Statistical learning for high-dimensional finite population sampling

Mehdi Dagdoug

Apprentissage statistique pour l'échantillonnage en grande dimension

Thèse de doctorat de
l'Université Bourgogne Franche-Comté

École doctorale CARNOT-PASTEUR

présentée et soutenue publiquement à Besançon le 12 juillet 2022
en vue de l'obtention du grade de

Docteur de l'Université de Franche-Comté

(mention Mathématiques Appliquées et Applications des Mathématiques)

par

Mehdi Dagdoug

Composition du jury:

Jean-Michel Poggi	Universités Paris Cité & Paris-Saclay	*Rapporteur*
Anne Ruiz-Gazen	Université Toulouse 1 Capitole	*Rapporteure*
Yves Tillé	Université de Neuchâtel	*Rapporteur*
Clément Dombry	Université Bourgogne Franche-Comté	*Président du jury*
Aurélie Vanheuverzwyn	Médiamétrie	*Invitée*
Camelia Goga	Université Bourgogne Franche-Comté	*Directrice de thèse*
David Haziza	University of Ottawa	*Directeur de thèse*

REMERCIEMENTS

Je souhaiterais en premier lieu exprimer ma profonde gratitude à mes directeurs de thèse, Camelia et David. Je ne pouvais rêver meilleure ouverture au monde de la recherche. J'espère que les trois ans passés à leurs côtés m'ont permis d'acquérir une part de leur rigueur, de leur intuition et de leur passion pour la statistique. Ils font partie des personnes les plus bienveillantes qu'il m'ait été donné de rencontrer; depuis le premier jour, ils n'ont cessé de me conseiller, de partager leur expérience, en bref, de faire tout leur possible pour que je puisse atteindre mes objectifs. À ce titre, je leur suis infiniment reconnaissant, et j'espère que nous continuerons à collaborer pendant encore de nombreuses années. Merci pour tout.

Je tiens aussi à remercier Jean-Michel Poggi, Anne Ruiz-Gazen et Yves Tillé de l'honneur qu'ils me font d'avoir accepté d'être les rapporteurs de cette thèse. Je remercie tout particulièrement Yves, pour m'avoir si bien accueilli à Neuchâtel et pour les nombreuses discussions que nous avons pu avoir; je remercie aussi Anne, pour ses retours et ses conseils prodigués à l'occasion de mon comité de suivi de thèse et lors des différentes conférences où l'on s'est rencontré.

Je tiens à exprimer ma reconaissance à Aurélie Vanheuverzwyn, tout d'abord pour avoir accepté de faire partie de ce jury, mais aussi pour avoir grandement soutenu et participé aux projets de cette thèse. Un grand merci aussi pour avoir organisé des échanges qui, bien que trop peu nombreux dû au contexte sanitaire actuel, m'ont permis de découvrir quelques-unes des problématiques concrètes auxquelles les statisticiens peuvent être confrontés.

Enfin, je remercie également Clément Dombry d'avoir accepté de faire partie de mon jury de thèse, ainsi que pour ses conseils et son soutien constant depuis le master.

J'adresse mes remerciements à l'intégralité des membres du Laboratoire de Mathématiques de Besançon, pour m'avoir si bien accueilli, intégré, et pour m'avoir permis de travailler dans une ambiance si chaleureuse. Je pense en particulier à Charlène et Claudia qui m'ont très fréquemment aidé, conseillé et soutenu dans les différentes démarches que j'ai eu à entreprendre.

Un grand merci aussi à l'ensemble des membres de l'équipe Probabilités et Statistiques. Je souhaite remercier tout particulièrement Yacouba, pour son soutien et ses conseils, ainsi que pour sa passion pour la recherche qui est communicative. Je tiens aussi à remercier Louis, pour ses nombreux conseils, ainsi que Christophe, pour son soutien et ses encouragements.

Bien sûr, un grand merci à tous mes collègues doctorants avec qui j'ai pu avoir la chance de partager un match de ping-pong, une partie d'échecs, ou tout simplement un café. Je pense tout particulièrement à Valentin, qui m'a suivi dans bon nombre de mes idées farfelues, et sans qui l'ambiance au bureau aurait été bien moins drôle; à Cécile, que j'ai probablement fatiguée plus d'une fois, mais qui, comme à son habitude, n'osait que très rarement le dire; et bien sûr, à Loris, Youssef, Benjamin, Yoël, Mathilde, Romain, Chifaa, Marsault, Audrey, Charles, Mathieu, et tous les autres.

Mes derniers remerciements s’adressent à ma famille et à mes amis, pour leur soutien constant, et pour avoir grandement contribué à rendre ces trois dernières années particulièrement agréables.

Merci à tous.

CONTENTS

Remerciements 2

Introduction 8

1 INTRODUCTION TO SURVEY SAMPLING AND REVIEW OF THE MAIN CONTRIBUTIONS 14

1.1 A basic introduction to survey sampling . . . 14
1.1.1 Superpopulation framework . . . 14
1.1.2 Probability sampling . . . 16
1.1.3 Sampling without auxiliary information . . . 19
1.1.4 Sampling with additional auxiliary information . . . 21
1.1.5 Asymptotic theory in survey sampling . . . 25

1.2 Use of auxiliary information at the estimation stage . . . 26
1.2.1 Model-assisted estimation in surveys . . . 26
1.2.2 Model-assisted estimation in high-dimensional settings for survey data 29
1.2.3 Random forest for model-assisted estimation in finite population sampling 32

1.3 Item nonresponse and imputation in surveys . . . 42
1.3.1 Basic framework and imputed estimators . . . 42
1.3.2 Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison . . . 45
1.3.3 Regression tree and random forest imputation in surveys with application to data integration . . . 46

2 MODEL-ASSISTED ESTIMATION IN HIGH-DIMENSIONAL SETTINGS FOR SURVEY DATA 52

2.1 Introduction . . . 52

2.2 The setup . . . 54

2.3 Least squares and penalized model-assisted estimators . . . 55
2.3.1 The GREG estimator . . . 55
2.3.2 Penalized least square estimators . . . 56
2.3.3 Consistency of the GREG and penalized GREG estimators in a high-dimensional setting . . . 58

2.4 Simulation study . . . 61
2.4.1 Simple random sampling without replacement . . . 63
2.4.2 Stratified simple random sampling with optimal allocation . . . 64
2.4.3 Stratified inclusion probability proportional-to-size sampling without replacement . . . 70
2.4.4 Stratified simple random sampling with proportional allocation . . . 71

2.5 Final remarks . . . 73
2.6 Supplementary material . . . 74

3 MODEL-ASSISTED ESTIMATION THROUGH RANDOM FORESTS IN FINITE POPULATION SAMPLING 84
3.1 Introduction . . . 84
3.2 Regression trees and random forests . . . 87
3.2.1 Regression trees . . . 87
3.2.2 Random forests . . . 89
3.3 Model-assisted estimation: Random forests . . . 90
3.3.1 Model-assisted estimation: Population-based partitions . . . 91
3.3.2 Model-assisted estimation: Sample-based partitions . . . 92
3.4 Asymptotic properties . . . 95
3.4.1 Asymptotic results . . . 97
3.5 A model calibration procedure for handling multiple survey variables . . . 99
3.6 Simulation study . . . 101
3.6.1 Performance of point estimators . . . 101
3.6.2 Performance of the proposed variance estimator . . . 104
3.6.3 Choice of hyper-parameters . . . 106
3.6.4 Real data application . . . 108
3.7 Final remarks . . . 109
3.8 Appendix . . . 110
3.9 Supplementary material . . . 112
3.9.1 Asymptotic results of the population RF model-assisted estimator $\widehat{t}^{*}_{rf}$ . . . 113
3.9.2 Asymptotic results: the sample RF model-assisted estimator $\widehat{t}_{rf}$ . . . 120

4 IMPUTATION PROCEDURES IN SURVEYS USING NONPARAMETRIC AND MACHINE LEARNING METHODS: AN EMPIRICAL COMPARISON 128
4.1 Introduction . . . 128
4.2 Preliminaries . . . 130
4.3 A description of imputation methods . . . 132
4.3.1 Parametric regression imputation . . . 132
4.3.2 Imputation classes : the score method . . . 133
4.3.3 K-nearest neighbours imputation . . . 133
4.3.4 B-splines and additive model nonparametric regression . . . 134
4.3.5 Regression trees . . . 136
4.3.6 Random forests . . . 137
4.3.7 Least square tree-boosting and other tree-boosting methods . . . 138
4.3.8 Cubist algorithm . . . 141
4.3.9 Support vector regression . . . 143
4.4 Simulation study: the case of population totals . . . 144

4.4.1 The setup . . . 145
4.4.2 Simulation results . . . 147
4.4.3 High-dimensional setting . . . 154
4.5 Simulation study: the case of population quantiles . . . 158
4.6 Final remarks . . . 161

5 REGRESSION TREE AND RANDOM FOREST IMPUTATION IN SURVEYS WITH APPLICATION TO DATA INTEGRATION 164
5.1 Introduction . . . 164
5.2 Mean square consistency of imputed estimators . . . 165
5.3 Tree imputation . . . 167
5.3.1 Trees and partitioning predictors . . . 168
5.3.2 Regression tree imputation . . . 169
5.3.3 Properties of the tree imputed estimator . . . 170
5.4 From trees to forest estimators . . . 171
5.4.1 Randomized predictors and random forests . . . 171
5.4.2 Random forest imputation . . . 172
5.4.3 From finite to infinite forests . . . 174
5.4.4 Convergence of random forest imputed estimators . . . 175
5.5 Variance estimation . . . 176
5.5.1 Variance estimation through the two-phase framework . . . 177
5.5.2 Variance estimation through the reverse framework . . . 178
5.6 Mass imputation for data integration . . . 178
5.7 Simulations . . . 179
5.7.1 Performances of point estimators . . . 180
5.7.2 Performance of variance estimators . . . 182
5.7.3 Empirical performances of tree and forest mass imputed estimators . . . 184
5.7.4 Empirical performances of variance estimators for mass imputed estimators . . . 185
5.8 Appendix . . . 186

6 CONCLUSION AND FUTURE WORKS 196
6.1 Some thoughts on the use of machine learning algorithms in surveys . . . 196
6.2 Open questions, extensions and future works . . . 198

Bibliography 213

Abstract 214

INTRODUCTION

One of the main goals of survey sampling is the estimation of finite population parameters. Common examples of these parameters of interest include population totals, population means or population proportions, among others. Auxiliary information is often available and can be incorporated in the estimation procedures to increase the precision of the resulting point estimators. If the sampled units respond to all items, model-assisted estimation (Cassel et al., 1976) and calibration (Deville and Särndal, 1992) provide flexible solutions for making use of auxiliary variables and construct efficient estimators of finite population totals. In presence of missing data, auxiliary information can also be used to reduce the undesirable effects of nonresponse. Popular methodologies include imputation or re-weighting the sampled elements.

This work was born from an idea of collaboration with *Médiamétrie*[1], the French audience company. Initially, we were interested in developing and analyzing matrix completion algorithms with survey data for estimating several finite population totals (i.e., multipurpose surveys). This topic is particularly promising, both from a theoretical and a practical point of view. Unfortunately, the COVID-19 pandemic made it difficult to meet and collaborate with Médiamétrie as much as we would have desired and this thesis also explored other research areas. Despite the sanitary complications, these three years have allowed us to have many enriching discussions with Médiamétrie. In particular, Médiamétrie continually supported us in the new research directions that the thesis had taken and, in particular, provided us with databases that were useful to assess the performances of the new methodologies. The project of matrix completion in surveys is still ongoing; more details about it are provided in Chapter 6. I hope that our future works in that area will be a useful contribution for Médiamétrie and that our collaboration will continue long after the time of this thesis.

Médiamétrie often has to deal with a large number of covariates. More generally, it is nowadays rather common for survey statisticians to face high-dimensional data. Therefore, in this thesis, the problem of estimating finite population totals in presence of a large number of auxiliary variables is considered. The scenarios of full response and missing data are both examined. In case of full response, I investigated the properties of existing model-assisted estimators in a high-dimensional asymptotic framework and I suggested a new class of model-assisted estimators based on random forests. In presence of missing data, I studied the use of statistical learning predictors based on a large number of covariates for imputation as well as the theoretical properties of imputed estimators based on regression trees and random forests.

Model-assisted estimation with high-dimensional data

Since the pioneering work of Särndal (1980), Robinson and Särndal (1983) and Särndal and Wright (1984), model-assisted estimation procedures have attracted a lot of attention in the literature; see also Särndal et al. (1992) for a comprehensive discussion of the model-assisted approach. The main idea of model-assisted estimation is to estimate the functional relationship

[1] Médiamétrie is leader of media and reference studies in audience measurements in France, see the website https://www.mediametrie.fr/en for more details.

between the survey variable and the set of covariates by means of a predictive model, and to incorporate its predictions in the definition of the estimator. When the predictions are close to the true values, the precision of the resulting estimator might increase. Many predictive models have been suggested in the literature, from parametric models (Robinson and Särndal, 1983) and penalized linear models such as the lasso (McConville et al., 2017) and ridge (Goga and Shehzad, 2010), to non-parametrics such as local polynomials (Breidt and Opsomer, 2000), B-splines (Goga, 2005, Goga and Ruiz-Gazen, 2014), penalized splines (Breidt et al., 2005, McConville and Breidt, 2013), neural networks (Montanari and Ranalli, 2005), generalized additive models (Opsomer et al., 2007) and regression trees (McConville and Toth, 2019).

Most of the aforementioned literature is gravitating around predictors which are especially effective when the number of covariates is relatively low; however, some of them tend to be relatively inefficient when used with a large number of covariates, a phenomenon known as the curse of dimensionality (Hastie et al., 2011); see also Giraud (2021), Györfi et al. (2006) for textbook discussions. Nowadays, it is no longer unusual for survey statisticians to face scenarios where a large number of auxiliary variables are available at the estimation stage. For example, Médiamétrie uses every day a panel of 7000 individuals and their television consumption is recorded every second, thus leading to 86400 covariates (Cardot et al., 2013a). Other applications on Médiamétrie data can also be found for instance in Goga et al. (2011). Similarly, in Cardot et al. (2013b), the authors considered samples of size 1 500 from a real data set collected by the Irish Commission for Energy Regulation Project concerning the electricity consumption of households and companies recorded every 30 min (for example, there are more than 300 variables over a week). In these examples, the assumption that the number of covariates is negligible with respect to the sample size may not be satisfied. Recently, to better account for these high-dimensional scenarios, Cardot et al. (2017), Chauvet and Goga (2022), Ta et al. (2020) investigated the asymptotic properties of linear model-assisted estimators when the number of covariates was allowed to increase to infinity.

In this thesis, I analyzed the performances of linear and penalized model-assisted estimators in a high-dimensional asymptotic framework (Dagdoug et al., 2022a) and suggested in Dagdoug et al. (2021b) the use of model-assisted estimator based on random forest algorithms (Breiman, 2001). Random forest predictors seem to remain relatively efficient in high-dimensional settings and seem to adapt well to sparsity (Biau, 2012, Scornet et al., 2015).

In Dagdoug et al. (2022a), we examined several model-assisted estimators from a design-based point of view and in a high-dimensional setting, including linear regression and penalized estimators. The consistency of model-assisted estimators based on linear regression, on the lasso and on ridge was established under relatively weak assumptions. We conducted an extensive simulation study using data from the Irish Commission for Energy Regulation Smart Metering Project, to assess the sensitivity of model-assisted estimators based on several statistical learning predictors to high-dimensional auxiliary information.

In Dagdoug et al. (2021b), we use random forests to estimate the functional relationship between the survey variable and the auxiliary variables. Several important results have been developed in this work. First, under regularity assumptions on the sampling design, on the survey variable and the random forest algorithms, we showed the L^1 convergence and the asymptotic normality of the estimator. The asymptotic variance of the estimator was derived and an L^1 consistent variance estimator was suggested. A nice and new non-asymptotic

property has also been shown: this estimator can be written according to the observations which did not participate in the prediction, observations called "out-of-bag" in the "machine learning" literature. The use of these units, which makes it possible to avoid overfitting, is new in the context of estimation for survey data. A model-calibration procedure for handling multiple survey variables was also discussed. The results of a simulation study suggested that the proposed point and estimation procedures perform well in terms of bias, efficiency and coverage of normal-based confidence intervals, in a wide variety of settings. We also suggested a new variance estimator based on K-fold cross-validation method. This new variance estimator constitutes an important advance in the theory of modern survey practice because it corrects the defect of overfitting of traditionally used variance estimators and thus makes it possible to construct reliable confidence intervals. This new variance estimation method is all the more important as it is completely general and can be used with other nonparametric estimation methods such as splines, local polynomials, etc. Finally, adaptations of random forest algorithms to different sampling plans (stratified, proportional to size) were proposed and the method was tested on audience data from Médiamétrie. More precisely, we had access to 3 882 auxiliary variables and used a sample size of 4 000 observations; the goal was to estimate the proportion of French individuals who listen to a radio of interest on a daily basis, both at the overall population level and for several domains of interest.

Treatment of item nonresponse with high-dimensional data

Missing data are present in most censuses and sample surveys. Nonresponse is particularly undesirable as unadjusted estimators might be biased and exhibit a substantial increase of variance (Rubin, 1976); as such, the use of unadjusted estimators should be avoided and nonresponse treated. Two cases are usually distinguished: unit nonresponse and item nonresponse. Unit nonresponse is defined by a complete lack of information of a given element, while item nonresponse characterizes elements for which some information is collected, but not all. The treatment of unit nonresponse is beyond the scope of this thesis. Usually, item nonresponse is handled using imputation, a procedure which consists in replacing missing values with artificial values. Most often, these artificial values are obtained by means of a predictive model. The theoretical properties of imputed estimators were investigated for several non-parametric methods such as the nearest neighbor (Chen and Shao, 2000, 2001, Yang and Kim, 2019), the score method (Haziza and Beaumont, 2007, Little, 1986), predictive mean matching (Yang and Kim, 2017), kernel regression (Zhong and Chen, 2014), to cite just a few. For a comprehensive review about the missing data literature in surveys, see Chen and Haziza (2019) or Haziza (2009). In the same way as for model-assisted estimators, most of the existing literature is centered around predictors which are sensitive to the number of auxiliary variables; that is, they are particularly efficient in low-dimensional settings, but might be particularly sensitive to the curse of dimensionality. For instance, using arguments of Abadie and Imbens (2006), it is shown in Yang and Kim (2019) that the nearest neighbor imputed estimator has a non-negligible bias whenever the number of covariates is strictly greater than one.

In this thesis, I also investigated the performances of imputed estimators based on predictors known for their high-dimensional efficiency. First, in Dagdoug et al. (2021a), we conducted a large-scale simulation study in which we compared imputed estimators based on numerous

statistical learning predictors commonly used by machine learning practitioners. In particular, we included the traditional linear regression, the score method, as well as support vector machines (Cortes and Vapnik, 1995, Smola and Schölkopf, 2004), k-nearest neighbors, regression trees (Breiman, 1984), random forest (Breiman, 2001), gradient boosting (Chen and Guestrin, 2016, Friedman, 2001), Bayesian additive models (BART, Chipman et al. (2010)), additive models with B-splines, Cubist Quinlan (1993), Quinlan et al. (1992). Various relationships between the survey variable and the covariates were considered, as well as several sampling designs and nonresponse models; high-dimensional scenarios were considered as well. We discovered that imputed estimators based on complex algorithms (e.g. Cubist, Boosting, Bayesian additive regression trees, Random forests) can often outperform traditional parametric models: when the parametric model was well specified, then the imputed estimator based on it was more efficient than imputed estimators based on nonparametric complex algorithms; however, in most cases, the loss of efficiency of nonparametric models versus parametric was relatively low. On the other hand, when the parametric model was misspecified, imputed estimators based on nonparametric models were much more efficient. Overall, we found that imputed estimators based on Cubist, XGBoost and BART were very efficient in most scenarios and substantially improved over parametric estimators.

In Dagdoug et al. (2022b), we investigated both theoretically and empirically the performances of regression tree and random forest imputed estimators. We gave a result exhibiting a set of conditions on the predictor used for imputation under which the imputed estimator based on this predictor is L^2 consistent with respect to the joint distribution induced by the model, the nonresponse mechanism and the sampling design. The conditions that we found revealed that, if the predicted values are based on a predictor consistent for the regression function and the L^2 prediction error is bounded, even for finite samples, then the resulting imputed estimator converges in L^2 towards the parameter of interest. Using this result, the L^2-convergence of the CART tree imputed is obtained. The L^2-convergence of the random forest imputed estimator is obtained using infinite forests as a tool. We suggested variance estimators which seemed to perform well (from a bias and coverage rate point of view) on simulation studies. An application to data integration was also considered.

Organization of the dissertation and list of publications

The rest of this thesis is organized as follows. Chapter 1 is a presentation of the basic concepts of survey sampling. These concepts are presented in a unified framework as in Rubin-Bleuer and Kratina (2005), Boistard et al. (2017), and Han and Wellner (2021). The mathematical formalism employed in this framework is rather different from the usual presentations of survey sampling theory given in Särndal et al. (1992), Tillé (2020) or Lohr (2021), but it has the advantage to include both design and model inferences commonly used in the model-assisted literature as well as the model-design-nonresponse inferences used in the imputation literature. Chapter 1 also describes briefly basic sampling designs, the usual Horvitz-Thompson estimator (Horvitz and Thompson, 1952) and summarizes the main contributions of this dissertation. The chapters 2 and 3 are devoted to model-assisted estimation in high-dimensional settings; Chapter 2 presents the article Dagdoug et al. (2022a) while Chapter 3 presents the article Dagdoug et al. (2021b). The chapters 4 and 5 are concerned with imputation in presence of a large number of covariates; Chapter 4 presents the article Dagdoug et al. (2021a) while Chapter 5 presents the work in progress Dagdoug et al. (2022b)

which will soon be submitted. Each of the chapters can be read independently one from another. Finally, the thesis ends with a conclusion and some perspectives in Chapter 6.

Publications

1. Dagdoug, M., Goga, C., & Haziza, D. (2022). Model-assisted estimation in high-dimensional settings for survey data. To appear in *Journal of Applied Statistics*. DOI: https://doi.org/10.1080/02664763.2022.2047905

2. Dagdoug, M., Goga, C., & Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. To appear in *Journal of the American Statistical Association*. DOI: https://doi.org/10.1080/01621459.2021.1987250

3. Dagdoug, M., Goga, C., & Haziza, D. (2021). Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. To appear in *Journal of Survey Statistics and Methodology*. DOI: https://doi.org/10.1093/jssam/smab004

Unpublished work

1. Dagdoug, M., Goga, C., & Haziza, D. (2022). Regression tree and random forest imputation in surveys with application to data integration.

2. Larbi, K., Haziza, D., & Dagdoug, M. (2022). Treatment of unit nonresponse in surveys through machine learning methods : an empirical comparison.[2]

Conference proceedings

1. Dagdoug, M., Goga, C., & Haziza, D. (2022). Arbres et forêts aléatoires : d'une approche par modélisation assistée au traitement de la nonréponse. *Journées de Méthodologie Statistique de l'Insee 2022*, Paris, France, 29-31st march 2022.

2. Dagdoug, M., Goga, C. & Haziza, D. (2021). Random forests imputation in surveys. *Forum des Jeunes Mathématicien.ne.s*, Besançon, France, 8-10th december 2021.

3. Dagdoug, M., Goga, C. & Haziza, D. (2021). Imputation par forêt aléatoire en théorie des sondages. *11e Colloque International Francophone sur les Sondages*, Bruxelles, Belgium, 6-8th october 2021.

4. Dagdoug, M., Goga, C., & Haziza, D. (2021). Convergence rates of model-assisted estimators in high-dimensional settings. *63rd ISI World Statistics Congress*, The Hague, Netherlands (virtual), 11-16th july 2021.

5. Dagdoug, M., Goga, C., & Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *52ème Journées de Statistiques de la Société Française de Statistique*, Nice, France (virtual), 7-11th june 2021.

6. Dagdoug, M., Goga, C. & Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *Congrès annuel 2021 de la Société de Statistique du Canada*, Ottawa, Canada, 7-11th june 2021.

[2] This work is a collaboration which is not part of this PhD and therefore will not be included in this manuscript.

1 INTRODUCTION TO SURVEY SAMPLING AND REVIEW OF THE MAIN CONTRIBUTIONS

1.1 A basic introduction to survey sampling

1.1.1 Superpopulation framework

Let Y be a *survey variable*, representing a characteristic of interest that the survey statistician wishes to study. Consider that the random variable Y is defined on a probability space $(\Omega, \mathcal{M}, \mathbb{P}_m)$ taking values in a measurable space $(E, \mathcal{E})$. The distribution of Y will be denoted by $\mathbb{P}_Y := \mathbb{P}_m \circ Y^{-1}$. Let $N \in \mathbb{N}^*$ be a positive integer. Define N independent and identically distributed (i.i.d.) random variables $\{Y_k\}_{k=1,...,N}$, with common law $\mathbb{P}_Y$. Then, the probability space $(\Omega, \mathcal{M}, \mathbb{P}_m)$ and the N-random vector $\boldsymbol{Y} := [Y_1, Y_2, ..., Y_N]^\top$ define what we call in the sequel a *superpopulation model* as described in Rubin-Bleuer and Kratina (2005) and denoted by $(\Omega, \mathcal{M}, \mathbb{P}_m, \boldsymbol{Y})$.

As stated in Shao and Tu (2012), "the basic objective of statistical analysis is extracting all the information from the data to deduce "properties" of the "population" that generated the data". This statement holds true for both *classical statistics* (sometimes referred to as *inferential statistics* in the literature) and *survey statistics*; yet, the terms "properties" and "population" have different meanings in these two scenarios. In the next paragraph, we highlight these differences and describe both approaches to inference.

In classical statistics, the term "population" is used to denote the probability distribution $\mathbb{P}_Y$ from which the data has been generated. It can be viewed as an *infinite population*, and in survey statistics, as a superpopulation.

In survey statistics, however, the term "population" refers to a *finite population*, that is, a set of finite cardinality, as opposed to the conceptual "infinite population" of classical statistics. The cardinality of a finite population will be referred to as its *size*. We thus now consider a finite population U of N labeled elements

$$U := \{u_1, u_2, ..., u_N\} = \{1, 2, ..., N\},$$

where u_k denotes the k-th element of the population of interest U, also called the *target population*. To each element u_k of U, we generate a realization y_k of Y_k (i.e., $Y_k(\omega) := y_k \in E$ for some $\omega \in \Omega$) and we say that U is generated by the superpopulation model $(\Omega, \mathcal{M}, \mathbb{P}_m, \boldsymbol{Y})$. For simplicity, in the sequel, we use $\{y_k\}_{k \in U}$ to denote the random variables $\{Y_k\}_{k \in U}$.

In classical statistics, the "properties" referred to above, can often be defined as functionals of the distribution $\mathbb{P}_Y$. To be more precise, let $(F, \mathcal{F})$ denote a measurable space and $\mathbb{M}_+^1(E)$ be the set of probability measures on E. An "infinite population parameter" or "superpopulation parameter" θ_M can often be represented as the image of an unknown element of $\mathbb{M}_+^1(E)$

through a known statistical functional $T : \mathbb{M}_+^1(E) \to F$, that is, $\theta_M := T(\mathbb{P}_Y)$, where $\mathbb{P}_Y$ is unknown. For instance, a typical example is given by the mean (whenever appropriate) of a real-valued distribution $\mathbb{P}_Y$, in which case

$$\theta_M = T(\mathbb{P}_Y) = \int_{\mathbb{R}} y d\mathbb{P}_Y.$$

In survey statistics, the "properties" that we wish to estimate are functions of the random variables $\{y_k\}_{k\in U}$. These functions are called *finite population parameters* in the sequel. Common examples include the population total t_y, the population mean $\bar{y}_U$ and population proportion p_c of elements having a particular characteristic denoted "C" ; these are defined as follows:

$$t_y := \sum_{k\in U} y_k, \qquad \bar{y}_U := \frac{1}{N}\sum_{k\in U} y_k, \qquad p_c := \frac{1}{N}\sum_{k\in U} \mathbb{1}_{\{u_k \text{ has characteristic "C"}\}}. \tag{1.1}$$

More generally, a finite population parameter θ_U is defined as the image of an unknown point $(y_1, ..., y_N)$ of E^N through a known measurable function $H : E^N \to F$, that is, $\theta_U := H(y_1, y_2, ..., y_N)$.

Therefore, in both classical and survey statistics, the aim is to estimate the image of a known function computed at an unknown point. However, the techniques to do so are sometimes quite different in both fields. Moreover, an additional important difference is that, in the model-based framework of survey statistics, the parameter of interest θ_U is a random variable taking values in F, whereas the parameter of interest θ_M of classical statistics is deterministic. Indeed, see that θ_U is measurable from Ω to F with realizations given by

$$\theta_U(\omega) = H(y_1(\omega), ..., y_N(\omega)), \qquad \omega \in \Omega.$$

Therefore, different realizations of the superpopulation model might lead to different values of the finite population parameters of interest.

Most often, as in the examples given in (1.1), the finite population parameters considered for practical purposes are real-valued and the survey variable is real-valued as well. In some cases, however, the spaces E and F might not be the set of real numbers. For instance, in Cardot et al. (2013c), the authors considered the case of functional data in which $E = \mathcal{F}([0;T], \mathbb{R})$, the set of functions from $[0;T]$ to $\mathbb{R}$, for some $T \in \mathbb{R}_+^*$. Similarly, statisticians might be interested in estimating more complex parameters such as the population distribution function

$$F_N(x) = \frac{1}{N}\sum_{k\in U} \mathbb{1}_{y_k \in]-\infty;x]},$$

in which case $F = \mathbb{D}(\mathbb{R}, [0;1])$, the Skorokhod space of cadlag functions with values in $[0;1]$.

In this work, we solely focus on the estimation of the total of real-valued random variables, so that $E = \mathbb{R}$ and $F = \mathbb{R}$. Indeed, since many population parameters can be expressed as functions of totals, the population total is an important population parameter. In the sequel, the aim is therefore to estimate the population total t_y as defined in (1.1).

1.1.2 Probability sampling

In order to estimate the unknown parameter t_y, two approaches may be considered: a *census* or a *sample survey*. A census consists of collecting the values of the variable of interest for all population elements, in which case the true value of t_y is known. However, due to practical constraints (e.g., cost, time, ...), it is usually impossible to have access to all population elements; a survey is used instead. A survey consists of selecting a subset s of the population, called a *sample*. An estimator based on the information gathered from the sampled elements is then used to infer on the parameter of interest. Throughout this work, several assumptions are made: 1) there is no coverage error[1]; 2) the sample s is selected according to a probability sampling design; 3) no measurement error[2] is present in the data collected. Moreover, in Chapter 3 and Chapter 4, we consider the idealistic scenario of full response. A framework for handling nonresponse is described in Section 1.3. We also restrict ourselves to the case of sampling without replacement, that is, we assume that population elements cannot be selected more than once in the samples.

Definition 1.1.1. *(Sampling design. Särndal et al. (1992))*
Let $\mathcal{S}$ be the collection of subsets of U. A sampling design is a probability mass function $p : \mathcal{S} \mapsto [0; 1]$ satisfying

$$\sum_{s \in \mathcal{S}} p(s) = 1. \tag{1.2}$$

For each sampling design, we can define a probability measure on the *design space*, called a *design probability*.

Definition 1.1.2. *(Design probability.)*
Let $\mathcal{D}$ be a sigma-algebra on $\mathcal{S}$. The measurable space $(\mathcal{S}, \mathcal{D})$ is often called the design space. A design probability $\mathbb{P}_p$ is a probability measure defined on the design space $(\mathcal{S}, \mathcal{D})$.

The measure

$$\mathbb{P}_p := \sum_{s \in \mathcal{S}} p(s)\delta_s$$

is a design probability associated to the sampling design p, where δ_s denotes the Dirac measure at $s \in \mathcal{S}$. Each design probability can be characterized by a sampling design. Note that the sample s selected from U can be viewed as the realization of a random sample S, defined on some probability space $(\Omega_S, \mathcal{A}_S, \mathbb{P}_S)$ taking values in the design space $(\mathcal{S}, \mathcal{D})$ such that $\mathbb{P}_p = \mathbb{P}_S \circ S^{-1}$. We summarize these ideas with a diagram illustrating the differences of classical and survey statistics, as well as sample selection, see Figure 1. The size of the sample S is denoted by n_S. A sampling design for which $n_s = n$ for all samples $s \in \mathcal{S}$ having a non-null probability of being selected is said to be of *fixed size*.

It is often convenient to represent samples without replacement as vectors in $\{0; 1\}^N$. We define the concept of *sample membership indicators*.

Definition 1.1.3 (Sample membership indicators)**.**
For all $k \in U$, we define the sampling membership indicator

$$I_k(S) := \mathbb{1}_{\{k \in S\}},$$

[1] A coverage error happens when the specification of sampling units in the population from which a sample was selected does not match the target population, see Lohr (2021) for details.
[2] A measurement error is a situation in which a response in the survey differs from the true value.

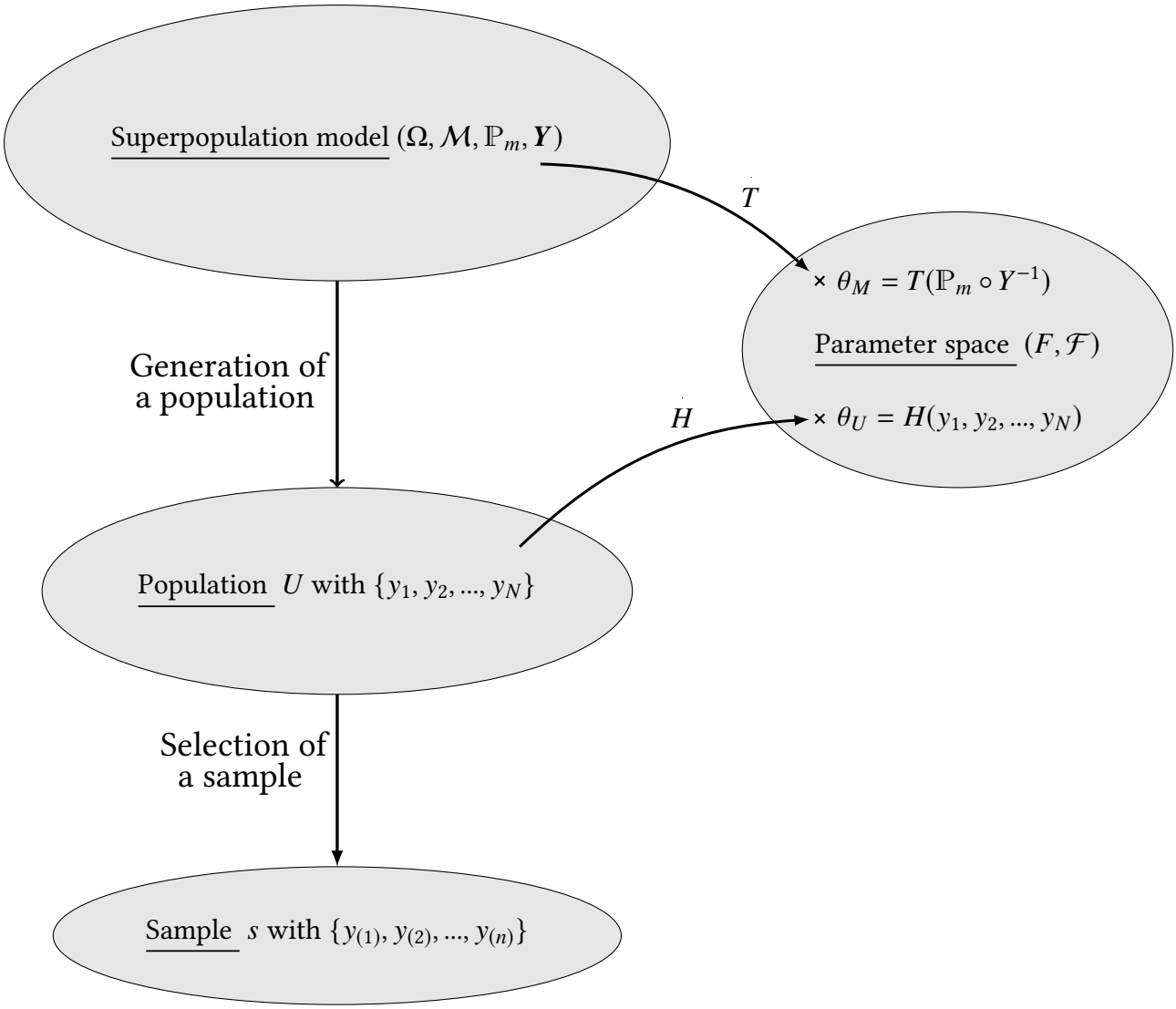


Figure 1: Diagram illustrating the differences between classical statistics and survey statistics. In the sample, $y_{(k)}$ denotes the k-th element selected in the sample, where the order is arbitrary.

which takes the value 1 *if the element* k *is selected in a sample and* 0 *otherwise. For simplicity, we write* I_k *for* $I_k(S)$.

Since the map

$$\phi : \begin{cases} \mathcal{S} \longrightarrow \{0;1\}^N, \\ s \longmapsto [I_1, I_2, ..., I_N]^\top . \end{cases}$$

is a bijection from $\mathcal{S}$ to $\{0;1\}^N$, it follows that each sample in $\mathcal{S}$ can be uniquely represented as a vector of $\{0;1\}^N$. Some authors (e.g., Bonnéry (2011), Conti and Mecatti (2022)) define sampling designs on $\{0;1\}^N$ rather than on $\mathcal{S}$; the two formulations are indeed equivalent.

The random variables $\{I_k\}_{k\in U}$ are, by construction, defined in the design space and follow a Bernoulli distribution $\{\mathcal{B}(\pi_k)\}_{k\in U}$, each, respectively, where π_k denotes the first-order inclusion probability of element k, a notion defined below. Note that $\{I_k\}_{k\in U}$ are not necessarily independent nor identically distributed.

Definition 1.1.4 (First and second order inclusion probabilities)**.**
The first-order inclusion probability π_k of element k is the probability that element k is selected in a sample; it is defined as:

$$\pi_k := \mathbb{P}_S(k \in S) = \mathbb{P}_p(\{k\}) = \sum_{\substack{s \in \mathcal{S}: \\ k \in s}} p(s), \qquad k \in U.$$

Similarly, the second-order inclusion probability $\pi_{k\ell}$ of elements k and ℓ is the probability that elements k and ℓ are both selected in a sample; it is defined by:

$$\pi_{k\ell} := \mathbb{P}_S((k,\ell) \in S \times S) = \mathbb{P}_p(\{k,\ell\}) = \sum_{\substack{s \in \mathcal{S}: \\ k,\ell \in s}} p(s), \qquad k,\ell \in U.$$

Throughout this dissertation, we assume that, for all $k \in U$, $\pi_k > 0$ and for all pairs $(k,\ell) \in U \times U$, $\pi_{k\ell} > 0$. The design covariance $\mathbb{C}ov_p(I_k, I_\ell)$ is denoted by $\Delta_{k\ell} := \mathbb{C}ov_p(I_k, I_\ell) = \pi_{k,\ell} - \pi_k \pi_\ell$.

To infer on the parameter θ_U based on a survey sample, the first step is to use a sampling design to select a sample. The second step is to use an estimator $\widehat{\theta}$ of θ_U and compute an estimate based on the data collected from the sample.

Definition 1.1.5 (Sampling strategy)**.**
A sampling strategy is defined as a pair $\left(p, \widehat{\theta}\right)$ which produces an estimate of the parameter of interest θ_U according to a sampling design p.

A strategy $(p, \hat{\theta})$ is efficient if the estimator $\hat{\theta}$ is precise under the sampling design p, namely if it has low variance computed with respect to p. The *Horvitz-Thompson estimator* (Horvitz and Thompson, 1952, Narain, 1951) $\widehat{t}_\pi$ of t_y is of particular interest.

Definition 1.1.6 (Narain (1951), Horvitz and Thompson (1952))**.**
The Horvitz-Thompson estimator of the population total t_y is defined as:

$$\widehat{t}_\pi := \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} d_k y_k, \tag{1.3}$$

where $d_k := \pi_k^{-1}$, denotes the sampling weight of element $k \in U$. It is also often referred to as the π-estimator or the expansion estimator.

The set $\{d_k\}_{k \in S}$ is often referred to as the *basic weighting system* and can be interpreted as follows: if element k of U is selected in the sample, then it represents d_k elements of the population. Next proposition displays the design properties of $\widehat{t}_\pi$.

Proposition 1.1.1 (Särndal et al. (1992))**.** *. Let p be an arbitrary sampling design. The Horvitz-Thompson estimator $\widehat{t}_\pi$ has the following design properties.*

i) Provided that $\pi_k > 0$ for all $k \in U$, the estimator $\widehat{t}_\pi$ is design-unbiased, that is, $\mathbb{E}_p\left[\widehat{t}_\pi\right] = t_y$, where $\mathbb{E}_p[\cdot]$ denotes the expectation with respect to the sampling design.

ii) The design variance of $\widehat{t_\pi}$ *is given by*

$$\mathbb{V}_p\left(\widehat{t_\pi}\right) = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_\ell},$$

where $\mathbb{V}_p(\cdot)$ *denotes the expectation with respect to the sampling design.*

iii) Provided that $\pi_{k\ell} > 0$ *for all* $(k,\ell) \in U \times U$*, the variance estimator*

$$\widehat{V}_\pi := \sum_{k\in S}\sum_{l\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell} \tag{1.4}$$

is design-unbiased for $\mathbb{V}_p\left(\widehat{t_\pi}\right)$.

From a design-based point of view, the quantities $\{y_k\}_{k\in U}$ in the expression of $\widehat{t_\pi}$ are not random, the only random quantities are the sampling membership indicators $\{I_k\}_{k\in U}$.

The unbiasedness property of the Horvitz-Thompson estimator is particularly attractive. Using arguments of Godambe (1955), it can be shown that the Horvitz-Thomspon estimator is the only unbiased homogeneous linear[3] estimator of t_y with weights independent of the sample.

1.1.3 Sampling without auxiliary information

As mentioned in the previous sections, two sources of randomness may be considered: the randomness due to the superpopulation model and the randomness due to the sampling design. Assuming that the sampling design and the survey variable are independent, functions of both sources of randomness might be considered using a product space representation with a product measure. This assumption is realistic if no auxiliary information[4] is available at the sampling stage. A more complex joint model-design probability will be defined in Section 1.1.4.

Definition 1.1.7 (Joint product model-design distribution)**.**
Consider a superpopulation probability space $(\Omega, \mathcal{M}, \mathbb{P}_m)$ *and a design probability space* $(\mathcal{S}, \mathcal{D}, \mathbb{P}_p)$*. Define by* $\mathcal{M}\times\mathcal{D} := \{A\times B;\ A\in\mathcal{M},\ B\in\mathcal{D}\}$ *the Cartesian product of* $\mathcal{M}$ *and* $\mathcal{D}$ *and by* $\mathcal{M}\otimes\mathcal{D} := \sigma(\mathcal{M}\times\mathcal{D})$ *the sigma-algebra product generated by the measurable rectangles of the form* $A\times B\in\mathcal{M}\times\mathcal{D}$*. The joint model-design probability induced by a sampling design without auxiliary information is defined as the product measure* $\mathbb{P}_{m,p} := \mathbb{P}_m\otimes\mathbb{P}_p$ *on the measurable space* $(\Omega\times\mathcal{S},\ \mathcal{M}\otimes\mathcal{D})$*, uniquely defined as*

$$\mathbb{P}_{m,p}(A\times B) := \mathbb{P}_m(A)\times\mathbb{P}_p(B),$$

for all measurable rectangles $A\times B\in\mathcal{M}\times\mathcal{D}$.

When no confusion arises, we omit the indexes and simply use $\mathbb{P}$ for $\mathbb{P}_{m,p}$ and we call this representation the joint distribution. Below, we describe two sampling designs which

[3] An estimator $\widehat{\theta}$ of θ_U is said to be linear in the survey variable Y if $\widehat{\theta}$ can be written $\widehat{\theta} = w_0 + \sum_{k\in S} w_k y_k$ with weights $\{w_k\}$ independents of the survey variable. If $w_0 = 0$, then $\widehat{\theta}$ is said to be an homogeneous linear estimator of θ_U; otherwise, if $w_0 \neq 0$, $\widehat{\theta}$ is said to be non-homogeneous (Cassel et al., 1976, Särndal et al., 1978).

[4] The term auxiliary information includes all information external to the survey itself which might be used to improve a survey strategy; see Tillé (2020).

can be included in this representation, that is, sampling designs which do not use any additional information. The descriptions given below are only brief, we refer the reader to the references Särndal (1992), Thompson (1997), Fuller (2009b) and Tillé (2020) for additional details.

Simple random sampling without replacement

Let n be the desired sample size. Simple random sampling without replacement (SRSWOR) of size n is the design which assigns the same probability to all without replacement samples of size equal to n and zero otherwise. That is, for $s \in \mathcal{S}$,

$$p(s) = \begin{cases} \binom{N}{n}^{-1} & \text{if } s \text{ is of size } n, \\ \\ 0 & \text{otherwise.} \end{cases}$$

The first-order inclusion probabilities are equal to $\pi_k = n/N$ for all $k \in U$ and the second-order inclusion probabilities are equal to

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

for all $k \neq \ell \in U$. It also follows that the π-estimator of t_y is given by

$$\widehat{t}_\pi = \frac{N}{n} \sum_{k \in S} y_k.$$

The implementation of simple random sampling without replacement is fairly easy (Fan et al., 1962); however, the sampling strategy $\left(SRSWOR, \widehat{t}_\pi\right)$ might lead to a large variance for the Horvitz-Thompson estimator; this phenomenon might happen when the population variability of the survey variable is large. This strategy may be improved if additional auxiliary information is used at the sampling stage, for instance with stratified or proportional to size sampling designs (described thereafter). As explained in Tillé (2020), "the use of simple random sampling is a way of selecting a sample without preconceptions on the studied population. However, if auxiliary information is known and the variable of interest Y is suspected of being linked to this auxiliary information, then it will be more interesting to use a design that integrates this auxiliary information".

Bernoulli sampling

Bernoulli sampling is another sampling design which gives equal inclusion probabilities to all units of the population. More precisely, let $\pi \in [0; 1]$. Bernoulli sampling design is the design for which the sample membership indicators $\{I_k\}_{k \in U}$ are independent and identically distributed random variables, each I_k follows a Bernoulli distribution of parameter π. The probability assigned to a sample s by a Bernoulli sampling is given by

$$p(s) = \pi^{n_s} (1-\pi)^{1-n_s}, \qquad s \in \mathcal{S},$$

where n_s is the size of the sample s. Its implementation is simple as it suffices to draw N independent realizations of a random variable with uniform distribution $\mathcal{U}(]0;1[)$ and select elements for which the realizations lie in the interval $]0;\pi[$. The first order inclusion probabilities are all equal to $\pi_k = \pi$, for all $k \in U$ and, we have $\pi_{kl} = \pi^2$, for all $k \neq l \in U$. The π-estimator is then given by

$$\widehat{t}_\pi = \pi^{-1} \sum_{k \in S} y_k.$$

For a Bernoulli design, the sample size $n_s = \sum_{k \in U} I_k$ is random and follows a Binomial distribution $\mathcal{B}(N, \pi)$. The random size is an important drawback since it is impossible to know in advance the cost of the survey and the Horvitz-Thompson estimator tends to be inefficient due to the random sample size (Särndal et al., 1992). To remedy the issue of the random size, a natural idea would be to consider the conditional sampling design $p(\cdot \,|n_s = n)$. This design is in fact a simple random sampling design of size n as described before; a proof can be found in Tillé (2020).

The parameter π needs to be chosen in order to implement the Bernoulli sampling. One can choose for example π such that the expected sample size is equal to n, the desired sample size, namely:

$$\mathbb{E}_p[n_s] = N\pi = n, \tag{1.5}$$

which gives $\pi = n/N$.

1.1.4 Sampling with additional auxiliary information

In some cases, additional auxiliary information is known prior to sampling, either for all population elements or in a aggregated form. On the probability space $(\Omega, \mathcal{M}, \mathbb{P}_m)$, we define a generic random vector $\mathbf{z} : \Omega \mapsto \mathbb{R}^q$. The distribution of $\mathbf{z}$ is denoted by $\mathbb{P}_\mathbf{z} := \mathbb{P}_m \circ \mathbf{z}^{-1}$ and we define N i.i.d. copies $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N$ on $(\Omega, \mathcal{M}, \mathbb{P}_m)$ with distribution $\mathbb{P}_\mathbf{z}$. The superpopulation model then becomes $(\Omega, \mathcal{M}, \mathbb{P}_m, \boldsymbol{Y}, \boldsymbol{Z}_U)$, where $\boldsymbol{Z}_U := [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N]^\top \in \mathbb{R}^{N \times q}$. We assume that, conditionally on the auxiliary variables $\boldsymbol{Z}_U$, the sampling design is independent of the survey variable. In that case, the sampling design is said to be *noninformative* (Pfeffermann, 1993). On the other hand, if conditional independence is not satisfied, the sampling design is said to be *informative*, as considered in Pfeffermann and Sverchkov (2009) or Bonnéry (2011), among others.

Definition 1.1.8. *(Sampling design with auxiliary information.)*
Consider the design space $(\mathcal{S}, \mathcal{D})$. A sampling design p using $\boldsymbol{Z}_U$ as auxiliary information is a function $p : \mathcal{S} \times \mathbb{R}^{N \times q} \to [0,1]$, such that

i) for all $s \in \mathcal{S}$, the map $\boldsymbol{Z}_U \mapsto p(s, \boldsymbol{Z}_U)$ is measurable.

ii) for all $\boldsymbol{Z}_U \in \mathbb{R}^{N \times q}$, the map $s \mapsto p(s, \boldsymbol{Z}_U)$ is a sampling design.

Note that Definition 1.1.1 is a special case of Definition 1.1.8, by taking sampling designs such that $p(s, \boldsymbol{Z}_U) = p(s)$, for all $s \in \mathcal{S}$ and $\boldsymbol{Z}_U \in \mathbb{R}^{N \times q}$. To each $\omega \in \Omega$, we can define a design probability $\mathbb{P}_p$ as

$$\mathbb{P}_p(\cdot, \omega) := \sum_{s \in \mathcal{S}} p(s, \boldsymbol{Z}_U(\omega)) \delta_s(\cdot).$$

Note that a sampling design with auxiliary information $p(\cdot, \boldsymbol{Z}_U)$ is a random variable in $(\Omega, \mathcal{M}, \mathbb{P}_m)$, as for different realizations of $\boldsymbol{Z}_U$, we might obtain different probability measures on $(\mathcal{S}, \mathcal{D})$. The joint distribution of the sampling design and the superpopulation model cannot be described anymore as a product measure on the product space; we adopt the construction detailed for instance in Rubin (1976), Boistard et al. (2012) and Han and Wellner (2021).

