_1 r_{1J} r_{2J} x_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n_m} r_{n_m 1} r_{n_r 1} x_{n_r 1} & d_{n_m} r_{n_m 2} r_{n_r 2} x_{n_r 2} & \cdots & d_{n_m} r_{n_m J} r_{n_r J} x_{n_r J} \end{pmatrix}.$$

3: Vector of totals for calibration equations $\widehat{\mathbf{X}}_m = \left(\widehat{X}_{1m}, \cdots, \widehat{X}_{Jm}\right)^\top$, where $\widehat{X}_{jm} = \sum_{k \in S_m} d_k r_{kj} x_{kj}$, for $j = 1, \cdots, J$.

Iterate For $t = 1, 2, \ldots$

1: *Calibrate* with the vector of initial weights

$$\mathbf{d}_v = \boldsymbol{\psi}_v^{(2t-1)} = \left(\psi_{11}^{(2t-1)}, \cdots, \psi_{n_r n_m}^{(2t-1)}\right)^\top,$$

to obtain the vector of calibrated weights

$$\mathbf{w}_v = \left[\psi_{11}^{(2t-1)} \exp\left(\sum_{j=1}^{J} d_k r_{1j} r_{1j} x_{1j} \lambda_j\right), \cdots, \psi_{n_r n_m}^{(2t-1)} \exp\left(\sum_{j=1}^{J} d_k r_{n_m j} r_{n_r j} x_{n_r j} \lambda_j\right)\right]^\top,$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)^\top$ is the solution to $\widehat{\mathbf{A}}^\top \mathbf{w_v} = \widehat{\mathbf{X}}_m$.

2: *Reweight* the vector of imputation probabilities; $\boldsymbol{\psi}_v^{(2t)} = \mathbf{w}_v$.

3: *Normalize* the vector of imputation probabilities;

$$\boldsymbol{\psi}_v^{(2t+1)} = \left(\psi_{11}^{(2t)} / \sum_{i \in S_r} \psi_{i1}^{(2t)}, \cdots, \psi_{n_r n_m}^{(2t)} / \sum_{i \in S_r} \psi_{i n_m}^{(2t)} \right)^\top .$$

4: *Repeat*, for some small tolerance δ, until

$$\max_{1 \le j \le J} \left| \frac{\sum_{i \in S_r} \sum_{k \in S_m} d_k r_{kj} \psi_{ik}^{(2t+1)} x_{ij} - \widehat{X}_{jm}}{\widehat{X}_{jm}} \right| \le \delta.$$

End with the final imputation probability $\psi_{ik} = \psi_{ik}^{(2t+1)}$ for $(i,k) \in S_r \times S_m$

Algorithm 5.2 Imputation probabilities for J variables sequentially

Initialize For variables $j = 1, \ldots, J$

1: Initial imputation probability $\psi_{ikN}^{(0)} = \psi_{ik}^0$.

2: Vector of the calibration variable $\mathbf{x}_{rj} = (x_{1j}, \ldots, x_{n_r j})^\top$.

3: Total for calibration equations $\sum_{k \in S_m} d_k r_{kj} x_{kj}$.

Iterate For $v = 1, 2, \ldots$

1: *Calibrate*, for $t = 1, 2, \ldots$

(i) Set $\psi_{ik}^{(1)} = \psi_{ikN}^{(v-1)}$.

(ii)For $j = 1$ and for any $i \in S_r$, the initial weight is

$$d_{ij} = \sum_{k \in S_m} d_k r_{kj} \psi_{ik}^{((t-1)J+j)},$$

the calibrated weight is $w_{ij} = d_{ij} \exp(\lambda x_{ij})$, where λ is the solution to

$$\sum_{i \in S_r} w_{ij} x_{ij} = \sum_{k \in S_m} d_k r_{kj} x_{kj},$$

and the new imputation probability of $(i,k) \in S_r \times S_m$ is

$$\psi_{ik}^{((t-1)J+j+1)} = (1 - r_{kj}) \psi_{ik}^{((t-1)J+j)} + r_{kj} \psi_{ik}^{((t-1)J+j)} \exp(\lambda x_{ij}).$$

(iii) Repeat (i) for $j = 2, \ldots, J$

(iv) Repeat for $t = 1, 2, \ldots$, for some small *tolerance*, until

$$\max_{1 \le j \le J} \left| \frac{\sum_{i \in S_r} \sum_{k \in S_m} r_{kj} \psi_{ik}^{((t-1)J+j+1)} x_{ij} - \sum_{k \in S_m} d_k r_{kj} x_{kj}}{\sum_{k \in S_m} d_k r_{kj} x_{kj}} \right| \le tolerance.$$

2: *Normalize* the imputation probability;

$$\psi_{ikN}^{(v)} = \psi_{ik}^{((t-1)J+j+1)} / \sum_{i \in S_r} \psi_{ik}^{((t-1)J+j+1)}$$

3: *Repeat* steps 1 and 2 for $v = 1, 2, \ldots,$ for some small tolerance δ, until

$$\max_{1 \leq j \leq J} \left| \frac{\sum_{i \in S_r} \sum_{k \in S_m} r_{kj} \psi_{ikN}^{(v)} x_{ij} - \sum_{k \in S_m} d_k r_{kj} x_{kj}}{\sum_{k \in S_m} d_k r_{kj} x_{kj}} \right| \leq \delta.$$

End with the final imputation probability for $\psi_{ik} = \psi_{ikN}^{(v)}$ for $(i,k) \in S_r \times S_m$.

5.4. Imputation

5.4.1. Imputation matrix

Once the matrix of imputation probabilities $\boldsymbol{\psi}$ is completed, the donors can be randomly selected. Consider $\boldsymbol{\phi} = (\phi_{ik})$, where $(i,k) \in S_r \times S_m$, the imputation matrix. The element ϕ_{ik} is 1 if unit i is selected to donate its values to unit k, 0 otherwise. Only one donor is selected per nonrespondent, this means

$$\sum_{i \in S_r} \phi_{ik} = 1.$$

To respect requirement (iv), the donors should be selected so that

$$\sum_{k \in S_m} \sum_{i \in S_r} \frac{\phi_{ik}}{\psi_{ik}} d_k r_{kj} \psi_{ik} x_{ij} = \sum_{k \in S_m} \sum_{i \in S_r} d_k r_{kj} \psi_{ik} x_{ij}. \tag{5.4.1}$$

The cube method for balanced sampling (Deville and Tillé, 2004) is used to respect the balancing constraints (5.4.1). To ensure that only one donor is selected per nonrespondent and that the balancing constraints are respected, the matrix $\boldsymbol{\phi}$ is generated with stratified balanced sampling (Chauvet, 2009; Hasler and Tillé, 2014).

Consider a population of cells $U^* = \{(i,k) | i \in S_r, k \in S_m\}$ of size $n_r \cdot n_m$. The population is stratified in n_m strata $U_k^* = \{(i,k) | i \in S_r\}$, for $k \in S_m$. Each stratum corresponds to one nonrespondent. Stratum k contains the set of n_r possible donors of the nonrespondent k, $k \in S_m$. Each element i, or possible donor, of stratum k has a probability ψ_{ik} to be the selected donor of nonrespondent k, $i \in S_r$. Thus selecting one element in each stratum corresponds to selecting one donor for each nonrespondent.

In Algorithm 5.3, a sample of cells is selected. The inclusion probability of cell (i,k) is ψ_{ik}, for $(i,k) \in S_r \times S_m$. The selected sample of cells corresponds to the selected set of donors: if the cell (i,k) is selected in the sample, then $\phi_{ik} = 1$ and the respondent i is the donor for the nonrespondent k. If the cell (i,k) is not selected in the sample, $\phi_{ik} = 0$.