Definition 1.1.9 (Joint model-design distribution with auxiliary information)**.**
Consider a superpopulation probability space $(\Omega, \mathcal{M}, \mathbb{P}_m)$ and a design probability space $(\mathcal{S}, \mathcal{D}, \mathbb{P}_p)$. The joint model-design probability induced by a sampling design with auxiliary information is defined as the probability $\mathbb{P}_{mz,p}$ on the measurable space $(\Omega \times \mathcal{S}, \mathcal{M} \otimes \mathcal{D})$, uniquely defined as

$$\mathbb{P}_{mz,p}(A \times B) := \int_A \mathbb{P}_p(B, \omega)\, d\mathbb{P}_m(\omega)$$

for all measurable rectangles $A \times B \in \mathcal{M} \times \mathcal{D}$.

Sample membership indicators and inclusion probabilities are defined similarly as in the previous section, but are random variables in the product space. As such, we should write $\pi_k(\boldsymbol{Z}_U)$ for the first order inclusion probability of element k; for simplicity, however, we omit the dependence of $\boldsymbol{Z}_U$ in the notations of the inclusion probabilities.

Poisson sampling

Bernoulli sampling has been defined as a sampling design for which the sample membership indicators $\{I_k\}_{k \in U}$ are i.i.d. . Poisson sampling generalizes Bernoulli sampling, by inducing sample membership indicators $\{I_k\}_{k \in U}$ that are independent but no longer identically distributed; they are now Bernoulli variables with parameter $\pi_k \in]0; 1[$ (possibly dependent of $\boldsymbol{Z}_U$), which are not necessarily the same for all $k \in U$. The sampling design p satisfies

$$p(s, \boldsymbol{Z}_U) = \prod_{k \in s} \pi_k \prod_{k \in U - s} (1 - \pi_k), \qquad s \in \mathcal{S}.$$

As in Bernoulli sampling, the sample size n_s is random. The inclusion probabilities are $\{\pi_k\}_{k \in U}$, specified by the statistician, and $\pi_{k\ell} = \pi_k \pi_\ell$ for all $k \neq \ell \in U$ (by independence of the sample membership indicators).

The quantities $\{\pi_k\}_{k \in U}$ need to be determined in order to implement the Poisson sampling. In Bernoulli sampling, the inclusion probability π can be chosen so that the expected sample size $\mathbb{E}_p[n_s]$ matches the desired sample size. In Poisson sampling, the relation (1.5) is insufficient to determine $\{\pi_k\}_{k \in U}$. A natural criterion would be to minimize the sampling variance of the Horvitz-Thompson estimator under the constraint of a fixed expected sample size:

$$\begin{cases} \min\limits_{[\pi_1, \ldots, \pi_N]^\top \in \mathbb{R}^N} \mathbb{V}_p\left(\widehat{t}_\pi\right) = \min\limits_{[\pi_1, \ldots, \pi_N]^\top \in \mathbb{R}^N} \sum\limits_{k \in U} \left(\dfrac{1}{\pi_k} - 1\right) y_k^2, \\ \text{s.t } \mathbb{E}_p[n_s] = n. \end{cases} \tag{1.6}$$

This problem is easily solved either by using the Lagrangian or with the Cauchy-Schwartz inequality, a proof can be found in Särndal et al. (1992). One finds

$$\pi_k = \frac{n y_k}{\sum\limits_{k \in U} y_k}, \qquad k \in U. \tag{1.7}$$

It is impossible to define such inclusion probabilities since the values $\{y_k\}_{k \in U}$ of the survey variable are unknown prior to sampling. Let Z be an auxiliary variable positively correlated to the survey variable Y and known for all $k \in U$, prior to sampling. We can define

$$\pi_k = \frac{n z_k}{\sum\limits_{k \in U} z_k}, \qquad k \in U, \tag{1.8}$$

that is, replacing the unknown values $\{y_k\}_{k \in U}$ in (1.7) by $\{z_k\}_{k \in U}$. The inclusion probabilities (1.8) may be greater than one; let A be the set of n_A elements such that $n z_k > \sum\limits_{k \in U} z_k$. Then, we set $\pi_k = 1$ for all $k \in A$ and

$$\pi_k = (n - n_A) \frac{z_k}{\sum_{k \in U - A} z_k}$$

for all $\pi_k \in U$ such that $k \notin A$.

Designs that satisfy (1.8) are called *probability proportional to size* designs (Särndal et al., 1992). A Poisson sampling design with inclusion probabilities satisfying (1.8) will be almost as efficient as the strategy solving (1.6). However, the drawback of having a random sample size still remains. It was then suggested to consider fixed size designs without replacement satisfying (1.8), referred to as πps-sampling designs. However, for such designs, it is difficult to compute the second-order inclusion probabilities which are required for estimating the variance of the Horvitz-Thompson estimator. For some probability proportional to size designs[5], Hájek (1964) suggested approximating the second order inclusion probabilities as a function of the first-order inclusion probabilities. Moreover, recent advances in the field provide flexible algorithms to implement designs satisfying (1.8); in particular, we mention the cube method by Deville and Tillé (2004), widely used by national survey offices. Its description is beyond the scope of this work, see Deville and Tillé (2004), Berger and Tillé (2009) or Tillé (2011) for details.

Stratified sampling

Let $\mathcal{P} = \{U_1, U_2, ..., U_H\}$ be a partition of U, where the H elements of $\mathcal{P}$ are called strata. Let N_h denote the number of elements of U_h, so that $N = N_1 + N_2 + ... + N_H$. From each stratum U_h, we select independently a sample S_h of n_h elements according to a sampling design p_h. The final sample S of size $n = \sum_{h=1}^{H} n_h$ is:

$$S = S_1 \cup S_2 \cup ... \cup S_H.$$

[5] This approximation is valid for high entropy sampling designs, where the entropy $I(p)$ of a design p is defined by $I(p) := -\sum_{s \in \mathcal{S}} p(s) \log(s)$, see Tillé (2020).

The overall sampling design is defined as

$$p(S, \mathbf{Z}_U) = p_1(S_1, \mathbf{Z}_U) \times p_2(S_2, \mathbf{Z}_U) \times ... \times p_H(S_h, \mathbf{Z}_U).$$

Since the strata form a partition of U, the population total t_y may be written as :

$$t_y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \sum_{h=1}^{H} t_{yh},$$

where $t_{yh} := \sum_{k \in U_h} y_k$. The first and second-order inclusion probabilities with respect to the sampling design $p_h(\cdot, \mathbf{Z}_U)$ are defined as follows:

$$\begin{aligned} \pi_k^h &= \mathbb{P}_{p,h}(\{k\},\ \mathbf{Z}_U), \qquad && k \in U_h, \\ \pi_{kl}^h &= \mathbb{P}_{p,h}(\{k, \ell\},\ \mathbf{Z}_U), \qquad && k \neq l \in U_h, \end{aligned}$$

for $h = 1, 2, ..., H$. The first and second order inclusion probabilities with respect to $p(\cdot, \mathbf{Z})$ are given by:

$$\pi_k = \pi_k^h, \qquad \text{if } k \in U_h,$$

$$\pi_{kl} = \begin{cases} \pi_{kl}^h & \text{if } \ k \neq l \in U_h, \\ \\ \pi_k^h \pi_l^{h'} & \text{if } k \in U_h, l \in U_{h'} \text{ and } h \neq h' \in \{1, ..., H\}. \end{cases}$$

Using these inclusion probabilities, we can define the π-estimator of t_y as

$$\widehat{t_\pi} = \sum_{h=1}^{H} \widehat{t_h},$$

where $\widehat{t_h} := \sum_{k \in S_h} y_k / \pi_k^h$ is the Horvitz-Thompson estimator of t_{yh}. Because the selections in different strata are independent, the design variance of $\widehat{t_\pi}$ is given by

$$\mathbb{V}_p\left(\widehat{t_\pi}\right) = \sum_{h=1}^{H} \mathbb{V}_p\left(\widehat{t_h}\right).$$

The overall variance of the Horvitz-Thompson estimator is thus dependent of the variance of each of the H Horvitz-Thompson estimators, within each stratum. Therefore, the choice of the stratification variable is important for the efficiency of the overall strategy. The choice of the sample size within each stratum is of great importance as well. The larger n_h is, the better the inferences will be for U_h due to a reduced variance, but increasing n_h also increases the cost of the survey. It is therefore of practical interest to search for sample sizes $\{n_h\}_{h=1,...,H}$ that best control the balance between precision and cost, a problem introduced by Neyman (1934) and known as *optimal sample allocation*. The optimal sample size allocation, at a fixed cost equal for all strata is given by

$$n_h = n \times \frac{N_h S_{y_{Uh}}}{\sum_{h=1}^{H} N_h S_{Y_{Uh}}}, \tag{1.9}$$

for $h = 1, 2, ..., H$, where $S_{y_{Uh}} := \left\{(N_h - 1)^{-1} \sum_{k \in U_h} (y_k - \bar{y}_{Uh})^2\right\}^{1/2}$ is the standard deviation of the variable of interest Y within the population stratum. Naturally, we do not have this information; hence, if a variable Z, available to us, is correlated with the variable of interest Y, we define the Z-optimal allocation:

$$n_h = n \times \frac{N_h S_{z_{Uh}}}{\sum_{h=1}^{H} N_h S_{z_{Uh}}},$$

for $h = 1, 2, ..., H$, where $S_{z_{U_h}} := \left\{(N_h - 1)^{-1} \sum_{k \in U_h} \left(z_k - \bar{z}_{U_h}\right)^2\right\}^{1/2}$, the standard deviation of the variable Z within the stratum h.

If no auxiliary information other than the stratification variable is known other than the stratification variable, a popular strategy is to use proportional allocation, for which the sample sizes are given by

$$n_h = n \times \frac{N_h}{N}$$

for $h = 1, 2, ..., H$. It follows that the sampling fraction is the same for each stratum. If $S_{Y_{U_h}} = S^2$ for all $h = 1, ..., H$, then proportional allocation is optimal.

1.1.5 Asymptotic theory in survey sampling

An important part of theoretical statistics is concerned with the study of the *asymptotic theory*, sometimes also called *large sample theory*. Indeed, it is often difficult to obtain information about the behavior of an estimator for a given fixed sample size. Usually, more information can be deduced about statistics for large samples, i.e., when taking the limit of the sample size to infinity. The rationale behind the procedure is that the conclusions that we may establish in the limit might still hold (approximately) in practice for "large enough" sample sizes. A large part of this dissertation is devoted to investigate the asymptotic properties of survey estimators. However, defining a formal asymptotic framework in survey sampling is not as straightforward as it is in most areas of statistics. The main difference here is that the random sample S takes its values in the subsets of the finite population U. As such, it is not possible to let n, the sample size, increase to infinity if U is fixed. We adopt the asymptotic framework developed in Isaki and Fuller (1982), which we describe below.

We start by considering a sequence $\{N_v\}_{v \in \mathbb{N}}$, strictly increasing to infinity. Next, we consider a sequence of embedded finite populations $\{U_v\}_{v \in \mathbb{N}}$ of size $\{N_v\}_{v \in \mathbb{N}}$, each generated by the superpopulation model. In each finite population U_v, a sample S_v of size n_v is selected according to a sampling design $p_v(\cdot, \boldsymbol{Z}_{U_v})$. While the finite populations are assumed to be embedded, we do not require this property to hold for the samples $\{S_v\}_{v \in \mathbb{N}}$. This asymptotic framework assumes that v goes to infinity, so that both the finite population sizes and the sample sizes go to infinity. To improve readability, we will use the subscript v only in the quantities U_v, N_v and n_v; quantities such as $\pi_{k,v}$ shall be denoted simply as π_k.

This framework has been used to establish the asymptotic properties of various estimators in survey statistics. The asymptotic properties of the Horvitz-Thompson estimator are of particular

interest since many estimators can be described as functions of Horvitz-Thompson estimators. Conditions for the design-consistency[6] of the Horvitz-Thompson estimators were investigated by Isaki and Fuller (1982). The L^2 design and joint consistency[7] of the Horvitz-Thompson estimator were established by various authors, e.g. Robinson and Särndal (1983), Breidt and Opsomer (2000). Central limit theorems[8] were obtained by various authors Madow (1948), Erdos and Rényi (1959), Hájek (1964), Krewski and Rao (1981) and Bickel and Freedman (1984), for particular sampling designs only. A review of some of these results can be found in Fuller (2009b). A joint model-design framework was introduced by Rubin-Bleuer and Kratina (2005) in which central limit theorems and convergence results were obtained for survey estimators. More recently, several authors Bertail et al. (2013), Boistard et al. (2017), Han and Wellner (2021) considered the problem of establishing limit theorems for survey empirical processes; a thorough review on the subject is given in Han and Wellner (2021). The variance estimator of the Horvitz-Thompson estimator given in (1.4) has been shown to be L^1-consistent as well, see Breidt and Opsomer (2000) or Goga and Ruiz-Gazen (2014) for proofs.

1.2 Use of auxiliary information at the estimation stage

1.2.1 Model-assisted estimation in surveys

The previous section described several strategies for which auxiliary variables were incorporated in the sampling design, with the Horvitz-Thompson estimator. In this section, we describe how auxiliary information can also be used at the estimation stage to build more efficient estimators of t_y.

We place ourselves in the superpopulation model as before, in which we define an additional generic random vector $\mathbf{x} \in \mathbb{R}^p$ on $(\Omega, \mathcal{M}, \mathbb{P}_m)$ and denote by $\mathbb{P}_{\mathbf{x}}$ its distribution. For simplicity, we assume that the support of $\mathbb{P}_{\mathbf{x}}$ belongs to the unit hypercube $[0;1]^p$. For simplicity of exposition, we further assume that the survey variable is compactly supported in an interval $\left[C_{1,Y}; C_{2,Y}\right]$. We define a collection of covariates $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ i.i.d. with distribution $\mathbb{P}_{\mathbf{x}}$ and concatenated in a matrix $\boldsymbol{X}_U := [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]^\top$. Without loss of generality, we assume that the random vectors $\{\mathbf{x}_k\}_{k\in U}$ are to be used as auxiliary variables at the *estimation stage*, with techniques detailed in Section 1.3, while the random vectors $\{\mathbf{z}_k\}_{k\in U}$ are designed to be used at

[6] A sequence of estimators $\{\widehat{\theta}_v\}_{v\in\mathbb{N}}$ of $\theta_{U,v}$ is said to be design probability consistent if, for all $\epsilon > 0$,

$$\lim_{v\to\infty} \mathbb{P}_p\left(N_v^{-1}|\widehat{\theta}_v - \theta_{U,v}| > \epsilon\right) = 0.$$

[7] Let $1 \leqslant d < \infty$. A sequence of estimators $\{\widehat{\theta}_v\}_{v\in\mathbb{N}}$ of $\theta_{U,v}$ is said to be design L^d consistent if

$$\lim_{v\to\infty} \mathbb{E}_p\left[N_v^{-d}|\widehat{\theta}_v - \theta_{U,v}|^d\right] = 0, \qquad a.s.$$

The sequence is said to be L^d consistent (for the joint distribution) if

$$\lim_{v\to\infty} \mathbb{E}\left[N_v^{-d}|\widehat{\theta}_v - \theta_{U,v}|^d\right] = 0.$$

[8] A central limit theorem for $\{\widehat{\theta}_v\}_{v\in\mathbb{N}}$ estimators of $\theta_{U,v}$ is a result formalizing the existence of a normalizing sequence $\{\alpha_v\}_{v\in\mathbb{N}}$ and a real σ such that such that $\alpha_v \to \infty$ and $\alpha_v N_v^{-1}\left(\widehat{\theta}_v - \theta_{U,v}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right)$.

the *sampling stage*, as described in the previous section. Throughout the following sections, the aim is to build efficient strategies $(p, \widehat{t})$ for the estimation of t_y. We assume that the sampling design p is already determined and the sample data already collected; hence, the efficiency of $(p, \widehat{t})$ is measured by the mean squared error of the estimator $\widehat{t}$ computed with respect to the sampling design p. At our disposal, we have the information contained in the set

$$D_{ma} := \{(\mathbf{x}_k, y_k); k \in S\} \bigcup \{\mathbf{x}_k; k \in U \backslash S\}.$$

An important part of this work relies on the fact that the problem of estimating finite population totals is closely related to a prediction problem. Indeed, let $f : \mathbb{R}^p \to \mathbb{R}$ be an arbitrary function and define the *generalized difference estimator* of t_y (Cassel et al., 1976) based on f as follows

$$\widehat{t}_{gd}(f) := \sum_{k \in U} f(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - f(\mathbf{x}_k)}{\pi_k}.$$

If f is a function independent of the sample S, the estimator $\widehat{t}_{gd}(f)$ is design-unbiased, namely, $\mathbb{E}_p\left[\widehat{t}_{gd}(f)\right] - t_y = 0$ and its design-variance is given by

$$\mathbb{V}_p\left[\widehat{t}_{gd}(f)\right] = \sum_{k \in U} \sum_{\ell \in U} \Delta_{k\ell} \frac{y_k - f(\mathbf{x}_k)}{\pi_k} \frac{y_\ell - f(\mathbf{x}_\ell)}{\pi_\ell}.$$

The design-mean square error (which is equal, in that case, to the design-variance) of the difference estimator is therefore linked to the choice of the function f and to the quality of the covariates for predicting Y. If we assume that the survey variable can be represented as a function of the covariates as follows

$$y_k = m(\mathbf{x}_k), \qquad \forall k \in U, \tag{1.10}$$

for a function $m : \mathbb{R}^p \to \mathbb{R}$, then it follows that

$$\mathbb{V}_p\left(\widehat{t}_{gd}(m)\right) = 0.$$

In that case, the difference estimator is an optimal estimator for t_y in the sense that its design mean square error has the smallest value. We see therefore that the quality of the strategy $(p, \widehat{t}_{gd}(f))$ is closely related to the quality of the function f chosen in terms of predictor of Y, that is, how close are the quantities $f(\mathbf{x}_k)$ from y_k, for all population elements k. In practice, Assumption (1.10) is often too simplistic; it would be more realistic to assume that

$$y_k = m(\mathbf{x}_k) + \epsilon_k, \qquad \forall k \in U, \tag{1.11}$$

with ϵ_k denoting a sequence of i.i.d. random variables such that $\mathbb{E}[\epsilon_k | \mathbf{x}_k] = 0$ and $\mathbb{V}(\epsilon_k | \mathbf{x}_k) = \sigma^2$. The function m is often called the *regression function*. Under Model (1.11), the difference estimator is expected to be an efficient estimator of t_y; for instance, (Cassel et al., 1976) proved that, $\widehat{t}_{gd}(m)$ is optimal in the sense that it minimizes the mean squared error with respect to the joint distribution among the class of design unbiased estimators of t_y (Cassel et al., 1976).

In most cases, the difference estimator based on m cannot be used as the regression function m is unknown. The idea of *model-assisted estimation* is to estimate the unknown function m by a *regression function estimator* (also called a *predictor*) $\widehat{m}$, fitted on $\{(\mathbf{x}_k, y_k)\,; k \in S\}$. The fitted model is then used to construct the *model-assisted estimator* of t_y (Särndal et al., 1992):

$$\widehat{t}_{ma} = \sum_{k\in U} \widehat{m}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}(\mathbf{x}_k)}{\pi_k}, \tag{1.12}$$

where $\widehat{m}(\mathbf{x}_k)$ denotes the *prediction* of $\widehat{m}$ at the point $\mathbf{x}_k$. The model-assisted estimator is therefore defined as the population total of the estimated predictions $\{\widehat{m}(\mathbf{x}_k)\}_{k\in U}$ and the weighted sum of the sampled units of the estimated residuals $\{y_k - \widehat{m}(\mathbf{x}_k)\}_{k\in S}$. If the predictions $\{\widehat{m}(\mathbf{x}_k)\}_{k\in S}$ are close to the true values $\{y_k\}_{k\in S}$, then the first term will dominate; if the predictions $\{\widehat{m}(\mathbf{x}_k)\}_{k\in U}$ are inefficient, however, the second term will counterbalance the inefficiency of the first term and thus adds robustness to the estimator.

Whenever the predictor $\widehat{m}$ is sample dependent, the estimator $\widehat{t}_{ma}(\widehat{m})$ is no longer design-unbiased and its properties do not follow from the properties of the difference estimator. The properties of (1.12) has been investigated for many predictors, from parametric models (Robinson and Särndal, 1983), to non-parametrics such as local polynomials (Breidt and Opsomer, 2000), B-splines (Goga, 2005) and (Goga and Ruiz-Gazen, 2014), penalized splines (Breidt et al., 2005, McConville and Breidt, 2013), neural networks (Montanari and Ranalli, 2005), generalized additive models (Opsomer et al., 2007) and regression trees (McConville and Toth, 2019). In the aforementioned articles, the design properties (asymptotic design unbiaseness and consistency) of these estimators were established. The conclusions made about model-assisted estimator are therefore independent of the quality of the model predictions. As explained in Särndal et al. (1992), the approach is model-based, but not model-dependent. Several authors, however, also investigated the joint design-model properties of these estimators, including Särndal (1980), Robinson and Särndal (1983), Breidt and Opsomer (2000) and Goga (2005).

In most theoretical investigations of model-assisted estimators, the asymptotic properties were established in an asymptotic framework in which the number of covariates included in the model was kept fixed, thus implying that the ratio p/n is negligible. However, nowadays, it is no longer unusual to face high-dimensional auxiliary information. These practical scenarios are therefore *not included* in the scenario usually considered because the assumption that p/n is negligible may not be satisfied. The behavior of these estimators in such scenarios was therefore, until recently, unknown. Recently, increasing attention has been devoted to establishing asymptotics properties of model-assisted estimators in a framework in which the number of covariates is increasing to infinity as well. This framework is called *high-dimensional*; the asymptotic results established in this framework include practical situations in which both $p = p_v$[9] and n_v are large. Often, they can be seen as generalizations of the results obtained in a low-dimensional framework. In particular, Cardot et al. (2017) studied dimension reduction through principal component analysis and established the design consistency of the resulting estimator in a setting in which the number of principal components is allowed to increase. More recently, Ta et al. (2020) investigated the properties of model-assisted estimators based on linear regression and the Lasso, under the joint model-design distribution. Chauvet and

[9] Here, the notation p is used to denote the number of covariates, and not the sampling design.

Goga (2022) studied the design asymptotic properties of calibration estimators[10], when the number p_v of calibration variables is going to infinity.

In this dissertation, two articles focus on model-assisted estimation. First, we present the article Dagdoug et al. (2022a) entitled *Model-assisted estimation in high-dimensional settings for survey data* which investigates the asymptotic design behavior of linear and penalized linear model-assisted estimators in a high-dimensional framework. The complete article is presented in Chapter 3. Next, the article Dagdoug et al. (2021b) entitled *Model-assisted estimation through random forests in finite population sampling* deals with model-assisted estimators based on random forests. The finite sample properties of the resulting estimator were thoroughly investigated and its asymptotic properties were established (L^1-consistency, determination of the asymptotic variance, suggestion of a L^1-consistent variance estimator, asymptotic distribution). The complete article is provided in Chapter 4.

1.2.2 Model-assisted estimation in high-dimensional settings for survey data

For simplicity of exposure, we describe the predictors as if the available data was the population data $D_U := \{(\mathbf{x}_k, y_k)\,; k \in U\}$. Extension to their definitions at the sample level will be detailed subsequently.

A description of linear and penalized linear models

In practice, it is often common and convenient for practitioners to assume that the regression function in (1.11) is a linear function of the covariates, that is, that there exists $\boldsymbol{\beta}$ in $\mathbb{R}^p$ such that $m(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$. The unknown vector $\boldsymbol{\beta}$ can be estimated by $\widetilde{\boldsymbol{\beta}}_{lr}$ through the ordinary least square criterion at the population level:

$$\widetilde{\boldsymbol{\beta}}_{lr} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{k\in U} (y_k - \mathbf{x}_k^\top \boldsymbol{\beta})^2. \tag{1.13}$$

Provided that the matrix $\boldsymbol{X}_U$ is of full rank, the solution of (1.13) is unique, a closed-form solution exists and is given by:

$$\widetilde{\boldsymbol{\beta}}_{lr} = \left(\sum_{k\in U} \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \sum_{k\in U} \mathbf{x}_k y_k.$$

The prediction at $\mathbf{x}_k$ based on $\widetilde{\boldsymbol{\beta}}_{lr}$ is given by $\widetilde{m}(\mathbf{x}_k) := \mathbf{x}_k^\top \widetilde{\boldsymbol{\beta}}_{lr}$, for all $k \in U$.

[10] Calibration is a procedure suggested by Deville and Särndal (1992), widely used in practice. The main idea is to search for weights $\{w_k\}_{k\in S}$, which are as close as possible to the original weighting system $\{d_k\}_{k\in S}$, from a distance point of view, while satisfying the constraint that the w-weighted estimator of the covariates totals perfectly estimates their totals, whatever the sample S is. The reader is referred to Särndal (2007) for a review on the subject.

In statistical learning, a popular method for improving the estimation of $\boldsymbol{\beta}$ in case of a large number of covariates is to have recourse to penalization to estimate the unknown vector $\boldsymbol{\beta}$. More precisely, we define

$$\widetilde{\boldsymbol{\beta}}_{pen} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{k\in U} \left(y_k - \mathbf{x}_k^\top \boldsymbol{\beta}\right)^2 + \sum_{\ell=1}^{t} \lambda_\ell ||\boldsymbol{\beta}||_{\nu_\ell}^{\gamma_\ell}, \tag{1.14}$$

with $t \in \mathbb{N}$, $\lambda_\ell \in \mathbb{R}^+$, $\nu_\ell \in \mathbb{N}$ and $\gamma_\ell \in \mathbb{R}^+$ are hyper-parameters to be chosen before estimation. Common choices include $t = 1$, $\gamma_1 = 1$ and $\eta_1 = 1$ for the lasso; $t = 1$, $\gamma_1 = 2$ and $\eta_1 = 2$ for ridge; $t = 2$, $\gamma_1 = 1$, $\eta_1 = 1$, $\gamma_2 = 2$ and $\eta_2 = 2$ for the elastic-net. The effect of penalization is to decrease the norm of the vector of estimated coefficients. Some choices such as the Lasso or the Elastic-net are able to put some coefficients down to zero, and therefore can be seen as variable selection methods as well. The prediction at the point $\mathbf{x}_k$ with a penalized linear model is given as

$$\widetilde{m}_{pen}(\mathbf{x}_k) := \mathbf{x}_k^\top \widetilde{\boldsymbol{\beta}}_{pen}, \qquad k \in U.$$

The generalized linear regression estimator and its penalized counterparts

In practice, the vectors $\widetilde{\boldsymbol{\beta}}_{lr}$ and $\widetilde{\boldsymbol{\beta}}_{pen}$ cannot be computed as the y-values are recorded for the sample units only. An estimator of $\widetilde{\boldsymbol{\beta}}_{lr}$, denoted by $\widehat{\boldsymbol{\beta}}_{lr}$, is obtained using the following weighted least square criterion at the sample level:

$$\widehat{\boldsymbol{\beta}}_{lr} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{k\in S} \frac{(y_k - \mathbf{x}_k^\top \boldsymbol{\beta})^2}{\pi_k}. \tag{1.15}$$

Again, the solution to (1.15) is unique provided that $\boldsymbol{X}_S := (\mathbf{x}_k^\top)_{k\in S}$ is of full rank and is given by:

$$\widehat{\boldsymbol{\beta}}_{lr} = \left(\sum_{k\in S} \frac{\mathbf{x}_k \mathbf{x}_k^\top}{\pi_k}\right)^{-1} \sum_{k\in S} \frac{\mathbf{x}_k y_k}{\pi_k}. \tag{1.16}$$

Plugging $\widehat{m}_{lr}(\mathbf{x}_k) := \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_{lr}$ in (1.12) leads to the model-assisted estimator based on linear regression, also called the Generalized regression estimator (GREG, Särndal et al. (1992)):

$$\widehat{t}_{greg} = \sum_{k\in U} \widehat{m}_{lr}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}_{lr}(\mathbf{x}_k)}{\pi_k}.$$

For the estimation of $\widetilde{\boldsymbol{\beta}}_{pen}$, we define the following sample criterion:

$$\widehat{\boldsymbol{\beta}}_{pen} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{k\in S} \frac{1}{\pi_k} \left(y_k - \mathbf{x}_k^\top \boldsymbol{\beta}\right)^2 + \sum_{\ell=1}^{t} \lambda_\ell ||\boldsymbol{\beta}||_{\nu_\ell}^{\gamma_\ell}.$$

A model-assisted estimator based on a penalized regression procedure is obtained from (1.12) by replacing $\widehat{m}(\mathbf{x}_k)$ with $\widehat{m}_{pen}(\mathbf{x}_k) = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_{pen}$, leading to

$$\widehat{t}_{pen} = \sum_{k\in U} \widehat{m}_{pen}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}_{pen}(\mathbf{x}_k)}{\pi_k}.$$

High-dimensional asymptotics for linear and penalized linear models

To state the high-dimensional consistency of the GREG estimator, mild regularity conditions on the sampling design, the survey variable and the covariates are needed. These are the extension of those used in Robinson and Särndal (1983), but adapted for the high-dimensional framework, see Chapter 3 for more details.

Result 1.2.1. *Consider a sequence of GREG estimators* $\{\widehat{t}_{greg}\}_{v\in\mathbb{N}}$. *Then,*

$$\frac{1}{N_v}(\widehat{t}_{greg} - t_y) = O_p\left(\sqrt{\frac{p_v^3}{n_v}}\right).$$

Result 1.2.1 shows that, to guarantee the consistency of the GREG estimator $\widehat{t}_{greg}$, the number of auxiliary variables must be relatively small with respect to the sample size, i.e., $p_v^3/n_v = o_p(1)$. Under stronger yet still realistic assumptions, the convergence rate can be improved to $\sqrt{p_v^2/n_v}$. Even then, the convergence rate is quite slow compared to the traditional asymptotic framework in which the GREG estimator is square-root consistent. With the same assumptions used in Result 1.2.1, we show that the penalized estimator $\widehat{t}_{pen}$ is consistent whenever the GREG estimator is and that, as a consequence, the penalized estimator cannot converge slower than the GREG estimator.

Result 1.2.2. *Consider a sequence of penalized model-assisted estimators* $\{\widehat{t}_{pen}\}_{v\in\mathbb{N}}$ *of* t_y *obtained by either ridge, lasso or elastic-net. Then,*

$$\frac{1}{N_v}(\widehat{t}_{pen} - t_y) = O_p\left(\sqrt{\frac{p_v^3}{n_v}}\right).$$

Under relatively stronger regularity conditions than those needed for result 1.2.2 (see chapter 3 for more details), it is possible to improve on the convergence rate of some penalized estimators. For instance, consider the particular case of the ridge estimator, $\widehat{t}_{ridge}$.

Result 1.2.3. *Consider a sequence of penalized ridge estimators* $\{\widehat{t}_{ridge}\}_{v\in\mathbb{N}}$. *Then,*

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{ridge} - t_y\right|^2 = O\left(\frac{p_v}{n_v}\right).$$

Simulation study

In Dagdoug et al. (2022a), we conducted a large simulation study that included many scenarios. We chose to compare a wide range of model-assisted estimators, such as based on linear regression, penalized regression with lasso, ridge and elastic-net, regression trees, random forests, Cubist, gradient boosting, k-nearest neighbors and principal component regression, see Chapter 3 and 5 for details. We were interested in estimating the finite population total of four survey variables; some of them were linear in the auxiliary variables while others were not. To isolate the influence of the dimension, we added an increasing number of noise variables that were not related to the survey variables and repeated our simulations. We used both equal and unequal sampling designs in the simulations.

Overall, Cubist and penalized regression estimators were the most efficient. Random forests, XGBoost, principal component regression, and, to a lesser extent, k-nearest neighbors, also improved on the Horvitz-Thompson estimator in most cases. The results of linear regression were very dependent of the survey variable considered and of the number of covariates considered. Our results also illustrated a few notable facts. First, whether or not the survey variable was linear in the auxiliary variables, the estimator based on linear regression was the most impacted by the addition of noise variables. For instance, in one scenario, its relative efficiency with respect to the Horvitz-Thompson estimator increased by almost 600%. For that same scenario, the other estimators had a relative efficiency which increased of only 14%, on average. The same observation was revealed in most scenarios tested, no matter the sampling design. Another interesting finding is that, when using unequal sampling designs with random forests, the estimator may exhibit a small sample bias if the hyper-parameters are not well-chosen (more precisely, if the covariates used in the sampling design are not sufficiently taken into account).

1.2.3 Random forest for model-assisted estimation in finite population sampling

Regression trees and random forests are algorithms suggested for estimating the unknown regression function m in (1.11) and making predictions; these are non-parametric *prediction methods* or *predictors*. We begin by describing regression trees and random forests defined at the population level.

A description of regression trees

A regression tree is a prediction method that can be viewed as an algorithm composed of two parts: a *partitioning algorithm* and a *prediction rule*. Let D_N denote the set of N-tuples of vectors of $[0;1]^p \times [C_{1,Y}; C_{2,Y}]$.

A *partitioning algorithm* is an algorithm which, given data points, defines a partition of the space of covariates. That is, this is a deterministic function $P : D_N \longrightarrow \mathcal{P}([0;1]^p)$ where $\mathcal{P}([0;1]^p)$ denotes the set of partitions of the unit hypercube of $\mathbb{R}^p$, see Nobel (1996) for more details. Generally, partitions are created by successive splits with the objective of optimizing a certain criterion. The elements of the resulting partition $\mathcal{P} := \{A_1, A_2, ..., A_T\}$ will be called the *leaves* or *nodes* of the tree.

A *prediction rule* is an algorithm which, takes as input a partition $\mathcal{P} := \{A_1, A_2, ..., A_T\}$ and a dataset D_U, and returns a prediction. In the case of regression trees, the prediction rule traditionally used returns the empirical average of the set $\{y_k;\ k \in U \text{ such that } \mathbf{x}_k \in A(\mathbf{x})\}$, where $A(\mathbf{x})$ denotes the leaf of the tree containing point $\mathbf{x}$. More precisely, the prediction $\widetilde{m}_{tree}(\cdot, \mathcal{P}, D_U) := \widetilde{m}_{tree}(\cdot)$ is defined by

$$\widetilde{m}_{tree}(\mathbf{x}) = \sum_{k \in U} \frac{\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in U} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} y_k, \tag{1.17}$$

where, $\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})} = 1$ if $\mathbf{x}_k \in A(\mathbf{x})$ and 0 otherwise. In this article, unless otherwise stated, the term tree or regression tree will designate any regression tree built from any partitioning algorithm.

Figure 2 illustrates a regression tree based on two covariates, leading to the corresponding partition of $\mathbb{R}^2$. Below we describe the CART partitioning algorithm, commonly used in practice.

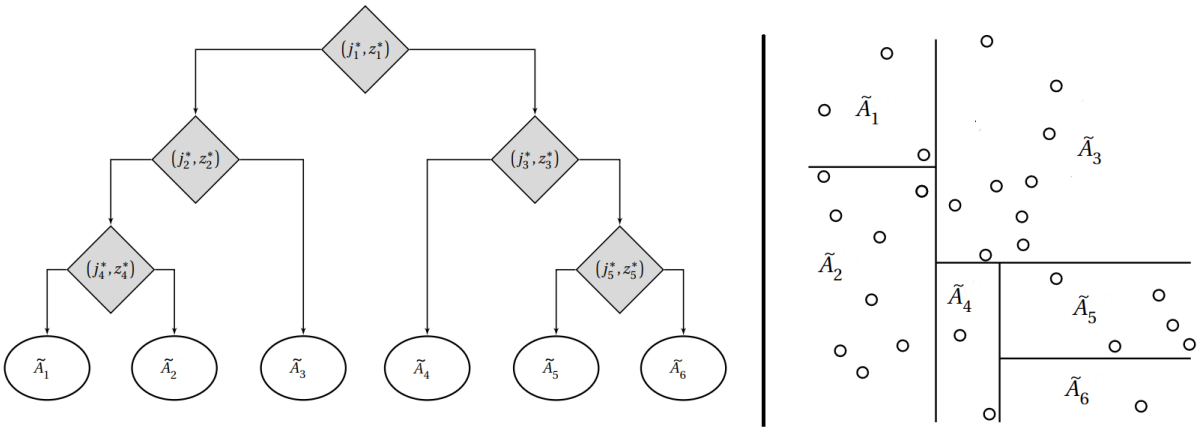


Figure 2: Regression tree on $\mathbb{R}^2$ (left) and its corresponding partition (right).

Example 1.2.1. *CART partitioning algorithm (Breiman, 1984).*
In the CART partitioning algorithm, the partition is obtained by successive splits. More precisely, let A be a leaf of cardinality $\#(A)$ considered for the next split and C_A be the set of all possible splits in the leaf A, which corresponds to all the pairs $(j, z) = (\text{variable}, \text{position})$. Define

$$mse(A) := \frac{1}{\#(A)} \sum_{k \in U} \mathbb{1}_{\mathbf{x}_k \in A} (y_k - \bar{y}_A)^2 \qquad \text{and} \qquad \bar{y}_A := \frac{1}{\#(A)} \sum_{k \in U} \mathbb{1}_{\mathbf{x}_k \in A} y_k .$$

The splitting procedure is performed by finding the best split (j^*, z^*), that is, the one maximizing the following criterion

$$L(j, z) = mse(A) - mse(A_L) - mse(A_R) \tag{1.18}$$

where $A_L = \{k \in A; x_{kj} < z\}$, $A_R = \{k \in A; x_{kj} \geqslant z\}$. Equivalently, note that maximizing (1.18) is similar to minimizing

$$L(j, z) = mse(A_L) + mse(A_R) .$$

This criterion therefore searches for the split which would generate child nodes as homogeneous as possible, in terms of mean square error. Splits are always performed in the middle of two points. The procedure continues as long as a stopping criterion is not reached. The usual stopping criteria are obtained by specifying a minimum number of elements (n_0) in the terminal nodes, or a maximum depth (K) for the tree.

Example 1.2.2. *Partitioning rule proposed by McConville and Toth (2019).*
The algorithm proposed by McConville and Toth can be described as follows:

1. *Consider $n_0 := n^{11/20}$, the minimum number of units in each node and choose $\alpha \in]0; 0.5[$, a confidence level.*

2. *If the chosen leaf A contains less than $2 \times n_0$ elements, then A is a terminal node. In this case, return to step 1. for the next node.*

3. *Among the available p covariates, choose the one that has the test statistic with the lowest p-value in the hypothesis test H_0 : "$\exists C \in \mathbb{R}$ such as $\mathbb{E}\left[Y|X_j \in A\right] = C$" for $j = 1, ..., p$. If none of these test statistics is significant (with respect to the α threshold set in step 1.), then A is a terminal leaf. In this case, return to step 1. for the next node.*

4. *Perform the split at a position $z^* \in \arg\max_z L(j, z)$, with L defined in the same way as for the CART criterion. This criterion is only optimized on the positions leading to child leaves containing at least n_0 elements in each child leaf.*

For more details about trees and partitioning procedures, the reader is referred to Hastie et al. (2011) or Györfi et al. (2006), for comprehensive treatments of the topic.

Heuristic motivation and description of random forests

In practice, regression trees are particularly popular because their predictions can be easily understood and interpreted. However, their predictive efficiency may be low in some cases; see Figure 3 for an illustration motivating the use of random forests instead of regression trees. We have generated 100 observations with a covariate X_1 from a distribution $\mathcal{U}([0; 1])$ and a survey variable $Y = 4 + 2X_1^2 + \mathcal{N}(0; 0.2)$. The green curve is the real regression function m and we computed two estimations of m: the red curve is the tree-regression estimate of m based on the CART criterion with $n_0 = 20$ (see Exemple 1.2.1 from below) and the yellow curve is the random forest estimation of m. We can remark that the random forest predictor provides a much better estimation of m and we give below an heuristic explanation of this fact.

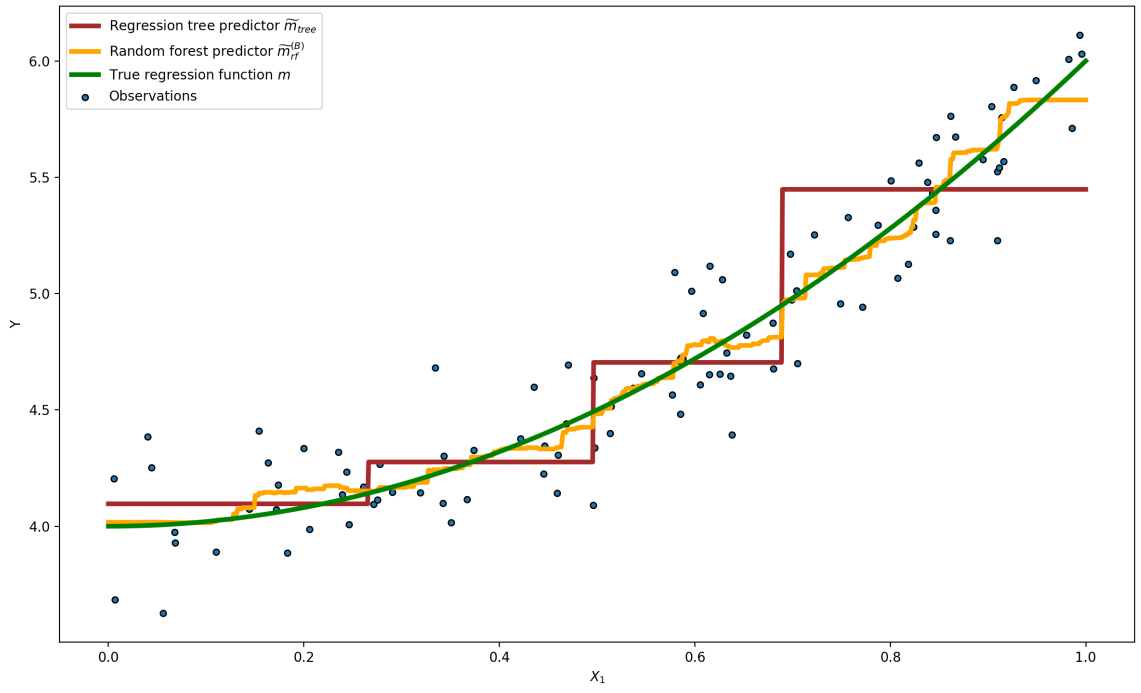


Figure 3: Estimation of a regression function with a regression tree and a random forest algorithm.

By construction, a regression tree belongs to the set of piecewise functions, a set of functions with finite complexity[11]. Furthermore, the complexity of trees is necessarily lower or equal to N; thus, $\widetilde{m}_{tree}$ belongs to the set of piecewise functions from $\mathbb{R}^p$ to $\mathbb{R}$ with, at most, N different values. If the regression function m belongs to this function space, then $\widetilde{m}_{tree}$ may be a good estimator of m. However, if m is a smooth function, say continuous, then $\widetilde{m}_{tree}$ may be far from m as the complexity of m is $+\infty$ if m is not constant. Yet, every continuous function can be seen as the uniform limit of a sequence of piecewise functions. The construction of such a sequence of piecewise functions converging to a continuous function is based on two facts: 1) the complexity must increase to infinity; 2) the diameter of each interval on which the piecewise function is constant must decrease to zero. It is thus possible to see N as an indicator of maximal complexity that a tree can reach. It follows that, asymptotically (for large N), it is possible for a tree to be an efficient estimator of a continuous regression function. For small samples, however, this may be another story.

While piecewise constant functions only have a small complexity, the following idea permits to increase it substantially. It is possible to show that, if $\{f_b\}_{b=1}^B$ is a list of B piecewise functions, each with maximal complexity N, then the average function

$$f_{ave} := \frac{1}{B}\sum_{b=1}^{B} f_b,$$

belongs to the set of piecewise functions with maximal complexity $B \times N$. Hence, the maximal complexity of f_{ave} can be much greater than the complexity of the piecewise functions $\{f_b\}_{b=1}^B$. Naturally, for the gain of complexity to be important, the functions $\{f_b\}_{b=1}^B$ must be different one from another (the intervals on which they are constant should be different).

A random forest is a predictor using this principle to estimate m: it is defined as an average of B regression trees (thus, an average of piecewise functions). We see that the random forest represented by the orange curve in Figure 3 is still a piecewise constant function, but with a larger complexity than the red curve of the regression tree. Since the prediction rules described in Examples 1.2.1 and 1.2.2 are deterministic, it is clear that, for a fixed set of elements, using the same partitioning rule to construct B trees would simply result in constructing the same tree B times. In this case, there would be no gain in complexity. Breiman thus suggested (Breiman, 1996, 2001) to introduce an additional randomness in the partitioning algorithm and/or prediction rule. The additional randomness introduced in the predictors can be defined using the concept of *stochastic predictors*. Let Θ be a random variable defined in a measurable space $(J, \mathcal{J})$. A stochastic predictor $\widetilde{m}$ is a measurable function such that $\widetilde{m} : \mathbb{R}^p \times J \to \mathbb{R}$. In other words, the predictor $\widetilde{m}$ may use a random variable Θ to make its predictions. It follows that the prediction method $\widetilde{m}$ is random with respect to Θ and, as such, an additional randomness is present.

Example 1.2.3. *Let $q \in]0;1[$ and Θ be a random variable with Bernoulli distribution $\mathcal{B}(q)$; define $\widetilde{m}(\mathbf{x}, \Theta) := \Theta||\mathbf{x}||_2$, where $||\cdot||_2$ denotes the Euclidean norm. Then, $\widetilde{m}$ is a stochastic prediction model, meaning that, for two different realizations of Θ, the predictor $\widetilde{m}$ may generate different predicted values. An additional source of randomness is present, i.e., $\mathbb{V}_\Theta(\widetilde{m}(\mathbf{x}, \Theta)) > 0$, where $\mathbb{V}_\Theta(\cdot)$ denotes the variance with respect to the random variable Θ.*

[11] For a function $f : \mathcal{X} \to \mathbb{R}$, we call complexity of f the number of different values that f can take, that is, the quantity $\#(\{f(x); x \in \mathcal{X}\})$.

Two more concrete examples of how the randomization procedure can be incorporated and used in regression trees are given below.

Example 1.2.4. *Breiman's random forests, (Breiman, 2001).*
The algorithm of Breiman can be described as follows:

1. *Select B bootstrap samples[12] in U, denoted by $\{U(\Theta_b)\}_{b=1}^{B}$.*
2. *In each bootstrap sample, $U(\Theta_b)$, build a stochastic regression tree $\widetilde{m}(\cdot, \Theta_b)$ by using the CART criterion, as described in Example 1.2.1, where the splitting criterion is optimized only on p_0 covariates among the p available. The p_0 covariates are chosen uniformly at random (without replacement) among the p covariates available, according to Θ_b, at each split.*

Example 1.2.5. *Uniform random forests (Biau et al., 2008, Scornet, 2016a).*
All the B trees of the forest have the same behavior; as such, we describe only the behavior of a generic tree among the B belonging in the forest. We begin by considering $[0;1]^p$ as the initial leaf. Then, recursively, the algorithm splits in the following fashion.

1. *A node G is chosen uniformly at random among the existing nodes.*
2. *A covariate X_j is chosen uniformly at random among the p covariates $X_1, X_2, ..., X_p$.*
3. *A split is performed in the node G on X_j at a position chosen uniformly at random.*

The process is repeated K times, with $K \in \mathbb{N}$, a parameter chosen by the user.

It is now possible to define the prediction of a random forest as an average of the predictions obtained from the B stochastic regression trees. More precisely, let $\{\Theta^{(b)}\}_{b=1}^{B}$ be a sequence of B i.i.d. random variables with distribution $\mathbb{P}_\Theta$ and $\{\widetilde{m}_{tree}(\cdot, \Theta^{(b)})\}_{b=1}^{B}$ be a sequence of stochastic regression trees; a random forest prediction is defined as

$$\widetilde{m}_{rf}(\cdot, \{\Theta^{(b)}\}_{b=1}^{B}) := \frac{1}{B}\sum_{b=1}^{B} \widetilde{m}_{tree}(\cdot, \Theta^{(b)}). \tag{1.19}$$

For simplicity of notations, we note $\widetilde{m}_{rf}^{(B)}$ the estimator in (1.19).

Remark 1.2.1. *Initially, the term "random forest" was used to denote the initial algorithm of Breiman (2001) and described in Example 1.2.4. However, the definition given in this section describes a class of random forests algorithms, rather than a particular algorithm. Indeed, to each partitioning rule and randomization process, it is possible to define a "random forest" algorithm. Therefore, the definition given above is quite general and includes many algorithms (including the original algorithm of Breiman (2001)). It is also important to note that (non-stochastic) regression trees as defined above are also part of this class; indeed, by taking $B = 1$ and a deterministic partitioning rule, we obtain a regression tree.*

For more details about random forests and their implementation, the reader is referred to Biau and Scornet (2016) and Genuer and Poggi (2019).

[12] A bootstrap sample of U is a sample of size N, selected from U with replacement.

Random forests model-assisted estimator

The (random forest) difference estimator is defined as

$$\widehat{t}_{dif}^{(B)} := \sum_{k\in U} \widetilde{m}_{rf}^{(B)}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widetilde{m}_{rf}^{(B)}(\mathbf{x}_k)}{\pi_k}, \tag{1.20}$$

where $\widetilde{m}_{rf}^{(B)}$ is given by (1.19). In practice, the estimator $\widehat{t}_{dif}^{(B)}$ is not feasible. Indeed, it is built from the prediction method $\widetilde{m}_{rf}$, itself built from D_U and therefore from unknown data. We propose to estimate the unknown prediction method $\widetilde{m}_{rf}$ by $\widehat{m}_{rf1}^{(B)}$ using the information in D_{ma}:

$$\widehat{m}_{rf1}^{(B)}(\mathbf{x}) := \frac{1}{B}\sum_{b=1}^{B} \sum_{k\in S(\Theta_b)} \frac{\pi_k^{-1}\mathbb{1}_{\mathbf{x}_k\in\widehat{A}^{(b)}(\mathbf{x})}}{\sum_{\ell\in S(\Theta_b)}\pi_\ell^{-1}\mathbb{1}_{\mathbf{x}_\ell\in\widehat{A}^{(b)}(\mathbf{x})}} y_k, \tag{1.21}$$

where $S(\Theta_b)$ denotes the sample which was used to build the b-th tree and $\widehat{A}^{(b)}(\mathbf{x})$ the leaf of the b-th tree containing the point $\mathbf{x}$ and obtained by applying a partitioning algorithm such as the CART algorithm given in Example 1.2.4 on the data D_{ma}. The sums over the population are therefore replaced by sums across the sample and a weighting process is applied. More precisely, the sampling weights are incorporated into the numerator and the denominator of (1.21), thus making it possible to better take into account sampling designs with unequal probabilities. If the sampling design considered induces equal inclusion probabilities for all the elements of the population, then the weights in (1.21) cancel themselves and the estimator $\widehat{m}_{rf}^{(B)}$ of $\widetilde{m}_{rf}^{(B)}$ is simply an estimator constructed by replacing population sums with sample sums and population partitions are replaced by sample partitions.