The sample of cells is selected by means of stratified balanced sampling. Hasler and Tillé (2014) proposed a method to select a stratified balanced sample when the number of strata is large. If the sum of the inclusion probabilities in each stratum is an integer, the method guaranties the selection of a fixed size sample in each stratum. The size of the sample in a stratum is the sum of the inclusion probabilities of the units in this stratum. The matrix $\boldsymbol{\psi}$ is such that $\sum_{i \in S_r} \psi_{ik} = 1$ for any $k \in S_m$,

thus the sum of the inclusion probabilities in each stratum is 1. Therefore, exactly one cell is selected per stratum, thus one donor is selected per nonrespondent. Moreover, by specifying the right balancing vectors in the method, the balancing equations (5.4.1) can be approximately respected. The balancing variable of cell (i,k) is defined as $d_k r_{kj} \psi_{ik} x_{ij}$ for $(i,k) \in S_r \times S_m$. The balancing equations might not be perfectly satisfied because of the complexity of the balancing problem. So requirements (5.3.1)-(5.3.4) are either perfectly or approximately fulfilled.

Algorithm 5.3 Selection of donors

1: Define the population of cells $U^* = \{(i,k) | i \in S_r, k \in S_m\}$;
2: Stratify population U^*, $U_k^* = \{(i,k) | i \in S_r\}$, for $k \in S_m$;
3: Define the balancing vector $\mathbf{x}_{ik}^* = (x_{ik,1}^* \dots x_{ik,J}^*)^\top$, where $x_{ik,j}^* = d_k r_{kj} \psi_{ik} x_{ij}$, for cell $(i,k) \in S_r \times S_m$;
4: Select a stratified balanced sample of cells respecting Equation (5.4.1), see Hasler and Tillé (2014);
5: If a cell is selected in the sample, then $\phi_{ik} = 1$ and $\phi_{ik} = 0$ otherwise.

5.4.2. Imputation

The imputation of the dataset is based on the matrix $\boldsymbol{\phi}$. The missing value of unit k for variable j such that $r_{kj} = 0$ is imputed by

$$x_{kj}^* = \sum_{i \in S_r} \phi_{ik} x_{ij}.$$

Requirements (i)-(iii) are perfectly respected. Requirement (iv) is either perfectly of approximately respected, according to the limitations of the balancing problem.

It is also possible to use a deterministic version of the proposed imputation method. The expectation of ϕ_{ik} is used for $(i,k) \in S_r \times S_m$. Then the missing value of unit $k \in S_m$ for j such that $r_{kj} = 0$ is imputed by

$$x_{kj}^* = \sum_{i \in S_r} \psi_{ik} x_{ij}.$$

It is not a donor imputation method anymore, but requirement (iv) is perfectly respected.

5.5. Simulation study

A simulation study was conducted to compare the proposed imputation method to K-nearest neighbor imputation. The population of interest in this study is 'MU284' from Särndal et al. (1992). The population is available in the `R package` '`sampling`' (Tillé and Matei, 2007). The population is a set of $N = 284$ Swedish municipalities. The following $J = 4$ variables are considered:

- $\mathbf{x}_1$: 1985 population, in thousands (`P85`);

- $\mathbf{x}_2$: 1975 population, in thousands (`P75`);
- $\mathbf{x}_3$: Revenues from the 1985 municipal taxation, in millions of kronor (`RMT85`);
- $\mathbf{x}_4$: Number of Conservative seats in municipal council (`CS82`).

The distributions of the variables are skewed: the Pearson's moment coefficient of skewness of the variables are 8.232, 8.472, 8.786 and 1.243 respectively. The correlations between the variables lay between 0.523 and 0.998.

A census of the units was considered in this simulation study and Swiss cheese nonresponse was randomly generated in the complete dataset. The probability p_{ij} that unit i responds to item j was generated according to the model

$$p_{ij} = \frac{1}{1 + \exp\left(1 - \beta_j x_{it}\right)}$$

where $i \in U$, $j = 1, 2, 3, 4$, $t = 2$ if $j = 1, 3$ and $t = 1$ otherwise. Each β_j was defined so that the mean response rate of variable j was 0.70, $j = 1, \ldots, 4$. For each variable, a set of respondents was randomly selected by means of a Poisson sampling design with the response probabilities. The indicator r_{ij} is 1 if unit i responded to item j and 0 otherwise. When $r_{ij} = 0$, it means that the value x_{ij} is missing in the simulation. The dataset with missing values was imputed with four imputation methods:

- Random hot deck (HD): for each nonrespondent, one donor is randomly selected with equal probabilities in S_r,
- K-nearest neighbor (Knn): for each nonrespondent, one donor is randomly selected with equal probabilities in the set of its K-nearest neighbors,
- Balanced K-nearest neighbor (B-Knn): as proposed in Sections 5.3-5.4,
- Deterministic balanced K-nearest neighbor (DB-Knn): the deterministic version of balanced K-nearest neighbor, as in Equation (5.4.1).