We now define a random forest model-assisted estimator as follows:

$$\widehat{t}_{rf1}^{(B)} = \sum_{k\in U} \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k)}{\pi_k}. \tag{1.22}$$

Remark 1.2.2. *As mentioned previously, the random forest definition used before includes a wide class of algorithms. We can note that, given the definitions of $\widehat{m}_{rf1}^{(B)}$ in (1.21) and $\widehat{t}_{rf1}^{(B)}$ in (1.22), the equation $\widehat{t}_{rf1}^{(B)}$ defines a class of estimators rather than a particular estimator. More precisely, let $\mathcal{F}_{rf}(D_{ma}, B)$ denote the set of weighted random forest functions with B trees, fitted on $\{(\mathbf{x}_k, y_k); k \in S\}$. In this article, $\widehat{t}_{rf1}^{(B)}$ actually represents any element of the set*

$$\mathcal{T}_{rf}(D_{ma}, B) := \left\{ \widehat{t} = \sum_{k\in U} f(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - f(\mathbf{x}_k)}{\pi_k};\ f \in \mathcal{F}_{rf}(D_{ma}, B) \right\}.$$

Observe that $\mathcal{T}_{rf}(D_{ma}, 1)$ is the space of model-assisted estimators based on regression trees (stochastic or not); thus, the set $\mathcal{T}_{rf}(D_{ma}, 1)$ contains the estimator proposed by McConville and

Toth (2019). The results presented in this section being independent of B, the number of trees, the results presented below and in Chapter 4 are valid for any element $\widehat{t}_{rf1}^{(B)}$ *of*

$$\mathcal{T}_{rf}(D_{ma}) := \bigcup_{B\in\mathbb{N}^*} \mathcal{T}_{rf}(D_{ma}, B).$$

Finite sample properties of the random forest estimator

Proposition 1.2.1. *Consider a random forest model-assisted estimator* $\widehat{t}_{rf1}^{(B)}$.

1. *The estimator* $\widehat{t}_{rf1}^{(B)}$ *can be seen as an average of model-assisted estimators:*

$$\widehat{t}_{rf1}^{(B)} = \frac{1}{B}\sum_{b=1}^{B} \widehat{t}_{tree1}^{(b)},$$

where $\widehat{t}_{tree1}^{(b)}$ *denotes the model-assisted estimator based on the b-th tree in the forest, i.e.,*

$$\widehat{t}_{tree1}^{(b)} = \sum_{k\in U} \widehat{m}_{tree1}^{(b)}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}_{tree1}^{(b)}(\mathbf{x}_k)}{\pi_k},$$

and $\widehat{m}_{tree1}^{(b)}$ *is an estimator of (1.17).*

2. *The estimator* $\widehat{t}_{rf1}^{(B)}$ *can be written*

$$\widehat{t}_{rf1}^{(B)} = \sum_{k\in U} \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k) + \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in O_b^{(S)}} \frac{y_k - \widehat{m}_{tree1}^{(b)}(\mathbf{x}_k)}{\pi_k}, \tag{1.23}$$

where $O_b^{(S)} := S - S(\Theta_b)$ *denotes the so-called "out-of-bag" elements of the b-th tree.*

3. *If* $\widehat{m}_{rf1}^{(B)}$ *does not use a resampling mechanism*[13]*, then* $\widehat{t}_{rf1}^{(B)}$ *has the projection property*[14]*:*

$$\widehat{t}_{rf1}^{(B)} = \sum_{k\in U} \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k).$$

The point 1. above reveals that the estimator $\widehat{t}_{rf1}^{(B)}$ is actually an average of B model-assisted estimators. More generally, it is possible to show that an average of model-assisted estimator remains a model-assisted estimator. Random forest estimators also have the following property: if $\widehat{t}_{rf1}^{(B)}$ and $\widehat{t}_{rf1'}^{(B)}$ denote two forest estimators, each with B trees and built from the same algorithm, then their average $(\widehat{t}_{rf1}^{(B)} + \widehat{t}_{rf1}^{(B)})/2$ is a random forest estimator built on $2B$ trees. This property is no longer exact if we take the average of two forests with different numbers of trees. The point 2. shows that the random forest estimator computes its residuals only on

[13] In a random forest, a resampling mechanism is a process consisting of selecting, with or without replacement, observations from the original data before constructing the trees.

[14] A model-assisted estimator $\widehat{t}_{ma}(\widehat{m})$ is said to be a projection estimator if it can be written as the sum of predictions of all population elements, i.e., $\widehat{t}_{ma}(\widehat{m}) = \sum_{k\in U} \widehat{m}(x_k)$.

the elements not selected to build the model. This is an unexpected positive point specific to model-assisted estimators built on "bagging" (Breiman, 1996) type algorithms. Therefore, the second term to the right of (1.34) can be seen as a protection against inefficient predictions and against overfitting. In particular, this implies that the efficiency of forests that do not include a resampling mechanism relies entirely on the prediction model (a consequence of point 3.).

Note that it is also possible to write $\widehat{m}_{rf1}^{(B)}$ as a weighted sum of the values $\{y_k\}_{k\in S}$:

$$\widehat{m}_{rf1}^{(B)}(\mathbf{x}) = \sum_{k\in S} \widehat{W}_{k1}^{(B)}(\mathbf{x}) y_k, \tag{1.24}$$

where

$$\widehat{W}_{k1}^{(B)}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \frac{\pi_k^{-1}\psi_k^{(b)}\mathbb{1}_{\mathbf{x}_k\in\widehat{A}(\mathbf{x})}}{\sum_{\ell\in S}\pi_\ell^{-1}\psi_\ell^{(b)}\mathbb{1}_{\mathbf{x}_\ell\in\widehat{A}(\mathbf{x})}}, \qquad k\in S, \tag{1.25}$$

with $\psi_\ell^{(b)} = 1$ if $\ell \in S(\Theta_b)$ and 0 otherwise.

Proposition 1.2.2. *Consider a model-assisted random forest estimator $\widehat{t}_{rf1}^{(B)}$.*

1. *The estimator $\widehat{t}_{rf1}^{(B)}$ can be seen as a weighted sum of the values $\{y_k\}_{k\in S}$:*

$$\widehat{t}_{rf1}^{(B)} = \sum_{k\in S} w_{k1}^{(B)} y_k,$$

where

$$w_{k1}^{(B)} = \frac{1}{\pi_k}\left\{1 + \sum_{\ell\in U}\widehat{W}_{k1}^{(B)}(\mathbf{x}_\ell)\left(1-\frac{I_\ell}{\pi_\ell}\right)\right\}, \quad k\in S. \tag{1.26}$$

2. *For all sampling designs, we have $\sum_{k\in S} w_{k1}^{(B)} = N$, for all $s\in\mathcal{S}$.*

3. *We have*

$$w_{k1}^{(B)} = 1/\pi_k$$

for elements which are never selected in subsamples, i.e., $k\in\bigcap_{b=1}^{B} O_b^{(S)}$.

4. *If bootstrap is used as resampling mechanism, the probability that an element is never selected converges to 0 when B increases to infinity.*

5. *The weights $\{w_{k1}^{(B)}\}_{k\in S}$ are independent of the survey variable if and only if the partitioning rule used by the trees in the forest is independent of the survey variable.*

The estimator $\widehat{t}_{rf1}^{(B)}$ is therefore an estimator which can be written as a weighted sum of the values of the survey variable. On the other hand, the weights $\{w_{k1}^{(B)}\}_{k\in S}$ might be dependent of the survey variable if the partitioning rule is itself dependent of the survey variable (which is often the case, in practice). Therefore, the application of this weighting system to other survey variables must be done cautiously. On the other hand, when the splitting mechanism does not depend on the variable of interest, the estimator $\widehat{t}_{rf1}^{(B)}$ therefore belongs to the class of linear estimators (i.e., it can be written as a weighted sum of the measurements of Y

with weights independent of Y). The previous proposition also reveals that the sum of the weights is always equal to the size of the population and that it is possible that some of these weights are in fact equal to the initial survey weights. However, for large forests, this phenomenon occurs only very rarely. Even when this scenario occurs, it should be noted that these elements are still used in the construction of the estimator: they contribute to the correction term in the form 2. Finally, we can note that, when no resampling mechanism is used in the algorithm, then the weights (1.26) are always positive, which is an attractive property.

Asymptotic properties the random forest estimator

The results that we describe below require certain regularity assumptions concerning the survey design, the survey variable and the forest algorithm on which the estimator is built, see Dagdoug et al. (2021b) for more details. Most of these assumptions are commonly used in the literature and verified in practice, see for example Breidt and Opsomer (2000) and McConville and Toth (2019) for more details. In particular, in Chapter 4, we restrict our work to the case of random forests where the resampling mechanism used is sampling *without replacement*, a slight modification from the original algorithm of Breiman.

Result 1.2.4. *There exists constants $C_1 > 0$ and $C_2 > 0$ such that:*

$$\mathbb{E}_p\left|\frac{1}{N_v}\left(\widehat{t}_{rf1}^{(B)} - t_y\right)\right| \leqslant \frac{C_1}{\sqrt{N_v}} + \frac{C_2}{n_{0v}}, \quad a.s. \tag{1.27}$$

where n_{0v} is the minimal number of elements allowed per terminal node in each tree of the forest.

It is therefore possible to bound the L^1 error of each estimator in the class $\mathcal{T}_{rf}^*(D_{ma})$. Moreover, if n_{0v} tends to infinity, then this bound decreases to 0. Consequently, in this case, the estimators of $\mathcal{T}_{rf}^*(D_{ma})$ are asymptotically unbiased and consistent for t_y. In the rest of the section, we will therefore assume that n_{0v} tends to infinity as n tends to infinity. To get some of the following results, we will actually need to consider n_{0v} such that $\sqrt{n_v}/n_{0v}$ converges to 0.

The following equivalence allows us to guide our suggestion regarding the asymptotic variance of the forest estimator and to determine its asymptotic distribution.

Result 1.2.5. *The estimator $\widehat{t}_{rf1}^{(B)}$ is equivalent to the generalized difference estimator $\widehat{t}_{dif}^{(B)}$:*

$$\frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{rf1}^{(B)} - t_y\right) = \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{dif}^{(B)} - t_y\right) + o_{\mathbb{P}}(1),$$

where $\widehat{t}_{dif}^{(B)}$ is given in (1.20).

This result allows us to deduce the asymptotic variance of $\widehat{t}_{rf1}^{(B)}$, equal to

$$\mathbb{AV}_p\left(\frac{1}{N_v}\widehat{t}_{rf1}^{(B)}\right) = \frac{1}{N_v^2}\sum_{k\in U_v}\sum_{\ell\in U_v}(\pi_{kl} - \pi_k\pi_\ell)\frac{y_k - \widetilde{m}_{rf}^{(B)}(\mathbf{x}_k)}{\pi_k}\frac{y_\ell - \widetilde{m}_{rf}^{(B)}(\mathbf{x}_\ell)}{\pi_\ell}. \tag{1.28}$$

In practice, this variance cannot be calculated and we therefore propose to estimate it by

$$\widehat{V}_{rf1}^{(B)} = \frac{1}{N_v^2} \sum_{k \in S_v} \sum_{\ell \in S_v} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}} \frac{y_k - \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k)}{\pi_k} \frac{y_\ell - \widehat{m}_{rf1}^{(B)}(\mathbf{x}_\ell)}{\pi_\ell}. \tag{1.29}$$

This is a consistent and asymptotically unbiased estimator for the asymptotic variance of $\widehat{t}_{rf1}^{(B)}$, as guaranteed by the following result.

Result 1.2.6. *The variance estimator $\widehat{V}_{rf1}^{(B)}$ is convergent for $\mathbb{AV}_p\left(N_v^{-1}\widehat{t}_{rf1}^{(B)}\right)$, i.e. ,*

$$\lim_{v \to \infty} \mathbb{E}_p\left(\frac{n_v}{N_v^2}\left|\widehat{V}_{rf1}^{(B)} - \mathbb{AV}_p\left(N_v^{-1}\widehat{t}_{rf1}^{(B)}\right)\right|\right) = 0.$$

In order to be able to determine asymptotic confidence intervals, it is necessary to determine the asymptotic distribution of the proposed estimator which is obtained under the additional assumption that the generalized difference estimator $\widehat{t}_{dif}^{(B)}$ follows a normal distribution.

Result 1.2.7. *Assume that*

$$\frac{N_v^{-1}\left(\widehat{t}_{dif}^{(B)} - t_y\right)}{\sqrt{\mathbb{V}_p\left(N_v^{-1}\widehat{t}_{dif}^{(B)}\right)}} \xrightarrow[v \to \infty]{\mathcal{L}} \mathcal{N}(0,1),$$

then

$$\frac{N_v^{-1}\left(\widehat{t}_{rf1}^{(B)} - t_y\right)}{\sqrt{\widehat{V}_{rf1}^{(B)}}} \xrightarrow[v \to \infty]{\mathcal{L}} \mathcal{N}(0,1).$$

As illustrated in Example 1.2.3, stochastic predictors have an additional source of variation, introduced by the randomization variables. The variance estimator (1.29) does not take into account these additional variations (induced by Θ, see 1.2.3). However, it is possible to show that there is a positive constant C such that

$$\mathbb{V}_\Theta\left(\frac{\widehat{t}_{rf1}^{(B)}}{N_v}\right) \leqslant \frac{C}{B}.$$

Therefore, if B is chosen large enough, then the variations coming from the introduced randomization are negligible and do not need to be estimated.

Simulations and empirical investigations

In Dagdoug et al. (2021b), several simulation studies have been conducted. Simulations were performed in a large variety of scenarios in order to investigates the empirical performances of the random forest point estimator $\widehat{t}_{rf1}^{(B)}$. In most scenarios, the estimator was relatively efficient: when the relationship was linear in the covariates, $\widehat{t}_{rf1}^{(B)}$ was slightly less efficient than the GREG estimator, yet remained efficient and improved on the Horvitz-Thompson estimator. When the relationship was not linear, however, important improvements were obtained with

$\widehat{t}_{rf1}^{(B)}$ over the GREG estimator. Our simulations also included a few high-dimensional scenarios, in which $\widehat{t}_{rf1}^{(B)}$ was the only estimator included in the simulations to remain more efficient than the Horvitz-Thompson estimator.

We have also investigated the performances of the variance estimator $\widehat{V}_{rf1}^{(B)}$ defined in (1.29). We noted, in accordance with our theoretical results, that the estimator $\widehat{V}_{rf1}^{(B)}$ is nearly unbiased and efficient when large sample and population sizes were used and when the minimal number of elements in each node n_0 is large enough. When this assumption was not satisfied, however, the estimator suffered from an important negative bias leading to an undercoverage for the confidence intervals. The problem detected here is in fact more general and may happen to all model-assisted estimators based on predictors capable of interpolating the data (i.e., flexible enough). We suggested a cross-validated type variance estimator, which estimates efficiently the variance of the RF estimator, independently of the choice of n_0, see Chapter 6 for more about this procedure.

Lastly, we conducted a simulation study investigating the influence of the main hyper-parameters of a random forest algorithm on the efficiency of the resulting model-assisted estimators. We found that the number of trees in the forest can be chosen arbitrarily large without risk of overfitting; that the default number of variables considered for the splitting process $p_0 = \sqrt{p}$ led to satisfactory results in most cases; and that the most influent parameter to be chosen was n_0, which should be chosen neither too small nor too large. The cross-validated variance estimator discussed in the previous paragraph can also be used for guiding our choice of parameters. A more elaborate discussion is provided in the Conclusion chapter of this dissertation.

1.3 Item nonresponse and imputation in surveys

1.3.1 Basic framework and imputed estimators

In the previous sections, we made the assumption of full response, meaning that every information sought for the sampled elements could be collected. In practice, however, nonresponse happens in practically almost every survey. Chen and Haziza (2019) explains, "every time data are collected, it is virtually certain that one will face the problem of missing values." Throughout this work, two assumptions are made: 1) each element of the population has a strictly positive probability of response; 2) the response of an element is independent from the response of other elements. It is clear that in some cases, these assumptions may not be satisfied. For instance, it is possible that some elements may never want to respond (Kott, 1994). However, these assumptions seem to be satisfied in most cases and are needed in order to have efficient nonresponse treatments.

Interestingly, nonresponse is an undesirable phenomenon which can be modeled as a probability sampling design with unknown inclusion probabilities. Indeed, under the assumption that the response of each element is not affected by the response of other elements, nonresponse is essentially a Poisson sampling design with unknown parameters $\{p_k\}_{k\in U}$ where p_k denotes the probability that unit k responds. To be more precise, let U_r be a random variable taking value in the design space $(\mathcal{S}, \mathcal{D})$ with distribution $\mathbb{P}_r$ called a *nonresponse mechanism*. In the literature, nonresponse mechanisms are classified in three different categories (Rubin, 1976): missing completely at random (MCAR), missing at random

(MAR) and missing not at random (MNAR). In the first category, it is assumed that the random variable U_r is independent of the survey variable Y; in the second scenario, it is assumed that, conditionally on the covariates $\boldsymbol{X}_U$, the random variable U_r is independent of Y; lastly, MNAR is used for distributions satisfying neither MCAR nor MAR. In this work, we restrict our investigations to the case of MAR.

Proposition 1.3.1. *(Missing at random nonresponse mechanism.)*
Consider the design space $(\mathcal{S}, \mathcal{D})$. A nonresponse mechanism $\mathbb{P}_r$ satisfying the missing at random assumption is a function $\mathbb{P}_r : \mathcal{D} \times \mathbb{R}^{N\times p} \to [0,1]$, such that

i) for all $A \in \mathcal{D}$, the map $\boldsymbol{X}_U \mapsto \mathbb{P}_r(A, \boldsymbol{X}_U)$ is measurable.

ii) for all $\boldsymbol{X}_U \in \mathbb{R}^{N\times q}$, the map $A \mapsto \mathbb{P}_r(A, \boldsymbol{X}_U)$ is a probability measure on $(\mathcal{S}, \mathcal{D})$.

iii) The map

$$p_r := \begin{cases} \mathcal{S} \longrightarrow [0;1], \\ s \longmapsto \mathbb{P}_r(\{s\}). \end{cases}$$

is a Poisson sampling design.

Note that, given the covariates, S and U_r are independent. In the imputation literature, it is common practice to work with the joint distribution induced by the sampling design, the superpopulation model and the nonresponse mechanism. The product space considered is the following.

Definition 1.3.1 (Joint model-design distribution with auxiliary information)**.**
Consider the superpopulation probability space $(\Omega, \mathcal{M}, \mathbb{P}_m)$, the sampling design space $(\mathcal{S}, \mathcal{D}, \mathbb{P}_p)$ and the response probability space $(\mathcal{S}, \mathcal{D}, \mathbb{P}_r)$. To each $\omega \in \Omega$, we define a product measure $\mathbb{P}_{p,r}(A \times B, \omega) := \mathbb{P}_p(A, \omega)\,\mathbb{P}_r(B, \boldsymbol{X}_U(\omega))$ for all measurable rectangles $A \times B \in \mathcal{D} \otimes \mathcal{D}$. The joint model-design-nonresponse probability induced by a nonresponse mechanism satisfying the MAR assumption is defined as the probability $\mathbb{P}_{m,p,r}$ on the measurable space $(\Omega \times \mathcal{S} \times \mathcal{S},\ \mathcal{M} \otimes \mathcal{D} \otimes \mathcal{D})$, uniquely defined as

$$\mathbb{P}_{m,p,r}(A \times B \times C) := \int_A \mathbb{P}_{p,r}(B \times C, \omega)\, d\mathbb{P}_m(\omega)$$

for all measurable rectangles $A \times B \times C \in \mathcal{M} \otimes \mathcal{D} \otimes \mathcal{D}$.

Each realization of U_r defines a subset of elements of U, which would be *respondents* if they were selected in the sample. Thus, the observed elements are given by the realizations of the random variable $S_r := S \cap U_r$, of size n_r. The set of sampled nonrespondents is denoted by $S_m := S \backslash S_r$ with size n_m. As before, since each element of $\mathcal{S}$ can be represented as a vector in $\{0;1\}^N$, we define the random vector $[r_1, r_2, ..., r_N]^\top$ in the product space, where $r_k = 1$ if $k \in U_r$ and 0 otherwise. By construction, $\{r_k\}_{k\in U}$ is a set of independent Bernoulli random variables $\{\mathcal{B}(p_k)\}_{k\in U}$, where the set of *response probabilities* $\{p_k\}_{k\in U}$ is unknown, and may depend on the covariates under MAR.

In presence of nonresponse, the Horvitz-Thompson cannot be used as it is based on unknown values. Indeed, we have the decomposition

$$\widehat{t_\pi} = \sum_{k\in S_r} \frac{y_k}{\pi_k} + \sum_{k\in S_m} \frac{y_k}{\pi_k}, \tag{1.30}$$

where the second term of (1.30) is unknown. Instead, it is common to use an imputed estimator, defined as

$$\widehat{t}_{imp} = \sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_m} \frac{y_k^*}{\pi_k}, \tag{1.31}$$

where y_k^* is a proxy value used to replace missing y_k, called an *imputed value*. If we assume that the set of covariates $\{\mathbf{x}_k\}_{k \in S}$ is fully observed for all sampled elements, then it is customary to use a predictor $\widehat{m}$, fitted on $D_{n_r} := \{(\mathbf{x}_k, y_k)\,; k \in S_r\}$ to define the imputed values $y_k^* := \widehat{m}(\mathbf{x}_k)$. In that case, the imputed estimator (1.31) is given by

$$\widehat{t}_{\widehat{m}} = \sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_m} \frac{\widehat{m}(\mathbf{x}_k)}{\pi_k}. \tag{1.32}$$

In the article Dagdoug et al. (2022b), to establish our results, we laid out a set of conditions on $\widehat{m}$ under which the imputed estimator $\widehat{t}_{\widehat{m}}$ converges in L^2 with respect to the joint distribution. The conditions that we found on $\widehat{m}$ reveal that if the predicted values are based on a predictor $\widehat{m}$ consistent (in L^2) for the regression function m, and whose error is uniformly bounded, then $\widehat{t}_{\widehat{m}}$ converges in L^2 towards t_y.

For this result, we assumed, in addition to the conditions on $\widehat{m}$, similar conditions such as those described in Breidt and Opsomer (2000) for the consistency of the Horvitz-Thompson, see Chapter 5.

Result 1.3.1. *Consider a sequence of predictors $\{\widehat{m}\}$ fitted on D_{n_r} and its population counterparts $\{\widetilde{m}\}$ fitted on $D_N := \{(\mathbf{x}_k, y_k)\,; k \in U\}$. If*

i) The sequence of population predictors $\{\widetilde{m}\}$ satisfies

$$\lim_{v \to \infty} \mathbb{E}\left[\left(\widetilde{m}(\mathbf{x}) - m(\mathbf{x})\right)^2\right] = 0,$$

with a convergence rate denoted γ_v.

ii) There exists a positive constant C, independent of v, such that

$$\mathbb{E}\left\{\left(\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right)^2 \,\middle|\, \boldsymbol{r}, \boldsymbol{X}, \boldsymbol{I}\right\} \leqslant C. \qquad a.s.$$

Then, the sequence of imputed estimators $\{\widehat{t}_{\widehat{m}}\}$ is L^2-consistent with rate

$$\mathbb{E}\left[\left(\frac{1}{N_v}\left\{\widehat{t}_{\widehat{m}} - t_y\right\}\right)^2\right] = O\left(\gamma_v\right).$$

The conditions that we found to ensure the L^2 consistency of the imputed estimator have simple interpretations: i) requires that, for large samples, the predictor $\widehat{m}$ estimates efficiently the true regression function m and ii) requires that, even for small samples, the error of estimation is bounded. Omitting condition ii), the result therefore states that, for an imputed estimator to be L^2 consistent, it is enough that it is based on a consistent prediction method. In other words,

in that scenario, the problem of imputation is not harder to solve than the problem of regression.

The theoretical properties of imputed estimators were investigated for the nearest neighbor predictor (Chen and Shao, 2000, 2001, Yang and Kim, 2019), the score method (Haziza and Beaumont, 2007, Little, 1986), predictive mean matching (Yang and Kim, 2017), kernel regression (Zhong and Chen, 2014), to cite just a few. For more information about the missing data literature in surveys, see Chen and Haziza (2019) or Haziza (2009).

In this dissertation, two articles were focused on imputed estimators. First, the article Dagdoug et al. (2021a) entitled *Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison*, for which the main results are presented next. This article investigates the empirical performances of imputed estimators based on machine learning procedures, in a wide variety of scenarios. The complete article is presented in Chapter 4. Next, the article Dagdoug et al. (2022b) entitled *Regression tree and random forest imputation in surveys with application to data integration* is focused on the analysis of imputed estimators based on regression trees and random forests. The finite sample properties of the resulting estimators are thoroughly investigated and their L^2-consistency towards t_y is established.

1.3.2 Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison

In the article Dagdoug et al. (2021a), we conducted an extensive simulation study to compare several nonparametric and machine learning imputation procedures in terms of bias and efficiency. The imputation procedures were evaluated in the case of finite population totals of continuous and binary variables and for population quantiles under both simple random sampling without replacement and proportional-to-size Poisson sampling. The Cubist algorithm, BART and XGBoost performed very well in a wide variety of settings. In general, these methods seem to be highly robust to model misspecification and seem to have the ability to capture nonlinear trends in the data. Additive models based on B-splines performed well in the case of population totals when the number of explanatory variables was small but broke down for large values of p. Finally, random forests performed relatively well in a high-dimensional setting. In practice, the choice of an imputation procedure is not clear-cut and depends on the data at hand. If one is reasonably confident about the correct specification of the first moment of the imputation model (that includes the correct specification of the functional form and the correct specification of the vector of explanatory variables), parametric imputation procedures are expected to do well in terms of bias and efficiency. In addition, parametric imputation is simpler to understand and the results are easier to interpret, in general. In the case of complex/nonlinear relationships and/or in a high-dimensional setting, our empirical investigations suggest that machine learning procedures outperform traditional imputation procedures as they tend to be robust against model misspecification. However, these procedures require the specification of some regularization parameters. For instance, for XGBoost, one must specify the learning rate, the maximal depth and the coefficient of penalization. In support vector regression, the cost function and the kernel function must be selected, among others. In practice, the value for some of these parameters are determined through a cross-validation procedure. More details can be found in Chapter 5.

1.3.3 Regression tree and random forest imputation in surveys with application to data integration

Let $\widehat{m}_{rf2}^{(B)}$ be an estimator of m obtained according to a random forest (with any partitioning rule), unweighted, built on $\{(\mathbf{x}_k, y_k); k \in S_r\}$, that is,

$$\widehat{m}_{rf2}^{(B)}(\mathbf{x}) = \sum_{k \in S_r} \widehat{W}_{k2}^{(B)}(\mathbf{x}) y_k,$$

where

$$\widehat{W}_{k2}^{(B)}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \frac{\psi_k^{(b)} \mathbb{1}_{\mathbf{x}_k \in \widehat{A}(\mathbf{x})}}{\sum_{\ell \in S_r} \psi_\ell^{(b)} \mathbb{1}_{\mathbf{x}_\ell \in \widehat{A}(\mathbf{x})}}, \qquad k \in S_r.$$

The class of estimators imputed by random forests $\widehat{t}_{rf2}^{(B)}$ is defined by the set of elements of the form

$$\widehat{t}_{rf2}^{(B)} := \sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_m} \frac{\widehat{m}_{rf2}^{(B)}(\mathbf{x}_k)}{\pi_k}. \tag{1.33}$$

Finite sample properties of the random forest imputed estimator

Proposition 1.3.2. *Consider a random forest imputed estimator $\widehat{t}_{rf2}^{(B)}$.*

1. *The estimator $\widehat{t}_{rf2}^{(B)}$ can be seen as an average of imputed estimators:*

$$\widehat{t}_{rf2}^{(B)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{t}_{tree2}^{(b)},$$

   where $\widehat{t}_{tree2}^{(b)}$ denotes the imputed estimator based on the b-th tree in the forest.

2. *If the sampling design has equal inclusion probability, then the estimator $\widehat{t}_{rf2}^{(B)}$ can be written*

$$\widehat{t}_{rf2}^{(B)} = \sum_{k \in S} \frac{\widehat{m}_{rf2}^{(B)}(\mathbf{x}_k)}{\pi_k} + \frac{1}{B} \sum_{b=1}^{B} \sum_{k \in O_b(S_r)} \frac{y_k - \widehat{m}_{tree2}^{(b)}(\mathbf{x}_k)}{\pi_k}, \tag{1.34}$$

   where $O_b(S_r) := S_r - S_r(\Theta_b)$ and $\widehat{m}_{tree2}^{(b)}$ denotes the b-th tree in forest $\widehat{m}_{rf2}^{(B)}$.

3. *If the sampling design has equal inclusion probability and if $\widehat{m}_{rf2}^{(B)}$ does not use a resampling mechanism, then $\widehat{t}_{rf2}^{(B)}$ has the projection property:*

$$\widehat{t}_{rf2}^{(B)} = \sum_{k \in S} \frac{\widehat{m}_{rf2}^{(B)}(\mathbf{x}_k)}{\pi_k}.$$

Proposition 1.3.3. *Consider a random forest imputed estimator $\widehat{t}_{rf2}^{(B)}$.*

1. *The* $\widehat{t}_{rf2}^{(B)}$ *estimator can be written as a weighted sum of* $\{y_k\}_{k\in S_r}$*:*

$$\widehat{t}_{rf2}^{(B)} = \sum_{k\in S_r} w_{k2}^{(B)} y_k,$$

where

$$w_{k2}^{(B)} = \frac{1}{\pi_k} + \sum_{\ell\in S_m} \frac{\widehat{W}_{k2}^{(B)}(\mathbf{x}_\ell)}{\pi_\ell} = \frac{1}{\pi_k} + \frac{1}{B}\sum_{b=1}^{B} \psi_k^{(b)} \frac{\widehat{N}_b(\mathbf{x}_k, S_m)}{N_b(\mathbf{x}_k, S_r(\Theta_b))}, \qquad k \in S_r, \tag{1.35}$$

where $\widehat{N}_b(\mathbf{x}_k, S_m)$ *denotes the weighted sum of elements of* S_m *belonging to the node containing* $\mathbf{x}_k$ *and* $N_b(\mathbf{x}_k, S_r(\Theta_b))$ *denotes the number of elements of* $S_r(\Theta_b)$ *belonging to the node containing the point* $\mathbf{x}_k$*.*

2. *In the case of a deterministic tree, if the sampling design has equal inclusion probability, then we have*

$$w_{k2}^{(B)} = \frac{1}{\pi_k} \times \left(1 + \frac{1}{B}\sum_{b=1}^{B} \psi_k^{(b)} \frac{N(\mathbf{x}_k, S_m)}{N(\mathbf{x}_k, S_r(\Theta_b))}\right).$$

where $N(\mathbf{x}_k, S_m)$ *denotes the number of missing elements in the node containing* $\mathbf{x}_k$*.*

3. *If the initial weights are calibrated to the population size* $\sum_{k\in S} \frac{1}{\pi_k} = N$*, then* $\sum_{k\in S} w_{k2}^{(B)} = N$*.*

4. *We have*

$$w_{k2}^{(B)} = \frac{1}{\pi_k}$$

for elements $k \in \bigcap_{b=1}^{B} O_b(S_r)$*.*

5. *If there are at least* n_0 *elements in the leaves of each tree, then the weights are bounded as follows*

$$d_k \leqslant w_{k2}^{(B)} \leqslant d_k \left(1 + \frac{n_m}{n_0}\right), \quad a.s. \qquad k \in S_r. \tag{1.36}$$

These bounds can be reached.

6. *The weights* $\{w_{k2}^{(B)}\}_{k\in S_r}$ *are independent of the survey variable if and only if the partitioning rule used by the trees in the forest is independent of the survey variable.*

As with the model-assisted estimator, the forest imputed estimator can be written as a weighted sum of the values of the variable of interest. In the case of the imputed estimator, these weights reveal a lot of information about the behavior of the estimator. First of all, we observe that the imputation weights are always greater than or equal than the initial weights and are calibrated on the sum of the initial weights. If we consider the weights of an estimator based on a deterministic tree (e.g., CART, scoring method) with equal inclusion probabilities, we have

$$w_{k2}^{(1)} = d_k \times \left(1 + \frac{N(\mathbf{x}_k, S_m)}{N(\mathbf{x}_k, S_r)}\right) = d_k \times \left\{1 + R_{mr}(\mathbf{x}_k)\right\}, \qquad k \in S_r.$$

We therefore observe that, if most of the elements with characteristics similar to an element $k \in S_r$ have not responded, then a significant weight will be attributed to element k because the ratio $R_{mr}(\mathbf{x}_k)$ will be important. Otherwise, if almost all the elements with characteristics similar to the individual $k \in S_r$ have responded, the imputation weights will be very close to the initial weights. This is a desired behavior: when we have only a few elements similar to many other, the imputation weights of these elements need to be large; otherwise, if most elements of a given category are observed, it is enough to let the imputation weights of these remain close to its original weight. In particular, a responding element has an imputation weight equal to its initial weight if and only if all the elements in its leaf are respondents. The same interpretation is valid in the case of unequal inclusion probabilities. On the other hand, if we consider stochastic trees and random forests, these properties are lost for forests with only a few trees (i.e., low B). Indeed, for the elements that have not been selected (there may be many of them for low B), the imputation weights are equal to the initial weights. Regardless of that, the sum of the imputation weights remains equal to the sum of the initial weights: a compensation phenomenon is therefore necessarily introduced and these weights can be relatively unstable. When B is large, this instability disappears and the behavior of the weights of the forest is very close to the behavior of the weights of a tree.

Asymptotic properties of the random forest imputed estimator

To study the asymptotic properties of estimators imputed by forests, it is useful to consider the infinite predictor $\widehat{m}_{rf2}^{(\infty)}$ defined by

$$\widehat{m}_{rf2}^{(\infty)} := \lim_{B\to\infty} \widehat{m}_{rf2}^{(B)} = \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} \widehat{m}_{tree2}^{(b)}$$

as well as the infinite estimator

$$\widehat{t}_{rf2}^{(\infty)} := \sum_{k\in S_r} \frac{y_k}{\pi_k} + \sum_{k\in S_m} \frac{\widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k} = \mathbb{E}_\Theta\left[\widehat{t}_{rf2}^{(B)}\right]. \tag{1.37}$$

This estimator is of purely academic interest since, in practice, the number of trees that we can use is always finite. However, this infinite estimator is particularly interesting for three reasons: 1) it is simpler to study theoretically than the finite forest estimator; 2) it is more efficient than the finite forest estimator, as the proposition below reveals; 3) it is approachable to a given precision by the finite forest estimator.

Proposition 1.3.4. *There exists $C > 0$ such that*

$$0 \leqslant \mathbb{E}\left[\left(\frac{\widehat{t}_{rf2}^{(B)} - t_y}{N_v}\right)^2\right] - \mathbb{E}\left[\left(\frac{\widehat{t}_{rf2}^{(\infty)} - t_y}{N_v}\right)^2\right] \leqslant \frac{C}{B}. \tag{1.38}$$

Moreover, for all $\epsilon > 0$,

$$\mathbb{P}_\Theta\left(|\widehat{t}_{rf2}^{(B)} - \widehat{t}_{rf2}^{(\infty)}| > \epsilon\right) \leqslant 2\exp\left(\frac{-B\epsilon^2}{2n_m^2\left(\frac{C_{2,Y}-C_{1,Y}}{\min_{k\in U}\pi_k}\right)^2}\right).$$

This proposition therefore shows that it seems to be interesting to build large forests, in the sense that the infinite forest estimator is more efficient than the finite forest estimator. We now restrict our analysis to the case where, for the imputation in S_v, we choose B_v such that, if $v_1 < v_2$ then the number of trees used to impute, B_{v_1}, is strictly less than the number of trees used to impute, B_{v_2}. This allows the use of the result (1.38) implying, as soon as the estimator of infinite forests is consistent in L^2, the consistency L^2 of the estimator of finite forests. Moreover, we restrict ourselves to the case where $\widehat{t}_{rf2}^{(B)}$ is an imputed estimator based on the random forest algorithm in Breiman's original sense. More details about the assumptions for this result are given in Chapter 5.

Result 1.3.2. *The estimator* $\widehat{t}_{rf2}^{(B)}$ *converges in* L^2 *with respect to the joint distribution, that is,*

$$\lim_{v\to\infty}\mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{rf2}^{(B)} - t_y\right)\right)^2\right] = 0.$$

Finally, concerning the estimation of the variance, similarly as for model-assisted estimators, we show that the variations due to the randomization variables decrease when B increases:

$$\mathbb{V}_\Theta\left(\frac{\widehat{t}_{rf2}^{(B)}}{N_v}\right) \leqslant \frac{C}{B_v}.$$

It is therefore sufficient to estimate the variance of $\widehat{t}_{rf2}^{(B)}$ with respect to the joint distribution induced by the design, the model and the nonresponse mechanism. In most cases, the "naive" variance estimator

$$\widehat{V}_{naive} := \sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}\frac{r_k y_k + (1-r_k)y_k^*}{\pi_k}\frac{r_\ell y_\ell + (1-r_\ell)y_\ell^*}{\pi_\ell} \tag{1.39}$$

is a severely biased estimator; the use of specific variance estimators is therefore necessary. Two approaches are traditionally used: the "two-phase" approach (Särndal, 1992), and the "reverse" approach (Shao and Steel, 1999). The interested reader can refer to Haziza and Vallée (2020) for a review of concepts and tools related to variance estimation of imputed estimators.

We suggested two corresponding estimators. For the two-phase approach,

$$\widehat{V}_{sar} := \widehat{V}_{sam} + \widehat{V}_{nr} + 2\widehat{V}_{mix}, \tag{1.40}$$

where

$$\widehat{V}_{sam} := \widehat{V}_{naive} + \sum_{k\in S_m} d_k^2(1-\pi_k)\widehat{\sigma}^2 \qquad \widehat{V}_{nr} := \sum_{k\in S}\gamma_k^2\widehat{\sigma}^2, \qquad \widehat{V}_{mix} := \sum_{k\in S}\gamma_k(d_k-1)\widehat{\sigma}^2,$$

with $\widehat{\sigma}$ is an estimator of the variance of the model residuals and $\gamma_k := r_k w_{k2}^{(B)} - d_k$ for $k \in S$. Finally, for the reverse approach, assuming that the sampling fraction n_v/N_v is negligible, we suggested the following variance estimator:

$$\widehat{V}_{rev} := \sum_{k\in S}\sum_{\ell\in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\xi}_k^{(B)}}{\pi_k} \frac{\widehat{\xi}_\ell^{(B)}}{\pi_\ell}, \tag{1.41}$$

where

$$\widehat{\xi}_k^{(B)} := \widehat{m}_{rf2}^{(B)}(\mathbf{x}_k) + r_k \cdot \frac{1}{B}\sum_{b=1}^{B} \frac{\widehat{N}_b(\mathbf{x}_k, S)}{\widehat{N}_b(\mathbf{x}_k, S_r)} \cdot \left(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k)\right), \qquad k \in S.$$

We also performed several empirical studies aiming at assessing the performances of point and variance estimators. The simulations suggest that point estimators behave well both in terms of bias and efficiency. Several variance estimators were used in the simulations, including the estimators defined in (1.40) and (1.41). The results suggest that the estimators perform relatively well in terms of bias and coverage rate. An application to mass imputation is also included in this work; see Chapter 5 for details.

2 MODEL-ASSISTED ESTIMATION IN HIGH-DIMENSIONAL SETTINGS FOR SURVEY DATA

Abstract.[1] Model-assisted estimators have attracted a lot of attention in the last three decades. These estimators attempt to make an efficient use of auxiliary information available at the estimation stage. A working model linking the survey variable to the auxiliary variables is specified and fitted on the sample data to obtain a set of predictions, which are then incorporated in the estimation procedures. A nice feature of model-assisted procedures is that they maintain important design properties such as consistency and asymptotic unbiasedness irrespective of whether or not the working model is correctly specified. In this article, we examine several model-assisted estimators from a design-based point of view and in a high-dimensional setting, including linear regression and penalized estimators. We conduct an extensive simulation study using data from the Irish Commission for Energy Regulation Smart Metering Project, in order to assess the performance of several model-assisted estimators in terms of bias and efficiency in this high-dimensional data set.

Keywords: Design consistency; Elastic net; Lasso; Random Forest; Ridge regression; XGBoost.

2.1 Introduction

Surveys conducted by national Statistical Offices (NSO) aim at estimating finite population parameters, which are those describing some aspects of the finite population under study. In this article, the interest lies in estimating the population total of a survey variable Y. Population totals can be estimated unbiasedly using the well-known Horvitz-Thompson estimator (Horvitz and Thompson, 1952). In the absence of nonsampling errors, the Horvitz-Thompson estimator is unbiased with respect to the customary design-based inferential approach, whereby the properties of estimators are evaluated with respect to the sampling design; e.g., see Särndal et al. (1992). However, Horvitz-Thompson type estimators may exhibit a large variance in some situations. The efficiency of the Horvitz-Thompson estimator can be improved by incorporating some auxiliary information, capitalizing on the relationship between the survey variable Y and a set of auxiliary variables $\mathbf{x}$. The resulting estimation procedures, referred to as model-assisted estimation procedures, use a working model as a vehicle for constructing point estimators. Model-assisted estimators remain design-consistent even if the working model is misspecified, which is a desirable feature. When the working model provides an adequate description of the relationship between Y and $\mathbf{x}$, model-assisted estimators are expected to be more efficient than the Horvitz-Thompson estimator.

The class of model-assisted estimators include a wide variety of procedures, some of which have been extensively studied in the literature both theoretically and empirically. When the working model is the customary linear regression model, the resulting estimator is the well-known generalized regression estimator (GREG); e.g., Särndal (1980), Särndal and Wright (1984) and Särndal et al. (1992). Other works include model-assisted procedures based

[1] The article is accepted for publication in Journal of Applied Statistics, in Special Issue: Statistical Approaches for Big Data and Machine Learning.

on generalized linear models (Firth and Bennett, 1998, Lehtonen and Veijanen, 1998), local polynomial regression (Breidt and Opsomer, 2000), splines (Breidt et al., 2005, Goga, 2005, Goga and Ruiz-Gazen, 2014, McConville and Breidt, 2013), neural nets (Montanari and Ranalli, 2005), generalized additive models (Opsomer et al., 2007), nonparametric additive models (Wang and Wang, 2011), regression trees (McConville and Toth, 2019, Toth and Eltinge, 2011) and random forests (Dagdoug et al., 2021b).

Due to the recent advances of information technology, NSOs have now access to a variety of data sources, some of which may exhibit a large number of observations on a large number of variables. So far, the properties of model-assisted estimator have been established under the customary asymptotic framework in finite population sampling (Isaki and Fuller, 1982) for which both the population size N and the sample size n increase to infinity, while assuming that the number of auxiliary variables p is fixed. In other words, existing results require n to be large relative to p. This framework is generally not adequate in the context of high-dimensional data sets as p may be of the same order as n, or even larger, i.e., $p > n$. A more appropriate asymptotic framework would let p increase to infinity in addition to N and n. Cardot et al. (2017) studied dimension reduction through principal component analysis and established the design consistency of the resulting calibration estimator. More recently, Ta et al. (2020) investigated the properties of the GREG estimator from a model point of view and when p is allowed to diverge and Chauvet and Goga (2022) studied the asymptotic variance of the calibration estimator when the number p of calibration variables is going to infinity.

The aim of this paper is to give a general consistency result for a class of model-assisted estimators when the number p of auxiliary variables is allowed to grow to infinity. This class of model-assisted estimators includes the GREG estimator as well as model-assisted estimators based on penalization methods such as ridge, lasso and elastic net. The latter methods were proposed to cope with multicolinearity between predictors in a high-dimension setting. Under mild regularity assumptions, we show that these model-assisted estimators are design-consistent provided that p^3/n goes to zero. As we argue in Section 2.3, this rate can be improved if one is willing to make additional assumptions about the rate of convergence of the estimated regression coefficient. In particular, we lay out a set of additional conditions under which the model-assisted ridge estimator is consistent if p/n goes to zero and moreover, is $\sqrt{n}$-consistent if $p = O(n^a)$ with $a \in [0, 1/2)$. Also, provided that the predictors are orthogonal, we show that both the model-assisted lasso and elastic net estimators are consistent provided that p/n goes to zero.

To the best of our knowledge, an empirical comparison of penalized or nonparametric model-assisted estimators in terms of bias and efficiency in a high-dimensional setting is currently lacking. We aim to fill this gap in the article. To assess the performance of several model-assisted estimators in a high-dimensional setting, we conduct a large simulation study using data from the Irish Commission for Energy Regulation Smart Metering Project. The data set consists of electricity consumption recorded every half an hour for a two-year period and for more than 6000 households and businesses, leading to highly correlated data. Due to the high-dimensional feature, model-assisted estimators based on a linear model tend to breakdown and penalized and reduction dimension based estimators may provide good alternatives.

The paper is organized as follows. In Section 2.2, we introduce the theoretical setup. In Section 2.3, we investigate the asymptotic properties of several model-assisted estimators: the GREG estimator as well as estimators based on ridge regression, lasso and elastic net. Section 2.4 contains an empirical comparison to assess the performance of several model-assisted estimators

in terms of bias and efficiency. In our empirical experiments, we included model-assisted estimators based on ridge regression, lasso and elastic net, principal component regression as well as model-assisted estimators based on CART, random forests, XGBoost and CUBIST. We considered three sampling designs: simple random sampling without replacement, stratified simple random sampling without replacement and stratified fixed-size without replacement proportional to size sampling. We make some final remarks in Section 2.5. The technical details, including the proofs of some results, are relegated to the Supplementary Material.

2.2 The setup

Consider a finite population $U = \{1, 2, ..., N\}$ of size N. We are interested in estimating $t_y = \sum_{i \in U} y_i$, the population total of the survey variable Y. We select a sample S from U according to a sampling design $\mathcal{P}(S)$ with first-order and second-order inclusion probabilities $\{\pi_i\}_{i \in U}$ and $\{\pi_{i\ell}\}_{i,\ell \in U}$, respectively. In the absence of nonsampling errors, the Horvitz-Thompson estimator

$$\widehat{t}_\pi = \sum_{i \in S} \frac{y_i}{\pi_i} \tag{2.1}$$

is design-unbiased for t_y provided that $\pi_i > 0$ for all $i \in U$; that is, $\mathbb{E}_{\mathrm{p}}(\widehat{t}_\pi) = t_y$, where $\mathbb{E}_p(\cdot)$ denotes the expectation operator with respect to the sampling design $\mathcal{P}(S)$. In the sequel, unless stated otherwise, the properties of estimators are evaluated with respect to the design-based approach. Under mild conditions (Breidt and Opsomer, 2000, Robinson and Särndal, 1983), it can be shown that the Horvitz-Thompson estimator $\widehat{t}_\pi$ is design-consistent for t_y.

At the estimation stage, we assume that a collection of auxiliary variables, $X_1, X_2, \ldots, X_p$, is recorded for all $i \in S$. Moreover, we assume that the corresponding population totals are available from an external source (e.g., a census or an administrative file). Let $\mathbf{x}_i = \left[\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{ip}\right]^\top$ be the $\mathbf{x}$-vector associated with unit i. Also, we denote by $\boldsymbol{X}_U = (\mathbf{x}_i^\top)_{i \in U}$ the $N \times p$ design matrix and $\boldsymbol{X}_S = (\mathbf{x}_i^\top)_{i \in S}$ its sample counterpart.

Model-assisted estimation starts with postulating the following working model:

$$\xi: \quad y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i \in U, \tag{2.2}$$

where $f(\cdot)$ is an unknown function and the errors ϵ_i are independent random variables such that $\mathbb{E}_\xi[\epsilon_i|\mathbf{x}_i] = 0$ and $\mathbb{V}_\xi(\epsilon_i|\mathbf{x}_i) = \sigma^2$, where σ^2 is an unknown parameter. Although we assume an homoscedastic variance structure, our results can be easily extended to the case of unequal variances of the form $\mathbb{V}_\xi(\epsilon_i|\mathbf{x}_i) = \sigma^2 v(\mathbf{x}_i)$ for some known function $v(\cdot)$.

The unknown function $f(\cdot)$ is estimated by $\widehat{f}(\cdot)$ from the sample data $(\mathbf{x}_i, y_i)_{i \in S}$. The fitted model is then used to construct the model-assisted estimator

$$\widehat{t}_{ma} = \sum_{i \in U} \widehat{f}(\mathbf{x}_i) + \sum_{i \in S} \frac{y_i - \widehat{f}(\mathbf{x}_i)}{\pi_i}, \tag{2.3}$$

where $\widehat{f}(\mathbf{x})$ denotes the prediction at $\mathbf{x}$ under the working model (2.2). Whenever the predictor $\widehat{f}(\cdot)$ is sample dependent, the estimator $\widehat{t}_{ma}$ is design-biased, but can be shown to be asymptotically design-unbiased and design-consistent for a wide class of working models, as the population size N and the sample size n increase.

2.3 Least squares and penalized model-assisted estimators

2.3.1 The GREG estimator

Suppose that the regression function $f(\cdot)$ is approximated by a linear combination of $X_j, j = 1, \ldots, p$. The working model (2.2) reduces to

$$\xi: \quad y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i \in U, \tag{2.4}$$

where $\boldsymbol{\beta} = \left[\beta_1, \ldots, \beta_p\right]^\top \in \mathbb{R}^p$ is a vector of unknown coefficients. Under a hypothetical census, where we observe y_i and $\mathbf{x}_i$ for all $i \in U$, the vector $\boldsymbol{\beta}$ would be estimated by $\widetilde{\boldsymbol{\beta}}$ through the ordinary least square criterion at the population level:

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} ||\boldsymbol{y}_U - \boldsymbol{X}_U\boldsymbol{\beta}||_2^2 = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i\in U}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2, \tag{2.5}$$

where $\boldsymbol{y}_U = (y_i)_{i\in U}$. Provided that the matrix $\boldsymbol{X}_U$ is of full rank, the solution to (2.5) is unique and given by

$$\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}_U^\top\boldsymbol{X}_U\right)^{-1}\boldsymbol{X}_U^\top\boldsymbol{y}_U = \left(\sum_{i\in U}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\sum_{i\in U}\mathbf{x}_i y_i. \tag{2.6}$$

In practice, the vector $\widetilde{\boldsymbol{\beta}}$ in (2.6) cannot be computed as the y-values are recorded for the sample units only. An estimator of $\widetilde{\boldsymbol{\beta}}$, denoted by $\widehat{\boldsymbol{\beta}}$, is obtained from (2.6) by estimating each total separately using the corresponding Horvitz-Thompson estimator. Alternatively, the estimator $\widehat{\boldsymbol{\beta}}$ can be obtained using the following weighted least square criterion at the sample level:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left(\boldsymbol{y}_S - \boldsymbol{X}_S\boldsymbol{\beta}\right)^\top\boldsymbol{\Pi}_S^{-1}\left(\boldsymbol{y}_S - \boldsymbol{X}_S\boldsymbol{\beta}\right)^\top = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\sum_{i\in S}\frac{(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2}{\pi_i}, \tag{2.7}$$

where $\boldsymbol{\Pi}_S = \operatorname{diag}(\pi_i)_{i\in S}$ and $\boldsymbol{y}_S = (y_i)_{i\in S}$. Again, the solution to (2.7) is unique provided that $\boldsymbol{X}_S$ is of full rank and it is given by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_S^\top\boldsymbol{\Pi}_S^{-1}\boldsymbol{X}_S\right)^{-1}\boldsymbol{X}_S^\top\boldsymbol{\Pi}_S^{-1}\boldsymbol{y}_S = \left(\sum_{i\in S}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i}\right)^{-1}\sum_{i\in S}\frac{\mathbf{x}_i y_i}{\pi_i}. \tag{2.8}$$

The prediction of $f(\cdot)$ at $\mathbf{x}$ under the working model (2.4) is $\widehat{f}_{\text{lr}}(\mathbf{x}) = \mathbf{x}^\top\widehat{\boldsymbol{\beta}}$. Plugging $\widehat{f}_{\text{lr}}(\cdot)$ in (2.3) leads to the well-known GREG estimator (Särndal et al., 1992):

$$\begin{aligned}\widehat{t}_{\text{greg}} &= \sum_{i\in U}\widehat{f}_{\text{lr}}(\mathbf{x}_i) + \sum_{i\in S}\frac{y_i - \widehat{f}_{\text{lr}}(\mathbf{x}_i)}{\pi_i} \\ &= \sum_{i\in U}\mathbf{x}_i^\top\widehat{\boldsymbol{\beta}} + \sum_{i\in S}\frac{y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}}{\pi_i}.\end{aligned} \tag{2.9}$$

If the intercept is included in the working model, the GREG estimator reduces to the population total of the fitted values $\widehat{f}_{lr}(\mathbf{x}_i) = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}$; that is, $\widehat{t}_{\text{greg}} = \sum_{i\in U} \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}$. Also, the GREG estimator can be written as a weighted sum of the sample y-values:

$$\widehat{t}_{\text{greg}} = \sum_{i\in S} w_{iS} y_i, \tag{2.10}$$

where

$$w_{iS} = \frac{1}{\pi_i}\left\{1 - \mathbf{x}_i^\top \left(\sum_{i\in S} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i}\right)^{-1} \left(\sum_{i\in S} \frac{\mathbf{x}_i}{\pi_i} - \sum_{i\in U} \mathbf{x}_i\right)\right\}, \quad i \in S.$$

These weights can be also obtained as the solution of a calibration problem (Deville and Särndal, 1992). More specifically, the weights w_{iS} are such that the generalized chi-square distance $\sum_{i\in S}(w_{iS} - \pi_i^{-1})^2/\pi_i^{-1}$ is minimized subject to the calibration constraints $\sum_{i\in S} w_{iS}\mathbf{x}_i = \sum_{i\in U} \mathbf{x}_i$. This attractive feature may not be shared by model-assisted estimators derived under more general working models.

2.3.2 Penalized least square estimators

While model-assisted estimators based on linear regression working models are easy to implement, they tend to breakdown when the number of auxiliary variables p is growing large. Also, when some of the predictors are highly related to each other, a problem known as multicolinearity, the ordinary least square estimator $\widetilde{\boldsymbol{\beta}}$ given by (2.6) may be highly unstable. As noted by Hoerl and Kennard (2000), "the worse the conditioning of $\boldsymbol{X}_U^\top \boldsymbol{X}_U$, the more $\widetilde{\boldsymbol{\beta}}$ can be expected to be too long and the distance from $\widetilde{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}$ will tend to be large". In survey sampling, the effect of multicolinearity on the stability of point estimators has first been studied by Bardsley and Chambers (1984) under the model-based approach. Chambers (1996) and Rao and Singh (1997) studied this problem in the context of calibration. These authors noted that the use of a large number of calibration constraints may lead to highly dispersed calibration weights, potentially resulting in unstable estimators.

In a classical *iid* linear regression setting, penalization procedures such as ridge, lasso or elastic-net can be used to help circumvent some of the difficulties associated with the usual least squares estimator $\widetilde{\boldsymbol{\beta}}$. Let $\widetilde{\boldsymbol{\beta}}_{\text{pen}}$ be an estimator of $\boldsymbol{\beta}$ obtained through the penalized least square criterion at the population level:

$$\widetilde{\boldsymbol{\beta}}_{\text{pen}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i\in U} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \sum_{\ell=1}^{t} \lambda_\ell ||\boldsymbol{\beta}||_{\nu_\ell}^{\gamma_\ell}, \tag{2.11}$$

where λ_ℓ, ν_ℓ and γ_ℓ are positive real numbers, $||\cdot||_\nu$ is a given norm and t is a fixed positive integer representing the number of different norm constraints. The values of ν_ℓ, γ_ℓ and t are typically predetermined. The tuning parameter λ_ℓ controls the strength of the penalty that one wants to impose on the norm of $\boldsymbol{\beta}$. Most often, the value of λ_ℓ is selected through a cross-validation procedure. The coefficients γ_ℓ and ν_ℓ are specific to the penalization method. Hence, they affect the properties of the resulting estimator $\widetilde{\boldsymbol{\beta}}_{\text{pen}}$. Three special cases are considered below.

When $t = 1$, $\gamma_1 = 2$ and $\nu_1 = 2$, $\lambda_1 = \lambda$, the estimator is known as the ridge regression estimator (Hoerl and Kennard, 1970):

$$\widetilde{\boldsymbol{\beta}}_{\text{ridge}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i\in U} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda ||\boldsymbol{\beta}||_2^2,$$

where $||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^p \beta_j^2$ is the usual Euclidean norm of $\boldsymbol{\beta}$. The solution is given explicitly by

$$\widetilde{\boldsymbol{\beta}}_{\text{ridge}} = \left(\boldsymbol{X}_U^\top \boldsymbol{X}_U + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}_U^\top \boldsymbol{y}_U = \left(\sum_{i\in U} \mathbf{x}_i \mathbf{x}_i^\top + \lambda \boldsymbol{I}_p\right)^{-1} \sum_{i\in U} \mathbf{x}_i y_i, \tag{2.12}$$

where $\boldsymbol{I}_p$ denotes the $p \times p$ identity matrix.

When $t = 1$, $\nu_1 = 1$ and $\lambda_1 = \lambda$, the estimator $\widetilde{\boldsymbol{\beta}}_{\text{pen}}$ is known as the lasso estimator (Tibshirani, 1996):

$$\widetilde{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i\in U} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda ||\boldsymbol{\beta}||_1, \tag{2.13}$$

where $||\boldsymbol{\beta}||_1 = \sum_{j=1}^p |\beta_j|$ is the L^1-norm of $\boldsymbol{\beta}$. As for the ridge, the lasso has the effect of shrinking the coefficients but, unlike the ridge, it can set some coefficients β_j to zero. Except when the auxiliary variables are orthogonal, there is no closed-form formula for the lasso estimator $\widetilde{\boldsymbol{\beta}}_{\text{lasso}}$ (Hastie et al., 2011). In survey sampling, McConville et al. (2017) investigated the design-based properties of the lasso model-assisted estimator for fixed p.

The elastic-net estimator, that was suggested by Zou and Hastie (2005), combines two norms: the euclidean norm $||\cdot||_2$ and the L^1 norm, $||\cdot||_1$. If, in (2.11), we set $t = 2$, $\gamma_1 = 1$, $\nu_1 = 1$, $\gamma_2 = 2$, $\nu_2 = 2$, $\lambda_1 = \lambda\alpha$ and $\lambda_1 = \lambda(1-\alpha)$, the resulting estimator is the elastic-net estimator, which can be viewed as a trade-off between the ridge estimator and the lasso estimator, realizing variable selection and regularization simultaneously:

$$\widetilde{\boldsymbol{\beta}}_{\text{en}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i\in U} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \left[\alpha ||\boldsymbol{\beta}||_1 + (1-\alpha)||\boldsymbol{\beta}||_2^2\right],$$

for $\lambda > 0$ and $\alpha \in [0, 1]$ a parameter that is usually chosen using a grid of multiple values of α. The penalized regression estimator $\widetilde{\boldsymbol{\beta}}_{\text{pen}}$ in (2.11) is unknown as the y-values are not observed for the non-sample units. To overcome this issue, we use the following weighted penalized least square criterion at the sample level:

$$\widehat{\boldsymbol{\beta}}_{\text{pen}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i\in S} \frac{1}{\pi_i} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \sum_{\ell=1}^t \lambda_\ell ||\boldsymbol{\beta}||_{\nu_\ell}^{\gamma_\ell}. \tag{2.14}$$

A model-assisted estimator based on a penalized regression procedure is obtained from (2.3) by replacing $\widehat{f}(\mathbf{x})$ with $\widehat{f}_{\text{pen}}(\mathbf{x}) = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}_{\text{pen}}$, leading to

$$\begin{aligned} \widehat{t}_{\text{pen}} &= \sum_{i\in U} \widehat{f}_{\text{pen}}(\mathbf{x}_i) + \sum_{i\in S} \frac{y_i - \widehat{f}_{\text{pen}}(\mathbf{x}_i)}{\pi_i} \\ &= \left(\sum_{i\in U} \mathbf{x}_i^\top\right) \widehat{\boldsymbol{\beta}}_{\text{pen}} + \sum_{i\in S} \frac{y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{\text{pen}}}{\pi_i}, \end{aligned} \tag{2.15}$$

where $\widehat{\boldsymbol{\beta}}_{\text{pen}}$ is a generic notation used to denote the estimated regression coefficient obtained through either lasso, ridge or elastic net. Unlike the GREG estimator, $\widehat{t}_{\text{greg}}$, the penalized model-assisted estimator is sensitive to unit change of the X-variables because $\widehat{\boldsymbol{\beta}}_{\text{pen}}$ is sensitive to unit change. This is why, as in the classical regression setting, standardization of the X-variables is recommended before computing $\widehat{\boldsymbol{\beta}}_{\text{pen}}$. If the intercept is included in the model, then it is usually left un-penalized.

Remark 2.3.1. *In the case of ridge regression, the estimator $\widehat{\boldsymbol{\beta}}_{\text{ridge}}$ is given by*

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\boldsymbol{X}_S^\top \boldsymbol{\Pi}_S^{-1} \boldsymbol{X}_S + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}_S^\top \boldsymbol{\Pi}_S^{-1} \boldsymbol{y}_S = \left(\sum_{i\in S} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i} + \lambda \boldsymbol{I}_p\right)^{-1} \sum_{i\in S} \frac{\mathbf{x}_i y_i}{\pi_i}. \tag{2.16}$$

Using (2.16) in (2.15) leads to the ridge model-assisted estimator $\hat{t}_{\text{ridge}}$ that can be expressed as a weighted sum of sampled y-values, $\hat{t}_{\text{ridge}} = \sum_{i\in S} w_{iS}(\lambda) y_i$ with weights given by

$$w_{iS}(\lambda) = \frac{1}{\pi_i}\left\{1 - \mathbf{x}_i^\top \left(\sum_{i\in S} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\pi_i} + \lambda \boldsymbol{I}_p\right)^{-1} \left(\sum_{i\in S} \frac{\mathbf{x}_i}{\pi_i} - \sum_{i\in U} \mathbf{x}_i\right)\right\}, \quad i \in S.$$

These weights can also be obtained through a penalized calibration problem. It can be shown that they minimize the penalized generalized chi-square distance, $\sum_{i\in S}(w_{iS} - \pi_i^{-1})^2/\pi_i^{-1} + \lambda^{-1}||\sum_{i\in S} w_{iS}\mathbf{x}_i - \sum_{i\in U} \mathbf{x}_i||_2^2$ (Beaumont and Bocci, 2008, Chambers, 1996). If some X-variables are left un-penalized in (2.11), the resulting weights ensure consistency between the survey estimates and their corresponding population totals associated with these variables.

We end this section by noting that the penalized model-assisted estimator $\widehat{t}_{\text{pen}}$ is sensitive to the choice of the penalty parameter λ_ℓ. In the case of ridge regression, Bardsley and Chambers (1984) suggested the ridge trace method for selecting the penalty parameter λ. This method consists of plotting the weights $w_{iS}(\lambda), i \in S$ for values of λ from a pre-determined grid values and to choose the value of λ for which the weights $w_{iS}(\lambda)$ are positive for all $i \in S$ and $\sum_{i\in S} w_{iS}(\lambda)\mathbf{x}_i - \sum_{i\in U} \mathbf{x}_i$ is the smallest difference among all the differences considered for the grid values of λ. Using the fact that the modified penalty $\lambda^* = \lambda/(1+\lambda)$ lies between 0 and 1 and is an increasing function of λ, Beaumont and Bocci (2008) proposed a method based on the bisection algorithm to first determine λ^* and then, λ. Guggemos and Tillé (2010) implemented a Fisher scoring algorithm in order to find the value of λ which maximizes a design-based estimated log-likelihood criterion. In case of the lasso model-assisted estimator, McConville et al. (2017) used a cross-validation procedure to choose the best value of λ. More research is needed to suggest a unified criterion in order to find the best penalty in a sample-based framework. This is beyond the scope of the article. Most of the computer software use a cross-validation criterion to choose the best penalty parameter.

2.3.3 Consistency of the GREG and penalized GREG estimators in a high-dimensional setting

We adopt the asymptotic framework of Isaki and Fuller (1982) and consider an increasing sequence of embedded finite populations $\{U_v\}_{v\in\mathbb{N}}$ of size $\{N_v\}_{v\in\mathbb{N}}$. In each finite population U_v, a sample, of size n_v, is selected according to a sampling design $\mathcal{P}_v(S_v)$ with first-order inclusion

probabilities $\pi_{i,v}$ and second-order inclusion probabilities $\pi_{i\ell,v}$. While the finite populations are considered to be embedded, we do not require this property to hold for the samples $\{S_v\}_{v\in\mathbb{N}}$. This asymptotic framework assumes that v goes to infinity, so that both the finite population sizes $\{N_v\}_{v\in\mathbb{N}}$, the samples sizes $\{n_v\}_{v\in\mathbb{N}}$ and the number of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ go to infinity. To improve readability, we shall use the subscript v only in the quantities U_v, N_v, n_v and p_v; for instance, quantities such as $\pi_{i,v}$ shall be simply denoted by π_i.

The following assumptions are required to establish the consistency of the GREG and penalized GREG estimators in a high-dimensional setting.

(H1) We assume that there exists a positive constant C_1 such that $N_v^{-1}\sum_{i\in U_v} y_i^2 < C_1$.

(H2) We assume that $\lim_{v\to\infty} \frac{n_v}{N_v} = \pi \in (0,1)$.

(H3) There exist a positive constant c such that $\min_{i\in U_v}\pi_i \geqslant c > 0$; also, we assume that $\limsup_{v\to\infty} n_v \max_{i\neq\ell\in U_v} |\pi_{i\ell} - \pi_i\pi_\ell| < \infty$.

(H4) We assume that there exists a positive constant C_2 such that, for all $i \in U_v$, $||\mathbf{x}_i||_2^2 \leq C_2 p_v$, where $||\cdot||_2$ denotes the usual Euclidean norm.

(H5) We assume that $||\widehat{\boldsymbol{\beta}}||_1 = O_{\mathrm{p}}(p_v)$, where $\widehat{\boldsymbol{\beta}}$ is the least square estimator given in (2.8) and $||\cdot||_1$ denotes the L^1 norm.

The assumptions (H1), (H2) and (H3) were used by Breidt and Opsomer (2000) in a nonparametric setting and similar assumptions were used by Robinson and Särndal (1983) to establish the consistency of the GREG estimator in a fixed dimensional setting. These assumptions hold for many usual sampling designs such as simple random sampling without replacement, stratified designs (Breidt and Opsomer, 2000), or high-entropy sampling designs. Assumptions (H4) and (H5) can be viewed, respectively, as extensions of Assumption A.1 and Assumption A.3 in Robinson and Särndal (1983) to p_v-dimensional vectors with p_v growing to infinity. Assumption (H5) is not very restrictive in this high-dimensional setting as it requires that components of $\widehat{\boldsymbol{\beta}}$ are all bounded. When p_v is fixed, then our assumptions essentially reduce to those of Robinson and Särndal (1983).

Result 2.3.1. *Assume (H1)-(H5). Consider a sequence of GREG estimators $\{\widehat{t}_{greg}\}_{v\in\mathbb{N}}$ of t_y. Then,*

$$\frac{1}{N_v}(\widehat{t}_{\mathrm{greg}} - t_y) = O_{\mathrm{p}}\left(\sqrt{\frac{p_v^3}{n_v}}\right).$$

If the numbers of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ and the sample sizes $\{n_v\}_{v\in\mathbb{N}}$ satisfy $p_v^3/n_v = o(1)$, then $N_v^{-1}(\widehat{t}_{greg} - t_y) = o_{\mathrm{p}}(1)$.

The $\sqrt{n}$-consistency obtained by Robinson and Särndal (1983) is a special case of Result 3.1 with $p_v = O(1)$. Result 3.1 highlights the fact that the rate of convergence decreases as the number of auxiliary variables p_v increases. Yet, this result guarantees the existence of a consistent GREG estimator, even when the number of auxiliary variables is allowed to diverge. An improved consistency rate for $\widehat{t}_{\mathrm{greg}}$ may be obtained if, in (H5), the usual euclidean norm is used instead of L^1-norm. Establishing the rate of convergence of the sampling error $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$ may also be utilized to obtain a lower consistency rate for $\widehat{t}_{\mathrm{greg}}$; e.g., see Chauvet and Goga (2022) .

The next result establishes the design-consistency of model-assisted penalized regression estimators. The proof is similar to that of Result 3.1 and is given in the Supplementary Material.

Result 2.3.2. *Assume (H1)-(H5). Consider a sequence of penalized model-assisted estimators* $\{\widehat{t}_{\text{pen}}\}_{v\in\mathbb{N}}$ *of* t_y *obtained by either ridge, lasso or elastic-net. Then,*

$$\frac{1}{N_v}(\widehat{t}_{\text{pen}} - t_y) = O_{\text{p}}\left(\sqrt{\frac{p_v^3}{n_v}}\right).$$

If the numbers of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ *and the sample sizes* $\{n_v\}_{v\in\mathbb{N}}$ *satisfy* $p_v^3/n_v = o(1)$*, then* $N_v^{-1}(\widehat{t}_{pen} - t_y) = o_{\text{p}}(1)$.

The above result makes no use of the asymptotic convergence rate of $\widehat{\boldsymbol{\beta}}_{\text{pen}}$ which depends on the penalization method. For example, if one can establish that $||\widehat{\boldsymbol{\beta}}_{\text{pen}}||_1 = O_{\text{p}}(\gamma_v)$, then $N_v^{-1}(\widehat{t}_{\text{pen}} - t_y) = O_{\text{p}}(\gamma_v\sqrt{p_v/n_v})$. Alternatively, improved consistency rates of $\widehat{t}_{\text{pen}}$ may be obtained if one can establish the magnitude of the sampling error $\widehat{\boldsymbol{\beta}}_{\text{pen}} - \widetilde{\boldsymbol{\beta}}_{\text{pen}}$ in a high-dimension setting. In other words, obtaining these improved rates requires additional assumptions, unlike Result 3.2 which is obtained under relatively mild assumptions.

Next, we show that, under additional assumptions on the auxiliary variables, the model-assisted ridge estimator is L^1 design-consistent for t_y if p_v/n goes to zero and that it has the usual $\sqrt{n}$-consistency rate if $p_v = O(n_v^a)$ with $0 \leq a < 1/2$, which constitutes a significant improvement over Result 3.2.

Result 2.3.3. *Assume (H1)-(H4). Also, assume that there exists a positive constant* $\tilde{C}$ *such that* $\lambda_{max}(\boldsymbol{X}_{U_v}^\top \boldsymbol{X}_{U_v}) \leqslant \tilde{C}N_v$*, where* $\lambda_{max}(\boldsymbol{X}_{U_v}^\top \boldsymbol{X}_{U_v})$ *is the largest eigenvalue of* $\boldsymbol{X}_{U_v}^\top \boldsymbol{X}_{U_v}$*. Assume also that* $N_v/\lambda_v = O(1)$.

1. *Then, there exists a positive constant* C *such that* $\mathbb{E}_p\left[||\widehat{\boldsymbol{\beta}}_{\text{ridge}}||_2^2\right] \leqslant C$ *and*

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{ridge}} - t_y\right| = O\left(\sqrt{\frac{p_v}{n_v}}\right).$$

   If the numbers of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ *and the sample sizes* $\{n_v\}_{v\in\mathbb{N}}$ *satisfy* $p_v/n_v = o(1)$*, then* $N_v^{-1}\mathbb{E}_p|\widehat{t}_{\text{ridge}} - t_y| = o(1)$.

2. $\mathbb{E}_{\text{p}}(||\widehat{\boldsymbol{\beta}}_{\text{ridge}} - \tilde{\boldsymbol{\beta}}_{\text{ridge}}||_2^2) = O(p_v/n_v)$*. Thus, if* $p_v/n_v = o(1)$*, then* $\mathbb{E}_{\text{p}}(||\widehat{\boldsymbol{\beta}}_{\text{ridge}} - \tilde{\boldsymbol{\beta}}_{\text{ridge}}||_2^2) = o(1)$.

3. *We have the following asymptotic equivalence:*

$$\frac{1}{N_v}\left(\widehat{t}_{\text{ridge}} - t_y\right) = \frac{1}{N_v}\left(\widehat{t}_{\text{diff,}^\vee} - t_y\right) + O_{\text{p}}\left(\frac{p_v}{n_v}\right),$$

   where

$$\widehat{t}_{\text{diff,}^\vee} = \sum_{i\in S_v} y_i/\pi_i - \left(\sum_{i\in S_v}\mathbf{x}_i/\pi_i - \sum_{i\in U_v}\mathbf{x}_i\right)^\top \tilde{\boldsymbol{\beta}}_{\text{ridge}}$$

   and

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{ridge}} - t_y\right| = O\left(\frac{1}{\sqrt{n_v}}\right) + O\left(\frac{p_v}{n_v}\right).$$

If $p_v = O(n_v^a)$ *with* $0 \leq a < 1/2$, *then*

$$\frac{1}{N_v}\left(\widehat{t}_{\text{ridge}} - t_y\right) = \frac{1}{N_v}\left(\widehat{t}_{\text{diff},\breve{}} - t_y\right) + o_{\text{p}}(1)$$

and

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{ridge}} - t_y\right| = O\left(\frac{1}{\sqrt{n_v}}\right).$$

It follows from Result 3.3 that, for $p_v = O(n_v^a)$ with $0 \leq a < 1/2$, the asymptotic variance of the model-assisted ridge estimator $\widehat{t}_{\text{ridge}}$ is equal to the variance of the generalized difference estimator $\widehat{t}_{\text{diff},\breve{}}$. For $a = 1/2$, we note that the model-assisted estimator is still $\sqrt{n}$-design consistent but the remainder term is no longer negligible with respect to $\widehat{t}_{\text{diff},\breve{}}$ and the variability of this term should be consider to compute the asymptotic variance of $\widehat{t}_{\text{ridge}}$. The case of model-assisted estimators based on lasso and elastic-net is more intricate. This is due to the fact that both estimators involve the L^1-norm. As a result, a closed-form expression of these estimators cannot be obtained. However, if the predictors are orthogonal, a closed-form expression exists for the lasso and elastic-net estimators and improved consistency rates can be obtained; see Proposition 3.1 below. The case of non-orthogonal predictors is more challenging and is beyond the scope of this article.

Proposition 2.3.1. *Suppose assumptions (H1)-(H3) and that the sampling design and the X-variables are such that the columns of* $\boldsymbol{\Pi}_{S_v}^{-1/2}\mathbf{X}_{S_v}$ *are orthogonal. Suppose also that there exist positive quantities* C_3 *and* C_4 *such that* $max_{j=1,\ldots,p_v} N_v^{-1}\sum_{i\in U_v} x_{ij}^4 \leq C_3 < \infty$ *and* $min_{j=1,\ldots,p_v} N_v^{-1}\sum_{i\in U_v} x_{ij}^2 \geq C_4 > 0$. *Then,* $N_v^{-1}(\widehat{t}_{\text{greg}} - t_y) = O_{\text{p}}(\sqrt{p_v/n_v})$ *and* $N_v^{-1}(\widehat{t}_{\text{pen}} - t_y) = O_{\text{p}}(\sqrt{p_v/n_v})$, *where* $\widehat{t}_{\text{pen}}$ *denotes either the lasso or the elastic-net estimator.*

2.4 Simulation study

In this section, we provide an empirical comparison of several model-assisted estimators. In addition to the estimators discussed in Section 2.3. In addition, we considered model-assisted estimators based on principal component regression (Cardot et al., 2017), regression trees (Breiman, 1984), random forests (Breiman, 2001), k-nearest neighbors, XGBoost (Chen and Guestrin, 2016) and Cubist (Quinlan et al., 1992). For a description of these methods, see Hastie et al. (2011) and Dagdoug et al. (2021a) and the references therein.

We used data from the Irish Commission for Energy Regulation (CER) Smart Metering Project that was conducted in 2009-2010 (CER, 2011)[2] (Cardot et al., 2017). This project focused on energy consumption and energy regulation. About 6000 smart meters were installed to collect the electricity consumption of Irish residential and business customers every half an hour over a period of about two years.

We considered a period of 14 consecutive days and a population of $N = 6,291$ smart meters (households and companies). Each day consisted of 48 measurements, leading to 672 measurements for each household. We denote by $X_j = X(t_j), j = 1,\ldots,672$, the electricity consumption (in kW) at instant t_j and by x_{ij} the value of X_j recorded by the ith smart meter for $i = 1,\ldots,6,291$. It should be noted that the matrix $N^{-1}\mathbf{X}^\top\mathbf{X}$ was ill-conditioned with a

[2] The data are available on request at: `https://www.ucd.ie/issda/data/commissionforenergyregulationcer/`.

condition number equal to 254 753. This suggests that some of the X-variables were highly correlated with each other.

We generated four survey variables based on these auxiliary variables according to the following models:

$$
\begin{aligned}
Y_1 &= 400 + 2X_1 + X_2 + 2X_3 + \mathcal{N}(0, 1500);\\
Y_2 &= 500 + 2X_4 + 400\mathbb{1}\left(X_5 > 156\right) - 400\mathbb{1}\left(X_5 \leqslant 156\right) + 1000\mathbb{1}\left(X_2 > 190\right)\\
&\quad + 300\mathbb{1}\left(X_5 > 200\right) + \mathcal{N}(0, 1500);\\
Y_3 &= 1 + \cos(2X_1 + X_2 + 2X_3)^2 + \epsilon_1;\\
Y_4 &= 4 + 3 \cdot \mathbb{V}\left(\{X_1 + X_2\}^2\right)^{-1/2} \times \{X_1 + X_2\}^2 + \mathcal{N}(0, 0.01),
\end{aligned}
$$

where $\mathbb{V}(\cdot)$ denotes the empirical variance and the errors ϵ_1 in the model for Y_3 were generated from an $\mathcal{E}xp(10)$ and these errors were centered so as to obtain a mean equal to zero.

Our goal was to estimate the population totals $t_{y_j} = \sum_{i \in U} y_{ij}, j = 1, \ldots, 4$. From the population, we selected $R = 2,500$ samples, of size $n = 600$, which corresponds to a sampling fraction n/N of about 10%. We considered three sampling schemes: simple random sampling without replacement, stratified simple random sampling without replacement with optimal allocation and stratified without replacement proportional to size sampling with proportional allocation.

In each sample, we computed twelve model-assisted estimators of the form

$$
\widehat{t}_{ma}^{(j)} = \sum_{i \in U} \widehat{f}^{(j)}(\mathbf{x}_i) + \sum_{i \in S} \frac{y_i - \widehat{f}^{(j)}(\mathbf{x}_i)}{\pi_i}, \quad j = 1, 2, \ldots, 12,
$$

where the predictors $\widehat{f}^{(j)}(\mathbf{x}_i)$, $j = 1, 2, \ldots, 12$, were obtained using the following procedures:

Procedure 1: "LR" : Deterministic linear regression, leading to the GREG estimator.

Procedure 2: "CART": Classification and regression tree algorithm (Breiman, 1984), leading to an estimator closely related to that of McConville and Toth (2019) and implemented with the R-package `rpart`.

Procedure 3: "RF": Random forests with the algorithm of Breiman (2001) with $B = 1000$ trees, a minimal number of elements in each terminal node $n_0 = 5$ and $p_0 = \lfloor \sqrt{p} \rfloor$ variables selected randomly at each split, where $\lfloor \cdot \rfloor$ denotes the customary floor function. The algorithm leads to the estimator described in Dagdoug et al. (2021b). Simulations were implemented with the R-package `ranger`.

Procedure 4: "Ridge": Ridge regression with a regularization parameter determined by cross-validation and implemented with the R-package `glmnet`. The estimator was studied by Goga and Shehzad (2010).

Procedure 5: "Lasso": Lasso regression with a regularization parameter determined by cross-validation and implemented with the R-package `glmnet` (McConville et al., 2017).

Procedure 6: "EN": Elastic net regression with penalization coefficients determined by cross-validation with the R-package `glmnet`.

Procedure 7: "XGB": XGBoost algorithm (Hastie et al., 2011) with 50 trees in the additive model, each tree being with a depth of at most 6 and a learning rate $\lambda = 0.01$. Simulations were implemented with the *R*-package `XGBoost`.

Procedure 8: "5NN": 5-nearest neighbors predictor with the euclidean distance and implemented with the *R*-package `caret`.

Procedure 9: "Cubist": A cubist algorithm (Kuhn and Johnson, 2013) with 5 models in each predictor, implemented with the *R*-package `cubist`; the algorithm and its adaptation for survey data are described in Dagdoug et al. (2021a).

Procedure 10: "PCR1": Principal component regression based on the first $\lfloor p^{1/4} \rfloor$ components kept and implemented with the *R*-package `pls` (Cardot et al., 2017).

Procedure 11: "PCR2": Principal component regression based on the first $\lfloor p^{2/4} \rfloor$ components kept.

Procedure 12: "PCR3": Principal component regression based on the first $\lfloor p^{3/4} \rfloor$ components kept.

As a measure of bias of the model-assisted estimators $\widehat{t}_{ma}^{(j)}$, $j = 1, 2, ..., 12$, we computed the Monte Carlo percent relative bias defined as

$$RB_{MC}\left(\widehat{t}_{ma}^{(j)}\right) = 100 \times \frac{1}{R}\sum_{r=1}^{R}\frac{(\widehat{t}_{ma}^{(j,r)} - t_y)}{t_y}, \quad j = 1, 2, \ldots, 12,$$

where $\widehat{t}_{ma}^{(j,r)}$ denotes the estimator $\widehat{t}_{ma}^{(j)}$ at the rth iteration, $r = 1, \ldots, R$. As a measure of efficiency, we computed the relative of efficiency, using the Horvitz-Thompson estimator $\widehat{t}_{\pi}$ given by (2.1), as the reference. That is,

$$RE_{MC}\left(\widehat{t}_{ma}^{(j)}\right) = 100 \times \frac{MSE_{MC}(\widehat{t}_{ma}^{(j)})}{MSE_{MC}(\widehat{t}_{\pi})}, \quad j = 1, 2, ..., 12,$$

where $MSE_{MC}(\widehat{t}_{ma}^{(j)}) = R^{-1}\sum_{r=1}^{R}(\widehat{t}_{ma}^{(j,r)} - t_y)^2$ and $MSE_{MC}(\widehat{t}_{\pi})$ is defined similarly.

We were also interested in investigating to which extent the model-assisted estimators $\widehat{t}_{ma}^{(j)}$, $j = 1, \ldots, 12$ were affected by the inclusion of a large number of predictors in the working models. To that end, in addition to the variables $X_1, \ldots, X_5$, we included d_{noise} predictors in the working models. These predictors were available in the Irish data set. We used the following values for d_{noise}: 5, 10, 20, 50, 100, 200, 300 and 400.

2.4.1 Simple random sampling without replacement

In this section, we present the results obtained under simple random sampling without replacement (SRSWOR) of size $n = 600$. All the point estimators $\widehat{t}_{ma}^{(j)}$, $j = 1, \ldots, 12$, exhibited a negligible or small percent RB with a maximum value of about 3.1% (obtained in the case of the GREG estimator). For this reason, results pertaining to relative bias are not reported here.

Figures 4-7 display the relative efficiency of the model-assisted estimators $\widehat{t}_{ma}^{(j)}$, $j = 1, \ldots, 12$ as a function of the number of auxiliary variables incorporated in the working models. To improve readability, we have truncated some large values of RE, when applicable.

We begin by discussing the results on relative efficiency pertaining to the estimation of the total of the survey variable Y_1. For low-dimensional settings, the GREG estimator was very efficient with values of RE below 10%. These results can be explained by the fact that Y_1 was linearly related to the **x**-variables. However, as the number of variables d_{noise} increased, the efficiency of the GREG estimator rapidly deteriorated, suggesting that the performance of the GREG estimator is sensitive to the dimension of the **x**-vector. As expected, model-assisted estimators based on regularization methods such as ridge, lasso, elastic-net or dimension reduction methods such as principal components regression, performed generally very well. Unlike the GREG, these estimators were not much affected by the number of auxiliary variables incorporated in the model. Turning to the model-assisted estimator based on a 5-nn, we note that it was less efficient than most competitors and that its efficiency got worse as d_{noise} increased, a phenomenon referred to as the curse of dimensionality. The model-estimators based on XGBoost, Cubist and random forests performed quite well and did not seem to be affected by the number of auxiliary variables incorporated in the model. Finally, the estimators based on CART were less efficient than those obtained through the other machine learning methods.

The results pertaining to the survey variable Y_2 and displayed in Figure 5 were fairly consistent with those obtained for the survey variable Y_1 with one exception: the Cubist algorithm was significantly more efficient than the other procedures in all the scenarios.

Turning to the survey variable Y_3 (see Figure 6), the model-assisted estimator based on random forests was significantly more efficient than the Horvitz–Thompson estimator, especially for large values of d_{noise}. The other procedures led to estimators less efficient than the Horvitz-Thompson estimator with values of RE above 100. In particular, the GREG estimator broke down as the number of auxiliary variable increased. The performance of model-assisted estimators based on CART and XGBoost algorithms deteriorated as the dimension increased. In a high-dimension setting with highly correlated predictors, random forests improved over CART due to the random subsampling of p_0 variables among the p variables, generating then decorrelated trees (Hastie et al., 2011).

The results in Figure 7 about the survey variable Y_4 were similar to the ones in previous figures. Most estimators remained mostly unaffected by the number of auxiliary variables d_{noise}. Again, the model-assisted estimator based on the Cubist algorithm was the best in all the scenarios.

2.4.2 Stratified simple random sampling with optimal allocation

In the second simulation study, we partitioned the Irish residential and business customer population into four strata $U_1, \ldots, U_4$, using an equal quantile method with respect to the variable, X_1, the electricity consumption at instant t_1. From the population, we selected $R = 2,500$ stratified simple random samples, of size $n = 600$. The stratum sample sizes n_h were determined using an X_2-optimal allocation, where X_2 denotes the electricity consumption recorded at instant t_2. This led to $n_1 = 20, n_2 = 36, n_3 = 45$ and $n_4 = 499$. The first-order inclusion probabilities, $\pi_i = n_h/N_h, i \in U_h$ and the sampling weights $w_i = \pi_i^{-1}$ are shown in Table 1.

We confined to the survey variables Y_1 and Y_3 only and we aimed at estimating t_{y_1} and t_{y_3}. It is worth pointing out that the resulting sampling design was informative as the variables used at the design stage (X_1 and X_2) were also related to the survey variables Y_1 and Y_3. In fact, the

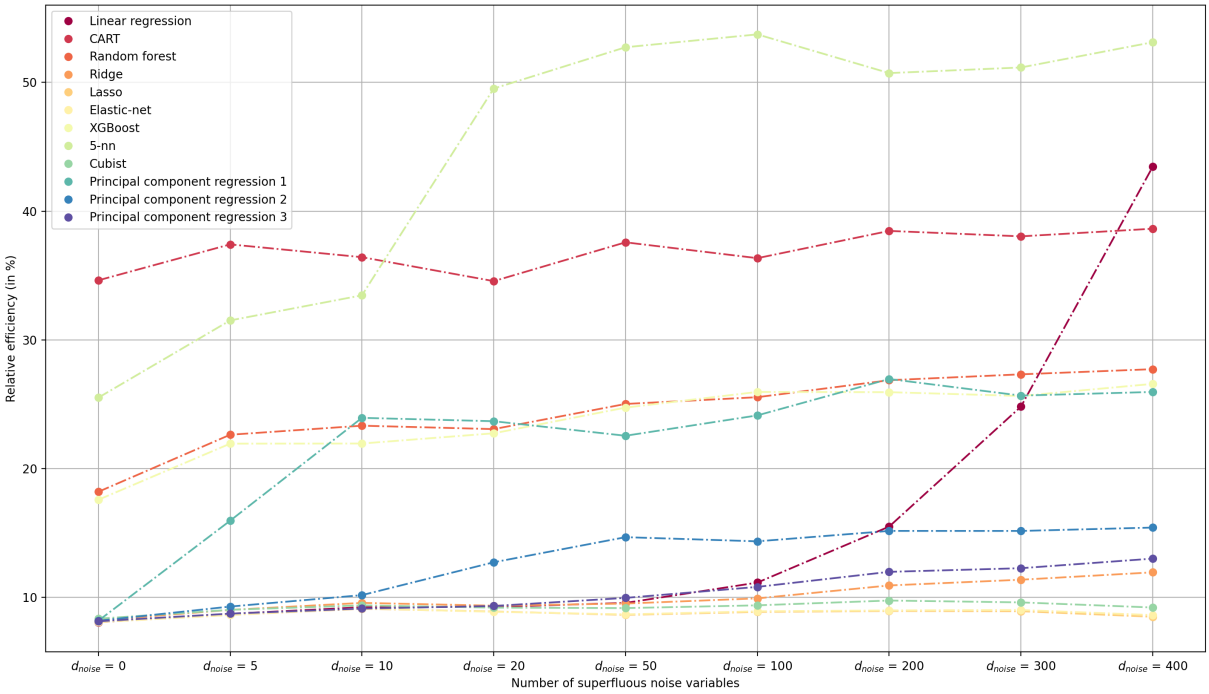


Figure 4: Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_1 with SRSWOR ($n = 600$) and increasing number of auxiliary variables

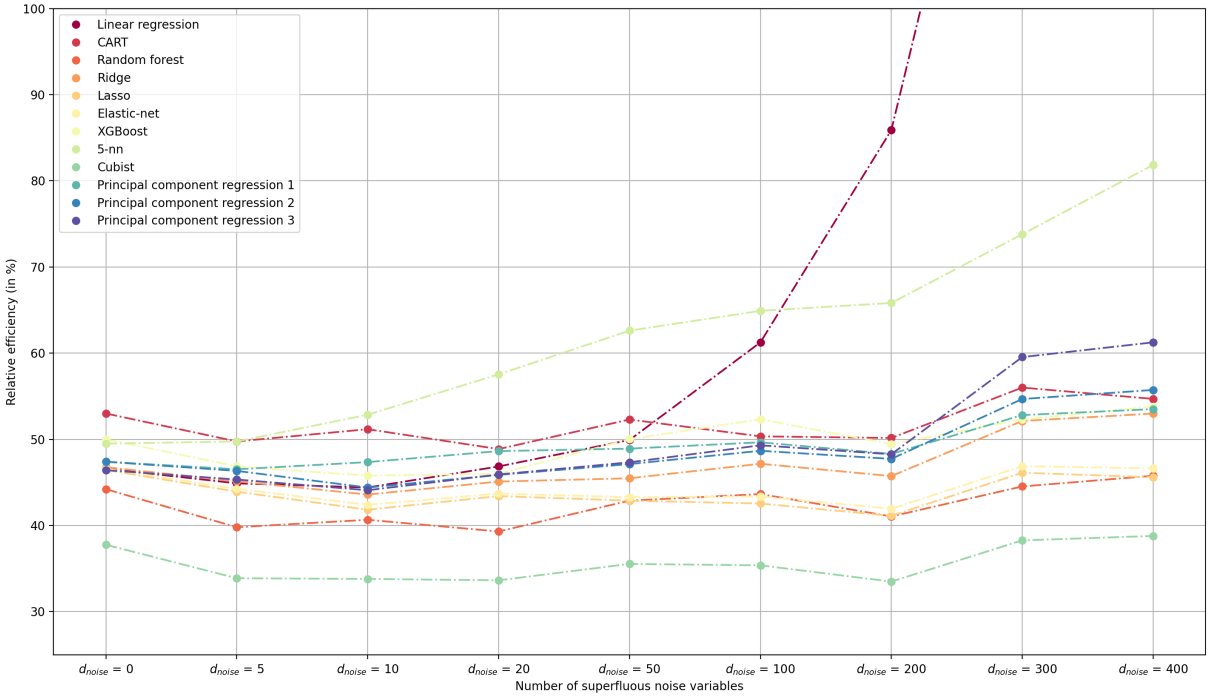


Figure 5: Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_2 with SRSWOR, $n = 600$ and increasing number of auxiliary variables

Monte Carlo coefficient of correlation between the sampling weights and Y_1 was approximately equal to 0.402. We do not report the coefficient of correlation between the sampling weights and Y_3 as the relationship between Y_3 and the set of predictors X_1, X_3 is not linear.

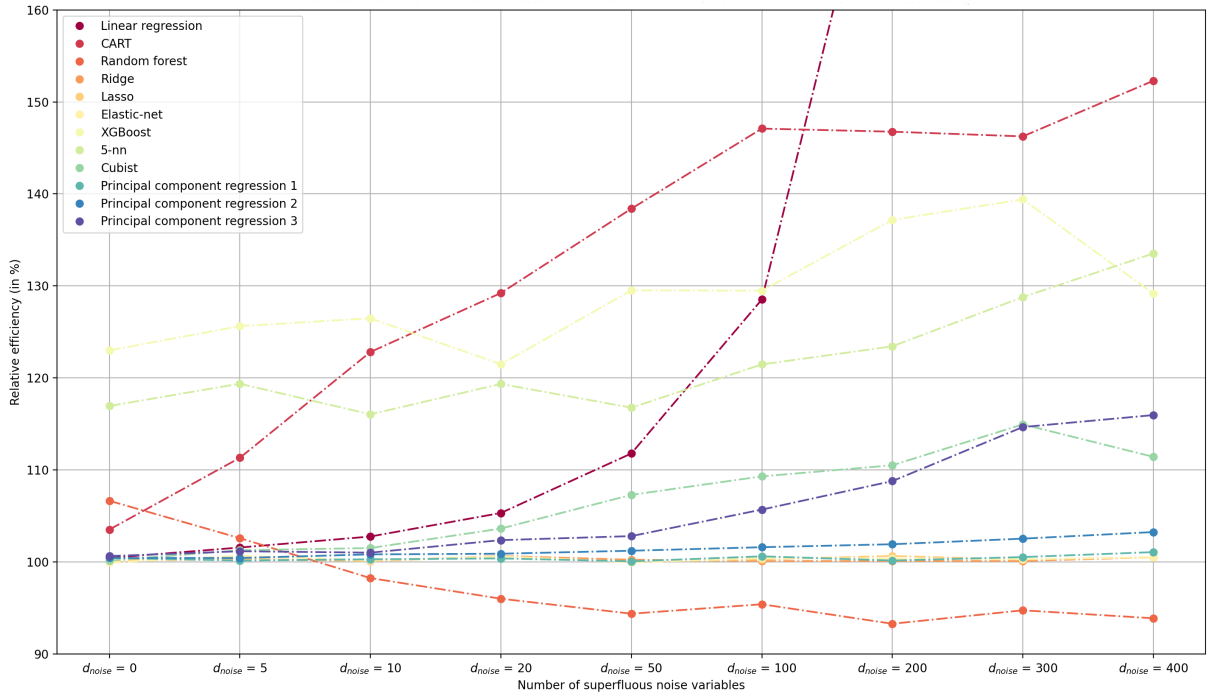


Figure 6: Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_3 with SRSWOR, $n = 600$ and increasing number of auxiliary variables

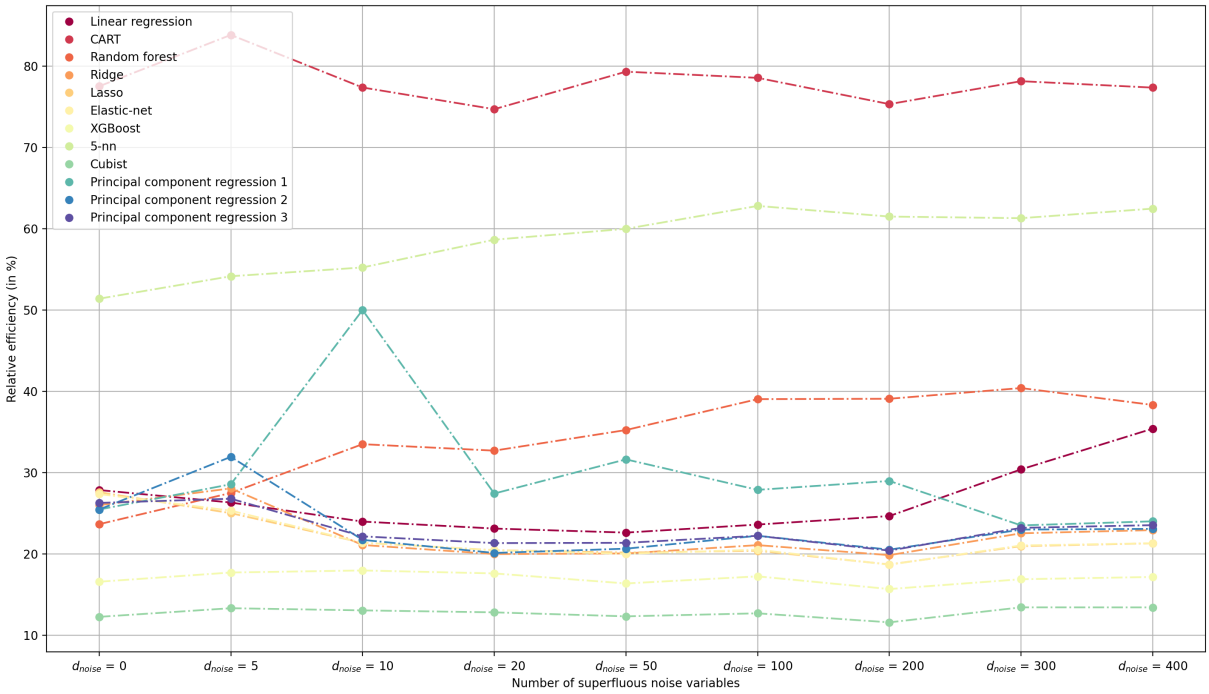


Figure 7: Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_4 with SRSWOR, $n = 600$ and increasing number of auxiliary variables

Again, in each sample we computed twelve model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for each of t_{y_1} and t_{y_3}. Since most machine learning software packages do not take the sampling weights into account, we have included the design variables X_1 and X_2 in the set of predictors.

Stratum	1	2	3	4
π_i	0.012	0.022	0.028	0.316
$w_i = \pi_i^{-1}$	77.85	43.83	35.11	3.16

Table 1: First-order inclusion probabilities and sampling weights within strata.

We begin by discussing the results pertaining to the estimation of the total of the survey variable Y_1. Figure 8 and Figure 9 display the Monte Carlo percent relative bias and the Monte Carlo relative efficiency as a function of the number of variables d_{noise}. Except for the model-assisted estimators based on 5-nn and random forest, the other estimators exhibit a small value of RB for all values of d_{noise}. Again, the 5-nn model-assisted estimator suffered from the curse of dimensionality. Turning to the estimator based on random forests, we note from Figure 8 that the bias increased as the number of predictors d_{noise} increased. For instance, for $d_{noise} = 400$, the value of RB was just above 10%. This significant bias may be explained by the fact that random forests is the only procedure among the ones considered in our simulation that randomly selects $p_0 = \sqrt{p}$ variables among the initial p predictors at each split. For instance, for $d_{noise} = 400$, only 20 variables are randomly selected at each split. As a result, most predictions obtained through a random forests algorithm were based on misspecified working models, leading to potentially bad fits and large residuals. Also, each prediction corresponds to a weighted mean computed within each node with $n_0 = 5$ observations only. Therefore, each predictions corresponds to a ratio-type estimate based on 5 observations only. This, together with the fact that the sampling weights are highly variable, constitutes a conducive ground for the occurrence of small sample bias. In terms of efficiency, except for the GREG, the 5-nn and the random forest estimators, the other procedures performed well with values of RE ranging from 60% to 80%. The best procedures were Cubist and Lasso.

We now turn to the survey variable Y_3. First, the Monte Carlo relative bias was negligible for all the estimation procedures and are not reported here. Results about relative efficiency are plotted in Figure 10. Random forests performed extremely well and their performance improved as d_{noise} increased. This suggests that the method was able to extract the information contained in the predictors. This was also true for Cubist and XGBoost, although to a lesser extent.

To get a better understanding of the performance of random forests for the estimation of the total of the survey variable Y_1, we conducted additional scenarios based on different values of the hyper parameters n_0, the number of observations within each terminal nodes, and p_0, the number of variables randomly selected at each split among the initial p model variables. We used the following values for n_0 and p_0:

- $n_0 = 5$ observations and $p_0 = \sqrt{p}$ variables which are the default choices in the *R*-package `ranger`;
- $n_0 = 5$ observations and $p_0 = p$ variables;
- $n_0 = 5$ observations and $p_0 = \sqrt{p}$ variables, with, in addition, the design variables X_1, X_2, as well as the vector of inclusion probabilities and the vector of strata that were selected with probability 1, at each split, besides the p_0 variables;
- $n_0 = n^{13/20}$ observations and $p_0 = \sqrt{p}$ variables.