Based on each imputed dataset, the total, the 10^{th}, the 50^{th} and the 90^{th} percentiles of each variable was estimated along with the variance-covariance matrix of $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$.

The nonresponse matrix was generated $M_r = 100$ times and each time, the imputation was repeated $M_I = 100$ times. This means that $M_r = 100$ datasets with different nonresponse patterns were created. For each dataset and for each imputation method, $M_I = 100$ imputed datasets were created. For each imputation method and parameter, the Monte Carlo bias of the imputed estimator was calculated,

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

where $\widehat{\theta}_{imp}^{r,i}$ is the value of the imputed estimator of a parameter θ at the simulation (r, i), $r = 1, \ldots, M_R$ and $i = 1, \ldots, M_I$. The subscript imp stands for one of the imputation method. The Monte Carlo mean squared error of the imputed estimator

was also calculated,

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

The Knn, B-Knn and DB-Knn imputation methods were compared to the random hot deck imputation method. The following ratios were calculated for the three methods and for each parameter,

$$\frac{\mathrm{Bias}(\widehat{\theta}_{imp})}{\mathrm{Bias}(\widehat{\theta}_{HD})} \times 100 \quad \text{and} \quad \frac{\mathrm{MSE}(\widehat{\theta}_{imp})}{\mathrm{MSE}(\widehat{\theta}_{HD})} \times 100,$$

where the subscript HD stands for the case of hot deck imputation and the subscript imp stands for the case of Knn, B-Knn or DB-Knn imputation.

The results for the totals, the percentiles and the covariance matrix are shown in Tables 5.1-5.3 respectively. A ratio close to 1 means that the imputation method produces results similar to the random hot deck imputation. When a ratio has a smaller value, it means that the imputation method has a better performance than the random hot deck imputation. For total estimation, the balanced K-nearest neighbor imputation (B-Knn) and its deterministic version (DB-Knn) seem to be equivalent in terms of bias and mean squared error (MSE). For the estimation of a total or a mean, deterministic imputation methods are expected to produce accurate estimations. The K-nearest neighbor (Knn) imputation method has greater ratios than the balanced methods. In this case, the balancing constraints clearly reduce the bias and the MSE of the estimators. In general, in Table 5.3, the B-Knn imputation method produces smaller or similar ratios than the Knn imputation for percentiles estimation. The ratios tend to be closer to 1 for the 10^{th} percentiles. This could be explain by the asymmetry in the distribution of the variables. There are many small units in the population, so the random hot deck is selecting many small donors to impute the missing values. This means that random hot deck imputation is already well estimating small percentiles, it is difficult to gain accuracy. The DB-Knn imputation is not performing well for small percentiles, but it has small biases and MSEs for greater percentiles. The quality of the estimation for the 90^{th} percentile might be explained by the skewness of the distributions. The imputation probabilities copy probably well the tail of the distributions. For the covariance estimations in Table 5.2, the B-Knn imputation seems to have smaller ratios than the Knn imputation. Globally, the balancing constraints seem to reduce the bias and the MSE of the estimators.

5.6. Discussion

Balanced K-nearest neighbor imputation utilizes neighbors to impute missing values. The method exploits the similarities between the units. If the neighbors are alike, the imputation method will gain in accuracy. The imputation method also

Tab. 5.1. Bias and mean squared errors (MSE) of total estimators, in the case of Knn, B-Knn and DB-Knn imputation, relative to the result in the case of hot deck imputation.