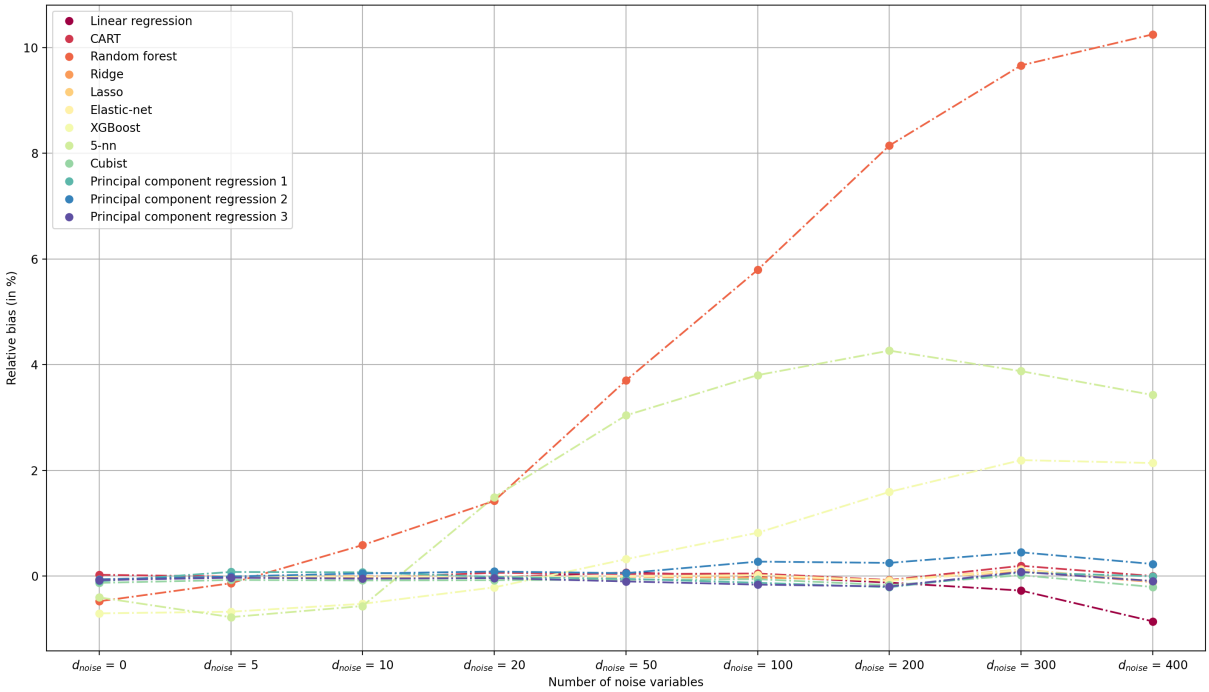


Figure 8: Relative bias of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_1 with stratified simple random sampling with X_2-optimal allocation, $n = 600$ with increasing number of auxiliary variables

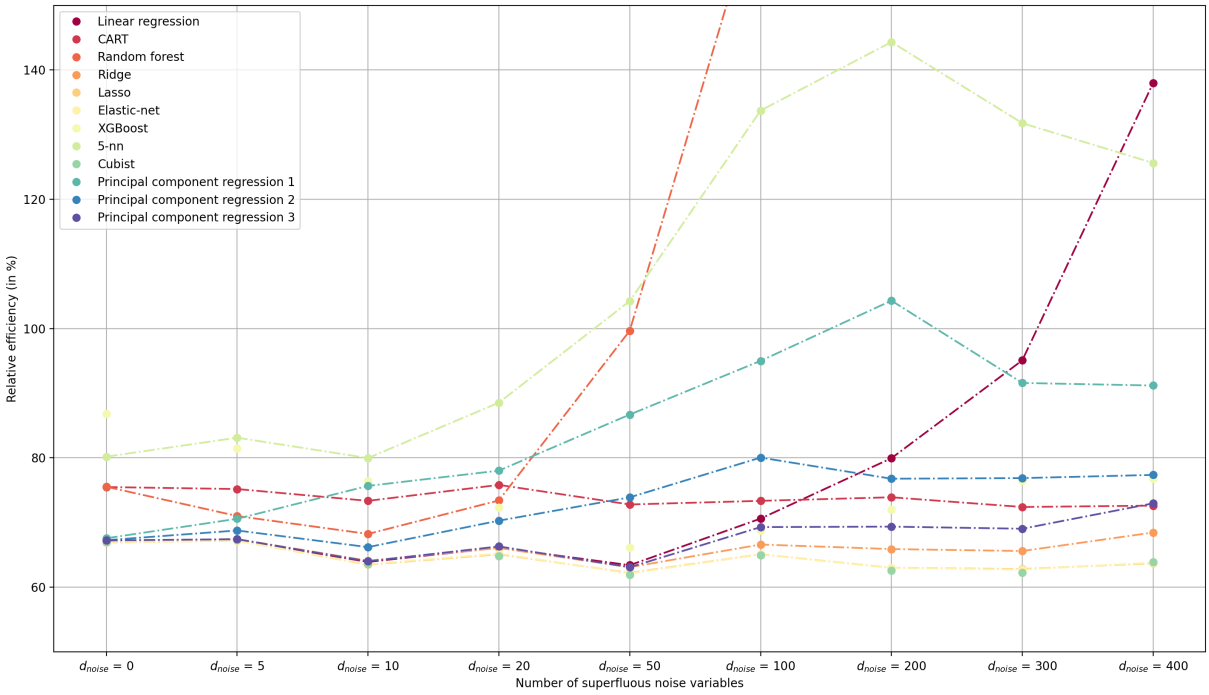


Figure 9: Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_1 with stratified simple random sampling with X_2-optimal allocation, $n = 600$ and increasing number of auxiliary variables

The Monte Carlo percent relative bias is displayed in Figure 11. We note that relative bias was much smaller (always less than 1%) when the design variables were considered besides p_0 variables at each split. To a lesser extent, the bias decreased when more observations

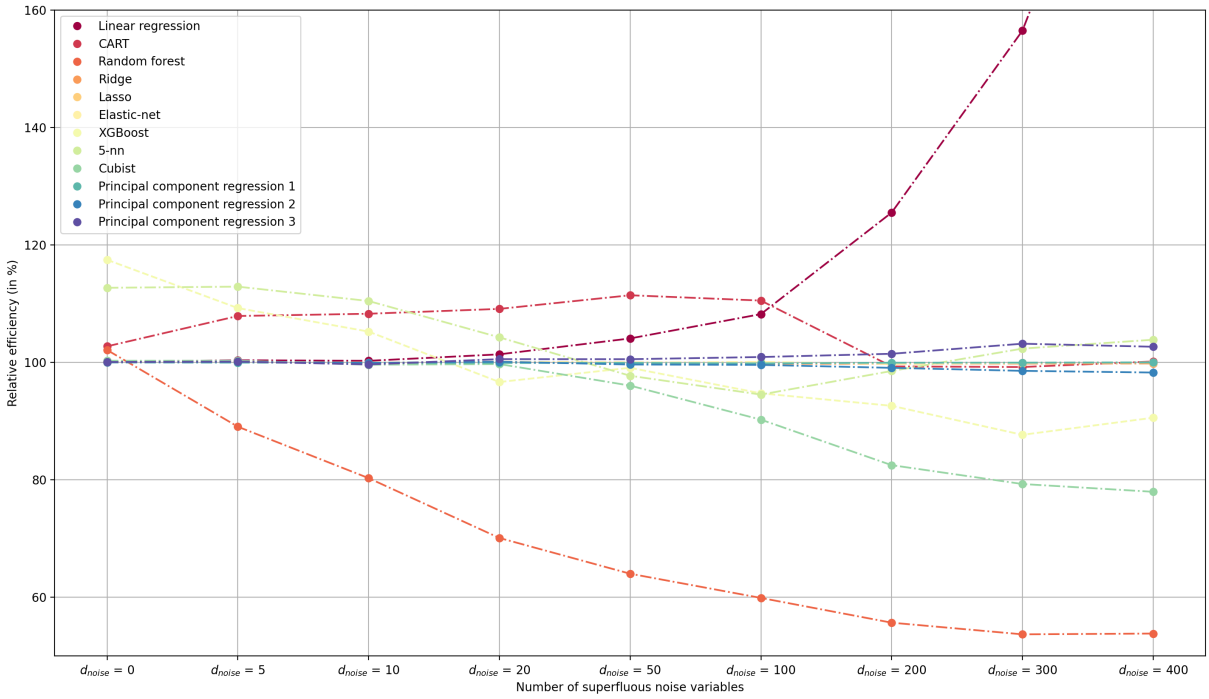


Figure 10: Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}$, $j = 1, \ldots, 12$ for the estimation of the total of Y_3 with stratified simple random sampling with X_2-optimal allocation, n=600 and increasing number of auxiliary variables

were allowed in each terminal node. These results suggest, that, when the sampling design is informative, in order to avoid significant small sample bias, we recommend to force the design variables to be selected at each split. This option is available in the *R* package `ranger`.

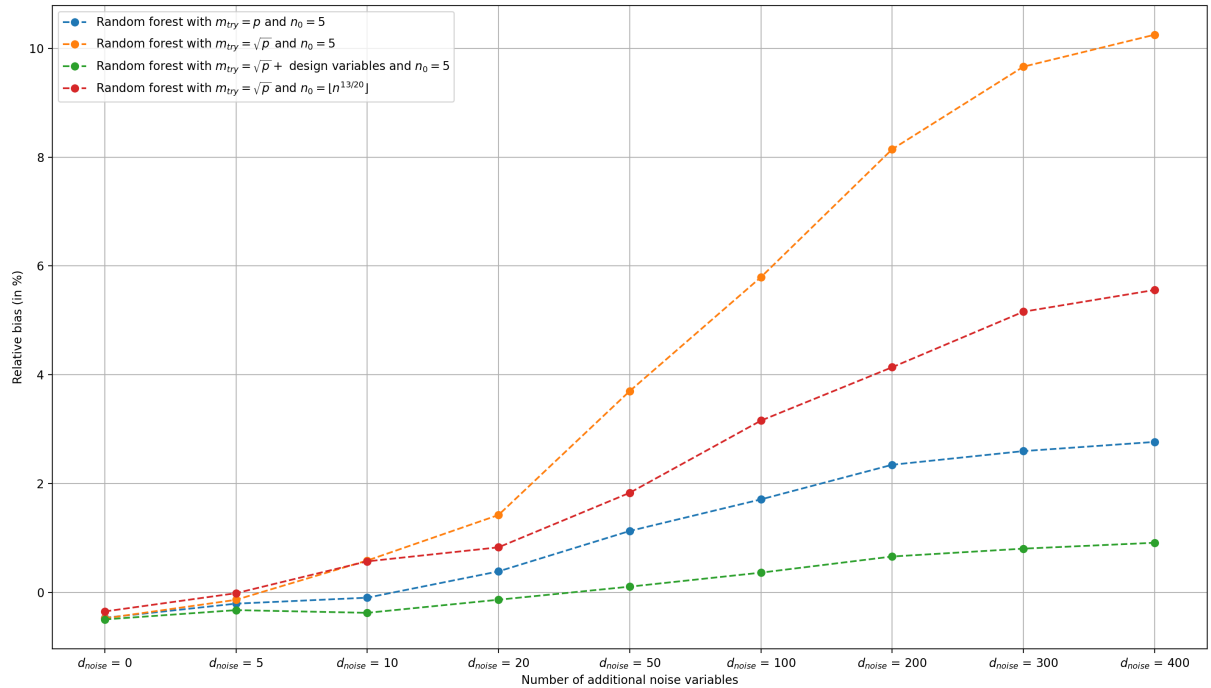


Figure 11: Comparison of different configurations of hyper-parameters for $\widehat{t}_{rf}$ for the estimation of the total of Y_1 with stratified simple random sampling and X_2-optimal allocation, $n = 600$.

2.4.3 Stratified inclusion probability proportional-to-size sampling without replacement

We consider the stratified population described in Section 2.4.2. In each stratum, we selected units according to a fixed-size inclusion probability proportional-to-size sampling without replacement using X_2, the electricity consumption at instant $t = 2$, as the size variable. In each stratum, we used the sample size n_h were determined according to proportional allocation; i.e., $n_h = n \cdot N_h/N$. The first-order inclusion probabilities were then given by

$$\pi_i = \frac{n_h x_{i2}}{\sum_{j \in U_h} x_{j2}}, \qquad i \in U_h, \qquad \text{and} \qquad h = 1, 2, 3, 4.$$

As in Section 2.4.2, we focused on estimating t_{y_1} and t_{y_3} and we computed the same twelve model-assisted estimators $\widehat{t}_{ma}^{(j)}$, $j = 1, \ldots, 12$. The inclusion probabilities were highly correlated with the survey variable Y_1, with a correlation coefficient of about 0.62; we do not report the coefficient of correlation in the case of Y_3 as the underlying relationship was nonlinear. Based on findings from the Section 2.4.2, we adopted the following configuration for the random forest algorithm: we considered $n_0 = 5$ observations in each terminal node and, at each split, we randomly selected $p_0 = \sqrt{p}$ variables. Note that the design variables X_1 and X_2 as well as the vector of inclusion probabilities and the vector of stratum indicators were selected with probability 1 at each split in addition to the p_0 variables.

All the estimators exhibited a negligible relative bias (less than 1%). Figure 12 and Figure 13 show the relative efficiency corresponding to t_{y1} t_{y3}, respectively.

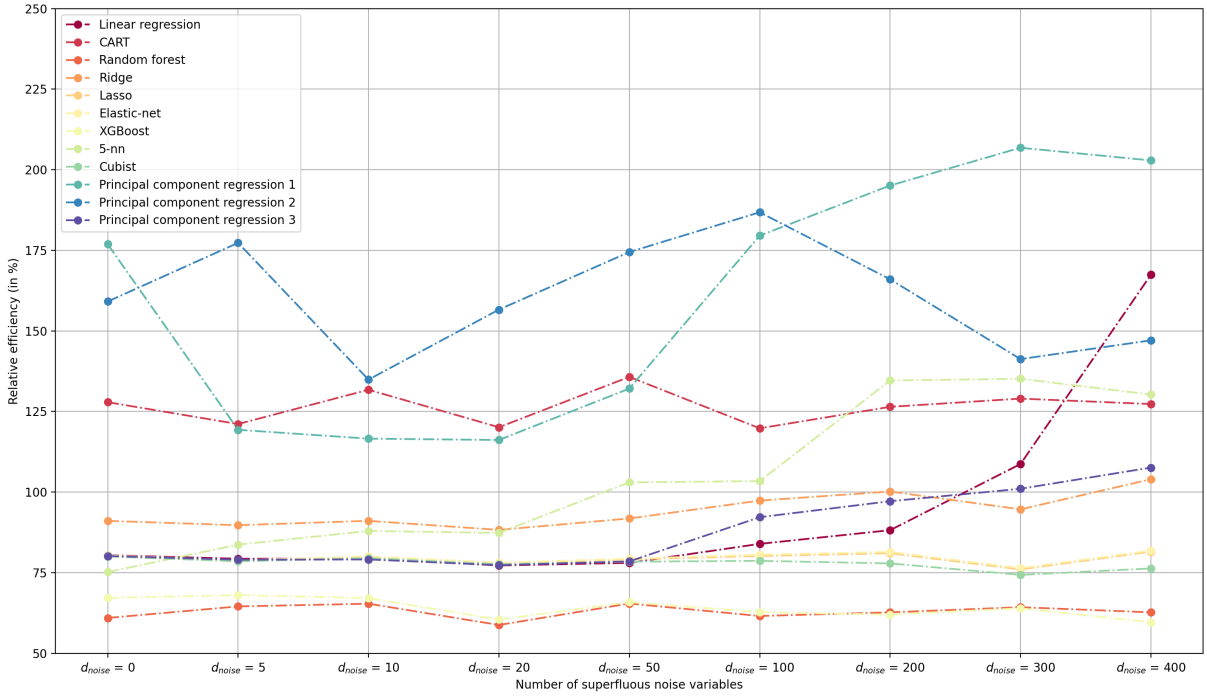


Figure 12: Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}$, $j = 1, \ldots, 12$ for the estimation of the total of Y_1 with stratified without replacement X_2-proportional to size sampling, $n = 600$ and increasing number of auxiliary variables

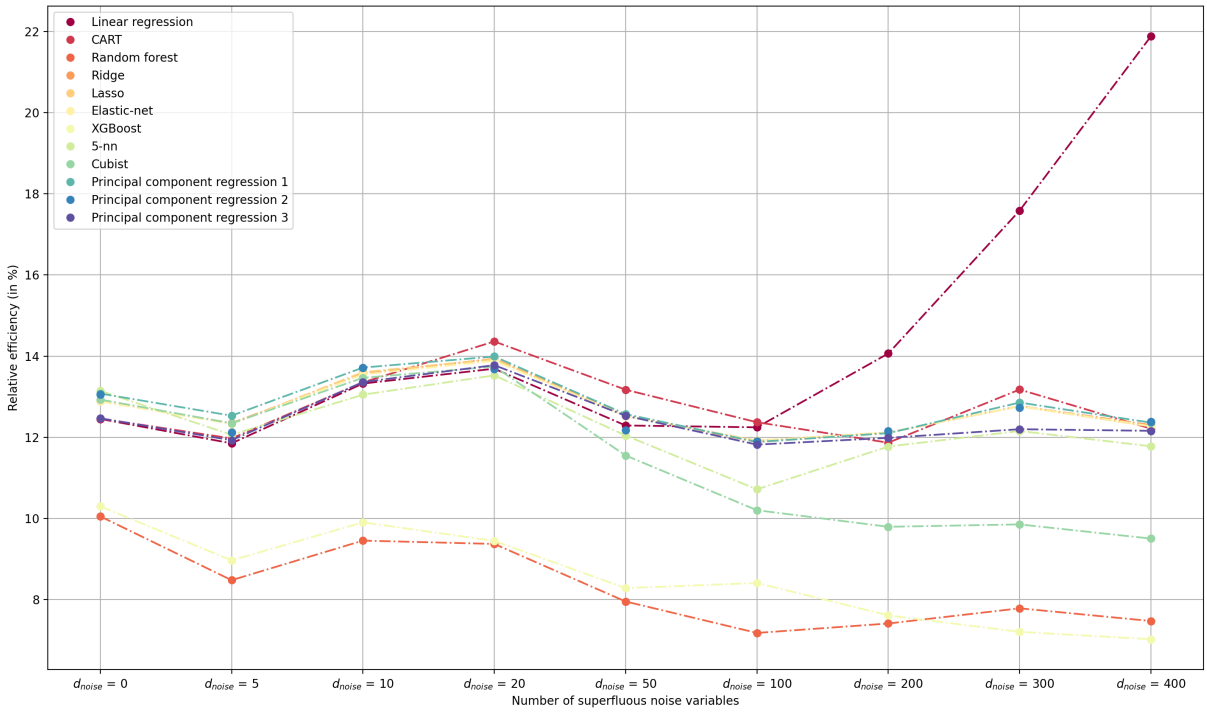


Figure 13: Relative efficiency of model-assisted estimators $\widehat{t}_{ma}^{(j)}, j = 1, \ldots, 12$ for the estimation of the total of Y_3 with stratified without replacement X_2-proportional to size sampling, $n = 600$ and increasing number of auxiliary variables

From Figure 12, we note that most estimators exhibited a behavior similar to that obtained in the case the stratified simple random sampling based on an X_2-optimal allocation (see Section 2.4.2). However, we note that the estimators PCR1 and PCR2 did poorly unlike in the case stratified simple random sampling based on an X_2-optimal allocation. This poor behaviour may be due to the fact that the sampling design was now much more informative and keeping a few principal components only may have led to a loss of information. The estimator PCR3 based on more principal components did better than PCR1 and PCR2. From Figure 13, we note that the use of model-assisted estimators led to significant improvement over the Horvitz-Thompson estimator, with value of relative efficiency ranging from 6% to 22%.

2.4.4 Stratified simple random sampling with proportional allocation

In this section, we consider a more realistic scenario based again on the Irish residential and business customer data. As a stratification variable, we used the mean electricity consumption recorded during the first week. Again, we constructed four strata using an equal-quantile method based, this time, on the mean electricity consumption; see also Cardot et al. (2013b) who used a similar design. The mean trajectories during the first week within each stratum are plotted in Figure 14. From Figure 14, we note that Stratum 1 corresponds to consumers with low global levels of electricity consumption, whereas Stratum 4 consists of consumers who have high levels of electricity consumption.

Our aim was to estimate the total electricity consumption recorded on the Monday of the second week and given by $t_y = \sum_{i=1}^{6291} \sum_{j=336}^{384} y_{ij}$, where y_{ij} is the electricity consumption recorded for the i-th unit at the j-th instant. Within each stratum, we selected a sample, of

size n_h, according to simple random sampling without replacement. The n_h's were determined according to proportional allocation; i.e, $n_h = n \times (N_h/N)$ with $n = 600$. In each of the 2,500 samples, we computed the same 12 model-assisted estimators as in the previous sections. Again, we computed the Monte Carlo percent relative bias and the relative efficiency for each the 12 estimators. The results are presented in Table 2.

Figure 14: Average electricity consumption on each stratum during first week

Estimator	Relative bias	Relative efficiency
LR	0.2	9.3
CART	-0.1	41.0
RF	-1.1	17.0
Ridge	0.1	4.0
Lasso	0.2	4.1
EN	0.2	4.1
XGB	-1.7	24.9
NN5	-4.0	65.6
Cubist	-0.0	4.3
PCR1	0.1	4.9
PCR2	0.1	4.2
PCR3	0.1	4.2

Table 2: Monte Carlo percent relative bias and relative efficiency of several model-assisted estimators under stratified simple random sampling with proportional allocation.

From Table 2, we note that the 5-nn model-assisted estimator was the only estimator to exhibit a non-negligible bias. Although it was less efficient than its competitors, it was more efficient than the Horvitz-Thompson estimator with a value of RE of about 65.6%. The ridge

estimator was the most efficient with a value of RE equal to 4% and was closely followed by lasso, elastic-net, Cubist and principal components model-assisted estimators. The GREG estimator performed very well with a value of RE of about 9.3%. Random forests led to considerable improvement over the CART model-assisted estimator with values of RE of 17% and 41%, respectively. Still, random forests were less efficient than the GREG estimator, which is not surprising as the relationship between the survey variable and the auxiliary variables was linear.

2.5 Final remarks

In this paper, we have examined a number of model-assisted estimation procedures in a high-dimensional setting both theoretically and empirically. If the relationship between the survey variable and the auxiliary information can be well described by a linear model, our results suggest that penalized estimators such as ridge, lasso and elastic net perform very well in terms of bias and efficiency, even in the case $p = n$. Model-assisted estimators based on random forests, Cubist and XGBoost methods were mostly unaffected by the number of predictors incorporated in the working model, even in the case of complex relationships between the study and the auxiliary variables. As expected, the GREG estimator suffered from poor performances in the case of a large number of auxiliary variables.

The procedure Cubist stood out from the other machine learning procedure with very good performances in virtually all the scenarios. Further work is needed to establish the theoretical properties of model-assisted estimators based on Cubist in both a low-dimensional and high-dimensional settings.

Variance estimation is an important stage of the estimation process. Further research includes identifying the regularity conditions under which the variance estimators are design-consistent in a high-dimensional setting.

We end this article by mentioning that virtually all the machine learning software packages cannot handle design features such as unequal weights and stratification. For instance, some random forests algorithms may involve a bootstrapping procedure and/or a cross-validation procedure. To fully account for the sampling design, both procedures must be modified so as to account for the design features. One notable exception is the R package RPMS (Toth, 2021) that has the ability to incorporate sampling weights for CART and random forests. Not fully accounting for the sampling design may be viewed as a form of model misspecification. However, model-assisted estimation procedures remain design-consistent even if the model is misspecified. In our experiments, several machine learning procedures (e.g., random forests, Cubist, XGboost) performed very well in most scenarios even though we did not modify the bootstrapping and cross-validation procedures to account for design features. In other words, it seems that, accounting for predictors that are highly predictive of the Y-variable, seems to be the preponderant factor with respect to the efficiency aspect of model-assisted estimators. We conjecture that fully accounting for the sampling design will likely lead to additional efficiency gains but that the predictive power of the model likely constitutes the "determining factor". Developing machine learning procedures that fully account for the sampling design is currently under investigation.

2.6 Supplementary material

Result 3.1. *Assume (H1)-(H5). Consider a sequence of GREG estimators* $\{\widehat{t}_{greg}\}_{v\in\mathbb{N}}$ *of* t_y*. Then,*

$$\frac{1}{N_v}(\widehat{t}_{\text{greg}} - t_y) = O_{\text{p}}\left(\sqrt{\frac{p_v^3}{n_v}}\right).$$

If the numbers of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ *and the sample sizes* $\{n_v\}_{v\in\mathbb{N}}$ *satisfy* $p_v^3/n_v = o(1)$*, then* $N_v^{-1}(\widehat{t}_{greg} - t_y) = o_{\text{p}}(1)$*.*

Proof. We adapt the proof of Robinson and Särndal (1983) to a high-dimensional setting. Let I_i be the sample membership indicator for unit i such that $I_i = 1$ if $i \in S$ and $I_i = 0$, otherwise. Let $\alpha_i := I_i/\pi_i - 1$ for all $i \in U_v$. We consider the following decomposition:

$$\frac{1}{N_v}\left(\widehat{t}_{\text{greg}} - t_y\right) = \frac{1}{N_v}\sum_{i\in U_v}\alpha_i y_i - \sum_{j=1}^{p_v} b_j\widehat{\beta}_j, \tag{2.17}$$

where $b_j = \frac{1}{N_v}\sum_{i\in U_v}\alpha_i x_{ij}$ for $j = 1, 2, ..., p_v$. Now, the first term does not depend on the auxiliary information and we have (Breidt and Opsomer, 2000, Robinson and Särndal, 1983):

$$\mathbb{E}_{\text{p}}\left(\frac{1}{N_v}\sum_{i\in U_v}\alpha_i y_i\right)^2 = \frac{1}{N_v^2}\sum_{i\in U} y_i^2\cdot\mathbb{E}_{\text{p}}(\alpha_i^2) + \frac{1}{N_v^2}\sum_{i\in U_v}\sum_{\ell\in U_v,\ell\neq i} y_i y_\ell\cdot\mathbb{E}_{\text{p}}(\alpha_i\alpha_\ell). \tag{2.18}$$

We have $\mathbb{E}_{\text{p}}(\alpha_i^2) = (1-\pi_i)/\pi_i \leqslant 1/c$ and for $i \neq \ell$, $\mathbb{E}_{\text{p}}(\alpha_i\alpha_\ell) = (\pi_{i\ell} - \pi_i\pi_\ell)/\pi_i\pi_\ell \leqslant \max_{i,\ell\in U_v, i\neq\ell}|\pi_{i\ell} - \pi_i\pi_\ell|/c^2$ by Assumption (H3). It follows from (H1), (H2) and (H3) that

$$\begin{aligned}\mathbb{E}_{\text{p}}\left(\frac{1}{N_v}\sum_{i\in U_v}\alpha_i y_i\right)^2 &\leqslant \frac{1}{cN_v^2}\sum_{i\in U_v} y_i^2 + \frac{n_v\max_{i,\ell\in U_v,i\neq\ell}|\pi_{i\ell} - \pi_i\pi_\ell|}{c^2 n_v N_v^2}\sum_{i\in U_v}\sum_{\ell\in U_v,\ell\neq i}|y_i y_\ell| \\ &\leqslant \left(\frac{1}{cN_v} + \frac{n_v\max_{i,\ell\in U_v,i\neq\ell}|\pi_{i\ell} - \pi_i\pi_\ell|}{c^2 n_v}\right)\frac{1}{N_v}\sum_{i\in U_v} y_i^2 = O\left(\frac{1}{n_v}\right)\end{aligned} \tag{2.19}$$

and so,

$$\left|\frac{1}{N_v}\sum_{i\in U_v}\alpha_i y_i\right| = O_{\text{p}}\left(\frac{1}{\sqrt{n_v}}\right). \tag{2.20}$$

Now, consider the second term from the right-side of (2.17):

$$\left|\sum_{j=1}^{p_v}\widehat{\beta}_j b_j\right| \leqslant \sqrt{\left(\sum_{j=1}^{p_v}\widehat{\beta}_j^2\right)\left(\sum_{j=1}^{p_v} b_j^2\right)} = ||\widehat{\boldsymbol{\beta}}||_2\sqrt{\sum_{j=1}^{p_v} b_j^2} \leqslant ||\widehat{\boldsymbol{\beta}}||_1\sqrt{\sum_{j=1}^{p_v} b_j^2}. \tag{2.21}$$

By Assumption (H5), we have that $||\widehat{\boldsymbol{\beta}}||_1 = O_{\mathrm{p}}(p_v)$. Furthermore,

$$\sqrt{\sum_{j=1}^{p_v} b_j^2} = \frac{1}{N_v}\left\|\sum_{i\in U_v}\alpha_i \mathbf{x}_i\right\|_2$$

and

$$\begin{aligned}
\frac{1}{N_v^2}\mathbb{E}_{\mathrm{p}}\left\|\sum_{i\in U_v}\alpha_i\mathbf{x}_i\right\|_2^2 &= \frac{1}{N_v^2}\sum_{i\in U_v}||\mathbf{x}_i||_2^2\mathbb{E}_{\mathrm{p}}(\alpha_i^2) + \frac{1}{N_v^2}\sum_{i\in U_v}\sum_{\ell\neq i\in U_v}\mathbf{x}_i^\top\mathbf{x}_\ell\mathbb{E}_{\mathrm{p}}(\alpha_i\alpha_\ell)\\
&\leqslant \frac{1}{cN_v^2}\sum_{i\in U_v}||\mathbf{x}_i||_2^2 + \frac{n_v\max_{i,\ell\in U_v,i\neq\ell}|\pi_{i\ell}-\pi_i\pi_\ell|}{c^2n_vN_v^2}\sum_{i\in U_v}\sum_{\ell\neq i\in U_v}|\mathbf{x}_i^\top\mathbf{x}_\ell|\\
&\leqslant \left(\frac{1}{cN_v} + \frac{n_v\max_{i,\ell\in U_v,i\neq\ell}|\pi_{i\ell}-\pi_i\pi_\ell|}{c^2n_v}\right)\frac{1}{N_v}\sum_{i\in U_v}||\mathbf{x}_i||_2^2\\
&= O\left(\frac{p_v}{n_v}\right), \qquad (2.22)
\end{aligned}$$

by Assumptions (H2)-(H4). It follows that

$$\sqrt{\sum_{j=1}^{p_v} b_j^2} = O_{\mathrm{p}}\left(\sqrt{\frac{p_v}{n_v}}\right). \qquad (2.23)$$

The result follows by using (2.17), (2.20), (2.21), (2.23) and Assumption (H5):

$$\frac{1}{N_v}\left|\widehat{t}_{\mathrm{greg}} - t_y\right| \leqslant \left|\frac{1}{N_v}\sum_{i\in U_v}\alpha_i y_i\right| + \left|\sum_{j=1}^{p_v}\widehat{\beta}_j b_j\right| = O_{\mathrm{p}}\left(\frac{1}{\sqrt{n_v}}\right) + O_{\mathrm{p}}\left(\sqrt{\frac{p_v^3}{n_v}}\right) = O_{\mathrm{p}}\left(\sqrt{\frac{p_v^3}{n_v}}\right).$$

■

Result 3.2. *Assume (H1)-(H5). Consider a sequence of penalized model-assisted estimators* $\{\widehat{t}_{\mathrm{pen}}\}_{v\in\mathbb{N}}$ *of* t_y *obtained by either ridge, lasso or elastic-net. Then,*

$$\frac{1}{N_v}(\widehat{t}_{\mathrm{pen}} - t_y) = O_{\mathrm{p}}\left(\sqrt{\frac{p_v^3}{n_v}}\right).$$

If the numbers of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ *and the sample sizes* $\{n_v\}_{v\in\mathbb{N}}$ *satisfy* $p_v^3/n_v = o(1)$, *then* $N_v^{-1}(\widehat{t}_{pen} - t_y) = o_{\mathrm{p}}(1)$.

Proof. From the proof of result (3.1), we only need to show that $||\hat{\boldsymbol{\beta}}_{\mathrm{pen}}||_2 = O_{\mathrm{p}}(p_v)$ or $||\hat{\boldsymbol{\beta}}_{\mathrm{pen}}||_1 = O_{\mathrm{p}}(p_v)$, where $\hat{\boldsymbol{\beta}}_{\mathrm{pen}}$ is one of the penalized regression coefficient: ridge, lasso and elastic-net. Consider first the ridge regression coefficient, $\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}$. The ridge regression estimator has the advantage of having an explicit expression. We will show that $||\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}||_2 < ||\hat{\boldsymbol{\beta}}||_2$ for $\lambda > 0$. Let denote $\hat{T}_\lambda = \boldsymbol{X}_{S_v}^\top\boldsymbol{\Pi}_{S_v}^{-1}\boldsymbol{X}_{S_v} + \lambda\boldsymbol{I}_{p_v} = \sum_{i\in S_v}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\pi_i} + \lambda\boldsymbol{I}_{p_v}$ sample counterpart of $T_\lambda = \boldsymbol{X}_{U_v}^\top\boldsymbol{X}_{U_v} + \lambda\boldsymbol{I}_{p_v} = \sum_{i\in U_v}\mathbf{x}_i\mathbf{x}_i^\top + \lambda\boldsymbol{I}_{p_v}$. Moreover, let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_{p_v}$ be the eigenvalues of $\sum_{i\in S_v}\mathbf{x}_i\mathbf{x}_i^\top/\pi_i$ in decreasing order and $\hat{\mathbf{v}}_j$ the orthonormal corresponding eigenvectors,

$j = 1, \ldots, p_v$. Then, the eigenvalues of the matrix $\hat{T}_\lambda$ are $\hat{\lambda}_1 + \lambda \geq \hat{\lambda}_2 + \lambda \geq \ldots \geq \hat{\lambda}_{p_v} + \lambda \geq \lambda > 0$ with the same eigenvectors $\hat{\mathbf{v}}_j, j = 1, \ldots, p_v$. Using the same arguments as those used in Hoerl and Kennard (1970), we obtain $\hat{\boldsymbol{\beta}}_{\text{ridge}} = \sum_{j=1}^{p_v} (\hat{\lambda}_j + \lambda)^{-1} \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j^\top \boldsymbol{X}_{S_v}^\top \boldsymbol{\Pi}_{S_v}^{-1} \mathbf{y}_{S_v}$ and $\hat{\boldsymbol{\beta}} = \sum_{j=1}^{p_v} (\hat{\lambda}_j)^{-1} \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j^\top \boldsymbol{X}_{S_v}^\top \boldsymbol{\Pi}_{S_v}^{-1} \mathbf{y}_{S_v}$. Let denote by $c_j = \hat{\mathbf{v}}_j^\top \boldsymbol{X}_{S_v}^\top \boldsymbol{\Pi}_{S_v}^{-1} \mathbf{y}_{S_v} \in \mathbf{R}$, then

$$||\hat{\boldsymbol{\beta}}_{\text{ridge}}||_2^2 = \sum_{j=1}^{p_v} \frac{c_j^2}{(\hat{\lambda}_j + \lambda)^2} < ||\hat{\boldsymbol{\beta}}||_2^2 = \sum_{j=1}^{p_v} \frac{c_j^2}{(\hat{\lambda}_j)^2} \quad \text{for} \quad \lambda > 0.$$

It follows that $||\hat{\boldsymbol{\beta}}_{\text{ridge}}||_2 < ||\hat{\boldsymbol{\beta}}||_2 \leq ||\hat{\boldsymbol{\beta}}||_1 = O_{\text{p}}(p_v)$ and we get $||\hat{\boldsymbol{\beta}}_{\text{ridge}}||_2 = O_{\text{p}}(p_v)$.
We now consider the lasso regression estimator, $\hat{\boldsymbol{\beta}}_{\text{lasso}}$, which minimizes the design-based version of the optimization problem given in (13) in the main article:

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i \in S_v} \frac{1}{\pi_i} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda ||\boldsymbol{\beta}||_1.$$

The lasso-estimator $\hat{\boldsymbol{\beta}}_{\text{lasso}}$ may be also obtained as the solution of a constrained optimization problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i \in S_v} \frac{1}{\pi_i} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

under the constraint

$$||\boldsymbol{\beta}||_1 \leq C,$$

for some small enough constant $C > 0$. If the ordinary least-square estimator $\widehat{\boldsymbol{\beta}}$ satisfies the constraint, namely if $||\widehat{\boldsymbol{\beta}}||_1 \leq C$, then the solution of the constrained optimization problem is $\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \widehat{\boldsymbol{\beta}}$; otherwise, if $||\widehat{\boldsymbol{\beta}}||_1 > C$, then the solution $\widehat{\boldsymbol{\beta}}_{\text{lasso}}$ will be different from the least-square estimator $\widehat{\boldsymbol{\beta}}$ and $||\widehat{\boldsymbol{\beta}}_{\text{lasso}}||_1 \leq C < ||\widehat{\boldsymbol{\beta}}||_1$. So, in both cases, we have $||\widehat{\boldsymbol{\beta}}_{\text{lasso}}||_1 \leq ||\widehat{\boldsymbol{\beta}}||_1 = O_{\text{p}}(p_v)$.
Finally, consider the elastic-net regression estimator, $\hat{\boldsymbol{\beta}}_{\text{en}}$. Consider the following objective functions:

$$\mathcal{L}_{ols}(\boldsymbol{\beta}) = \sum_{i \in S_v} \frac{1}{\pi_i} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

$$\mathcal{L}_{en}(\boldsymbol{\beta}) = \sum_{i \in S_v} \frac{1}{\pi_i} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||_2^2 = \mathcal{L}_{ols}(\boldsymbol{\beta}) + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||_2^2,$$

where $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1 - \alpha)$ with $\lambda > 0$ and $\alpha \in (0, 1)$. The cases $\alpha = 0$ and $\alpha = 1$ lead, respectively, to the ridge and lasso regression estimators which have been discussed above. The ordinary least squares estimator $\widehat{\boldsymbol{\beta}}$ verifies $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{L}_{ols}(\boldsymbol{\beta})$ and the elastic-net estimator verifies $\widehat{\boldsymbol{\beta}}_{en} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{L}_{en}(\boldsymbol{\beta})$. Since $\widehat{\boldsymbol{\beta}}$ minimizes $\mathcal{L}_{ols}(\boldsymbol{\beta})$, we have $\mathcal{L}_{ols}(\widehat{\boldsymbol{\beta}}) \leqslant \mathcal{L}_{ols}(\widehat{\boldsymbol{\beta}}_{en})$. Similarly, we have $\mathcal{L}_{en}(\widehat{\boldsymbol{\beta}}_{en}) \leqslant \mathcal{L}_{en}(\widehat{\boldsymbol{\beta}}_{ols})$. Therefore, the following inequalities hold:

$$\begin{aligned} \mathcal{L}_{ols}(\widehat{\boldsymbol{\beta}}) + \lambda_1 ||\widehat{\boldsymbol{\beta}}_{en}||_1 + \lambda_2 ||\widehat{\boldsymbol{\beta}}_{en}||_2^2 &\leqslant \mathcal{L}_{ols}(\widehat{\boldsymbol{\beta}}_{en}) + \lambda_1 ||\widehat{\boldsymbol{\beta}}_{en}||_1 + \lambda_2 ||\widehat{\boldsymbol{\beta}}_{en}||_2^2 = \mathcal{L}_{en}(\widehat{\boldsymbol{\beta}}_{en}) \\ &\leqslant \mathcal{L}_{ols}(\widehat{\boldsymbol{\beta}}) + \lambda_1 ||\widehat{\boldsymbol{\beta}}||_1 + \lambda_2 ||\widehat{\boldsymbol{\beta}}||_2^2 = \mathcal{L}_{en}(\widehat{\boldsymbol{\beta}}_{ols}), \end{aligned}$$

which implies

$$\lambda_1||\widehat{\boldsymbol{\beta}}_{en}||_1 + \lambda_2||\widehat{\boldsymbol{\beta}}_{en}||_2^2 \leqslant \lambda_1||\widehat{\boldsymbol{\beta}}||_1 + \lambda_2||\widehat{\boldsymbol{\beta}}||_2^2. \tag{2.24}$$

Furthermore, since $\lambda_1 > 0$, we can write

$$\lambda_2||\widehat{\boldsymbol{\beta}}_{en}||_2^2 \leqslant \lambda_1||\widehat{\boldsymbol{\beta}}_{en}||_1 + \lambda_2||\widehat{\boldsymbol{\beta}}_{en}||_2^2. \tag{2.25}$$

Using (2.24), (2.25) and the fact that $||\widehat{\boldsymbol{\beta}}||_2 \leqslant||\widehat{\boldsymbol{\beta}}||_1$, we obtain

$$\lambda_2||\widehat{\boldsymbol{\beta}}_{en}||_2^2 \leqslant \lambda_1||\widehat{\boldsymbol{\beta}}||_1 + \lambda_2||\widehat{\boldsymbol{\beta}}||_2^2 \leqslant \lambda_1||\widehat{\boldsymbol{\beta}}||_1 + \lambda_2||\widehat{\boldsymbol{\beta}}||_1^2$$

which implies

$$||\widehat{\boldsymbol{\beta}}_{en}||_2^2 \leqslant \frac{\alpha}{1-\alpha}||\widehat{\boldsymbol{\beta}}||_1 + ||\widehat{\boldsymbol{\beta}}||_1^2 = O_{\mathrm{p}}(p_v^2)$$

and so, $||\widehat{\boldsymbol{\beta}}_{en}||_2 = O_{\mathrm{p}}(p_v)$.

■

Result 3.3. *Assume (H1)-(H4). Also, assume that there exists a positive constant $\tilde{C}$ such that $\lambda_{max}(\boldsymbol{X}_{U_v}^\top \boldsymbol{X}_{U_v}) \leqslant \tilde{C}N_v$, where $\lambda_{max}(\boldsymbol{X}_{U_v}^\top \boldsymbol{X}_{U_v})$ is the largest eigenvalue of $\boldsymbol{X}_{U_v}^\top \boldsymbol{X}_{U_v}$. Assume also that $N_v/\lambda_v = O(1)$.*

1. *Then, there exists a positive constant C such that $\mathbb{E}_p\left[||\widehat{\boldsymbol{\beta}}_{\text{ridge}}||_2^2\right] \leqslant C$ and*

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{ridge}} - t_y\right| = O\left(\sqrt{\frac{p_v}{n_v}}\right).$$

   If the numbers of auxiliary variables $\{p_v\}_{v\in\mathbb{N}}$ and the sample sizes $\{n_v\}_{v\in\mathbb{N}}$ satisfy $p_v/n_v = o(1)$, then $N_v^{-1}\mathbb{E}_p|\widehat{t}_{\text{ridge}} - t_y| = o(1)$.

2. *$\mathbb{E}_{\mathrm{p}}(||\widehat{\boldsymbol{\beta}}_{\text{ridge}} - \tilde{\boldsymbol{\beta}}_{\text{ridge}}||_2^2) = O(p_v/n_v)$. Thus, if $p_v/n_v = o(1)$, then $\mathbb{E}_{\mathrm{p}}(||\widehat{\boldsymbol{\beta}}_{\text{ridge}} - \tilde{\boldsymbol{\beta}}_{\text{ridge}}||_2^2) = o(1)$.*

3. *We have the following asymptotic equivalence:*

$$\frac{1}{N_v}\left(\widehat{t}_{\text{ridge}} - t_y\right) = \frac{1}{N_v}\left(\widehat{t}_{\text{diff,}^\sim} - t_y\right) + O_{\mathrm{p}}\left(\frac{p_v}{n_v}\right),$$

   where

$$\widehat{t}_{\text{diff,}^\sim} = \sum_{i\in S_v} y_i/\pi_i - \left(\sum_{i\in S_v}\mathbf{x}_i/\pi_i - \sum_{i\in U_v}\mathbf{x}_i\right)^\top \tilde{\boldsymbol{\beta}}_{\text{ridge}}$$

   and

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{ridge}} - t_y\right| = O\left(\frac{1}{\sqrt{n_v}}\right) + O\left(\frac{p_v}{n_v}\right).$$

If $p_v = O(n_v^a)$ *with* $0 \leq a < 1/2$, *then*

$$\frac{1}{N_v}\left(\widehat{t}_{\text{ridge}} - t_y\right) = \frac{1}{N_v}\left(\widehat{t}_{\text{diff},\nu} - t_y\right) + o_{\text{p}}(1)$$

and

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{ridge}} - t_y\right| = O\left(\frac{1}{\sqrt{n_v}}\right).$$

Proof. 1. As in the proof of result (3.2), we consider the eigenvalues of the matrix $\hat{T}_\lambda$ in decreasing order: $\hat{\lambda}_1 + \lambda \geq \hat{\lambda}_2 + \lambda \geq \ldots \geq \hat{\lambda}_{p_v} + \lambda \geq \lambda > 0$. The matrix $\hat{T}_\lambda$ is always invertible and the eigenvalues of $\hat{T}_\lambda^{-1}$ are $0 < (\hat{\lambda}_1 + \lambda)^{-1} \leq (\hat{\lambda}_2 + \lambda)^{-1} \leq \ldots \leq (\hat{\lambda}_{p_v} + \lambda)^{-1} \leq \lambda^{-1}$. We then obtain

$$||\hat{T}_\lambda^{-1}||_2 \leqslant \lambda^{-1}, \tag{2.26}$$

where $||\cdot||_2$ is the spectral norm matrix defined for a squared $p \times p$ matrix $\mathbf{A}$ as $||\mathbf{A}||_2 = \sup_{\mathbf{x}\in\mathbf{R}^p, ||\mathbf{x}||_2\neq 0} ||\mathbf{A}\mathbf{x}||_2/||\mathbf{x}||_2$. For a symmetric and positive definite matrix $\mathbf{A}$, we have that $||\mathbf{A}||_2 = \lambda_{max}(\mathbf{A})$, where $\lambda_{max}(\mathbf{A})$ is the largest eigenvalue of $\mathbf{A}$. Now, we can write

$$\begin{aligned}\frac{1}{N_v^2}\left\|\sum_{i\in S_v}\frac{\mathbf{x}_i y_i}{\pi_i}\right\|_2^2 &= \frac{1}{N_v^2}\sum_{i\in U_v}\sum_{\ell\in U_v}\mathbf{x}_i^\top\mathbf{x}_\ell\frac{y_i I_i}{\pi_i}\frac{y_\ell I_\ell}{\pi_\ell} = \frac{1}{N_v^2}\boldsymbol{\mathcal{Y}}^\top\boldsymbol{X}_{U_v}\boldsymbol{X}_{U_v}^\top\boldsymbol{\mathcal{Y}} \\ &\leqslant \frac{1}{N_v}||\boldsymbol{\mathcal{Y}}||_2^2\frac{1}{N_v}||\boldsymbol{X}_{U_v}\boldsymbol{X}_{U_v}^\top||_2,\end{aligned}$$

where $\boldsymbol{\mathcal{Y}}^\top = \left(\frac{y_i I_i}{\pi_i}\right)_{i\in U_v}$. The symmetric and positive semi-definite $N_v \times N_v$ matrix $\boldsymbol{X}_{U_v}\boldsymbol{X}_{U_v}^\top$ has the same non-null eigenvalues as those of the positive definite $p_v \times p_v$ matrix $\boldsymbol{X}_{U_v}^\top\boldsymbol{X}_{U_v}$,. Therefore,

$$\frac{1}{N_v}||\boldsymbol{X}_{U_v}\boldsymbol{X}_{U_v}^\top||_2 = \frac{1}{N_v}\lambda_{max}(\boldsymbol{X}_{U_v}^\top\boldsymbol{X}_{U_v}) \leqslant \tilde{C}.$$

Using Assumptions (H1) and (H3), we have

$$\frac{1}{N_v^2}\left\|\sum_{i\in S_v}\frac{\mathbf{x}_i y_i}{\pi_i}\right\|_2^2 \leqslant \frac{\tilde{C}}{N_v}||\boldsymbol{\mathcal{Y}}||_2^2 = \frac{\tilde{C}}{N_v}\sum_{i\in U_v}\frac{y_i^2 I_i}{\pi_i^2} \leqslant \frac{\tilde{C}}{c^2 N_v}\sum_{i\in U_v}y_i^2 = O(1).$$

Finally, using also the fact that $N_v/\lambda = O(1)$, we have

$$||\widehat{\boldsymbol{\beta}}_{\text{ridge}}||_2^2 \leqslant ||\hat{T}_\lambda^{-1}||_2^2\left\|\sum_{i\in S_v}\frac{\mathbf{x}_i y_i}{\pi_i}\right\|_2^2 \leqslant N_v^2\lambda^{-2}\left\|\frac{1}{N_v^2}\sum_{i\in S_v}\frac{\mathbf{x}_i y_i}{\pi_i}\right\|_2^2 = O(1).$$

It follows that

$$\mathbb{E}_{\text{p}}\left[||\widehat{\boldsymbol{\beta}}_{\text{ridge}}||_2^2\right] = O(1). \tag{2.27}$$

To obtain the L^1 design-consistency of the ridge model-assisted estimator, we write as in the proof of Result 3.1:

$$\begin{aligned}\frac{1}{N_v}\left(\widehat{t}_{\text{ridge}} - t_y\right) &= \frac{1}{N_v}\sum_{i\in U_v}\alpha_i y_i - \sum_{j=1}^{p_v} b_j \widehat{\beta}_{j,\text{ridge}} \\ &= \frac{1}{N_v}\sum_{i\in U_v}\alpha_i y_i - \frac{1}{N_v}\left(\sum_{i\in U_v}\alpha_i \mathbf{x}_i\right)^\top \widehat{\boldsymbol{\beta}}_{\text{ridge}}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\text{p}}\left|\frac{1}{N_v}\left(\widehat{t}_{\text{ridge}} - t_y\right)\right| &\leqslant \mathbb{E}_{\text{p}}\left|\frac{1}{N_v}\sum_{i\in U_v}\alpha_i y_i\right| + \sqrt{\mathbb{E}_{\text{p}}\left(\frac{1}{N_v^2}\left\|\sum_{i\in U_v}\alpha_i \mathbf{x}_i\right\|_2^2\right)\mathbb{E}_{\text{p}}||\widehat{\boldsymbol{\beta}}_{\text{ridge}}||_2^2} \\ &= O\left(\sqrt{\frac{1}{n_v}}\right) + O\left(\sqrt{\frac{p_v}{n_v}}\right) = O\left(\sqrt{\frac{p_v}{n_v}}\right)\end{aligned}$$

by (2.19), (2.22), (2.27).