Ratio	Imputation	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
Bias	Knn	19.7	19.8	16.4	21.3
	B-Knn	15.5	15.5	12.8	11.5
	DB-Knn	15.5	15.5	12.8	11.6
MSE	Knn	19.9	19.9	16.3	25.6
	B-Knn	15.6	15.6	12.9	19.0
	DB-Knn	15.4	15.4	12.7	18.1

Tab. 5.2. Bias and mean squared errors (MSE) of estimated covariances, in the case of Knn, B-Knn and DB-Knn imputation, relative to the result in the case of hot deck imputation. Notation s_{ij} stands for the covariance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Ratio	Imputation	s_{11}	s_{12}	s_{31}	s_{41}	s_{22}	s_{23}	s_{24}	s_{33}	s_{34}	s_{44}
Bias	Knn	1.3	7.4	4.9	97.3	1.3	4.6	102.5	0.7	178.5	24.8
	B-Knn	0.4	2.7	1.7	62.9	0.4	1.7	63.4	0.2	110.1	17.0
	DB-Knn	2.2	9.3	6.4	83.9	2.0	5.7	83.3	1.1	149.6	75.8
MSE	Knn	1.4	3.8	3.0	34.3	1.3	2.8	31.9	0.7	28.3	32.1
	B-Knn	0.8	2.1	1.6	24.9	0.8	1.5	22.6	0.4	20.2	31.8
	DB-Knn	1.8	4.2	3.3	30.0	1.6	3.0	26.5	0.9	24.3	71.4

exploits the linear relations between the variables. If there is any linear relationship between the variables, this will improve the accuracy of the imputed values and reduce the variance of the estimated parameters. Other strengths of the method are the possibility to impute both categorical and continuous variables and the possibility to force the probability ψ_{ik} to be null if needed. Indeed, if the survey sampler does not want to allow a respondent i to be the donor of a nonrespondent k for any reason, it is possible.

The variance of estimated parameters is a complex matter when the datasets are imputed. The variance needs to take into account the variability caused by the sampling design, by the nonresponse and by the imputation method. The determination of an explicit variance estimator is a work in progress. At the present time, a Bootstrap variance estimator could utterly be used.

A more extensive simulation study could be of interest to evaluate the performance of the imputation method in different contexts and to test variance estimators. The choice of K could also be studied in an extended simulation study. Indeed, if K is small, the imputation method is expected to gain in performance. But there might be no solution to Algorithms 5.1 and 5.2. In this case, K should be increased until a solution is found. It would be important to further investigate the choice of K.

Tab. 5.3. Bias and mean squared errors (MSE) of estimated percentiles, in the case of Knn, B-Knn and DB-Knn imputation, relative to the result in the case of hot deck imputation. The 10^{th}, 50^{th} and 90^{th} percentiles are considered for $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ and $\mathbf{x}_4$.

		P_{10}				P_{50}				P_{90}			
Ratio	Imputation	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
Bias	Knn	82.0	85.2	82.0	37.1	29.1	32.1	26.7	39.0	14.5	24.6	19.2	10.6
	B-Knn	62.2	66.7	56.2	21.8	18.0	24.6	18.4	24.4	14.2	24.6	19.0	8.4
	DB-Knn	107.1	107.7	109.4	108.4	14.0	20.7	15.3	4.0	0.2	0.7	0.9	0.2
MSE	Knn	83.6	86.1	83.9	62.3	31.6	33.0	28.5	45.5	18.8	25.8	21.0	16.0
	B-Knn	66.7	70.4	62.9	65.8	19.6	25.2	18.9	36.0	18.6	25.9	20.9	13.9
	DB-Knn	106.4	107.6	109.0	105.4	15.2	21.6	16.0	25.1	2.1	1.6	2.5	1.8

Conclusion

In this thesis, we discussed the use of auxiliary information in survey sampling and the consequences that this has on the estimators. We have seen that auxiliary variables can contribute a lot to the efficiency and accuracy arising from a sampling design. It can also improve the corrections in a data set, but it complicates the measurement of the quality of the resulting estimators. The chapters on variance estimation clarify this matter and propose an estimation method even if the estimators are complex.

In Chapter 2, a sampling design for a forest inventory was suggested. We used stratified balanced sampling in a two-stage design to select the sample of trees. The proposed sampling design selects the right number of trees of each type of trees and it optimizes the work of the ground teams. The objectives of the inventory are therefore entirely satisfied. However, at the time of the implementation, it would not be surprising to be faced with nonresponse. Indeed, some regions of trees may be impossible to visit and the sampling design may need to be adjusted to take that into account. The estimators will also need to be adapted accordingly.

Chapter 3 is a key article to the understanding of the preferences on variance estimation in practice. The variance of an estimator calculated in the intersection of two independent samples is developed. The variance and its estimator are decomposed conditionally to one sample or conditionally to the other sample. The preferences in terms of conditioning for variance decomposition were already clear in practice, but this article helped explaining those habits.

Chapter 4 deals with the estimation of the variance in the presence of nonresponse. In this case, variances are typically difficult to calculate. It must account for the variability due to the sampling design, the nonresponse and the imputation. In this chapter, we proposed a linearization method to find explicit variance estimators even if the parameter or the treatment of the nonresponse are complex. This method has the advantage to produce an explicit estimator and it needs no computation effort. It does not necessarily lead to a greater precision than resampling methods. The choice between a linearization method and a resampling method for variance estimation is then up to the user. It is indeed difficult to conclude which technique is better, it is a matter of preferences. Some statisticians can prefer to avoid the

derivation of the variance estimator, but some others can prefer to have an explicit estimator. In the latter case, it is possible to identify the elements that could reduce or increase the variance. Overall, both methods can be seen as complementary rather than being rivals.

In Chapter 5, we suggested an imputation method for Swiss cheese nonresponse. This method is the extension of balanced K nearest neighbor imputation for the univariate case. The proposed donor imputation method is satisfying balancing constraints to reduce the variance of the estimators. We used stratified balanced sampling to select one donor for each nonrespondent while respecting the constraints. There is a certain workload remaining to deepen the properties of this imputation method. The properties of the estimators should be studied with respect to different assumptions on the model of the variable of interest. Furthermore, the limitations of the method should be explored in function of the number of nearest neighbors used in the algorithm and in function of the dimensions of the data set.

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.

Antal, E. and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106:534–543.

Ardilly, P. (2006). *Les Techniques de Sondage*. Technip, Paris.

Beaumont, J.-F., Béliveau, A., and Haziza, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology*, 3(4):524–542.

Beaumont, J.-F. and Haziza, D. (2016). A note on the concept of invariance in two-phase sampling designs. *Survey Methodology*, 42(2):319–323.

Beaumont, J.-F. and Patak, Z. (2012). On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling. *International Statistical Review*, 80(1):127–148.

Berger, Y. G. (2008). A note on asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. *Journal of Official Statistics*, 24:541–555.

Berger, Y. G. and Skinner, C. J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society*, B67:79–89.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex survey. *International Statistical Review*, 51:279–292.

Binder, D. A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*, 22:17–22.

Binder, D. A. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89:1035–1043.

Booth, J. G., Butler, R. W., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89:1282–1289.

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

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, 35:115–119.

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

Chen, S. and Haziza, D. (2018). Recent developments in dealing with item non-response in surveys: A critical review. *International Statistical Review*. `https://doi.org/10.1111/insr.12305`.

Demnati, A. and Rao, J. N. K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30:17–34.

Demnati, A. and Rao, J. N. K. (2010). Linearization variance estimators for model parameters from complex survey data. *Survey Methodology*, 36:193–201.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25:193–204.

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.

Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, 21:125–135.

Edwards, C. H. (1994). *Advanced calculus of several variables*. Courier Corporation, North Chelmsford.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26.

Estevao, V. M. and Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in twophase sampling. *Journal of Official Statistics*, 18:233–255.

Fattorini, L., Corona, P., Chirici, G., and Pagliarella, M. C. (2015). Design-based strategies for sampling spatial units from regular grids with applications to forest surveys, land use, and land cover estimation. *Environmetrics*, 26(3):216–228.

Fattorini, L., Marcheselli, M., and Pisani, C. (2006). A three-phase sampling strategy for large-scale multiresource forest inventories. *Journal of agricultural, biological, and environmental statistics*, 11(3):296.

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.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*, pages 114–127. Wiley, Chichester.

Graf, M. (2015). A simplified approach to linearization variance for surveys. Technical report, Institut de statistique, Université de Neuchâtel, Neuchâtel.