2. We can write

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} - \widetilde{\boldsymbol{\beta}}_{\text{ridge}} = \widehat{T}_\lambda^{-1}\left(\sum_{i\in S_v}\frac{E_{i\lambda}}{\pi_i} - \sum_{i\in U_v} E_{i\lambda}\right), \tag{2.28}$$

where $E_{i\lambda} = \mathbf{x}_i(y_i - \mathbf{x}_i^\top\tilde{\boldsymbol{\beta}}_{\text{ridge}})$ with $\sum_{i\in U_v} E_{i\lambda} = \lambda \boldsymbol{I}_{p_v}\tilde{\boldsymbol{\beta}}_{\text{ridge}}$. Using the same arguments as those used in the proof of Result 3.1, we get

$$\frac{1}{N_v^2}\mathbb{E}_{\text{p}}\left\|\sum_{i\in S_v}\frac{E_{i\lambda}}{\pi_i} - \sum_{i\in U_v} E_{i\lambda}\right\|_2^2 \leqslant \left(\frac{1}{cN_v} + \frac{n_v \max_{i,\ell\in U_v, i\neq\ell}|\pi_{i\ell} - \pi_i\pi_\ell|}{c^2 n_v}\right)\frac{1}{N_v}\sum_{i\in U_v}||E_{i\lambda}||_2^2. \tag{2.29}$$

Furthermore,

$$\frac{1}{N_v}\sum_{i\in U_v}||E_{i\lambda}||_2^2 \leqslant \frac{2C_2 p_v}{N_v}\left(\sum_{i\in U_v} y_i^2 + \sum_{i\in U_v}(\mathbf{x}_i^\top\tilde{\boldsymbol{\beta}}_{\text{ridge}})^2\right) = O(p_v) \tag{2.30}$$

by Assumptions (H1) and (H4) and the fact that

$$\frac{1}{N_v}\sum_{i\in U_v}(\mathbf{x}_i^\top\tilde{\boldsymbol{\beta}}_{\text{ridge}})^2 = \tilde{\boldsymbol{\beta}}_{\text{ridge}}^\top\left(\frac{1}{N_v}\sum_{i\in U_v}\mathbf{x}_i\mathbf{x}_i^\top\right)\tilde{\boldsymbol{\beta}}_{\text{ridge}} \leqslant ||\tilde{\boldsymbol{\beta}}_{\text{ridge}}||_2^2\frac{1}{N_v}||\boldsymbol{X}_{U_v}^\top\boldsymbol{X}_{U_v}||_2 = O(1). \tag{2.31}$$

To obtain the above inequality, we have also used the fact that $||\tilde{\boldsymbol{\beta}}_{\text{ridge}}||_2 = O(1)$ which can be proved by using the same arguments as the ones used for showing that $||\widehat{\boldsymbol{\beta}}_{\text{ridge}}||_2 = O(1)$ in point (1). Expressions (2.29) and (2.30) lead to

$$\frac{1}{N_v^2}\mathbb{E}_{\text{p}}\left\|\sum_{i\in S_v}\frac{E_{i\lambda}}{\pi_i}-\sum_{i\in U_v}E_{i\lambda}\right\|_2^2 = O\left(\frac{p_v}{n_v}\right). \tag{2.32}$$

The result follows from (2.28), (2.32) and the fact that $||N_v\hat{T}_\lambda^{-1}||_2 = O(1)$:

$$\mathbb{E}_{\text{p}}||\widehat{\boldsymbol{\beta}}_{\text{ridge}}-\tilde{\boldsymbol{\beta}}_{\text{ridge}}||_2^2 = O\left(\frac{p_v}{n_v}\right). \tag{2.33}$$

3. We use the following decomposition:

$$\frac{1}{N_v}\left(\widehat{t}_{\text{ridge}}-t_y\right) = \frac{1}{N_v}\left(\widehat{t}_{\text{diff,}\check{}}-t_y\right) - \frac{1}{N_v}\left(\sum_{i\in S_v}\frac{\mathbf{x}_i}{\pi_i}-\sum_{i\in U_v}\mathbf{x}_i\right)^\top\left(\widehat{\boldsymbol{\beta}}_{\text{ridge}}-\tilde{\boldsymbol{\beta}}_{\text{ridge}}\right),$$

and

$$\begin{aligned}\frac{1}{N_v}\left(\widehat{t}_{\text{diff,}\check{}}-t_y\right) &= \frac{1}{N_v}\left(\sum_{i\in S_v}\frac{y_i}{\pi_i}-\sum_{i\in U_v}y_i\right)-\frac{1}{N_v}\left(\sum_{i\in S_v}\frac{\mathbf{x}_i}{\pi_i}-\sum_{i\in U_v}\mathbf{x}_i\right)^\top\tilde{\boldsymbol{\beta}}_{\text{ridge}}\\ &= \frac{1}{N_v}\sum_{i\in U_v}\alpha_i y_i-\frac{1}{N_v}\sum_{i\in U_v}\alpha_i\mathbf{x}_i^\top\tilde{\boldsymbol{\beta}}_{\text{ridge}},\end{aligned}$$

where $\alpha_i = I_i/\pi_i - 1, i \in U_v$. From (2.19), we have that $N_v^{-2}\mathbb{E}_{\text{p}}(\sum_{i\in U_v}\alpha_i y_i)^2 = O(n_v^{-1})$ and we can get $N_v^{-2}\mathbb{E}_{\text{p}}\left(\sum_{i\in U_v}\alpha_i\mathbf{x}_i^\top\tilde{\boldsymbol{\beta}}_{\text{ridge}}\right)^2 = O(n_v^{-1})$ by using similar arguments as those used in the proof of Result 3.1 and (2.31). We obtain

$$\frac{1}{N_v^2}\mathbb{E}_{\text{p}}\left(\widehat{t}_{\text{diff,}\check{}}-t_y\right)^2 = O\left(\frac{1}{n_v}\right).$$

The result follows since

$$\begin{aligned}\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{ridge}}-t_y\right| &\leqslant \frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{\text{diff,}\check{}}-t_y\right| + \sqrt{\frac{1}{N_v^2}\mathbb{E}_p\left\|\sum_{i\in S_v}\frac{\mathbf{x}_i}{\pi_i}-\sum_{i\in U_v}\mathbf{x}_i\right\|_2^2\mathbb{E}_p\left\|\widehat{\boldsymbol{\beta}}_{\text{ridge}}-\tilde{\boldsymbol{\beta}}_{\text{ridge}}\right\|_2^2}\\ &= O\left(\frac{1}{\sqrt{n_v}}\right)+O\left(\frac{p_v}{n_v}\right)\end{aligned}$$

by using (2.22) and (2.33).

■

Proposition 3.1. *Suppose assumptions (H1)-(H3) and that the sampling design and the X-variables are such that the columns of* $\boldsymbol{\Pi}_{S_v}^{-1/2}\mathbf{X}_{S_v}$ *are orthogonal. Suppose also that there exist positive quantities* C_3 *and* C_4 *such that* $max_{j=1,\ldots,p_v}N_v^{-1}\sum_{i\in U_v}x_{ij}^4 \leq C_3 < \infty$

and $min_{j=1,\ldots,p_v} N_v^{-1} \sum_{i\in U_v} x_{ij}^2 \geq C_4 > 0$. *Then,* $N_v^{-1}(\widehat{t}_{\text{greg}} - t_y) = O_{\text{p}}(\sqrt{p_v/n_v})$ *and* $N_v^{-1}(\widehat{t}_{\text{pen}} - t_y) = O_{\text{p}}(\sqrt{p_v/n_v})$, *where* $\widehat{t}_{\text{pen}}$ *denotes either the lasso or the elastic-net estimator.*

Proof. From the proof of Result 3.1 (more specifically, Equations 2.21 and 2.22), we need to show that $\sum_{i\in U_v} ||\mathbf{x}_i||_2^2/N_v = O(p_v)$ and that $||\widehat{\boldsymbol{\beta}}||_2 = O_{\text{p}}(1)$. The same result holds for $\widehat{\boldsymbol{\beta}}_{\text{lasso}}$ and $\widehat{\boldsymbol{\beta}}_{\text{en}}$. We have $\sum_{i\in U_v} ||\mathbf{x}_i||_2^2/N_v = \sum_{j=1}^{p_v} \sum_{i\in U_v} x_{ij}^2/N_v \leq p_v\sqrt{C_3} = O(p_v)$ under the assumption of uniformly bounded forth moment of X_j, $j = 1, \ldots, p_v$.

We first show that, under the assumed orthogonality condition, $||\widehat{\boldsymbol{\beta}}_{\text{lasso}}||_2 \leq ||\widehat{\boldsymbol{\beta}}||_2$, $||\widehat{\boldsymbol{\beta}}_{\text{en}}||_2 \leq ||\widehat{\boldsymbol{\beta}}||_2$ and also $||\widehat{\boldsymbol{\beta}}||_2 = O_{\text{p}}(1)$.
Consider again the objective function $\mathcal{L}_{ols}(\boldsymbol{\beta})$ as in the proof of Result 3.2. We can write

$$\mathcal{L}_{ols}(\boldsymbol{\beta}) = \sum_{i\in S_v} \frac{1}{\pi_i}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2 = \sum_{i\in S_v}(\tilde{y}_i - \tilde{\mathbf{x}}_i^\top\boldsymbol{\beta})^2 \tag{2.34}$$

where $\tilde{y}_i = y_i/\sqrt{\pi_i}$ and $\tilde{\mathbf{x}}_i = (\tilde{x}_{ij})_{j=1}^{p_v} = \mathbf{x}_i/\sqrt{\pi_i}$ for all $i \in S_v$. Let $\tilde{\mathbf{X}}_{S_v} = \boldsymbol{\Pi}_{S_v}^{-1/2}\mathbf{X}_{S_v} = (\tilde{\mathbf{x}}_i^\top)_{i\in S_v} = (\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_{p_v})$. The columns of $\tilde{\mathbf{X}}_{S_v}$, denoted by $\tilde{\mathbf{X}}_j$, $j = 1, \ldots, p_v$ are assumed to be orthogonal. This means that $\tilde{\mathbf{X}}_j^\top\tilde{\mathbf{X}}_k = 0$ for $j \neq k$. The ordinary least-square estimator $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}_{S_v}^\top\tilde{\mathbf{X}}_{S_v})^{-1}\tilde{\mathbf{X}}_{S_v}^\top\tilde{\mathbf{y}}_{S_v}.$$

Under the orthogonality condition, $\tilde{\mathbf{X}}_{S_v}^\top\tilde{\mathbf{X}}_{S_v}$ is a diagonal matrix with diagonal elements given by $||\tilde{\mathbf{X}}_j||_2^2 = \sum_{i\in S_v} \tilde{x}_{ij}^2 = \sum_{i\in S_v} \frac{x_{ij}^2}{\pi_i}$, which corresponds to the Horvitz-Thompson estimator of $\sum_{i\in U_v} x_{ij}^2$. Therefore, $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_j)_{j\in S_v}$ and the j-th coordinate is given by $\hat{\beta}_j = (\sum_{i\in S_v} \tilde{x}_{ij}^2)^{-1} \sum_{i\in S_v} \tilde{x}_{ij}\tilde{y}_i$.

The lasso estimator $\hat{\boldsymbol{\beta}}_{\text{lasso}} = (\hat{\beta}_{j,\text{lasso}})_{j=1}^{p_v}$ as well as the elastic-net estimator $\hat{\boldsymbol{\beta}}_{\text{en}} = (\hat{\beta}_{j,\text{en}})_{j=1}^{p_v}$ are obtained by using the cyclic soft-thresholding algorithm (Hastie et al., 2011):

$$\hat{\beta}_{j,\text{lasso}} = \frac{\mathcal{S}_\lambda(\sum_{i\in S_v} r_{ij}\tilde{x}_{ij})}{\sum_{i\in S_v} \tilde{x}_{ij}^2}$$

and

$$\hat{\beta}_{j,\text{en}} = \frac{\mathcal{S}_{\lambda\alpha}(\sum_{i\in S_v} r_{ij}\tilde{x}_{ij})}{\sum_{i=1}^{n_v} \tilde{x}_{ij}^2 + \lambda(1-\alpha)},$$

where $r_{ij} = \tilde{y}_i - \sum_{k\neq j} \tilde{x}_{ik}\hat{\beta}_k$ and $\mathcal{S}_\lambda(z) = sign(z)(|z| - \lambda)_+$ is the soft-thresholding function with $(|z| - \lambda)_+ = |z| - \lambda$ if $|z| \geq \lambda$, and zero otherwise. If the columns of $\tilde{\mathbf{X}}_{S_v}$ are orthogonal, then $\sum_{i\in S_v} r_{ij}\tilde{x}_{ij} = \sum_{i\in S_v} \tilde{x}_{ij}\tilde{y}_i$ and $\hat{\beta}_{j,\text{lasso}}$ is the soft-threshold estimator of the least-square estimator $\widehat{\beta}_j$:

$$\hat{\beta}_{j,\text{lasso}} = \frac{\mathcal{S}_\lambda(\sum_{i\in S_v} \tilde{x}_{ij}\tilde{y}_i)}{\sum_{i\in S_v} \tilde{x}_{ij}^2}.$$

The elastic-net estimator is given by

$$\hat{\beta}_{j,\text{en}} = \frac{\mathcal{S}_{\lambda\alpha}(\sum_{i\in S_v} \tilde{x}_{ij}\tilde{y}_i)}{\sum_{i\in S_v} \tilde{x}_{ij}^2 + \lambda(1-\alpha)}.$$

It follows that

$$|\hat{\beta}_{j,\text{lasso}}| = \frac{|(|\sum_{i\in S_v} \tilde{x}_{ij}\tilde{y}_i| - \lambda)_+|}{\sum_{i\in S_v} \tilde{x}_{ij}^2} \leq \frac{|\sum_{i\in S_v} \tilde{x}_{ij}\tilde{y}_i|}{\sum_{i\in S_v} \tilde{x}_{ij}^2} = |\hat{\beta}_j|, \quad j = 1, \ldots, p_v$$

and $||\widehat{\boldsymbol{\beta}}_{\text{lasso}}||_2 \leq ||\widehat{\boldsymbol{\beta}}||_2$. Similarly, $||\widehat{\boldsymbol{\beta}}_{\text{en}}||_2 \leq ||\widehat{\boldsymbol{\beta}}||_2$.

We now show that $||\widehat{\boldsymbol{\beta}}||_2 = O_{\text{p}}(1)$. We have

$$||\widehat{\boldsymbol{\beta}}||_2 \leq ||N_v(\tilde{\mathbf{X}}_{S_v}^\top \tilde{\mathbf{X}}_{S_v})^{-1}||_2 \left\| \frac{1}{N_v} \tilde{\mathbf{X}}_{S_v}^\top \tilde{\mathbf{y}}_{S_v} \right\|_2 .$$

The matrix $\tilde{\mathbf{X}}_{S_v}^\top \tilde{\mathbf{X}}_{S_v}$ is diagonal with diagonal elements equal to $\sum_{i\in S_v} \frac{x_{ij}^2}{\pi_i}$. Then,

$$||N_v(\tilde{\mathbf{X}}_{S_v}^\top \tilde{\mathbf{X}}_{S_v})^{-1}||_2 = \max_{j=1,\ldots,p_v} \left(\frac{1}{N_v^{-1} \sum_{i\in S_v} \frac{x_{ij}^2}{\pi_i}} \right)$$

and for all $j = 1, \ldots, p_v$:

$$\frac{1}{N_v^{-1} \sum_{i\in S_v} \frac{x_{ij}^2}{\pi_i}} = \frac{1}{N_v^{-1} \sum_{i\in U_v} x_{ij}^2} + O_{\text{p}}\left(\frac{1}{\sqrt{n_v}}\right) = O_{\text{p}}(1)$$

by using (H2), (H3) and the assumption of uniformly bounded fourth moment of $X_j, j = 1, \ldots, p_v$. We have also used the fact that $1/(N_v^{-1} \sum_{i\in U_v} x_{ij}^2) \leq 1/(\min_{j=1,\ldots,p_v} N_v^{-1} \sum_{i\in U_v} x_{ij}^2) \leq 1/C_4 = O(1)$ for all $j = 1, \ldots, p_v$. Then,

$$||N_v(\tilde{\mathbf{X}}_{S_v}^\top \tilde{\mathbf{X}}_{S_v})^{-1}||_2 = O_{\text{p}}(1). \tag{2.35}$$

Now,

$$\left\| \frac{1}{N_v} \tilde{\mathbf{X}}_{S_v}^\top \tilde{\mathbf{y}}_{S_v} \right\|_2^2 \leq \frac{1}{N_v} \left\| \tilde{\mathbf{y}}_{S_v} \right\|_2^2 \left\| \frac{1}{N_v} \tilde{\mathbf{X}}_{S_v} \tilde{\mathbf{X}}_{S_v}^\top \right\|_2 .$$

We have

$$\left\| \frac{1}{N_v} \tilde{\mathbf{X}}_{S_v} \tilde{\mathbf{X}}_{S_v}^\top \right\|_2 = \left\| \frac{1}{N_v} \tilde{\mathbf{X}}_{S_v}^\top \tilde{\mathbf{X}}_{S_v} \right\|_2 = \max_{j=1,\ldots,p_v} \left(\frac{1}{N_v} \sum_{i\in S_v} \frac{x_{ij}^2}{\pi_i} \right) \leq \max_{j=1,\ldots,p_v} \left(\frac{1}{N_v} \sum_{i\in U_v} x_{ij}^2 \right) \leq \sqrt{C_3}$$

and

$$\frac{1}{N_v} ||\tilde{\mathbf{y}}_{S_v}||_2^2 = \frac{1}{N_v} \sum_{i\in S_v} \frac{y_i^2}{\pi_i^2} \leq \frac{1}{c^2 N_v} \sum_{i\in U_v} y_i^2 \leq \frac{C_1}{c^2}$$

by Assumption (H1). So, $||\frac{1}{N_v} \tilde{\mathbf{X}}_{S_v} \tilde{\mathbf{y}}_{S_v}||_2 = O(1)$ and combined with (2.35), we obtain $||\widehat{\boldsymbol{\beta}}||_2 = O_{\text{p}}(1)$.

■

3 MODEL-ASSISTED ESTIMATION THROUGH RANDOM FORESTS IN FINITE POPULATION SAMPLING

Abstract[1]**.** In surveys, the interest lies in estimating finite population parameters such as population totals and means. In most surveys, some auxiliary information is available at the estimation stage. This information may be incorporated in the estimation procedures to increase their precision. In this article, we use random forests to estimate the functional relationship between the survey variable and the auxiliary variables. In recent years, random forests have become attractive as National Statistical Offices have now access to a variety of data sources, potentially exhibiting a large number of observations on a large number of variables. We establish the theoretical properties of model-assisted procedures based on random forests and derive corresponding variance estimators. A model-calibration procedure for handling multiple survey variables is also discussed. The results of a simulation study suggest that the proposed point and estimation procedures perform well in term of bias, efficiency and coverage of normal-based confidence intervals, in a wide variety of settings. Finally, we apply the proposed methods using data on radio audiences collected by Médiamétrie, a French audience company.

Keywords: Model-assisted approach; Model-calibration; Nonparametric regression; Random forest; Survey data; Variance estimation.

3.1 Introduction

Since the pioneering work of Särndal (1980), Robinson and Särndal (1983) and Särndal and Wright (1984), model-assisted estimation procedures have attracted a lot of attention in the literature; see also Särndal et al. (1992) for a comprehensive discussion of the model-assisted approach. At the estimation stage, auxiliary information is often available and can be incorporated in the estimation procedures to increase the precision of the resulting point estimators. The model-assisted approach starts with postulating a working model, describing the relationship between a survey variable Y and a set of p auxiliary variables $X_1, X_2, \ldots, X_p$. The model is fitted to the sample observations to obtain predicted values, which then serve to build point estimators of population means/totals. Model-assisted estimators are asymptotically design-unbiased and design consistent, irrespective of whether or not the working model is correctly specified, which is an attractive feature; see Särndal et al. (1992) and Breidt and Opsomer (2017), among others. When the working model holds, model-assisted estimators are expected to be highly efficient. However, when the sample size is small, the use of model-assisted estimators requires some caution as they may suffer from small sample bias. In this article, we use random forests to estimate the functional relationship between Y and $X_1, X_2, \ldots, X_p$. In recent years, random forests have become attractive as National Statistical Offices have now access to a variety of data sources, potentially exhibiting a large number of observations on a large number of variables.

Consider a finite population $U = \{1, ..., k, ..., N\}$ of size N. We are interested in estimating the population total of a survey variable Y, $t_y = \sum_{k \in U} y_k$. We select a sample S, of size n, according

[1] The article is accepted for publication in Journal of the American Statistical Association.

to a sampling design $\mathcal{P}(S \mid \mathbf{Z}_U)$, where $\mathbf{Z}_U$ denotes the matrix of design information, available prior to sampling for all the population units. Let $\mathbf{I}_U = (I_1, ..., I_k, ..., I_N)^\top$ be the N-vector of sample selection indicators such that $I_k = 1$ if $k \in S$ and $I_k = 0$, otherwise. The first-order and second-order inclusion probabilities are given by $\pi_k = \mathbb{E}[I_k \mid \mathbf{Z}_U]$ and $\pi_{kl} = \mathbb{E}[I_k I_\ell \mid \mathbf{Z}_U]$, respectively.

A basic estimator of t_y is the well-known Horvitz-Thompson estimator given by

$$\widehat{t}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}. \tag{3.1}$$

Provided that $\pi_k > 0$ for all $k \in U$, the estimator (3.1) is design-unbiased for t_y in the sense $\mathbb{E}\left[\widehat{t}_\pi \mid \mathbf{y}_U, \mathbf{Z}_U\right] = t_y$, where $\mathbf{y}_U = (y_1, y_2, ..., y_N)^\top$. The Horvitz-Thompson estimator makes no use of auxiliary information beyond what is already contained in the matrix $\mathbf{Z}_U$.

We assume that a vector $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp})^\top$ of auxiliary variables is available for all $k \in U$. We also assume that y_k, $k \in U$, are independent realizations from a working model ξ, often referred to as a superpopulation model:

$$\begin{aligned} \mathbb{E}[y_k \mid \mathbf{X}_k = \mathbf{x}_k] &= m(\mathbf{x}_k), \\ \mathbb{V}(y_k \mid \mathbf{X}_k = \mathbf{x}_k) &= \sigma^2 v(\mathbf{x}_k), \end{aligned} \tag{3.2}$$

where $m(\cdot)$ and $v(\cdot)$ are two unknown functions and σ^2 is an unknown parameter.

Suppose that Model (3.2) is fitted at the population level and let $\widetilde{m}(\mathbf{x}_k)$ be the population-level fit associated with unit k obtained by fitting a parametric or nonparametric procedure. This leads to the pseudo generalized difference estimator

$$\widehat{t}_{pgd} = \sum_{k \in U} \widetilde{m}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \widetilde{m}(\mathbf{x}_k)}{\pi_k}. \tag{3.3}$$

Because the values $\widetilde{m}(\mathbf{x}_k)$ do not involve the sample selection indicators $I_1, \ldots, I_N$, if follows that $\mathbb{E}\left[\widehat{t}_{pgd} \mid \mathbf{y}_U, \mathbf{Z}_U, \mathbf{X}_U\right] = t_y$, where $\mathbf{X}_U$ is the $N \times p$ matrix whose N rows are the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$. That is, the pseudo generalized difference estimator (3.3) is design-unbiased for t_y. In the sequel, we use the simpler notation $\mathbb{E}_p[\cdot]$ instead of $\mathbb{E}[\cdot|\mathbf{Z}_U, \mathbf{X}_U, \mathbf{y}_U]$ to denote the expectation operator with respect to the sampling design $\mathcal{P}(S|\mathbf{Z}_U)$. Similarly, the notation $\mathbb{V}_p[\cdot]$ is used to denote the design variance of an estimator.

Most often, the estimator (3.3) is unfeasible as the population-level fits $\widetilde{m}(\mathbf{x}_k)$ are unknown. Using the sample observations, we fit the working model and obtain the sample-level fits $\widehat{m}(\mathbf{x}_k)$. Replacing $\widetilde{m}(\mathbf{x}_k)$ with $\widehat{m}(\mathbf{x}_k)$ in (3.3), we obtain the so-called model-assisted estimator of t_y:

$$\widehat{t}_{ma} = \sum_{k \in U} \widehat{m}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \widehat{m}(\mathbf{x}_k)}{\pi_k}. \tag{3.4}$$

Unlike (3.3), the estimator (3.4) is no longer design-unbiased, but can be shown to be design-consistent for t_y for a relatively wide class of procedures $\widehat{m}(\cdot)$. The model-assisted estimator (3.4) is expressed as the sum of the population total of the predictions $\widehat{m}(\mathbf{x}_k)$ and an adjustment term that can be viewed as a protection against model-misspecification.

If $\widehat{m}(\mathbf{x}_k) = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}$ with coefficients estimated by weighted least squares, the estimator (3.4) reduces to the well-known generalized regression (GREG) estimator; e.g., see Särndal et al.

(1992, Chap. 6). Model-assisted estimators based on generalized linear models were considered by Lehtonen and Veijanen (1998) and Firth and Bennett (1998), among others. There are some practical issues associated with the use of a parametric model such as linear and generalized linear models: they may lead to inefficient estimators if the function $m(\cdot)$ is misspecified or if the model fails to include interactions or predictors that account for curvature (e.g., quadratic and cubic terms). In contrast, nonparametric procedures are robust to model misspeficiation, which is a desirable property. A number of nonparametric model-assisted estimation procedures have been studied in the last two decades: local polynomial regression (Breidt and Opsomer, 2000), B-splines (Goga, 2005) and penalized B-splines (Goga and Ruiz-Gazen, 2014), penalized splines (Breidt et al., 2005, McConville and Breidt, 2013), neural nets (Montanari and Ranalli, 2005), generalized additive models (Opsomer et al., 2007), nonparametric additive models (Wang and Wang, 2011) and regression trees (McConville and Toth, 2019, Toth and Eltinge, 2011).

In this paper, we propose a new class of model-assisted estimators of t_y based on random forests (RF). Generally speaking, RF is an ensemble method that trains a (large) number of trees and combines them to produce more accurate predictions than a single regression tree would. Trees define a class of algorithms that recursively split the p-dimensional predictor space into distinct and non-overlapping regions. In other words, a tree algorithm generates a partition of regions or hyperrectangles of $\mathbb{R}^p$. For an observation belonging to a given region, the prediction is simply obtained by averaging the y-values associated with the units belonging to the same region. While regression trees are easy to interpret and allow the user to visualize the partition (Hastie et al., 2011, pp. 306), they may suffer from a high model variance, hence their qualification of "weak learners". A number of tree-based procedures have been proposed with the aim of improving the predictive performances of regression trees, including pruning (Breiman, 1984), Bayesian regression trees (Chipman et al., 1998), gradient boosting (Friedman, 2001) and RF (Breiman, 2001).

Several empirical studies suggest that RF can outperform state-of-the-art prediction models; see e.g. Han et al. (2018), Hamza and Larocque (2005), Díaz-Uriarte and de Andrés (2006). RF are widely used due to their predictive performances and their ability to handle small sample sizes with a large number of predictors (Scornet, 2016b). Also, RF algorithms can be parallelized, leading to a decrease in the training time. RF have been applied in a wide variety of fields, including medicine (Fraiwan et al., 2012), time series analysis (Kane et al., 2014), agriculture (Grimm et al., 2008), missing data (Stekhoven and Buhlmann, 2011), genomics (Qi, 2012) and pattern recognition (Rogez et al., 2008). In recent years, neural networks and deep learning algorithms have attracted a lot of attention and have been shown to be effective in a wide range of applications involving mostly unstructured data, such as speech recognition, image reconstruction and text translation; see Najafabadi et al. (2015) and the references therein for a review on the topic. However, to exhibit high levels of performance, deep learning algorithms typically require huge amounts of data (Arnould et al., 2020, Najafabadi et al., 2015). This is seldom the case in surveys as most data sets consist of structured data consisting of (at most) a few hundred thousand observations and a few hundred survey variables. For an empirical comparison of RF and neural networks, see Han et al. (2018). Finally, unlike RF algorithms that require the specification of a small number of hyper-parameters (see Section 3.6.3), gradient boosting, Bayesian regression trees or deep learning approaches depend upon the complex choice of a large number of hyper-parameters (Bergstra et al., 2011).

To the best of our knowledge, only little is known about the theoretical properties of RF based on the original algorithm of Breiman (2001). Often, the theoretical investigations are

made at the expense of simplifying assumptions; see for instance Biau et al. (2008) and Biau (2012). Two notable exceptions are Wager (2014) and Scornet et al. (2015) who established the theoretical properties of an algorithm closely related to that of Breiman (2001). In a finite population setting, the theoretical properties of RF algorithms have yet to be established, even in the ideal situation of 100% response. This paper aims to fill this important gap. While we are mostly concerned with RF for regression, we can easily extend our methods to the case of RF for classification. Some recent empirical studies on the performance of RF for complex survey data can be found in Tipton et al. (2013), Buskirk and Kolenikov (2015), De Moliner and Goga (2018) and Kern et al. (2019).

The rest of the paper is organized as follows. Regression trees and RF are presented in Section 3.2. In Section 3.3, we suggest two classes of model-assisted estimators based on random forests: the first is based on partitions built at the population level, while the second class is based on partitions built at the sample level. In Section 3.4, we establish the theoretical properties of model-assisted estimators based on RF and derive corresponding variance estimators. In Section 3.5, we describe a model-calibration procedure for handling multiple survey variables. In Sections 3.6.1-3.6.3, the finite sample properties of the proposed point and variance estimation procedures are evaluated through a simulation study, and in Section 3.6.4, we apply the proposed methods using data on radio audiences collected by Médiamétrie, a French audience company. The paper ends with some final remarks in Section 3.7. Proofs of major results and further technical details are relegated to the Appendix and the Supplementary Material.

3.2 Regression trees and random forests

3.2.1 Regression trees

The original RF uses regression trees based on the classification and regression tree algorithm (CART) of Breiman (1984), whereby the partition of the predictor space is generated by a greedy recursive algorithm. In this paper, we focus on the CART algorithm for regression, designed for handling quantitative survey variables Y, but our methods also applies to the case of binary survey variables. With regression trees, these estimated probabilities always lie between 0 and 1, which is a desirable feature. Alternative criteria may be used with binary variables, such as the Gini impurity or the entropy instead of the CART regression criterion (Hastie et al., 2011, Chapter 9). The CART algorithm for regression searches for the splitting variable and the splitting position (i.e., the coordinates on the predictor space where to split) for which the difference in empirical variance in the node before and after splitting is maximized. As a starting point, we consider the hypothetical situation, where y_k and $\mathbf{x}_k$ are observed for all $k \in U$ and assume that the regression tree is fitted at the population level. We use the generic notation A to denote a node with cardinality $\#(A)$ considered for the next split, and C_A to denote the set of possible splits in the node A, which corresponds to the set of all possible pairs $(j, z) = (\text{variable}, \text{position})$. This splitting process is performed by searching for the best split (j^*, z^*) for which the following empirical CART population criterion is maximized:

$$L_N(j,z) = \frac{1}{\#(A)} \sum_{k \in U} \mathbb{1}_{\mathbf{x}_k \in A} \left\{ (y_k - \bar{y}_A)^2 - \left(y_k - \bar{y}_{A_L} \mathbb{1}_{x_{kj} < z} - \bar{y}_{A_R} \mathbb{1}_{x_{kj} \geq z} \right)^2 \right\}, \tag{3.5}$$

where $A_L = \{k \in A; x_{kj} < z\}$, $A_R = \{k \in A; x_{kj} \geqslant z\}$ and $\bar{y}_A$ is the average of the y-values of units belonging to A. The best cut is always performed in the middle of two consecutive data points. In practice, it is common to impose a minimal number of observations N_0 (say) in each terminal node. In this case, the splitting process is performed until an additional split generates a terminal node with fewer observations than N_0.

The splitting process leads to the set

$$\mathcal{P}_U = \left\{A_1^{(U)}, \ldots, A_j^{(U)}, \ldots, A_{J_U}^{(U)}\right\} \tag{3.6}$$

of J_U hyperrectangles of $\mathbb{R}^p$ such that $A_j^{(U)} \bigcap A_{j'}^{(U)} = \emptyset$, for all $j \neq j' \in \{1, 2, \ldots, J_U\}$ and $\bigcup_{j=1}^{J_U} A_j^{(U)} = \mathbb{R}^p$. Thus, the set $\mathcal{P}_U$ defines a partition of $\mathbb{R}^p$, whose elements are called the terminal nodes. We use the generic notation $A^{(U)}(\mathbf{x}_k)$ to denote a terminal node belonging to the partition $\mathcal{P}_U$ given in (3.6) and that contains $\mathbf{x}_k$.

Figure 15 below illustrates how the recursive splitting procedure creates a partition in the simple case of two auxiliary variables X_1 and X_2, based on 5 splits. Each grey rotated square represents a split (variable, position) performed at some position along one of the two auxiliary variables, X_1 or X_2. The white ellipses represent the 6 terminal nodes, also represented by the scatter plot on the right; see also Biau and Devroye (2014) for a similar illustration.

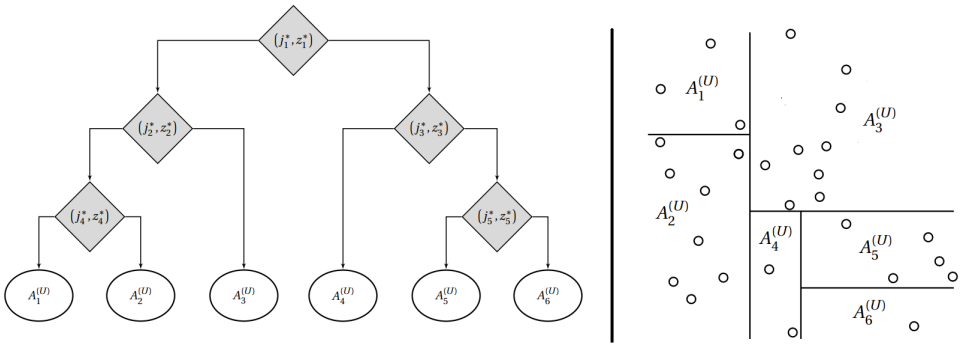


Figure 15: A regression tree (left) and the corresponding partition of $\mathbb{R}^2$ (right).

The prediction $\widetilde{m}_{tree}(\mathbf{x}_k)$ at the point $\mathbf{x}_k$ is simply defined as the average of the y-values of population individuals ℓ such that $\mathbf{x}_\ell$ belongs to $A^{(U)}(\mathbf{x}_k)$:

$$\widetilde{m}_{tree}(\mathbf{x}_k) = \sum_{\ell \in U} \frac{\mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}(\mathbf{x}_k)} y_\ell}{\widetilde{N}(\mathbf{x}_k)}, \tag{3.7}$$

where $\widetilde{N}(\mathbf{x}_k) = \sum_{\ell \in U} \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}(\mathbf{x}_k)}$ denotes,the number of units belonging to the terminal node $A^{(U)}(\mathbf{x}_k)$. Given the partition $\mathcal{P}_U$, the population-level fit $\widetilde{m}_{tree}(\mathbf{x}_k)$ may be viewed as the least squares type prediction obtained by fitting a one-way ANOVA model with Y as the response variable and the node membership indicators $\{\mathbb{1}_{\mathbf{x}_k \in A_j^{(U)}}\}_{j=1}^{J_U}$ as the set of explanatory variables; see (Hastie et al., 2011, Chapter 9) and the Supplementary Material for more details.

3.2.2 Random forests

To introduce random forests (RF) in a finite population setting, we again assume that y_k and $\mathbf{x}_k$ are observed for all $k \in U$. RF are based on a (large) number B (say) of regression trees. The prediction attached to unit k is defined as the average of the predictions produced by each of the B regression trees. That is,

$$\widetilde{m}_{rf}(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \widetilde{m}_{tree}^{(b)}(\mathbf{x}_k),$$

where $\widetilde{m}_{tree}^{(b)}(\mathbf{x}_k)$ is the predicted value attached to unit k obtained from the bth regression tree, $b = 1, \ldots, B$.

Obviously, if $\widetilde{m}_{tree}^{(1)}(\mathbf{x}_k) = \ldots = \widetilde{m}_{tree}^{(B)}(\mathbf{x}_k)$, then $\widetilde{m}_{rf}(\mathbf{x}_k) = \widetilde{m}_{tree}^{(1)}(\mathbf{x}_k)$. Such a situation would occur if each regression tree uses a deterministic splitting criterion in (3.5), which would lead to B identical partitions of $\mathbb{R}^p$. To cope with this issue, some amount of randomization is introduced in the tree building process, leading to B different predictions of $m(\cdot)$. The original algorithm of Breiman (2001) is implemented as follows:

1. Select B bootstrap data sets with replacement from the population data set, $D_U = \{(\mathbf{x}_k, y_k)\}_{k\in U}$, each data set containing N pairs of the form $(\mathbf{x}_k, y_k)$;

2. Fit a regression tree on each bootstrap data set. Before each split is performed, m_{try} predictors are selected randomly and without replacement from the full set of p predictors. The m_{try} selected predictors are the split candidates to be considered for searching the best split in (3.5).

The algorithm stops when each terminal node contains less than a predetermined number of observations. This procedure leads to a set $\widetilde{\mathcal{P}}_U = \left\{\mathcal{P}_U^{(1)}, \mathcal{P}_U^{(2)}, \ldots, \mathcal{P}_U^{(B)}\right\}$ of B different partitions of $\mathbb{R}^p$, each of the form (3.6). The randomization used in the tree building process is denoted by the random variable $\theta^{(U)}$, assumed to belong to some measurable space $(\Theta, \mathcal{F})$ and independent of the data (Biau and Scornet, 2016). Let $\theta_b^{(U)}$ be the random variable associated with the bth tree. The random variables $\theta_b^{(U)}$, $b = 1, \ldots, B$, are assumed to be independent and their distribution is identical to that of the generic random variable $\theta^{(U)}$. In the RF algorithm of Breiman, the randomization is induced by the selection (with replacement) of observations in Step 1 of the above algorithm and the random selection of split variables in Step 2 of the above algorithm. A number of RF algorithms algorithms have been considered in the literature. For example, (Biau et al., 2008, Scornet, 2016a) considered a simple RF algorithm called the uniform random forest (URF) algorithm. In the URF algorithm, a variable is selected with equal probability among the initial p predictors at each node and a split position is chosen uniformly in the node along the direction of the selected variable. The algorithm stops when each terminal node has a predetermined number of cuts. In this case, the randomization $\theta_b^{(U)}$ is characterized by the random selections of the node, the split variable and the location. For more details on RF algorithms, the reader is referred to Geurts et al. (2006), Biau et al. (2008), Biau (2012), Genuer (2012), Scornet (2016a), among others. In the sequel, unless stated otherwise, we assume that the observations in Step 1 of the above algorithm are selected without replacement (Scornet, 2017), which we will refer to as subsampling. Also, for more generality, the splitting criterion is left unspecified.

Let $\widetilde{m}^{(1)}_{tree}(\cdot, \theta^{(U)}_1), \ldots, \widetilde{m}^{(B)}_{tree}(\cdot, \theta^{(U)}_B)$, denote the predictions obtained with the B stochastic or randomized regression trees. The RF prediction attached to unit k is defined as a bagged estimator of B trees:

$$\widetilde{m}_{rf}(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \widetilde{m}^{(b)}_{tree}(\mathbf{x}_k, \theta^{(U)}_b). \tag{3.8}$$

It is worth pointing out that considering a new set of predictors at each split leads to B trees which are less correlated with each other; that is, trees that are quite different from one another. As a result, the RF may lead to substantial gains in precision compared to a single tree (James et al., 2015, Chapter 8). The number of predictors selected at each split, denoted by m_{try}, is thus an important tuning parameter in the RF algorithm. In practice, the choice $m_{try} = \sqrt{p}$ seems to give good results, in general. In Section 3.6.3, we assess the impact of m_{try} through a simulation study.

For any RF algorithm, the prediction at the point $\mathbf{x}_k$ in (3.8) can also be expressed as

$$\widetilde{m}_{rf}(\mathbf{x}_k) = \sum_{\ell \in U} \widetilde{W}_\ell(\mathbf{x}_k) y_\ell, \tag{3.9}$$

where

$$\widetilde{W}_\ell(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \frac{\psi^{(b,U)}_\ell \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}\left(\mathbf{x}_k, \theta^{(U)}_b\right)}}{\widetilde{N}(\mathbf{x}_k, \theta^{(U)}_b)} \tag{3.10}$$

is a prediction weight attached to unit k with $\widetilde{N}(\mathbf{x}_k, \theta^{(U)}_b) = \sum_{\ell \in U} \psi^{(b,U)}_\ell \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}(\mathbf{x}_k, \theta^{(U)}_b)}$ denoting the number of observations belonging to the terminal node $A^{(U)}$ containing $\mathbf{x}_k$ in the bth regression tree. The random variables $\psi^{(b,U)}_\ell$ in (3.10) depend on the resampling mechanism used in the RF algorithm and depend on $\theta^{(U)}_b$, but are independent of the sampling design $\mathcal{P}(S \mid \mathbf{Z}_U)$. In the case of subsampling, the random variables $\psi^{(b,U)}_\ell$ follow a Bernoulli distribution, $\psi^{(b,U)}_\ell \sim \mathcal{B}(N'/N)$, where N' denotes the number of units in each subsample. Note that the prediction $\widetilde{m}_{rf}$ in (3.9) can be computed for either a continuous or a categorical y-variable. In the latter case, the prediction $\widetilde{m}_{rf}$ in (3.9) corresponds to the population proportion of units who belong to a given category computed over the B trees.

Proposition 3.2.1. *Consider the predictor weights $\widetilde{W}_\ell(\mathbf{x}_k)$ given in (3.10).*

i) The weights $\widetilde{W}_\ell(\mathbf{x}_k)$ are uniformly bounded. That is,

$$0 < \widetilde{W}_\ell(\mathbf{x}_k) \leqslant cN_0^{-1}$$

for all $\ell \in U$ and all $\mathbf{x}_k \in \mathbb{R}^p$, where c is a positive constant that does not depend either on k,ℓ, or N_0, the minimal number of observations in the terminal nodes.

ii) The weight functions sum up to one; that is, $\sum_{\ell \in U} \widetilde{W}_\ell(\mathbf{x}_k) = 1$ for all $\mathbf{x}_k \in \mathbb{R}^p$.

The proof of Proposition 3.2.1 is given in the Appendix.

3.3 Model-assisted estimation: Random forests

In Section 3.2, we assumed that y_k and $\mathbf{x}_k$ were observed for all $k \in U$, which led to the population-level fits $\widetilde{m}_{tree}(\mathbf{x}_k)$ and $\widetilde{m}_{rf}(\mathbf{x}_k)$ given by (3.7) and (3.9), respectively. However,

both (3.7) and (3.9) cannot be computed in practice as the y-values are observed only for $k \in S$. Moreover, the regression trees in Sections 3.2.1 and 3.2.2 were based on partitions built recursively at the population level so as to optimize the population criterion (3.5). As a result, these partitions depend on the vector of predictors $\{\mathbf{x}_k\}_{k\in U}$ but also on the unknown population values $\{y_k\}_{k\in U}$. While the former type of dependency is inherent to most parametric and nonparametric procedures, the latter is absent in many commonly used parametric and nonparametric procedures such as spline procedures (Breidt et al., 2005, Breidt and Opsomer, 2000, Goga, 2005, Goga and Ruiz-Gazen, 2014, McConville and Breidt, 2013). Due to the dependency on the unknown population values $\{y_k\}_{k\in U}$, establishing the theoretical properties of model-assisted estimators based on RF is more challenging.

For these reasons, in Section 3.3.1, we start by considering the simpler case of population partitions obtained using a variable Y^*, assumed to be closely related to Y and available for all $k \in U$. While this assumption is somehow strong and not tenable in many practical situations, it provides some insights on how to tackle the problem in the presence of Y-dependency. Algorithms allowing to get rid of the Y-dependency have been suggested in the random-forest literature; see e.g. Biau et al. (2008), Biau (2012) or Devroye et al. (2013, Chap. 20). Sample-based partitions are considered in Section 3.3.2.

3.3.1 Model-assisted estimation: Population-based partitions

In this section, we consider the case of a splitting criterion that does not depend on the data $\{y_k\}_{k\in s}$. We consider a variable Y^* assumed to be closely related to Y and such that the values y_k^* are available for all $k \in U$. We seek population partitions $\widetilde{\mathcal{P}}_U^*$, independent of the survey variable Y, that maximize the following criterion:

$$L_N^*(j,z) = \frac{1}{\#(A)} \sum_{k\in U} \mathbb{1}_{\mathbf{x}_k\in A} \left\{ \left(y_k^* - \overline{y}_A^*\right)^2 - \left(y_k^* - \overline{y}_{A_L}^* \mathbb{1}_{x_{kj}<z} - \overline{y}_{A_R}^* \mathbb{1}_{x_{kj}\geq z}\right)^2 \right\}, \tag{3.11}$$

where A_R, A_L are as in (3.5) and $\overline{y^*}_A$ is the average of the y^*-values for the units belonging to a node A.

Based on (3.11), the population-level fit at the point $\mathbf{x}_k$ is given by

$$\widetilde{m}_{rf}^*(\mathbf{x}_k) = \sum_{\ell\in U} \widetilde{W}_\ell^*(\mathbf{x}_k) y_\ell, \tag{3.12}$$

where the weights $\widetilde{W}_\ell^*(\mathbf{x}_k)$ in (3.12) are obtained from (3.10) by replacing $A^{(U)}$ with $A^{*(U)}$, a generic member of the partition $\widetilde{\mathcal{P}}_U^*$.

The weights $\{\widetilde{W}_\ell^*(\cdot)\}_{\ell\in U}$ in (3.12) are known for all $\ell \in U$ and are independent of Y. Since $\widetilde{m}_{rf}^*(\mathbf{x}_k)$ in (3.12) requires the y-values for all the population units, it cannot be computed. A simple solution consists of replacing the population total on the right hand-side of (3.12) by its corresponding Horvitz–Thompson estimator, which leads to

$$\widehat{m}_{rf}^*(\mathbf{x}_k) = \sum_{\ell\in S} \frac{\widetilde{W}_\ell^*(\mathbf{x}_k) y_\ell}{\pi_\ell}. \tag{3.13}$$

A model-assisted estimator of t_y based on population RF is obtained by plugging $\widehat{m}^*_{rf}(\mathbf{x}_k)$ in (3.3):

$$\widehat{t}^*_{rf} = \sum_{k\in U} \widehat{m}^*_{rf}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}^*_{rf}(\mathbf{x}_k)}{\pi_k}. \tag{3.14}$$

Proposition 3.3.1. *The RF estimator given in (3.14) can be expressed as*

$$\widehat{t}^*_{rf} = \sum_{k\in S} w_{ks} y_k,$$

where the weights w_{ks} are given by

$$w_{ks} = \frac{1}{\pi_k}\left\{1 + \sum_{\ell\in U} \widetilde{W}^*_k(\mathbf{x}_\ell)\left(1 - \frac{I_\ell}{\pi_\ell}\right)\right\}. \quad k \in S \tag{3.15}$$

Proof. By rearranging the sums, we get:

$$\begin{aligned}\widehat{t}^*_{rf} &= \sum_{k\in S}\frac{y_k}{\pi_k} + \sum_{\ell\in U}\left(1 - \frac{I_\ell}{\pi_\ell}\right)\widehat{m}^*_{rf}(\mathbf{x}_\ell) = \sum_{k\in S}\frac{y_k}{\pi_k} + \sum_{\ell\in U}\left(1 - \frac{I_\ell}{\pi_\ell}\right)\left(\sum_{k\in S}\widetilde{W}^*_k(\mathbf{x}_\ell)\frac{y_k}{\pi_k}\right)\\ &= \sum_{k\in S}\left\{1 + \sum_{\ell\in U}\left(1 - \frac{I_\ell}{\pi_\ell}\right)\widetilde{W}^*_k(\mathbf{x}_\ell)\right\}\frac{y_k}{\pi_k}.\end{aligned}$$

■

Since the partitions $\widetilde{\mathcal{P}}^*_U$, are independent of both the survey variable Y and the sample S, the weights w_{ks} given by (3.15) depend on the sample only through the sample selection indicators $I_\ell, \ell \in U$, but are independent of Y. As a result, these weights may be used to estimate the population total of any survey variable, which is an attractive feature in multipurpose surveys. However, for RF algorithms based on the splitting criterion in (3.11), we expect the weights w_{ks} to be efficient whenever the survey variable Y is highly correlated to the variable Y^*. In multipurpose surveys where the survey variables are not necessarily correlated with one another, it may be preferable to use a splitting criterion that depends on the data $\{\mathbf{x}_k\}_{k\in U}$ as done in quantile random forests (Devroye et al., 2013, Scornet, 2016a).

3.3.2 Model-assisted estimation: Sample-based partitions

In this section, we seek sample partitions $\widehat{\mathcal{P}}_S = \left\{\widehat{\mathcal{P}}^{(1)}_S, \ldots, \widehat{\mathcal{P}}^{(b)}_S, \ldots, \widehat{\mathcal{P}}^{(B)}_S\right\}$ using

$$L_n(j,z) \quad = \quad \frac{1}{\#(A)}\sum_{k\in S}\mathbb{1}_{\mathbf{x}_k\in A}\left\{(y_k - \bar{y}_A)^2 - \left(y_k - \bar{y}_{A_L}\mathbb{1}_{x_{kj}<z} - \bar{y}_{A_R}\mathbb{1}_{x_{kj}\geq z}\right)^2\right\}. \tag{3.16}$$

Based on the partition $\widehat{\mathcal{P}}_S$, we obtain the sample-level fits

$$\widehat{m}_{rf}(\mathbf{x}_k) = \sum_{\ell\in S}\frac{\widehat{W}_\ell(\mathbf{x}_k)y_\ell}{\pi_\ell}, \tag{3.17}$$

where

$$\widehat{W}_\ell(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \frac{\psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)}}{\widehat{N}(\mathbf{x}_k, \theta_b^{(S)})}, \quad \ell \in S, \tag{3.18}$$

and $\widehat{N}(\mathbf{x}_k, \theta_b^{(S)}) = \sum_{\ell \in U} I_\ell \pi_\ell^{-1} \psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)}$ denotes the estimated number of observations in the terminal node $A^{(S)}$ containing $\mathbf{x}_k$ in the bth regression tree. The variable $\psi_\ell^{(b,S)}$ indicates whether or not unit ℓ has been selected in the bth sub-sample and is such that $\psi_\ell^{(b,S)} \sim \mathcal{B}(n'/n)$ for RF based on subsampling, where n' denotes the number of units in each sub-sample.

Plugging $\widehat{m}_{rf}(\cdot)$ in (3.4) leads to the RF model-assisted estimator

$$\widehat{t}_{rf} = \sum_{k \in U} \widehat{m}_{rf}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k}. \tag{3.19}$$

Using similar arguments to those used in the proof of Proposition 3.3.1, we can show that $\widehat{t}_{rf}$ can be expressed as

$$\widehat{t}_{rf} = \sum_{k \in S} w'_{ks} y_k,$$

where the weights w'_{ks} are given by

$$w'_{ks} = \frac{1}{\pi_k}\left\{1 + \sum_{\ell \in U} \widehat{W}_k(\mathbf{x}_\ell)\left(1 - \frac{I_\ell}{\pi_\ell}\right)\right\}, \quad k \in S. \tag{3.20}$$

Noting that $\sum_{k \in S} \widehat{W}_k(\mathbf{x}_\ell)\pi_k^{-1} = 1$ for all $\ell \in U$, it follows from (3.20) that $\sum_{k \in S} w'_{ks} = N$ for every sample S. That is, the sum of the weights w'_{ks} match the population size N perfectly, a desirable property shared by other nonparametric model-assisted estimators (Breidt et al., 2005, Goga, 2005, Goga and Ruiz-Gazen, 2014). Unlike the weights w_{ks} in (3.15), the weights w'_{ks} depend on both the sample selection indicators $I_\ell, \ell \in U$, and the partition $\widehat{\mathcal{P}}_S$ that varies from one sample to another. This is due to the fact that the nodes $A^{(S)}$ are constructed so as to optimize the sample criterion (3.16). For this reason, the weights $w'_{ks}, k \in S$, are variable specific in the sense that depend on the survey variable Y. To cope with this issue, we describe a model calibration procedure in Section 3.5 for handling multiple survey variables while producing a single set of weights.

Remark 3.3.1. *In practice, the variables $\psi_k^{(b,S)}$ in (3.18) are not generated for the units outside the sample. However, at least conceptually, nothing precludes defining these variables for $k \in U \setminus S$. For $k \in U \setminus S$, we set $\psi_k^{(b,S)} \sim \mathcal{B}((N' - n')/(N - n))$ so that $\sum_{k \in U} \psi_k^{(b,S)} = N'$. Defining the variables $\psi_k^{(b,S)}$ for units outside the sample will have no effect on the predictions $\widehat{m}_{rf}(\cdot)$ associated with the sample units since $I_k = 0$ for $k \in U \setminus S$. This construction will prove useful in establishing the asymptotic properties of the proposed procedures; see Section 3.4.*

As for the RF prediction built at the population level described in Section 3.2.2, the prediction $\widehat{m}_{rf}(\mathbf{x}_k)$ in (3.17) can be expressed as a bagged predictor (Hastie et al., 2011). That is,

$$\widehat{m}_{rf}(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}),$$

where $\widehat{m}^{(b)}_{tree}(\mathbf{x}_k, \theta^{(S)}_b) = \sum_{\ell \in S} \pi_\ell^{-1} \psi^{(b,S)}_\ell \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x}_k, \theta^{(S)}_b\right)} y_\ell / \widehat{N}(\mathbf{x}_k, \theta^{(S)}_b)$ is the prediction associated with unit k based on the bth stochastic regression tree. The model-assisted estimator $\widehat{t}_{rf}$ given by (3.19) can thus be viewed as a bagged estimator:

$$\widehat{t}_{rf} = \frac{1}{B} \sum_{b=1}^{B} \widehat{t}^{(b)}_{tree}(\theta^{(S)}_b),$$

where

$$\widehat{t}^{(b)}_{tree}(\theta^{(S)}_b) = \sum_{k \in U} \widehat{m}^{(b)}_{tree}(\mathbf{x}_k, \theta^{(S)}_b) + \sum_{k \in S} \frac{y_k - \widehat{m}^{(b)}_{tree}(\mathbf{x}_k, \theta^{(S)}_b)}{\pi_k}$$

is the model-assisted estimator of t_y based on the bth stochastic regression tree. As in the case of regression trees built at the population level (see Section 3.2.1), given the partition $\widehat{\mathcal{P}}^{(b)}_S = \{A^{(bS)}_j\}_{j=1}^{J_{bS}}$, the predictions $\widehat{m}^{(b)}_{tree}(\mathbf{x}_k, \theta^{(S)}_b)$ are least squares type predictions obtained by fitting the one-way ANOVA model with Y as the response and the node membership indicators $\{\mathbb{1}_{\mathbf{x}_k \in A^{(bS)}_j}\}_{j=1}^{J_{bS}}$ as the explanatory variables; see the proof of Proposition 3.3.3 and the Supplementary Material for more details. As a result, the estimator $\widehat{t}^{(b)}_{tree}(\theta^{(S)}_b)$ is related to the customary post-stratified estimator (Särndal et al., 1992).