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

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

Grafström, A. and Ringvall, A. H. (2013). Improving forest field inventories by using remote sensing data in novel sampling designs. *Canadian Journal of Forest Research*, 43(11):1015–1022.

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

Grafström, A. and Schelin, L. (2014). How to select representative samples? *Scandinavian Journal of Statistics*, 41:277–290.

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

Gregoire, T. G. and Valentine, H. T. (2007). *Sampling strategies for natural resources and the environment*. Chapman and Hall/CRC, Boca Raton.

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

Hasler, C. and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, 74:81–94.

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

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C. R., editors, *Sample surveys: Design, methods and applications*, pages 215–246. Elsevier, Amsterdam.

Hidiroglou, M. A. and Särndal, C.-E. (1998). Use of auxiliary information for twophase sampling. *Survey Methodology*, 24:11–20.

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.

Judkins, D. R. (1997). Imputing for Swiss cheese patterns of missing data. In *Proceedings of Statistics Canada Symposium*, page 97, Statistics Canada.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods. *Journal of official statistics*, 19(2):81.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35:501–514.

Kim, J. K. and Rao, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96:917–932.

Kim, J. K. and Shao, J. (2013). *Statistical methods for handling incomplete data*. Chapman and Hall/CRC, Boca Raton.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32:133–142.

Langel, M. and Tillé, Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society*, A176(2):521–540.

Lawrence, M., McRoberts, R. E., Tomppo, E., Gschwantner, T., and Gabler, K. (2010). *Comparisons of national forest inventories*. Springer, New York.

Mandallaz, D. (2008). *Sampling Techniques for Forest Inventories*. Chapman and Hall/CRC, Boca Raton.

McRoberts, R. E. and Tomppo, E. O. (2007). Remote sensing support for national forest inventories. *Remote Sensing of Environment*, 110(4):412–419.

Metla (2015). Natural Resources Institute Finland, National Forest Inventory (NFI). `http://www.metla.fi/ohjelma/vmi/vmi-mvarat-en.htm`. Accessed: 2015-08-20.

Minister of Forests, Wildlife and Parks (2003). Croissance de la forêt publique du Québec sous aménagement : évolution mesurée à partir des placettes-échantillons permanentes. `http://www.mffp.gouv.qc.ca/publications/forets/connaissances/croissance-foret-publique.pdf`. Accessed: 2015-08-20.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3:169–174.

Opsomer, J. D., Breidt, F. J., Moisen, G. G., and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*, 102(478):400–409.

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, pages 355–375.

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

Rao, J. N. K. (1990). Variance estimation under imputation for missing data. Technical report, Statistics Canada, Ottawa.

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*, 82:453–460.

Saarela, S., Grafström, A., Ståhl, G., Kangas, A., Holopainen, M., Tuominen, S., Nordkvist, K., and Hyyppä, J. (2015). Model-assisted estimation of growing stock volume using different combinations of lidar and landsat data as auxiliary

information. *Remote Sensing of Environment*, 158:431–440.

Särndal, C.-E. (1982). Implication of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7:155–170.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18:241–252.

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 non-response. *International Statistical Review*, 55:279–294.

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

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.

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

Stuart, A. and Ord, J. K. (1994). *Kendall's Advanced Theory of Statistics, vol. 1*. Edward Arnold, London.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies*. Dunod, Paris.

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

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.

Tillé, Y. and Vallée, A.-A. (2017). Revisiting variance decomposition when independent samples intersect. *Statistics & Probability Letters*, 130:71–75.

Vallée, A.-A., Ferland-Raymond, B., Rivest, L.-P., and Tillé, Y. (2015). Incorporating spatial and operational constraints in the sampling designs for forest inventories. *Environmetrics*, 26(8):557–570.

Vallée, A.-A. and Tillé, Y. (2019). Linearisation for variance estimation by means of sampling indicators: application to non-response. *International Statistical Review*. `https://doi.org/10.1111/insr.12313`.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66:411–414.

Zeileis, A. (2014). *ineq: Measuring Inequality, Concentration, and Poverty*. R package version 0.2-13.