Under mild assumptions, Proposition 3.3.2 below shows that bagging improves the efficiency of model-assisted estimators. This is similar to what is encountered in the classical RF literature (Hastie et al., 2011).

Proposition 3.3.2. *Let $\widehat{t}^{(1)}, \ldots, \widehat{t}^{(b)}, \ldots, \widehat{t}^{(B)}$ be a sequence of model-assisted estimators of t_y and let $\widehat{t} = B^{-1} \sum_{b=1}^{B} \widehat{t}^{(b)}$ be a bagged estimator. Assuming that the $\widehat{t}^{(b)}$'s have approximately the same design bias and design variance, then, for B large enough:*

$$MSE_p(\widehat{t}) - MSE_p(\widehat{t}^{(1)}) \leqslant \mathbb{V}_p(\widehat{t}^{(1)}) \left(\max_{b \neq b'} \left| \mathbb{C}or_p\left(\widehat{t}^{(b)}, \widehat{t}^{(b')}\right) \right| - 1 \right) \leqslant 0,$$

where $MSE_p(\cdot)$ and $\mathbb{C}or_p(\cdot)$ denote the mean squared error and correlation operators with respect to the sampling design.

The proof of Proposition 3.3.2 is given in the Appendix. We end this section by giving an alternative expression for $\widehat{t}_{rf}$.

Proposition 3.3.3. *The RF estimator $\widehat{t}_{rf}$ given by (3.19) can be written as*

$$\widehat{t}_{rf} = \sum_{k \in U} \widehat{m}_{rf}(\mathbf{x}_k) + \frac{1}{B} \sum_{b=1}^{B} \sum_{k \in S} \frac{\left(1 - \psi^{(b,S)}_k\right) \left(y_k - \widehat{m}^{(b)}_{tree}(\mathbf{x}_k, \theta^{(S)}_b)\right)}{\pi_k}, \tag{3.21}$$

where $\widehat{m}^{(b)}_{tree}(\mathbf{x}_k, \theta^{(S)}_b)$ is the predictor associated with unit k based on the bth stochastic regression tree.

The proof of Proposition 3.3.3 is given in the Appendix. It follows from Proposition 3.3.3, that the second term on the right hand-side of (3.21) vanishes if $\psi^{(b,S)}_k = 1$ for all $k \in S$. That is, the estimator $\widehat{t}_{rf}$ reduces to the so-called projection form (Breidt et al., 2005, Goga, 2005, Särndal et al., 1992)

$$\widehat{t}_{rf} = \sum_{k \in U} \widehat{m}_{rf}(\mathbf{x}_k)$$

if the RF algorithm does not involve a resampling mechanism. In addition, the second term on the right hand-side of (3.21) vanishes if $y_k = c$ for all k, for some $c \in \mathbb{R}$ or if the trees in the forest are fully grown (i.e., each terminal node contains a single observation), which implies that the observations y_k and the corresponding prediction $\widehat{m}^{(b)}_{tree}(\mathbf{x}_k, \theta^{(S)}_b)$ coincide. When the estimator $\widehat{t}_{rf}$ can be expressed in the projection form, the weights w'_{ks} given by (3.20) are always positive and cannot exceed the number of terminal nodes from the largest tree of the forest.

In practice, a resampling mechanism is typically used with RF algorithms. In this case, the second term on the right hand-side of (3.21) does not vanish and is equal to the weighted sum of residuals computed for the non-resampled units, also called the *out-of-bag* individuals (James et al., 2015, Chapter 8), from each of the B trees. The second term on the right hand-side of (3.21) can then be viewed as a correction term which brings additional information from the units not used in computing the predictions $\widehat{m}^{(b)}_{tree}(\cdot, \theta^{(S)}_b)$, $b = 1, \ldots, B$.

3.4 Asymptotic properties

To establish the asymptotic properties of the proposed estimators and to derive the associated variance estimators, we consider the asymptotic framework of Isaki and Fuller (1982). We start with an increasing sequence of embedded finite populations $\{U_v\}_{v\in\mathbb{N}}$ of size $\{N_v\}_{v\in\mathbb{N}}$. In each finite population U_v, a sample of size n_v is selected according to a sampling design $\mathcal{P}_v(S_v = s_v \mid \mathbf{Z}_U)$. While the finite populations are assumed to be embedded, we do not require this property to hold for the samples $\{S_v\}_{v\in\mathbb{N}}$. This asymptotic framework assumes that v goes to infinity, so that both the finite population sizes and the samples sizes go to infinity. To improve readability, we shall use the subscript v only in the quantities U_v, N_v and n_v; quantities such as $\pi_{k,v}$ shall be denoted simply as π_k.

Assumptions: RF model-assisted estimator $\widehat{t}^{*}_{rf}$

We make the following assumptions:

(H6) There exists a positive constant C such that $\sup_{k\in U_v} |y_k| \leqslant C < \infty$.

(H7) We assume that $\lim_{v\to\infty} \frac{n_v}{N_v} = \pi \in (0, 1)$.

(H8) There exist positive constants λ and λ^* such that $\min_{k\in U_v} \pi_k \geqslant \lambda > 0$ and $\min_{k,\ell\in U_v} \pi_{k\ell} \geqslant \lambda^* > 0$. Also, we assume that $\limsup_{v\to\infty} n_v \max_{k\neq\ell\in U_v} |\pi_{k\ell} - \pi_k\pi_\ell| < \infty$.

Assumptions (H13)-(H15) have been extensively used in parametric, nonparametric and functional model-assisted estimation (Breidt et al., 2005, Breidt and Opsomer, 2000, Cardot et al., 2013c, Goga, 2005, Goga and Ruiz-Gazen, 2014, Robinson and Särndal, 1983). Assumption (H13) implies that the survey variable Y is uniformly bounded (Breidt and Opsomer, 2000, Cardot et al., 2010). Assumptions (H14) and (H15) deal with the first and second order inclusion probabilities and they are satisfied for the classical fixed-size sampling designs; see for example, Robinson and Särndal (1983) and Breidt and Opsomer (2000). Furthermore, we assume that the minimum number of observations N_{0v} in a terminal nodes is growing to infinity and we make the following additional assumption

(C1) The number of subsampled elements N'_v is such that $\lim_{v\to\infty} N'_v/N_v \in (0;1]$.

This assumption requires that the number N'_v of elements in each subsample increases at the same speed as the population size N_v, allowing each terminal node to have at least N_{0v} observations.

Assumptions: RF model-assisted estimator $\widehat{t}_{rf}$

In addition to the above assumptions, we make the following assumptions to establish the asymptotic properties of $\hat{t}_{rf}$ given by (3.19).

(H9) There exists a positive constant C_1 such that $n_v \max_{k\neq\ell\in U_v} \left|\mathbb{E}_p\left\{(I_k-\pi_k)(I_\ell-\pi_\ell)|\widehat{\mathcal{P}}_S\right\}\right| \leqslant C_1$.

(H10) The random forests based on population partitions and those based on sample partitions are such that, for all $\mathbf{x}\in\mathbb{R}^p$:

$$\mathbb{E}_p\left(\widehat{\widetilde{m}}_{rf}(\mathbf{x})-\widetilde{m}_{rf}(\mathbf{x})\right)^2 = o(1),$$

where $\widehat{\widetilde{m}}_{rf}(\mathbf{x}) = \sum_{\ell\in U_v}\frac{1}{B}\sum_{b=1}^{B}\frac{\psi_\ell^{(b,S)}\mathbb{1}_{\mathbf{x}_\ell\in A^{(S)}\left(\mathbf{x},\theta_b^{(S)}\right)}y_\ell}{\widehat{\widetilde{N}}(\mathbf{x},\theta_b^{(S)})}$ with $\widehat{\widetilde{N}}(\mathbf{x},\theta_b^{(S)}) = \sum_{\ell\in U_v}\psi_\ell^{(b,S)}\mathbb{1}_{\mathbf{x}_\ell\in A^{(S)}\left(\mathbf{x},\theta_b^{(S)}\right)}$.

Assumption (H18) is similar to that used by Toth and Eltinge (2011) and McConville and Toth (2019); it requires that, as the sample and population size grow, the influence of extreme observations on the sample partitions decreases. Assumption (H16) requires that the average number of elements at the population level in the sample partitions converges to the average number of population elements in the population partitions. It implicitly assumes that the sample partitions converge to the population partitions. A similar result was established in Toth and Eltinge (2011) in the case of regression trees. Toth and Eltinge (2011) evaluated the properties of point estimators with respect to the joint distribution induced by the superpopulation model and the sampling design. In a *iid* setting, Scornet et al. (2015) showed that the population partitions converge to the theoretical partitions. Assumption (H16) can thus be viewed as a design-based version of the result from Scornet et al. (2015). In the Supplementary Material, we conduct a simulation study, whose results suggest that Assumption (H16) seems to be verified, at least in our experiments. More research is needed to provide a rigorous proof of Assumption (H16) in the design-based approach and is beyond the scope of this article.

As in the case of model-assisted estimators based on RF with population-based partitions, we assume that the minimum number of observations, n_{0v}, in the terminal nodes is also growing to infinity and we assume the following additional assumption about the RF resampling algorithm :

(C2) The number of subsampled elements n'_v is such that $\lim_{v\to\infty} n'_v/n_v \in (0;1]$.

This assumption requires that the number n'_v of elements in each subsample increases at the same speed as the sample size n_v, allowing each terminal node to have at least n_{0v} observations.

3.4.1 Asymptotic results

In this section, we state some results pertaining to sequences of RF model-assisted estimators $\{\widehat{t}_{rf}\}$. The corresponding results for the model-assisted estimators $\{\widehat{t}^{*}_{rf}\}$ can be found in the Supplementary Material.

Result 3.4.1. *Consider a sequence of RF model-assisted estimators $\{\widehat{t}_{rf}\}$. Then, there exist positive constants $\tilde{C}_1, \tilde{C}_2$ such that*

$$\mathbb{E}_p\left|\frac{1}{N_v}\left(\widehat{t}_{rf} - t_y\right)\right| \leqslant \frac{\tilde{C}_1}{\sqrt{n_v}} + \frac{\tilde{C}_2}{n_{0v}}, \quad \textit{with } \xi\textit{-probability one.}$$

If $\dfrac{n_v^u}{n_{0v}} = O(1)$ with $1/2 \leqslant u \leqslant 1$, then there exists a positive constant $\tilde{C}$ such that

$$\mathbb{E}_p\left|\frac{1}{N_v}\left(\widehat{t}_{rf} - t_y\right)\right| \leqslant \frac{\tilde{C}}{\sqrt{n_v}}, \quad \textit{with } \xi\textit{-probability one.}$$

Result 3.9.1 implies that the RF model-assisted estimator $\{\widehat{t}_{rf}\}$ is asymptotically design-unbiased, i.e.,

$$\lim_{v\to\infty} \mathbb{E}_p\left[\frac{1}{N_v}\left(\widehat{t}_{rf} - t_y\right)\right] = 0, \quad \text{with } \xi\text{-probability one,}$$

and design-consistent in the sense that

$$\lim_{v\to\infty} \mathbb{E}_p\left[\mathbf{1}_{\{N_v^{-1}|\widehat{t}_{rf}-t_y|>\eta\}}\right] = 0, \quad \text{with } \xi\text{-probability one}$$

for all $\eta > 0$. Moreover, if n_{0v} is large enough with respect to the sample size n_v, the RF estimator $\widehat{t}_{rf}$ is $\sqrt{n_v}$-consistent. For a given partition, note that the number of terminal nodes is of order $O(n_v/n_{0v})$, and if n_{0v} satisfies the condition from the Result 3.9.1, the number of terminal nodes is of order $O(n^{1-u})$ for $1/2 \leqslant u \leqslant 1$.

The next result shows that the RF model-assisted estimator $\widehat{t}_{rf}$ is asymptotically equivalent to the pseudo-generalized difference estimator:

$$\widehat{t}_{pgd} = \sum_{k\in U} \widetilde{m}_{rf}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widetilde{m}_{rf}(\mathbf{x}_k)}{\pi_k}, \tag{3.22}$$

where $\widetilde{m}_{rf}(\mathbf{x}_k)$ is given by (3.9).

Result 3.4.2. *Consider a sequence of RF estimators $\{\widehat{t}_{rf}\}$. Assume also that $\dfrac{n_v^u}{n_{0v}} = O(1)$ with $1/2 < u \leqslant 1$. Then, $\{\widehat{t}_{rf}\}$ is asymptotically equivalent to the pseudo-generalized difference estimator $\widehat{t}_{pgd}$ in the sense that*

$$\frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{rf} - t_y\right) = \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{pgd} - t_y\right) + o_{\mathbb{P}}(1).$$

From Proposition 3.9.2, it follows that the asymptotic variance of $\widehat{t}_{rf}$ can be approximated by the variance of (3.22). That is,

$$\begin{aligned} \mathbb{AV}_p\left(\frac{1}{N_v}\widehat{t}_{rf}\right) &= \mathbb{V}_p\left(\frac{1}{N_v}\widehat{t}_{pgd}\right) \\ &= \frac{1}{N_v^2}\sum_{k\in U_v}\sum_{\ell\in U_v}(\pi_{kl}-\pi_k\pi_\ell)\frac{y_k-\widetilde{m}_{rf}(\mathbf{x}_k)}{\pi_k}\frac{y_\ell-\widetilde{m}_{rf}(\mathbf{x}_\ell)}{\pi_\ell}. \end{aligned} \tag{3.23}$$

While the RF model-assisted estimator $\widehat{t}_{rf}$ is design-consistent as long as n_{0v} and n_v grow to infinity (Result 3.9.1), the asymptotic equivalence of $\widehat{t}_{rf}$ with the pseudo-generalized difference estimator $\widehat{t}_{pgd}$ is obtained only for n_{0v} satisfying a certain rate. Stronger assumptions on higher-order inclusion probabilities (Breidt and Opsomer, 2000, McConville and Toth, 2019) are required in order to show that the asymptotic mean squared error of $\widehat{t}_{rf}$ is equivalent to the variance of the pseudo-generalized difference estimator. We do not pursue this further.

Expression (3.23) suggests that $\widehat{t}_{rf}$ is efficient if the residuals $y_k - \widetilde{m}_{rf}(\mathbf{x}_k)$ are small for all $k \in U_v$. The asymptotic variance given in (3.23) cannot be computed in practice because the residuals, $y_k - \widetilde{m}_{rf}(\mathbf{x}_k)$, $k \in U$, are unknown. Assuming that $\pi_{k\ell} > 0$ for all pairs $(k,\ell) \in U_v \times U_v$, a design-consistent estimator of $\mathbb{AV}_p\left(\frac{1}{N_v}\widehat{t}_{rf}\right)$ is given by

$$\widehat{\mathbb{V}}_{rf}\left(\frac{1}{N_v}\hat{t}_{rf}\right) = \frac{1}{N_v^2}\sum_{k\in U_v}\sum_{\ell\in U_v} I_k I_\ell \frac{\pi_{k\ell}-\pi_k\pi_\ell}{\pi_{k\ell}}\frac{y_k-\widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k}\frac{y_\ell-\widehat{m}_{rf}(\mathbf{x}_\ell)}{\pi_\ell}, \tag{3.24}$$

where $\widehat{m}_{rf}(\mathbf{x}_k)$ is given by (3.17). To establish the design consistency of (3.24), we require the following additional assumption:

(H11) We assume that $\lim_{v\to\infty}\max_{i,j,k,\ell\in D_{4,N_v}} |\mathbb{E}_p\left\{\left(I_iI_j-\pi_i\pi_j\right)\left(I_kI_\ell-\pi_k\pi_\ell\right)\right\}| = 0$, where D_{4,N_v} denotes the set of distinct 4-tuples from U_v.

Assumption (H17) was suggested by Breidt and Opsomer (2000) and, together with (H14)-(H15), is used to establish the design consistency of the unbiased estimator of the variance of the Horvitz-Thompson estimator $\sum_{k\in S_v} y_k/\pi_k$, assuming that the survey variable Y has finite fourth moment. Assumption (H17) is satisfied for simple random sampling without replacement and stratified simple random sampling without replacement. It is also satisfied for high entropy sampling designs (Boistard et al., 2012, Cardot et al., 2013c).

Result 3.4.3. *Consider a sequence of RF model-assisted estimators $\{\widehat{t}_{rf}\}$. Assume also that $\frac{n_v^u}{n_{0v}} = O(1)$ with $1/2 < u \leqslant 1$. Then, the variance estimator $\widehat{\mathbb{V}}_{rf}(\widehat{t}_{rf})$ is asymptotically design-consistent for the asymptotic variance $\mathbb{AV}_p\left(\widehat{t}_{rf}\right)$. That is,*

$$\lim_{v\to\infty}\mathbb{E}_p\left(\frac{n_v}{N_v^2}\left|\widehat{\mathbb{V}}_{rf}(\widehat{t}_{rf})-\mathbb{AV}_p(\widehat{t}_{rf})\right|\right) = 0.$$

Finally, we establish the central limit theorem that can be used to obtain asymptotically normal confidence intervals of t_y. To that end, we assume that $\widehat{t}_{pgd}$ is normally distributed, an assumption that is satisfied in many classical sampling designs; e.g., see Fuller (2009a).

(H12) The sequence of pseudo-generalized difference estimators $\{\widehat{t}_{pgd}\}$ satisfies

$$\frac{N_v^{-1}\left(\widehat{t}_{pgd}-t_y\right)}{\sqrt{\mathbb{V}_p\left(N_v^{-1}\widehat{t}_{pgd}\right)}} \xrightarrow[v\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

where $\mathbb{V}_p\left(N_v^{-1}\widehat{t}_{pgd}\right)$ is given by (3.23).

Result 3.4.4. *Consider the sequence of RF estimators* $\{\widehat{t}_{rf}\}$. *Then,*

$$\frac{N_v^{-1}\left(\widehat{t}_{rf}-t_y\right)}{\sqrt{\widehat{\mathbb{V}}_{rf}(N_v^{-1}\widehat{t}_{rf})}} \xrightarrow[v\to\infty]{\mathcal{L}} \mathcal{N}\left(0,1\right).$$

The proof of Result 3.4.4 is a direct application of Results 3.9.2 and 3.9.3, and is thus omitted.

3.5 A model calibration procedure for handling multiple survey variables

In practice, most surveys conducted by national statistical offices (NSO) collect information on multiple survey variables. The collected data are stored in rectangular data files. A column of weights, referred to as a weighting system, is made available on the data file. This weighting system can then be applied to obtain an estimate for any survey variable. However, applying a RF algorithm yield the variable-specific weights (3.20). In other words, the weights were derived to obtain an estimate of the total for a specific survey variable Y. Hence, applying the weights (3.20) to other survey variables may produce inefficient estimators. A solution to this issue consists of developing multiple sets of weights, one for each survey variable. This is usually deemed undesirable by data users who are used to work with a single set of weights. In this section, we describe a model calibration procedure (Wu and Sitter, 2001), originally proposed by Montanari and Ranalli (2009), that yields a single weighting system while accounting for multiple survey variables that are deemed important.

Suppose that we can identify a subset of survey variables $Y_1,\ldots,Y_q$, that are deemed important. We postulate the following working model for each variable:

$$\mathbb{E}\left[Y_{jk} \mid \boldsymbol{X}_k = \mathbf{x}_k\right] = m^{(j)}(\mathbf{x}_k^{(j)}), \quad j = 1,\cdots,q, \tag{3.25}$$

where $m^{(j)}(\cdot)$ is an unknown function and $\mathbf{x}_k^{(j)}$ is a vector of auxiliary variable associated with unit k for the variable Y_j. We allow a different link functions $m(\cdot)$ and different sets of explanatory variables for each of the survey variables $Y_1,\ldots,Y_q$. The interest lies in estimating the population totals $t_{y_1},\ldots,t_{y_q}$. We assume that each of these totals is estimated using a model-assisted estimator of the form (3.4) but with possibly different methods. For instance, some of the estimates may be based on a parametric working model, while others may be based on a nonparametric working model (e.g., RF). We can construct the set of q predicted values $\widehat{m}^{(1)}(\mathbf{x}_k^{(1)}),\ldots,\widehat{m}^{(q)}(\mathbf{x}_k^{(q)})$, for $k\in U$.

In addition, we assume that, at the estimation stage, a vector $\mathbf{v}_k$ of size q' of calibration variables is available for $k\in S$ and that the corresponding vector of population totals $\mathbf{t}_{\mathbf{v}} = \sum_{k\in U}\mathbf{v}_k$ is known. In practice, survey managers often want to ensure consistency between

survey estimates and known population totals for important variables such as gender and age group.

Given these predictions $\widehat{m}^{(1)}(\mathbf{x}_k^{(1)}), \ldots, \widehat{m}^{(q)}(\mathbf{x}_k^{(q)})$, and the vector calibration variables $\mathbf{v}$, we seek calibrated weights w_k^C, $k \in S$, as close as possible to the initial weights π_k^{-1} subject to the following $q + q' + 1$ calibration constraints:

$$\sum_{k \in S} w_k^C = N, \tag{3.26}$$

$$\sum_{k \in S} w_k^C \widehat{m}^{(j)}(\mathbf{x}_k^{(j)}) = \sum_{k \in U} \widehat{m}^{(j)}(\mathbf{x}_k^{(j)}), \quad j = 1, \ldots, q, \tag{3.27}$$

$$\sum_{k \in S} w_k^C \mathbf{v}_k = \sum_{k \in U} \mathbf{v}_k. \tag{3.28}$$

More specifically, we seek calibrated weights w_k^C such that

$$\sum_{k \in S} G(w_k^C / \pi_k^{-1})$$

is minimized subject to (3.26)–(3.28), where $G(\cdot)$ is a pseudo-distance function measuring the closeness between two sets of weights, such that $G(w_k^C/\pi_k^{-1}) \geq 0$, differentiable with respect to w_k^C, strictly convex, with continuous derivatives $g(w_k^C/\pi_k^{-1}) = \partial G(w_k^C/\pi_k^{-1})/\partial w_k^C$ such that $g(1) = 0$; see Deville and Särndal (1992).

The weights w_k^C are given by

$$w_k^C = \pi_k^{-1} F(\widehat{\boldsymbol{\lambda}}^\top \mathbf{h}_k),$$

where $F(.)$ is the calibration function defined as the inverse of $g(.)$, $\widehat{\boldsymbol{\lambda}}$ is a $q + q' + 1$-vector of estimated coefficients and

$$\mathbf{h}_k = \left(1, \widehat{m}_k^{(1)} - \widehat{\overline{m}}^{(1)}, \ldots, \widehat{m}_k^{(q)} - \widehat{\overline{m}}^{(q)}, v_{1k}, \ldots, v_{q'k}\right)^\top \tag{3.29}$$

with $\widehat{m}_k^{(j)} \equiv m^{(j)}(\mathbf{x}_k^{(j)})$ and $\widehat{\overline{m}}^{(j)} \equiv \sum_{k \in S} \pi_k^{-1} \widehat{m}_k^{(j)} / \sum_{k \in S} \pi_k^{-1}$, $j = 1, \cdots, q$.

The calibrated weights w_k^C may be viewed as a compressed score summarizing the information contained in the q working models (3.25) and the vector of calibration variables $\mathbf{v}$. The weighting system $\{w_k^C; k \in S\}$ may be then applied to any survey variable Y, which leads to the model calibration type estimator

$$\widehat{t}_{y,mc} = \sum_{k \in S} w_k^C y_k.$$

If the number of calibration constraints $q + q' + 1$ is large, the resulting weights w_k^C may be highly dispersed leading to potentially unstable estimates $\widehat{t}_{y,mc}$. A number of pseudo-distance functions such as the truncated linear and the logit methods may be used to limit the variability of the weights w_k^C; see Deville and Särndal (1992) for a description of these methods. A simple alternative is to use additional constraints on the weights as part of the calibration constraints. For instance, we may impose that $w_k^C < w_0$, where w_0 is a threshold set by the survey statistician; see also Santacatterina and Bottai (2018) for alternative constraints on the weights. Finally,

we can relax the calibration constraints (3.26)-(3.28) by considering a L^2-penalized criterion, leading to a ridge-type model calibration estimator; see Montanari and Ranalli (2009). Montanari and Ranalli (2009) reports the results of a simulation study, assessing the performance of point estimators obtained through multiple and ridge model calibration methods.

3.6 Simulation study

3.6.1 Performance of point estimators

We conducted a simulation study to assess the performance of several model-assisted estimators, in terms of bias and efficiency. We generated a finite population of size $N = 10,000$, consisting of a set of auxiliary variables and 8 survey variables. We first generated 7 auxiliary variables $X_0, \cdots, X_6$, according to the following distributions: $X_0 \sim \mathcal{U}(0,1)$; $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim$ Beta$(3,1)$, $X_3 \sim 2 \times$ Gamma$(3,2)$, $X_4 \sim$ Bernoulli(0.7), $X_5 \sim$ Multinomial$(0.4, 0.3, 0.3)$ and $X_6 \sim \mathcal{E}(1)$. The variables X_1, X_2, X_3, and X_6 have been standardized so as to have a mean and a variance equal to 0 and 1, respectively. To assess the performance of the proposed method in a high-dimensional setting, we also generated 100 additional auxiliary variables $V_1, V_2, \cdots, V_{100}$, from a uniform distribution $\mathcal{U}(-1,1)$. Given the X-variables and the V-variables, we generated the survey variables according to the following models:

Model 1: $Y_1 = 1 + 2(X_0 - 0.5) + \mathcal{N}(0, 0.1)$;

Model 2: $Y_2 = 1 + 2(X_0 - 0.5)^2 + \mathcal{N}(0, 0.1)$;

Model 3: $Y_3 = 2 + X_6 + X_2 + X_3 + X_4 + X_5 + \mathcal{N}(0,1)$;

Model 4: $Y_4 = 2 + (X_6 + X_2 + X_3)^2 + \mathcal{N}(0,1)$;

Model 5: $Y_5 = 0.5X_5 + \exp(-X_1) + 3X_4 + \exp(-X_6) + \mathcal{E}(1)$;

Model 6: $Y_6 = V_1^2 + \exp(-V_2^2) + \mathcal{N}(0, 0.3)$;

Model 7: $Y_7 = V_1^2 + \exp(-V_2^2) + \mathcal{N}(0, 0.3)$;

Model 8: $Y_8 = 3 + V_1V_2 + V_3^2 - V_4V_7 + V_8V_{10} - V_6^2 + \mathcal{N}(0, 0.5)$.

The errors in Model 5 have been scaled and centered so as to have a mean and a variance equal to 0 and 1, respectively. Models 1 and 2 were used in Breidt and Opsomer (2000), while Models 7 and 8 were introduced in Scornet (2017). Models 1-8 were generated so as to include a relatively wide range of relationships between the Y-variable and the set of explanatory variables: linear/non-linear relationships, presence/absence of quadratic terms and presence/absence of interactions. Our scenarios also included low, medium and high-dimensional settings. From the population, we selected $R = 5,000$ samples, of size n, according to simple random sampling without replacement. We used $n = 250$ and $n = 1,000$. In each sample, we computed the following estimators: (i) The Horvitz-Thompson (HT) estimator given by (3.1); (ii) The generalized regression (GREG) estimator given by (3.4) with $\widehat{m}(\mathbf{x}_k) = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}$; (iii) The model-assisted estimator (3.4) with $\widehat{m}(\mathbf{x}_k)$ obtained through regression trees (CART); and (iv) The model-assisted estimator (3.4) based on RF, where $\widehat{m}(\mathbf{x}_k)$ is given by (3.17). We considered three RF algorithms, each based on 1,000 trees. The first (RF1) was based on

bootstrap. The second algorithm (RF2) was based on subsampling with a sampling fraction equal to 0.63 (Scornet, 2017). For both RF1 and RF2, the minimum number of observations per terminal node was set to $n_0 = 5$. Finally, the third algorithm (RF3) was based on bootstrap with $n_0 = \sqrt{n}$ observations in each terminal node. In RF1-RF3, we used $m_{try} = \sqrt{p}$ as it is the default number of variables considered for the splitting process in most software packages dealing with RF for regression.

For the estimators GREG, CART, RF1, RF2 and RF3, the predictions $\widehat{m}(\mathbf{x}_k)$ were obtained using the working models described in Table 3. For the survey variables Y_7 and Y_8, the working models were based on a large number of superfluous explanatory variables (50 and 100, respectively), which allowed us to assess the behavior of the resulting estimators in a medium/high dimensional setting.

Table 3: The working models

Survey variable	Vector of explanatory variable $\mathbf{X}$ used in the working model
Y_1	X_0
Y_2	X_0
Y_3	$X_1 - X_6$
Y_4	$X_1 - X_6$
Y_5	$X_1 - X_6$
Y_6	$V_1 - V_{10}$
Y_7	$V_1 - V_{50}$
Y_8	$V_1 - V_{100}$

We were interested in estimating the population totals $t_{y_j} = \sum_{k \in U} y_{kj}$, $j = 1, \ldots, 8$. As a measure of bias of an estimator $\hat{t}_{y_j}$, we used the Monte Carlo percent relative bias defined as

$$RB(\widehat{t}_{y_j}) = 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{(\widehat{t}_{y_j}^{(r)} - t_{y_j})}{t_{y_j}},$$

where $\widehat{t}_{y_j}^{(r)}$ denotes the estimator $\widehat{t}_{y_j}$ in the rth iteration, $r = 1, ..., R$. As a measure of efficiency of an estimator $\hat{t}_{y_j}$, we used the relative efficiency, using the Horvitz-Thompson estimator, $\widehat{t}_{y_j,\pi}$, as the reference:

$$RE(\widehat{t}_{y_j}) = 100 \times \frac{MSE(\widehat{t}_{y_j})}{MSE(\widehat{t}_{y_j,\pi})},$$

where

$$MSE(\widehat{t}_{y_j}) = \frac{1}{R} \sum_{r=1}^{R} (\widehat{t}_{y_j}^{(r)} - t_{y_j})^2$$

and $MSE(\widehat{t}_{y_j,\pi})$ is defined similarly. The results are displayed in Tables 4 and 5. The simulations were performed using the R software with the package *ranger* (Wright and Ziegler, 2015).

We start by noting that all the estimators displayed a negligible bias in all the scenarios, as expected. Also, both RF1 and RF2 showed very similar performances in terms of bias and efficiency in all the scenarios. This is consistent with the empirical results of Scornet (2017);

Table 4: Monte Carlo percent relative bias (RB) and Monte Carlo efficiency (RE) of several model-assisted estimators for $n = 250$

Population		GREG	CART	RF1	RF2	RF3
Y_1	RB	-0.0	-0.0	-0.0	-0.0	0.0
	RE	3.0	3.5	3.7	3.6	3.4
Y_2	RB	-0.0	0.0	0.0	0.0	0.0
	RE	101.0	37.6	39.4	38.3	35.0
Y_3	RB	0.0	-0.0	-0.1	-0.1	-0.0
	RE	19.6	55.2	33.8	34.0	35.4
Y_4	RB	-0.7	-1.2	-1.2	-1.5	-0.7
	RE	81.1	61.1	49.7	49.0	53.1
Y_5	RB	-0.1	0.1	-0.0	-0.0	-0.0
	RE	37.9	32.7	25.8	26.5	30.7
Y_6	RB	-0.0	0.3	-0.0	-0.0	-0.0
	RE	105.2	72.2	57.5	57.5	58.3
Y_7	RB	-0.0	0.2	0.1	0.0	0.0
	RE	127.6	84.3	75.8	75.5	76.8
Y_8	RB	0.0	0.0	0.0	0.0	0.0
	RE	127.0	135.6	92.7	92.5	95.6

i.e., the strategy based on bootstrap and the strategy based on subsampling with a sampling fraction of 0.63 led to similar performances. The results for RF3 were similar to those obtained for RF1 and RF2, which suggests that the number of observations in each terminal node did not seem to affect the behavior of the point estimator, at least in our experiments. This may not be the case in other scenarios as we illustrate in Section 3.6.3.

In the case of a linear relationship (which corresponds to the survey variables Y_1 and Y_3), the GREG estimator was the most efficient, as expected. For instance, for the survey variables Y_3, the value of RE for the GREG estimator was about 19.6%, whereas the RF1, RF2 and RF3 estimators showed a value of RE of about 34%. In the case of a nonlinear relationship (which corresponds to the survey variables Y_2 and $Y_4, \ldots, Y_8$), the GREG estimator was less efficient than RF1, RF2 and RF3. For instance, in the case of the variable Y_4, the GREG showed a value of RE of about 81.1%, whereas the RE of RF estimators lied between 49.0% and 53.1%. For the variables Y_6, Y_7, Y_8, the GREG estimator was even less efficient than the Horvitz-Thompson estimator with values of RE ranging from 105% to 127%.

In the case of a single explanatory variable (which corresponds to the survey variables Y_1 and Y_2), RF and regression trees displayed very similar performances. In contrast, the estimators

Table 5: Monte Carlo percent relative bias (RB) and Monte Carlo efficiency (RE) of several model-assisted estimators for $n = 1000$.

Population		GREG	CART	RF1	RF2	RF3
Y_1	RB	0.0	0.0	0.0	0.0	0.0
	RE	2.8	3.5	3.6	3.5	3.0
Y_2	RB	0.0	0.0	0.0	0.0	0.0
	RE	100.1	38.7	40.5	39.6	33.3
Y_3	RB	0.0	0.0	-0.1	-0.1	0.0
	RE	20.4	41.1	28.1	27.8	31.6
Y_4	RB	-0.1	-1.1	-0.9	-0.7	-0.2
	RE	78.9	52.3	36.7	36.1	44.5
Y_5	RB	-0.0	0.0	0.0	0.0	-0.0
	RE	37.3	24.5	20.9	21.2	24.8
Y_6	RB	0.0	0.0	-0.0	-0.0	-0.0
	RE	101.1	65.5	49.1	49.2	50.3
Y_7	RB	0.0	0.0	0.0	0.0	0.0
	RE	105.5	73.2	63.3	63.2	65.0
Y_8	RB	-0.0	-0.0	-0.0	-0.0	0.0
	RE	166.6	137.6	96.0	95.7	89.5

RF1, RF2 and RF3 were more efficient than the CART estimator when the vector of explanatory variables was multi-dimensional (i.e., variables $Y_3, \ldots, Y_9$). In a high-dimensional setting (which corresponds to the survey variables Y_7 and Y_8), the RF estimators were more efficient than the Horvitz-Thompson estimator, even for $n = 250$.

3.6.2 Performance of the proposed variance estimator

We have also investigated the performance of the variance estimator $\widehat{\mathbb{V}}_{rf}$ given by (3.24) in the case of RF with subsampling, in terms of relative bias and coverage of normal-based confidence intervals. We generated a population of size $N = 100,000$ according to Model 5. The sample size was set to $n = 500; 1,000; 5,000; 10,000; 20,000$ and $50,000$. Here, we present the results for $B = 1$ but other values of B led to similar results and are not shown here. As we suspected that the number of observations in each terminal node, n_0, may have an impact on the behavior of

$\widehat{\mathbb{V}}_{rf}$, we used different values for n_0 : $n_0 = \lfloor n^{a/20} \rfloor$ for $a = 1; 3; 5; 7; 9; 11; 13; 15; 17$. The choice $n_0 = \lfloor n^{11/20} \rfloor$ was advocated by McConville and Toth (2019). Figure 16 shows the Monte Carlo percent relative bias of $\widehat{\mathbb{V}}_{rf}$ for different values of n and n_0. Figure 17 shows the Monte Carlo coverage rate of the confidence interval, $\widehat{t}_{rf} \pm 1.96\sqrt{\widehat{\mathbb{V}}_{rf}}$, for different values of n and n_0.

From Figure 16, we note that $\widehat{\mathbb{V}}_{rf}$ is severely biased for small values of n_0 and as a consequence, the confidence intervals (see Figure 17) perform poorly for small values of n_0 because of the substantial underestimation of the true variance in these scenarios. For a given value of n_0, we note that the bias decreases as n increases and for a given value of n, the bias decreases as n_0 increases. For $n_0 = \lfloor 13/20 \rfloor$, the confidence intervals perform relatively well with coverage rates close to the nominal rate. The significant bias for small values of n_0 is most likely due to overfitting, which is characterized by the presence of artificially small residuals $y_k - \widehat{m}(\mathbf{x}_k)$ in each terminal node, which in turn, leads to underestimation. This issue was raised by Opsomer and Miller (2005) in the context of local polynomial regression. To cope with this issue, we suggest a variance estimator based on a K-fold criterion. More specifically, we randomly split the sample S into K groups S_κ, $\kappa = 1, \ldots, K$, of approximately equal size. For $k \in S_\kappa$, let $\widehat{m}^{(-\kappa)}(\mathbf{x}_k)$ denote the prediction at the point $\mathbf{x}_k$ built on $S - S_\kappa$ and $\widehat{\epsilon}_k^{(-\kappa)} = y_k - \widehat{m}^{(-\kappa)}(\mathbf{x}_k)$ the associated residual. The proposed K-fold variance estimator is given by $\widehat{V}^{(K)} = \sum_{\kappa_1=1}^{K} \sum_{\kappa_2=1}^{K} \sum_{k \in S_{\kappa_1}} \sum_{\ell \in S_{\kappa_2}} (\Delta_{k\ell}/\pi_{k\ell})(\widehat{\epsilon}_k^{(-\kappa_1)}/\pi_k)(\widehat{\epsilon}_\ell^{(-\kappa_2)}/\pi_\ell)$. In practice, the number of groups (or folds) is often set to $K = 5$ or $K = 10$. We tested the performance of $\widehat{V}^{(5)}$ in terms of bias and coverage probability, using the same scenarios as above. The bias was almost negligible for all sizes n and n_0 and the coverage rates lied between 93% and 96%, which constitutes a significant improvement over the results displayed in Figures 16 and 17. More research is needed in order to establish the theoretical properties of the variance estimator based on a K-fold criterion evaluate. It will be treated elsewhere.

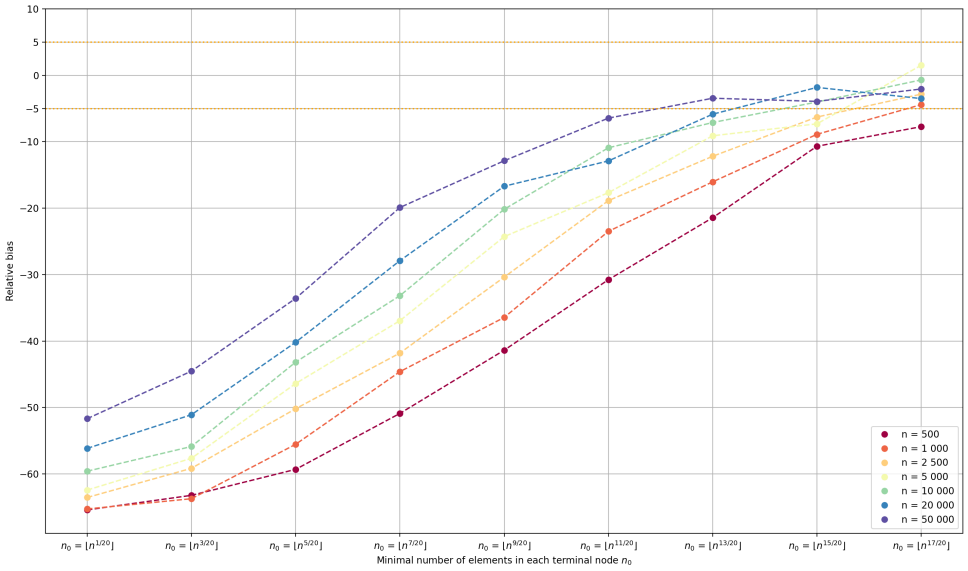


Figure 16: Evolution of the relative bias with respect to n_0.

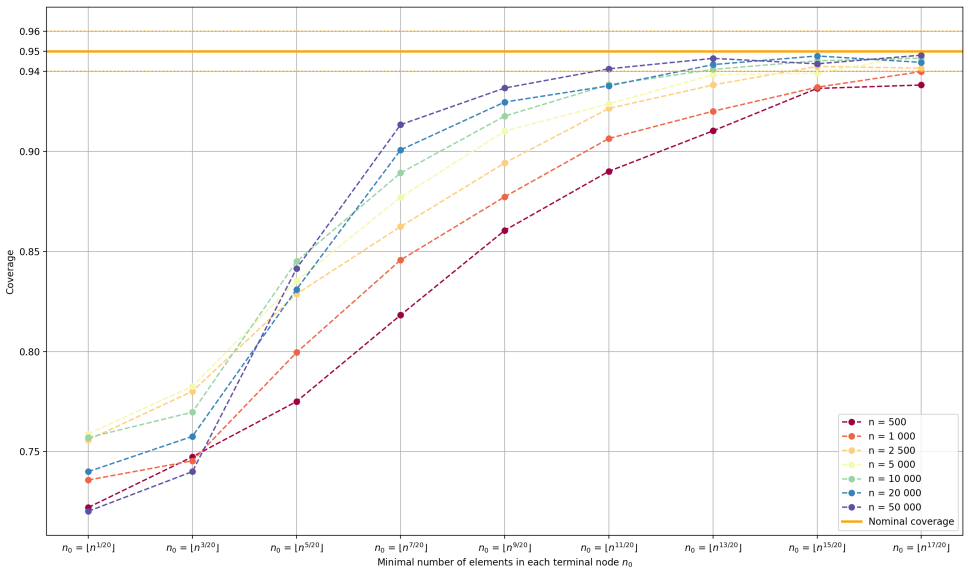


Figure 17: Evolution of the effective coverage with respect to n_0.

3.6.3 Choice of hyper-parameters

To get a better understanding of how the choice of hyper-parameters impacts the behavior of model-assisted estimators based on RF, we conducted additional scenarios. We first identified the following important hyper-parameters involved in the RF algorithm of Breiman (2001):

i) The minimal number of observations, n_0, in each terminal node;

ii) The number of trees in the forest B;

iii) The number of variables considered for the search of the best split in the optimization criterion (3.16);

iv) The resampling mechanism.

The additional scenarios were conducted using a finite population of size $N = 10,000$ consisting of the survey variables Y_5 and Y_8 described in Section 3.6.1. Recall that the working model for the survey variables Y_5 included the predictors $X_1, \ldots, X_6$, whereas it included the predictors $V_1, \ldots, V_{100}$ for the variable Y_8 (see Table 3).

From the population, we generated $R = 10,000$ samples, of size $n = 1,000$, according to simple random sampling without replacement. Figure 18 and Figure 19 show, respectively, the relative efficiency of the model-assisted estimators based on RF, $\widehat{t}_{rf}$, corresponding to Y_5 and Y_8, respectively, for several values of n_0. Figure 18 suggests that $\widehat{t}_{rf}$ was much more efficient than the Horvitz-Thompson estimator for small values of n_0 and that the value of RE approached 100 as n_0 increased. This result can be explained by the fact that small values of n_0 led to homogeneous terminal nodes, which in turn led to small residuals $y_k - \widehat{m}_{rf}(\mathbf{x}_k)$. For the survey variable Y_8, we note from Figure 19 that the value of n_0 did not seem to affect the efficiency of the corresponding model-assisted estimator.

Figure 20 display the relative efficiency for several values of B, the number of trees in the forest for the survey variable Y_8. As expected, a small value of B causes the estimator $\widehat{t}_{rf}$ to loose some efficiency. Figure 20 suggests that $B = 50$ led to good results and that the efficiency

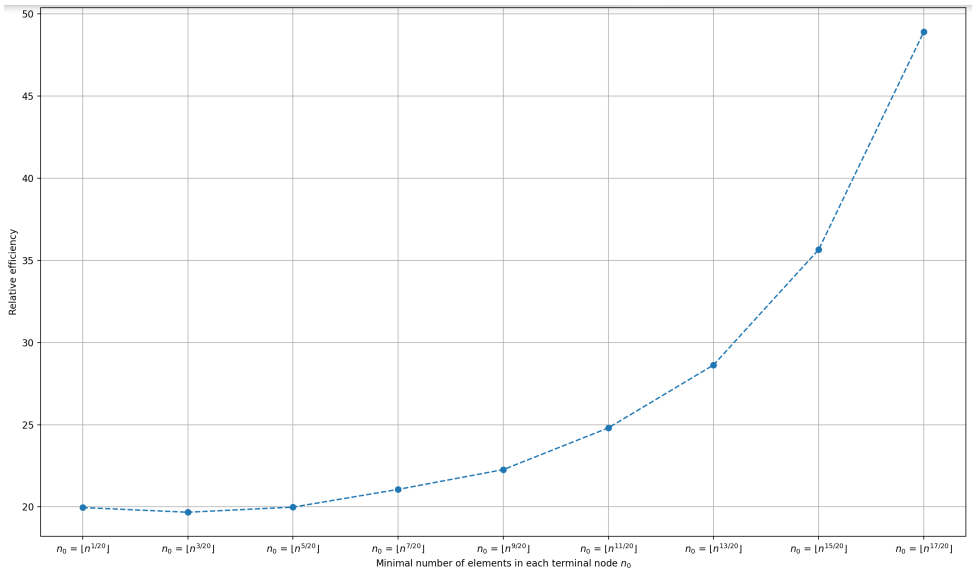


Figure 18: Relative efficiency of $\widehat{t}_{rf}$ for the survey variable Y_5 and for several values of n_0.

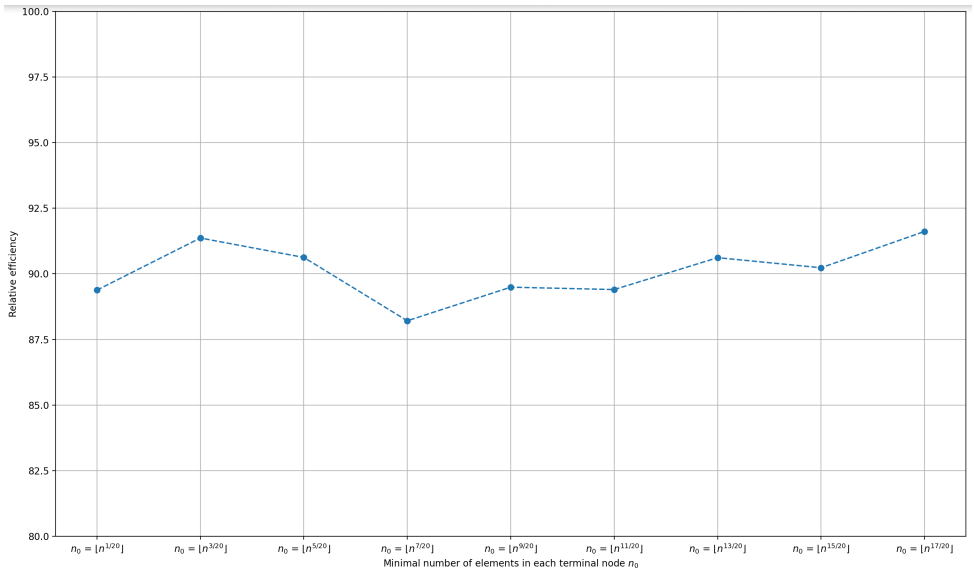


Figure 19: Relative efficiency of $\widehat{t}_{rf}$ for the survey variable Y_8 and for several values of n_0

of $\widehat{t}_{rf}$ was not much affected by the number of trees B for $B \geq 50$. Nevertheless, it is advisable to choose a large value of B if the computational capacity permits. The results for the survey variable Y_5 were very similar and so we omit them.

In most software packages, the default number of variables considered for the splitting process is $m_{try} = \sqrt{p}$ in case of regression. In our simulations, this choice led to satisfactory results in most scenarios. Figure 21 shows the relative efficiency of $\widehat{t}_{rf}$ for the survey variable Y_8 and for several values of m_{try}. Since the working model for Y_8 contained $p = 100$ explanatory variables, the default value $\sqrt{p}$ was equal to 10. Although the value $\sqrt{p} = 10$ was not the best choice for optimal performances, it led efficient model-assisted estimators. Furthermore, the relative efficiency did not vary much for values B larger than 30.

Turning to the resampling mechanism, a common choice is to use bootstrap (with replacement), for which some of the results presented in the paper do not apply. However, as noted by several authors (see e.g. Scornet et al. (2015), Wager (2014) and the references therein) and as shown in

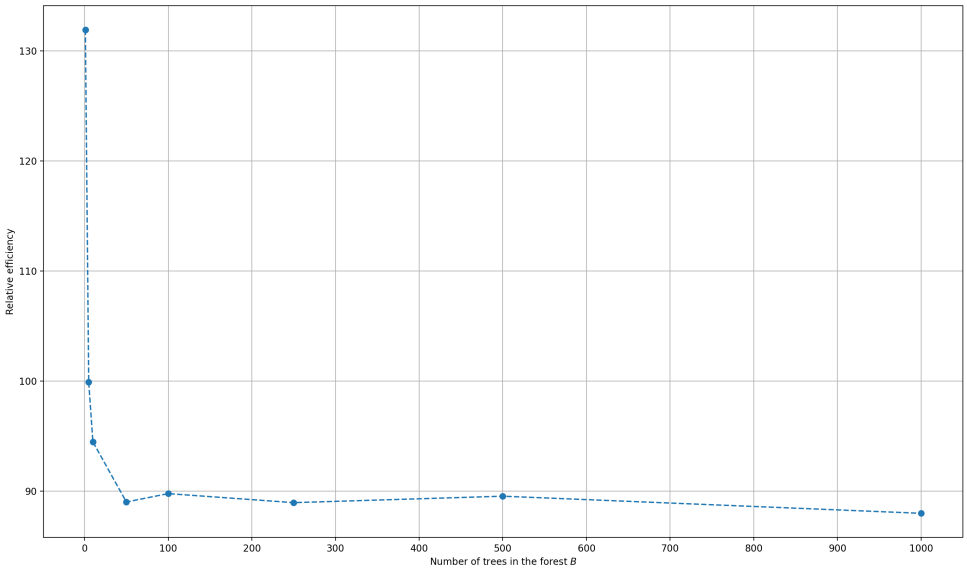


Figure 20: Relative efficiency of $\widehat{t}_{rf}$ for the survey variable Y_8 and for several values of B.

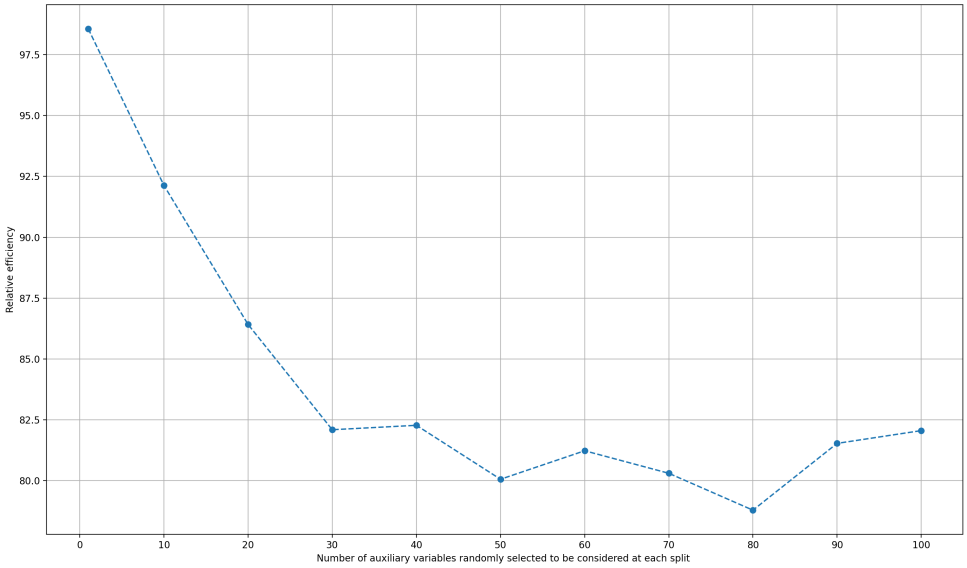


Figure 21: Relative efficiency of $\widehat{t}_{rf}$ for the survey variable Y_8 and for several values of m_{try}.

our simulations, selecting the points without replacement rather than with replacement does not seem to affect the performance of the resulting model-assisted estimators in most cases.

3.6.4 Real data application

In this section, we apply the proposed methods using data collected by Médiamétrie, the company that measures the media audience in France. In this application, we focus on radio audiences. Each year, Médiamétrie conducts a survey aiming at gathering detailed information about French individuals 13 years of age and over, including socio-demographic variables and radio listening habits. We used the 2019 radio audience data that consisted of $N = 26,293$ individuals. As a survey variable, we considered the binary variable Y, such that $y_k = 1$ if an individual in the kth individual listens to the radio of interest on a daily basis, and $y_k = 0$

otherwise. For confidentiality reasons, we omit the name of the radio broadcaster. We aimed at estimating the proportion of French individuals who listen to the radio of interest on a daily basis, both at the overall population level and for several domains of interest. For each individual, we had access to 43 socio-demographic variables (e.g., number of individuals in the household, age of each member of the household, gender, internet habits, occupation, etc.) and their listening habits of 21 other radios. For each individual, we also knew whether or not the individual listens to any of these 21 radios, for each interval of 7.5 minutes on a typical day. This led to a data set with $p = 3,882$ variables, among which $3,839$ were binary.

From the data set, we selected a single sample of size $n = 4,000$ according to a stratified sampling design with 5 strata, each stratum corresponding to a French region: North-East, North-West, Île-de-France, South-east and South-West. The strata sample sizes were determined according to proportional allocation. We considered the following domains of interest: the sub-population of individuals who connects to the internet everyday, almost every day, once or twice per week, once to three times per month, very rarely, never, the sub-population of individuals with/without children, the sub-populations of individuals living in cities of size (less than $20,000$, between $20,000$ and $50,000$, between $50,000$ and $100,000$, between $100,000$ and $200,000$ and larger than $200,000$) and the sub-population of individuals living in households of size $1, 2, 3, 4, 5$ and $5+$

We computed the following estimates both at the overall level and at the domain level: (i) The Horvitz-Thompson estimator; (ii) the GREG estimator and (iii) the model-assisted estimator based on RF with hyper-parameters $B = 1,000$, $n_0 = \lfloor n^{11/20} \rfloor$ and $m_{try} = \sqrt{p}$. The working models used for the GREG estimator and the model-assisted RF estimator included 3882 explanatory variables. In each scenario, we also computed a 95% confidence interval for the proportion in the population of individuals who listen to the radio of interest. Finally, we computed the ratio of the estimated variances, using the estimated variance of the Horvitz-Thompson estimator, as the reference. Note that the "true value" was known for each domain of interest. The results (in percentage) are given in Table **??**.

From Table **??**, we note that the Horvitz-Thompson estimator performed relatively well in most scenarios. Because of the large number of predictors, the GREG estimator suffered from significant small sample bias. For instance, the estimate based on the GREG estimate at the overall level was equal to 27.7%, far from the true value of about 13.5%. In terms of point estimation, the RF model-assisted estimator led to very similar results than those obtained with the Horvitz-Thompson estimator. However, RF led to substantial improvement in terms of estimated variance. Indeed, out of the 22 domains, the value of RV(RF) was smaller than 0.65 for 20 domains. The results suggest that, unlike the GREG estimator, the model-assisted estimator based on RF was not affected by the large number of explanatory variables in the working model. The median length of the confidence intervals was equal to 5.4% for the Horvitz-Thompson estimator, 4.2% for the RF estimator and 3.8% for the GREG estimator.

3.7 Final remarks

In this paper, we have introduced a new class of model-assisted estimators based on random forests and derived corresponding variance estimators. We have established the theoretical properties of point and variance estimators obtained through a RF algorithm based on subsampling. The results of an empirical study suggest that the proposed estimators perform well in a wide variety of settings, unlike the GREG and CART estimators. In practice, this

robustness property is especially attractive when the data and the underlying relationships are complex. The application on radio audience data recorded by the French company Médiamétrie showed that the RF proposed estimator performed well in this high-dimension setting. We have also described a model calibration procedure for handling multiple survey variables, yet producing a single set of weights, which is attractive from a data user's perspective.

In practice, virtually all survey face the problem of missing values. Survey statisticians distinguish unit nonresponse (when no information is collected on a sampled unit) from item nonresponse (when the absence of information is limited to some variables only). The treatment of unit nonresponse starts with postulating a nonresponse model describing the relationship between the response indicators (equal to 1 for respondents and 0 for nonrespondents) and a vector of explanatory variables. The treatment of item nonresponse starts with postulating an imputation model describing the relationship between the variable requiring imputation and a set of explanatory variables. In both unit and item nonresponse, determining a suitable model is crucial. Therefore, regression trees and RF may prove useful for obtaining accurate estimated response propensities and predicted values. To the best of our knowledge, a theoretical treatment of regression trees and RF in the context of either unit nonresponse or item nonresponse in a finite population setting is lacking. These topics are currently under investigation.

Traditionally, survey samples have been collected through probability sampling procedures and inferences were conducted with respect to the customary design-based framework. In recent years, there has been a shift of paradigm that can be explained by three main factors: (i) a dramatic decrease of response rates; (ii) a rapid increase in data collection costs; and (iii) the proliferation of nonprobabilistic data sources (e.g., administrative files, web survey panels, social media data, satellite information, etc.). To meet these new challenges, survey statisticians face increasing pressure to utilize these convenient but often uncontrolled data sources. While such sources provide timely data for a large number of variables and population elements, they often fail to represent the target population of interest because of inherent selection biases. The integration of data from a nonprobability source with data from a probability survey is a topic that is currently being scrutinized by National Statistical Offices. An approach to data integration is statistical matching or mass imputation; see Yang and Kim (2020) for a very recent review on the topic. Again, regression trees and RF algorithms may prove useful in the context of integration of survey data. This topic is currently under investigation.

In a high-dimensional setting, RF may be used to select the most predictive predictors, which in turn may be used in the construction of model-assisted estimators of population totals/means. In this context, issues such as variable selection bias (Strobl et al., 2007) in a finite population setting need to be investigated. This will be treated elsewhere.

3.8 Appendix

Proof of Proposition 3.2.1 Since

$$\widetilde{W}_\ell(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \frac{\psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}\left(\mathbf{x}_k, \theta_b^{(U)}\right)}}{\widetilde{N}(\mathbf{x}_k, \theta_b^{(U)})}, \tag{3.30}$$

involves positive quantities only, the weights $\widetilde{W}_\ell(\mathbf{x}_k)$ are nonnegative. Since $\psi_\ell^{(b,U)} \in \{0,1\}$ for all $\ell \in U$ and for all $b \in 1,2,\ldots,B$, the weight can be bounded as follows:

$$\widetilde{W}_\ell(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \frac{\psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}\left(\mathbf{x_k}, \theta_b^{(U)}\right)}}{\widetilde{N}(\mathbf{x}_k, \theta_b^{(U)})} \leqslant \frac{1}{B}\sum_{b=1}^{B}\left(\widetilde{N}\left(\mathbf{x}_k, \theta_b^{(U)}\right)\right)^{-1} \leqslant cN_0^{-1}.$$

where c does not depend on b nor on k or ℓ. To show ii), fix $b \in 1,2,...,B$. The result follows by noting that $\widetilde{W}_\ell(\mathbf{x}_k) = \left(\widetilde{N}\left(\mathbf{x}_k, \theta_b^{(U)}\right)\right)^{-1}$ exactly $\widetilde{N}\left(\mathbf{x}_k, \theta_b^{(U)}\right)$ times.

Proof of Proposition 3.3.2 Let $\{\widehat{t}^{(b)}\}$ be a sequence of estimators of t_y. Then,

$$\mathbb{V}_p\left(\frac{1}{B}\sum_{b=1}^{B}\widehat{t}^{(b)}\right) = \frac{1}{B^2}\sum_{b=1}^{B}\left(\mathbb{V}_p\left(\widehat{t}^{(b)}\right) + \sum_{b=1}^{B}\sum_{b\neq b'=1}^{B}\mathbb{C}or_p\left(\widehat{t}^{(b)}, \widehat{t}^{(b')}\right)\mathbb{V}_p^{1/2}\left(\widehat{t}^{(b)}\right)\mathbb{V}_p^{1/2}\left(\widehat{t}^{(b')}\right)\right).$$
$$\leqslant \frac{\mathbb{V}_p(\widehat{t}^{(1)})}{B} + \mathbb{V}_p(\widehat{t}^{(1)}) \max_{b\neq b'}\left|\mathbb{C}or_p\left(\widehat{t}^{(b)}, \widehat{t}^{(b')}\right)\right|$$

and $\text{Bias}_p^2(B^{-1}\sum_{b=1}^{B}\widehat{t}^{(b)}) = \text{Bias}_p^2(\widehat{t}^{(1)}) = MSE_p(\widehat{t}^{(1)}) - \mathbb{V}_p(\widehat{t}^{(1)})$. So, for B large enough:

$$MSE_p\left(\frac{1}{B}\sum_{b=1}^{B}\widehat{t}^{(b)}\right) \leqslant \mathbb{V}_p(\widehat{t}^{(1)}) \max_{b\neq b'}\left|\mathbb{C}or_p\left(\widehat{t}^{(b)}, \widehat{t}^{(b')}\right)\right| - \mathbb{V}_p(\widehat{t}^{(1)}) + MSE_p(\widehat{t}^{(1)}).$$

Proof of Proposition 3.3.3 Consider the B partitions build at the sample level $\widehat{\mathcal{P}}_S = \{\widehat{\mathcal{P}}_S^{(b)}\}_{b=1}^{B}$. For a given $b = 1,\ldots,B$, the partition $\widehat{\mathcal{P}}_S^{(b)}$ is composed by disjointed regions as follows $\widehat{\mathcal{P}}_S^{(b)} = \{A_j^{(bS)}\}_{j=1}^{J_{bS}}$ and for each b, consider the J_{bS} dimensional vector $\hat{\mathbf{z}}_k^{(b)} = \left(\mathbb{1}_{\mathbf{x}_k \in A_1^{(bS)}}, \ldots, \mathbb{1}_{\mathbf{x}_k \in A_{J_{bS}}^{(bS)}}\right)^\top$ where $\mathbb{1}_{\mathbf{x}_k \in A_j^{(bS)}} = 1$ if $\mathbf{x}_k$ belongs to the region $A_j^{(bS)}$ and zero otherwise for all $j = 1,\ldots,J_{bS}$. Since $\{A_j^{(bS)}\}_{j=1}^{J_{bS}}$ is a partition, then $\mathbf{x}_k$ will belong to only one region and so, the vector $\hat{\mathbf{z}}_k^{(b)}$ will contain only one non zero component. We have $\widehat{m}_{rf}(\mathbf{x}_k) = B^{-1}\sum_{b=1}^{B}\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})$ and $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})$ can be written as $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}) = (\hat{\mathbf{z}}_k^{(b)})^\top\widehat{\boldsymbol{\beta}}^{(b)}$ where $\widehat{\boldsymbol{\beta}}^{(b)} = \left(\sum_{\ell\in S}\pi_\ell^{-1}\psi_\ell^{(b,S)}\hat{\mathbf{z}}_\ell^{(b)}(\hat{\mathbf{z}}_\ell^{(b)})^\top\right)^{-1}\sum_{\ell\in S}\pi_\ell^{-1}\psi_\ell^{(b,S)}\hat{\mathbf{z}}_\ell^{(b)}y_\ell$ (see also the supplementary materiel for more details). Now,

$$\sum_{k\in S}\frac{y_k - \widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k} = \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S}\frac{y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})}{\pi_k} = \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S}\frac{(1-\psi_k^{(b,S)})(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}))}{\pi_k}$$
$$+\frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S}\frac{\psi_k^{(b,S)}(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}))}{\pi_k}.$$

For each b, consider the J_{bS} dimensional vector $\mathbf{1}_{J_{bS}}$ whose elements are all equal to one, we have then $\mathbf{1}_{J_{bS}}^\top\hat{\mathbf{z}}_k^{(b)} = 1$ for all k, so

$$\sum_{k\in S}\frac{\psi_k^{(b,S)}}{\pi_k}\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}) = \sum_{k\in S}\frac{\psi_k^{(b,S)}}{\pi_k}(\hat{\mathbf{z}}_k^{(b)})^\top\widehat{\boldsymbol{\beta}}^{(b)} = \mathbf{1}_{J_{bS}}^\top\sum_{k\in S}\frac{\psi_k^{(b,S)}}{\pi_k}\hat{\mathbf{z}}_k^{(b)}(\hat{\mathbf{z}}_k^{(b)})^\top\widehat{\boldsymbol{\beta}}^{(b)} = \sum_{\ell\in S}\frac{\psi_\ell^{(b,S)}}{\pi_\ell}y_\ell.$$

3.9 Supplementary material

Assumptions: population-based RF model-assisted estimator $\hat{t}^*_{rf}$

To establish the properties of the proposed estimators, we will consider three categories of assumptions[2]: assumptions on the sampling design, assumptions on the survey variable and, finally, assumptions on the random forests.

(H13) We assume that there exists a positive constant C such that $\sup_{k\in U_v} |y_k| \leqslant C < \infty$.

(H14) We assume that $\lim_{v\to\infty} \dfrac{n_v}{N_v} = \pi \in (0;1)$.

(H15) There exist positive constants λ and λ^* such that $\min_{k\in U_v} \pi_k \geqslant \lambda > 0$, $\min_{k,\ell\in U_v} \pi_{k\ell} \geqslant \lambda^* > 0$ and $\limsup_{v\to\infty} n_v \max_{k\neq\ell\in U_v} |\pi_{k\ell} - \pi_k\pi_\ell| < \infty$.

(C3) The number of subsampled elements N'_v is such that $\lim_{v\to\infty} N'_v/N_v \in (0;1]$.

Assumptions: sample-based RF model-assisted estimator $\hat{t}_{rf}$

(H16) We assume that there exists a positive constant $C_1 > 0$ such that

$$n_v \max_{k\neq\ell\in U_v} \left|\mathbb{E}_p\left\{(I_k - \pi_k)(I_\ell - \pi_\ell)|\widehat{\mathcal{P}}_S\right\}\right| \leq C_1.$$

(H17) The random forests based on population partitions and those based on sample partitions are such that, for all $\mathbf{x} \in \mathbb{R}^p$:

$$\mathbb{E}_p\left(\widehat{\widetilde{m}}_{rf}(\mathbf{x}) - \widetilde{m}_{rf}(\mathbf{x})\right)^2 = o(1).$$

where $\widehat{\widetilde{m}}_{rf}(\mathbf{x})$ is given by

$$\widehat{\widetilde{m}}_{rf}(\mathbf{x}) = \sum_{\ell\in U_v} \frac{1}{B}\sum_{b=1}^{B} \frac{\psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x},\theta_b^{(S)}\right)}}{\widehat{\widetilde{N}}(\mathbf{x},\theta_b^{(S)})} y_\ell$$

with $\widehat{\widetilde{N}}(\mathbf{x},\theta_b^{(S)}) = \sum_{k\in U_v} \psi_k^{(b,S)} \mathbb{1}_{\mathbf{x}_k \in A^{(S)}\left(\mathbf{x},\theta_b^{(S)}\right)}$ and

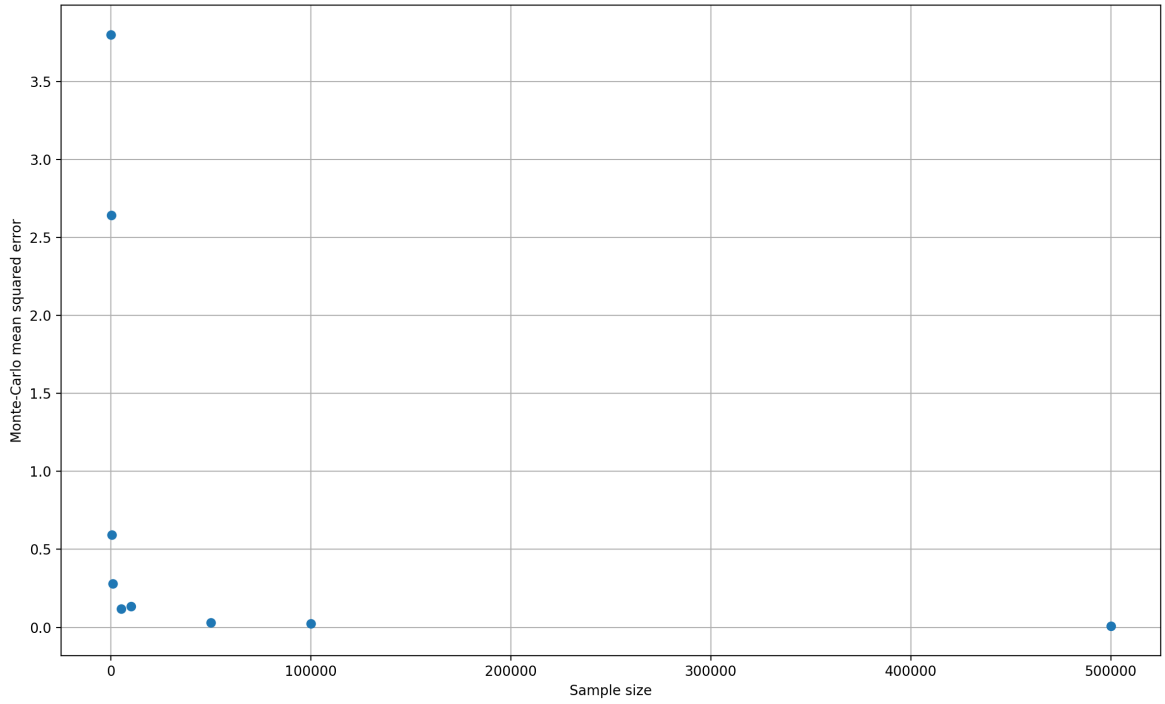
$$\widetilde{m}_{rf}(\mathbf{x}) = \sum_{\ell\in U_v} \frac{1}{B}\sum_{b=1}^{B} \frac{\psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}\left(\mathbf{x},\theta_b^{(U)}\right)}}{\widetilde{N}(\mathbf{x},\theta_b^{(U)})} y_\ell$$

with $\widetilde{N}(\mathbf{x},\theta_b^{(U)}) = \sum_{k\in U_v} \psi_k^{(b,U)} \mathbb{1}_{\mathbf{x}_k \in A^{(U)}\left(\mathbf{x},\theta_b^{(U)}\right)}$.

Below, we include a graph illustrating the convergence of the difference $\widehat{\widetilde{m}}_{rf} - \widetilde{m}_{rf}$ towards 0 in L^2 where the regression function was defined as $m(X) = 2 + 2X_1 + X_2 + X_3$, with X_1, X_2 and X_3 defined as in Section 6 from the main paper. The population sizes were such that the sampling fraction was of 10%.

[2] The assumptions presented here are the same as those presented in the main document; they are only recalled here for simplicity of exposition.

Similar results can be obtained using other simulation parameters.



(C4) The number of subsampled elements n'_v is such that $\lim_{v\to\infty} n'_v/n_v \in (0;1]$.

Consistency of the Horvitz-Thompson variance estimator

(H18) Assume that $\lim\limits_{v\to\infty} \max\limits_{i,j,k,\ell\in D_{4,N_v}} |\mathbb{E}_p\left\{\left(I_iI_j-\pi_i\pi_j\right)\left(I_kI_\ell-\pi_k\pi_\ell\right)\right\}| = 0$, where D_{4,N_v} denotes the set of distinct 4-tuples from U_v.

3.9.1 Asymptotic results of the population RF model-assisted estimator $\widehat{t}^*_{rf}$

The population RF model-assisted estimator is given by

$$\widehat{t}^*_{rf} = \sum_{k\in U_v} \widehat{m}^*_{rf}(\mathbf{x}_k) + \sum_{k\in S_v} \frac{y_k - \widehat{m}^*_{rf}(\mathbf{x}_k)}{\pi_k},$$

where $\widehat{m}^*_{rf}$ is the sample-based estimator of m by using RF built at the population level (for more details, see relation (13) from the main paper):

$$\widehat{m}^*_{rf}(\mathbf{x}_k) = \sum_{\ell\in S_v} \frac{1}{\pi_\ell} \widetilde{W}^*_\ell(\mathbf{x}_k) y_\ell, \tag{3.31}$$

where

$$\widetilde{W}^*_\ell(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \frac{\psi^{(b,U)}_\ell \mathbb{1}_{\mathbf{x}_\ell\in A^{*(U)}\left(\mathbf{x}_k,\theta^{(U)}_b\right)}}{\widetilde{N}^*(\mathbf{x}_k,\theta^{(U)}_b)}$$

and $\widetilde{N}^*(\mathbf{x}_k, \theta_b^{(U)}) = \sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A^{*(U)}\left(\mathbf{x_k}, \theta_b^{(U)}\right)}$ is the number of units falling in the terminal node $A^{*(U)}\left(\mathbf{x_k}, \theta_b^{(U)}\right)$ containing $\mathbf{x}_k$. The estimator $\widehat{m}_{rf}^*(\mathbf{x}_k)$ can be written as a bagged estimator as follows:

$$\widehat{m}_{rf}^*(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \widehat{m}_{tree}^{*(b)}(\mathbf{x}_k, \theta_b^{(U)}),$$

where $\widehat{m}_{tree}^{*(b)}(\mathbf{x}_k, \theta_b^{(U)})$ is the sample-based estimation of m based on the b-th stochastic tree:

$$\widehat{m}_{tree}^{*(b)}(\mathbf{x}_k, \theta_b^{(U)}) = \sum_{\ell \in S_v} \frac{1}{\pi_\ell} \frac{\psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A^{*(U)}\left(\mathbf{x_k}, \theta_b^{(U)}\right)}}{\widetilde{N}^*(\mathbf{x}_k, \theta_b^{(U)})} y_\ell. \tag{3.32}$$

For more readability, we will use in the sequel $\widehat{m}_{tree}^{*(b)}(\mathbf{x}_k)$ instead of $\widehat{m}_{tree}^{*(b)}(\mathbf{x}_k, \theta_b^{(U)})$. Consider the pseudo-generalized difference estimator:

$$\widehat{t}_{pgd} = \sum_{k \in U_v} \widetilde{m}_{rf}^*(\mathbf{x}_k) + \sum_{k \in S_v} \frac{y_k - \widetilde{m}_{rf}^*(\mathbf{x}_k)}{\pi_k},$$

where $\widetilde{m}_{rf}^*(\mathbf{x}_k)$ is the population-based estimator of m by using RF built at the population level (for more details, see relation (12) from the main paper):

$$\widetilde{m}_{rf}^*(\mathbf{x}_k) = \sum_{\ell \in U_v} \widetilde{W}_\ell^*(\mathbf{x}_k) y_\ell.$$

The estimator $\widetilde{m}_{rf}^*$ can be written as a bagged estimator as follows:

$$\widetilde{m}_{rf}^*(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \widetilde{m}_{tree}^{*(b)}(\mathbf{x}_k)$$

and $\widetilde{m}_{tree}^{*(b)}(\mathbf{x}_k)$ is the predictor associated with unit k and based on the b-th stochastic tree:

$$\widetilde{m}_{tree}^{*(b)}(\mathbf{x}_k) = \sum_{\ell \in U_v} \frac{\psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A^{*(U)}\left(\mathbf{x_k}, \theta_b^{(U)}\right)}}{\widetilde{N}^*(\mathbf{x}_k, \theta_b^{(U)})} y_\ell. \tag{3.33}$$

We remark that $\widehat{m}_{tree}^{*(b)}(\mathbf{x}_k)$ is the Horvitz-Thompson estimator of $\widetilde{m}_{tree}^{*(b)}(\mathbf{x}_k)$. As before, $\widetilde{m}_{tree}^{*(b)}$ depends on $\theta_b^{(U)}$ but, for more readability, we drop $\theta_b^{(U)}$ from the expression of $\widetilde{m}_{tree}^{*(b)}(\mathbf{x}_k)$.

We give in the next equivalent expressions of $\widetilde{m}_{tree}^{*(b)}$ and $\widehat{m}_{tree}^{*(b)}$. Consider for that the B partitions built at the population level: $\widetilde{\mathcal{P}}_U^* = \{\widetilde{\mathcal{P}}_U^{*(b)}\}_{b=1}^B$. For a given $b = 1, \ldots, B$, the partition $\widetilde{\mathcal{P}}_U^{*(b)}$ build in the b-th stochastic tree is composed by the J_{bU}^* disjointed regions: $\widetilde{\mathcal{P}}_U^{*(b)} = \{A_j^{*(bU)}\}_{j=1}^{J_{bU}^*}$. Consider $\mathbf{z}_k^{*(b)} = (\mathbb{1}_{\mathbf{x}_k \in A_1^{*(bU)}}, \ldots, \mathbb{1}_{\mathbf{x}_k \in A_{J_{bU}^*}^{*(bU)}})^\top$ where $\mathbb{1}_{\mathbf{x}_k \in A_j^{*(bU)}} = 1$ if $\mathbf{x}_k$ belongs to the region $A_j^{*(bU)}$ and zero otherwise for all $j = 1, \ldots, J_{bU}^*$. We drop the exponent U from the expression of $\mathbf{z}_k^{*(b)}$ for more readability. Since $\widetilde{\mathcal{P}}_U^{*(b)}$ is a partition, then $\mathbf{x}_k$ belongs to only one region of the b-th tree, so the vector $\mathbf{z}_k^{*(b)}$ will

contain only one non-null component. Consider for example that $\mathbf{x}_k \in A_j^{*(bU)}$, then $\widetilde{m}_{tree}^{*(b)}(\mathbf{x}_k)$ is the mean of y-values of individuals ℓ for which $\mathbf{x}_\ell \in A_j^{*(bU)}$:

$$\widetilde{m}_{tree}^{*(b)}(\mathbf{x}_k) = \sum_{\ell \in U_v} \frac{\psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A_j^{*(bU)}}}{\widetilde{N}_j^{*(b)}} y_\ell, \quad \text{for} \quad \mathbf{x}_k \in A_j^{*(bU)},$$

where $\widetilde{N}_j^{*(b)}$ is the number of units belonging to the region $A_j^{*(bU)}$:

$$\widetilde{N}_j^{*(b)} = \sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A_j^{*(bU)}}, \quad j = 1, \ldots, J_{bU}^*. \tag{3.34}$$

Then, $\widetilde{m}_{tree}^{*(b)}(\mathbf{x}_k)$ can be written as follows:

$$\widetilde{m}_{N,rf}^{*(b)}(\mathbf{x}_k) = (\mathbf{z}_k^{*(b)})^\top \widetilde{\boldsymbol{\beta}}^{*(b)}, \quad k \in U_v \tag{3.35}$$

where

$$\widetilde{\boldsymbol{\beta}}^{*(b)} = \left(\sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{*(b)} (\mathbf{z}_\ell^{*(b)})^\top \right)^{-1} \sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{*(b)} y_\ell.$$

Remark that $\widetilde{\boldsymbol{\beta}}^{*(b)}$ may be obtained as solution of the following weighted estimating equation:

$$\sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{*(b)} (y_\ell - (\mathbf{z}_\ell^{*(b)})^\top \boldsymbol{\beta}^{*(b)}) = 0.$$

Since the regions $A_j^{*(bU)}, j = 1, \ldots, J_{bU}^*$, form a partition, then the matrix $\sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{*(b)} (\mathbf{z}_\ell^{*(b)})^\top$ is diagonal with diagonal elements equal to $\widetilde{N}_j^{*(b)}$, the number of units falling in the region $A_j^{*(bU)}$ for all $j = 1, \ldots, J_{bU}^*$. By the stopping criterion, we have that all $\widetilde{N}_j^{*(b)} \geq N_{0v} > 0$ for all j, so the matrix $\sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{*(b)} (\mathbf{z}_\ell^{*(b)})^\top$ is always invertible and $\widetilde{\boldsymbol{\beta}}^{*(b)}$ is well-defined.

Consider now $\widehat{m}_{tree}^{*(b)}(\mathbf{x}_k)$, the estimator of the unknown $\widetilde{m}_{tree}^{*(b)}(\mathbf{x}_k)$. Then, $\widehat{m}_{tree}^{*(b)}(\mathbf{x}_k)$ is the weighted mean of y-values for sampled individuals ℓ belonging to the same region $A_j^{*(bU)}$ as unit k :

$$\widehat{m}_{tree}^{*(b)}(\mathbf{x}_k) = \sum_{\ell \in S_v} \frac{1}{\pi_\ell} \frac{\psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A_j^{*(bU)}}}{N_{j,N}^{*(b)}} y_\ell \quad \text{for} \quad \mathbf{x}_k \in A_j^{*(bU)}$$

and we can write:

$$\widehat{m}_{tree}^{*(b)}(\mathbf{x}_k) = (\mathbf{z}_k^{*(b)})^\top \widehat{\boldsymbol{\beta}}^{*(b)}, k \in U_v \tag{3.36}$$

where

$$\widehat{\boldsymbol{\beta}}^{*(b)} = \left(\sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{*(b)} (\mathbf{z}_\ell^{*(b)})^\top \right)^{-1} \sum_{\ell \in S_v} \frac{1}{\pi_\ell} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{*(b)} y_\ell.$$

In the expression of $\widehat{\boldsymbol{\beta}}^{*(b)}$, we do not estimate the matrix $\sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{*(b)} (\mathbf{z}_\ell^{*(b)})^\top$ since it is known and besides, we guarantee in this way that we will always have non-empty terminal nodes at the population level. So, $\widehat{\boldsymbol{\beta}}^{*(b)}$ will be always well-defined whatever the sample S is.

Let denote by $\alpha_k = \pi_k^{-1} I_k - 1$ for all $k \in U_v$, where I_k is the sample membership, $I_k = 1$ if $k \in S$ and zero otherwise. In order to prove the consistency of $\widehat{t}^*_{rf}$ as well as its asymptotic equivalence to the pseudo-generalized difference estimator $\widehat{t}_{pgd}$, we use the following decomposition:

$$\begin{aligned}\frac{1}{N_v}\left(\widehat{t}^*_{rf} - t_y\right) &= \frac{1}{N_v}\left(\widehat{t}_{pgd} - t_y\right) - \frac{1}{N_v}\sum_{k\in U_v}\alpha_k(\widehat{m}^*_{rf}(\mathbf{x}_k) - \widetilde{m}^*_{rf}(\mathbf{x}_k)) \\ &= \frac{1}{N_v}\left(\widehat{t}_{pgd} - t_y\right) - \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{N_v}\sum_{k\in U_v}\alpha_k\left(\widehat{m}^{*(b)}_{tree}(\mathbf{x}_k) - \widetilde{m}^{*(b)}_{tree}(\mathbf{x}_k)\right)\right]. \end{aligned} \tag{3.37}$$

We will prove that each term form the decomposition (3.37) is convergent to zero. We give first several useful lemmas.

Lemma 1. *There exists a positive constant $\tilde{c}_1$ such that:*

$$\frac{n_v}{N_v^2}\mathbb{E}_p(\widehat{t}_{pgd} - t_y)^2 \leqslant \tilde{c}_1.$$

Proof. First of all, from relation (3.31), $\widetilde{m}^*_{rf}(\mathbf{x}_k)$ is a weighted sum at the population level of y-values with positive weights summing to one (see proposition 2.1. from the main paper). Then, we get that $\sup_{k\in U_v}|\widetilde{m}^*_{rf}(\mathbf{x}_k)| \leqslant C$ by using also assumption (H13). We have:

$$\frac{1}{N_v}(\widehat{t}_{pgd} - t_y) = \frac{1}{N_v}\sum_{k\in U_v}\alpha_k(y_k - \widetilde{m}^*_{rf}(\mathbf{x}_k))$$

and

$$\begin{aligned} n_v\mathbb{E}_p\left(\frac{\widehat{t}_{pgd} - t_y}{N_v}\right)^2 &= \frac{n_v}{N_v^2}\mathbb{V}_p\left(\sum_{k\in S_v}\frac{(y_k - \widetilde{m}^*_{rf}(\mathbf{x}_k))}{\pi_k}\right) \\ &\leqslant \left(\frac{n_v}{N_v}\cdot\frac{1}{\lambda} + \frac{n_v \max_{k\neq\ell\in U_v}|\pi_{k\ell} - \pi_k\pi_\ell|}{\lambda^2}\right)\cdot\frac{2}{N_v}\sum_{k\in U_v}\left(y_k^2 + (\widetilde{m}^*_{rf}(\mathbf{x}_k))^2\right) \\ &\leqslant \tilde{c}_1 \end{aligned}$$

by assumptions (H13)-(H15). ∎

Lemma 2. *There exists a positive constant $\tilde{c}_2$ not depending on $b = 1, \ldots, B$, such that*

$$\mathbb{E}_p||\widehat{\boldsymbol{\beta}}^{*(b)} - \widetilde{\boldsymbol{\beta}}^{*(b)}||_2^2 \leqslant \frac{\tilde{c}_2 N_v}{N_{0v}^2} \quad \textit{for all} \quad b = 1, \ldots, B.$$

Proof. We can write

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{*(b)} - \widetilde{\boldsymbol{\beta}}^{*(b)} &= \left(\sum_{\ell\in U_v}\psi_\ell^{(b,U)}\mathbf{z}_\ell^{*(b)}(\mathbf{z}_\ell^{*(b)})^\top\right)^{-1}\left(\sum_{\ell\in S_v}\frac{1}{\pi_\ell}\psi_\ell^{(b,U)}\mathbf{z}_\ell^{*(b)}y_\ell - \sum_{\ell\in U_v}\psi_\ell^{(b,U)}\mathbf{z}_\ell^{*(b)}y_\ell\right) \\ &= \left(\sum_{\ell\in U_v}\psi_\ell^{(b,U)}\mathbf{z}_\ell^{*(b)}(\mathbf{z}_\ell^{*(b)})^\top\right)^{-1}\left(\sum_{\ell\in U_v}\alpha_\ell\psi_\ell^{(b,U)}\mathbf{z}_\ell^{*(b)}y_\ell\right) \end{aligned} \tag{3.38}$$

Let denote by $\widetilde{\mathbf{T}}^{*(b)} = \sum_{\ell\in U_v}\psi_\ell^{(b,U)}\mathbf{z}_\ell^{*(b)}(\mathbf{z}_\ell^{*(b)})^\top$. As already mentioned before, the matrix $\widetilde{\mathbf{T}}^{*(b)}$ is diagonal with positive diagonal elements given by $\widetilde{N}_j^{*(b)}$ the number of units falling in the region $A_j^{*(bU)}$ (see

relation 3.34) for $j = 1, \ldots, J^*_{bU}$ and by the stopping criterion, we have that $\widetilde{N}_j^{*(b)} \geq N_{0v} > 0$. We obtain then

$$||(\widetilde{\mathbf{T}}^{*(b)})^{-1}||_2 = \max_{j=1,\ldots,J_{bU}} \left(\frac{1}{\widetilde{N}_j^{*(b)}} \right) \leq N_{0v}^{-1}, \quad \text{for all} \quad b = 1, \ldots, B \tag{3.39}$$

where $||\cdot||_2$ is the spectral norm matrix defined for a squared $p \times p$ matrix $\mathbf{A}$ by $||\mathbf{A}||_2 = \sup_{\mathbf{x} \in \mathbf{R}^p, ||\mathbf{x}||_2 \neq 0} ||\mathbf{A}\mathbf{x}||_2 / ||\mathbf{x}||_2$. For a symmetric and positive definite matrix $\mathbf{A}$, we have that $||\mathbf{A}||_2 = \lambda_{max}(\mathbf{A})$ where $\lambda_{max}(\mathbf{A})$ is the largest eigenvalue of $\mathbf{A}$. We get, for $b = 1, \ldots, B$:

$$\begin{aligned} \mathbb{E}_p ||\widehat{\boldsymbol{\beta}}^{*(b)} - \widetilde{\boldsymbol{\beta}}^{*(b)}||_2^2 &\leqslant \mathbb{E}_p \left[||N_v (\widetilde{\mathbf{T}}^{*(b)})^{-1}||_2^2 \cdot \left\| \frac{1}{N_v} \sum_{\ell \in U_v} \alpha_\ell \psi_\ell^{(b,U)} \mathbf{z}_\ell^{*(b)} y_\ell \right\|_2^2 \right] \\ &\leqslant \frac{N_v^2}{N_{0v}^2} \mathbb{E}_p \left\| \frac{1}{N_v} \sum_{\ell \in U_v} \alpha_\ell \psi_\ell^{(b,U)} \mathbf{z}_\ell^{*(b)} y_\ell \right\|_2^2 \end{aligned} \tag{3.40}$$

and

$$\begin{aligned} & \mathbb{E}_p \left\| \frac{1}{N_v} \sum_{k \in U_v} \alpha_k \psi_k^{(b,U)} \mathbf{z}_k^{*(b)} y_k \right\|_2^2 \\ = & \frac{1}{N_v^2} \left(\sum_{k \in U_v} (\psi_k^{(b,U)})^2 y_k^2 ||\mathbf{z}_k^{*(b)}||_2^2 \mathbb{E}_p(\alpha_k^2) + \sum_{k \in U_v} \sum_{\substack{\ell \in U_v \\ \ell \neq k}} \psi_k^{(b,U)} \psi_\ell^{(b,U)} y_k y_\ell (\mathbf{z}_k^{*(b)})^\top \mathbf{z}_\ell^{*(b)} \mathbb{E}_p(\alpha_k \alpha_\ell) \right) \\ \leqslant & \frac{1}{n_v} \left(\frac{n_v}{\lambda N_v} + \frac{n_v \max_{k,\ell \in U_v, k \neq \ell} |\pi_{k\ell} - \pi_k \pi_\ell|}{\lambda^2} \right) \left(\frac{1}{N_v} \sum_{k \in U_v} (\psi_k^{(b,U)})^2 y_k^2 ||\mathbf{z}_k^{*(b)}||_2^2 \right) \\ \leqslant & \frac{C_0}{n_v} \end{aligned} \tag{3.41}$$

by assumptions (H13)-(H15) and the fact that $||\mathbf{z}_k^{*(b)}||_2^2 = 1$ for all $k \in U_v$ and $b = 1, \ldots, B$. From (3.40), (3.41) and assumption (H13), we obtain that it exists a positive constant $\tilde{c}_2$ such that

$$\mathbb{E}_p ||\widehat{\boldsymbol{\beta}}^{*(b)} - \widetilde{\boldsymbol{\beta}}^{*(b)}||_2^2 \leqslant \frac{\tilde{c}_2 N_v}{N_{0v}^2}.$$

■

Result 3.9.1. *Consider a sequence of population RF estimators* $\{\widehat{t}^*_{rf}\}$. *Then, there exist positive constants* $\tilde{C}_1, \tilde{C}_2$ *such that*

$$\mathbb{E}_p \left| \frac{1}{N_v} \left(\widehat{t}^*_{rf} - t_y \right) \right| \leqslant \frac{\tilde{C}_1}{\sqrt{n_v}} + \frac{\tilde{C}_2}{N_{0v}}, \quad \textit{with } \xi\textit{-probability one.}$$

If $\dfrac{N_v^a}{N_{0v}} = O(1)$ *with* $1/2 \leqslant a \leqslant 1$, *then*

$$\mathbb{E}_p \left| \frac{1}{N_v} \left(\widehat{t}^*_{rf} - t_y \right) \right| \leqslant \frac{\tilde{C}}{\sqrt{n_v}}, \quad \textit{with } \xi\textit{-probability one.}$$

Proof. We get from relation (3.37) :

$$\frac{1}{N_v} \mathbb{E}_p \left| \widehat{t}^*_{rf} - t_y \right| \leqslant \frac{1}{N_v} \mathbb{E}_p \left| \widehat{t}_{pgd} - t_y \right| + \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_v} \mathbb{E}_p \left| \sum_{k \in U_v} \alpha_k (\widehat{m}_{tree}^{*(b)}(\mathbf{x}_k) - \widetilde{m}_{tree}^{*(b)}(\mathbf{x}_k)) \right|.$$

Lemma 1 gives us that there exists positive constant $\tilde{C}_1$ such that

$$\frac{1}{N_v}\mathbb{E}_p\left|\widehat{t}_{pgd} - t_y\right| \leqslant \frac{\tilde{C}_1}{\sqrt{n_v}}. \tag{3.42}$$

Now, by using relations (3.35) and (3.36), we can then write for any $b = 1, \ldots, B$:

$$\sum_{k\in U_v} \alpha_k(\widehat{m}^{*(b)}_{tree}(\mathbf{x}_k) - \widetilde{m}^{*(b)}_{tree}(\mathbf{x}_k)) = \sum_{k\in U_v} \alpha_k(\mathbf{z}^{*(b)}_k)^\top(\widehat{\boldsymbol{\beta}}^{*(b)} - \widetilde{\boldsymbol{\beta}}^{*(b)})$$

and

$$\frac{1}{N_v}\mathbb{E}_p\left|\sum_{k\in U_v} \alpha_k(\widehat{m}^{*(b)}_{tree}(\mathbf{x}_k) - \widetilde{m}^{*(b)}_{tree}(\mathbf{x}_k))\right| \leqslant \left(\mathbb{E}_p\left\|\frac{1}{N_v}\sum_{k\in U_v}\alpha_k\mathbf{z}^{*(b)}_k\right\|_2^2\right)^{1/2}\left(\mathbb{E}_p||\widehat{\boldsymbol{\beta}}^{*(b)} - \widetilde{\boldsymbol{\beta}}^{*(b)}||_2^2\right)^{1/2} \tag{3.43}$$

and

$$\begin{aligned}
\mathbb{E}_p\left\|\frac{1}{N_v}\sum_{k\in U_v}\alpha_k\mathbf{z}^{*(b)}_k\right\|_2^2 &= \frac{1}{N_v^2}\left(\sum_{k\in U_v}\mathbb{E}_p(\alpha_k^2)||\mathbf{z}^{*(b)}_k||_2^2 + \sum_{k\in U_v}\sum_{\substack{\ell\in U_v\\ \ell\neq k}}\mathbb{E}_p(\alpha_k\alpha_\ell)(\mathbf{z}^{*(b)}_k)^\top\mathbf{z}^{*(b)}_\ell\right)\\
&\leqslant \frac{1}{n_v}\left(\frac{n_v}{\lambda N_v} + \frac{n_v\max_{k\neq\ell\in U_v}|\pi_{k\ell} - \pi_k\pi_\ell|}{\lambda^2}\right)\cdot\frac{1}{N_v}\sum_{k\in U_v}||\mathbf{z}^{*(b)}_k||_2^2\\
&\leqslant \frac{C_2}{n_v}
\end{aligned} \tag{3.44}$$

by assumptions (H13)-(H15) and the fact that $||\mathbf{z}^{*(b)}_k||_2^2 = 1$ for all $k \in U_v$ and $b = 1, \ldots, B$. Then, from relations (3.43), (3.44) and lemma 2, we get that there exists a positive constant $\tilde{C}_2$ such that, for any $b = 1, \ldots, B$, we have:

$$\frac{1}{N_v}\mathbb{E}_p\left|\sum_{k\in U_v}\alpha_k(\widehat{m}^{*(b)}_{rf}(\mathbf{x}_k) - \widetilde{m}^{*(b)}_{N,rf}(\mathbf{x}_k))\right| \leqslant \sqrt{\frac{C_2}{n_v}\frac{\tilde{c}_2 N_v}{N_{0v}^2}} \leqslant \frac{\tilde{C}_2}{N_{0v}} \tag{3.45}$$

by using also the assumption (H13). The result follows then from relations (3.42) and (3.45). ∎

Result 3.9.2. *Consider a sequence of RF estimators* $\{\widehat{t}^*_{rf}\}$. *If* $\frac{N_v^a}{N_{0v}} = O(1)$ *with* $1/2 < a \leqslant 1$, *then*

$$\frac{\sqrt{n_v}}{N_v}\left(\widehat{t}^*_{rf} - t_y\right) = \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{pgd} - t_y\right) + o_{\mathbb{P}}(1).$$

Proof. We get from relation (3.37) and lemmas (1) and the proof of result 3.9.1 (relation 3.45) that

$$\begin{aligned}
\frac{\sqrt{n_v}}{N_v}\left(\widehat{t}^*_{rf} - t_y\right) &= \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{pgd} - t_y\right) + \frac{1}{B}\sum_{b=1}^{B}\left[\frac{\sqrt{n_v}}{N_v}\sum_{k\in U_v}\alpha_k(\widehat{m}^{*(b)}_{tree}(\mathbf{x}_k) - \widetilde{m}^{*(b)}_{tree}(\mathbf{x}_k))\right].\\
&= \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{pgd} - t_y\right) + O_{\mathbb{P}}\left(\frac{\sqrt{n_v}}{N_{0v}}\right)\\
&= \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{pgd} - t_y\right) + o_{\mathbb{P}}(1)
\end{aligned}$$

provided that $\frac{N_v^a}{N_{0v}} = O(1)$ with $1/2 < a \leqslant 1$. ∎

Result 3.9.3. *Consider a sequence of population RF estimators* $\{\widehat{t}^*_{rf}\}$. *Assume that* $\frac{N_v^a}{N_{0v}} = O(1)$ *with* $1/2 < a \leqslant 1$, *then the variance estimator* $\widehat{\mathbb{V}}_{rf}(\widehat{t}^*_{rf})$ *is design-consistent for the asymptotic variance* $\mathbb{AV}_p\left(\widehat{t}^*_{rf}\right)$. *That is,*

$$\lim_{v\to\infty} \mathbb{E}_p\left(\frac{n_v}{N_v^2}\left|\widehat{\mathbb{V}}_{rf}(\widehat{t}^*_{rf}) - \mathbb{AV}_p(\widehat{t}^*_{rf})\right|\right) = 0.$$

Proof Consider the following decomposition

$$\begin{aligned} & n_v\left(\widehat{\mathbb{V}}_p\left(N_v^{-1}\widehat{t}^*_{rf}\right) - \mathbb{AV}_p\left(N_v^{-1}\widehat{t}^*_{rf}\right)\right) \\ = \; & n_v\left(\widehat{\mathbb{V}}_p\left(N_v^{-1}\widehat{t}^*_{rf}\right) - \widehat{\mathbb{V}}_p\left(N_v^{-1}\widehat{t}_{pgd}\right)\right) + n_v\left(\widehat{\mathbb{V}}_p\left(N_v^{-1}\widehat{t}_{pgd}\right) - \mathbb{AV}_p\left(N_v^{-1}\widehat{t}^*_{rf}\right)\right) \end{aligned}$$

where $\widehat{\mathbb{V}}_p\left(N_v^{-1}\widehat{t}_{pgd}\right)$ is the pseudo-type variance estimator of $\mathbb{V}_p\left(N_v^{-1}\widehat{t}_{pgd}\right) = \mathbb{AV}_p\left(N_v^{-1}\widehat{t}^*_{rf}\right)$ given by

$$\widehat{\mathbb{V}}_p\left(N_v^{-1}\widehat{t}_{pgd}\right) = \frac{1}{N_v^2}\sum_{k\in U_v}\sum_{\ell\in U_v}\frac{\pi_{k\ell}-\pi_k\pi_\ell}{\pi_{k\ell}}\frac{y_k - \widetilde{m}^*_{rf}(\mathbf{x}_k)}{\pi_k}\frac{y_\ell - \widetilde{m}^*_{rf}(\mathbf{x}_\ell)}{\pi_\ell}I_k I_\ell.$$

Now, to prove that the consistency of the first term from right of (3.46), we use the same decomposition as in Goga and Ruiz-Gazen (2014). Denote $\tilde{e}_k = y_k - \widetilde{m}^*_{rf}(\mathbf{x}_k)$, $\hat{e}_k = y_k - \widehat{m}^*_{rf}(\mathbf{x}_k)$ and $c_{k\ell} = \frac{\pi_{k\ell}-\pi_k\pi_\ell}{\pi_{k\ell}\pi_k\pi_\ell}I_k I_\ell$. Then,

$$\begin{aligned} n_v(\widehat{\mathbb{V}}_p\left(N_v^{-1}\widehat{t}^*_{rf}\right) - \widehat{\mathbb{V}}_p\left(N_v^{-1}\widehat{t}_{pgd}\right)) &= \frac{n_v}{N_v^2}\sum_{k\in U_v}\sum_{\ell\in U_v} c_{k\ell}\left(\hat{e}_k\hat{e}_\ell - \tilde{e}_k\tilde{e}_\ell\right) \\ &= \frac{n_v}{N_v^2}\sum_{k\in U_v}\sum_{\ell\in U_v} c_{k\ell}\left[(\hat{e}_k-\tilde{e}_k)(\hat{e}_\ell-\tilde{e}_\ell) + \tilde{e}_k(\hat{e}_\ell-\tilde{e}_\ell) + \tilde{e}_\ell(\hat{e}_k-\tilde{e}_k)\right] \\ &= A_1 + A_2 + A_3. \end{aligned}$$

For all $k \in U_v$, $\hat{e}_k - \tilde{e}_k = \widetilde{m}^*_{rf}(\mathbf{x}_k) - \widehat{m}^*_{rf}(\mathbf{x}_k)$ and thus,

$$\mathbb{E}_p|A_1| \leq \left(\frac{n_v}{\lambda^2 N_v} + \frac{n_v \max_{k\neq\ell\in U_v}|\pi_{k\ell}-\pi_k\pi_\ell|}{\lambda^*\lambda^2}\right)\frac{1}{N_v}\sum_{k\in U_v}\mathbb{E}_p(\hat{e}_k-\tilde{e}_k)^2,$$

by assumptions (H14)-(H15). Therefore, it suffices to show that, for all $k \in U_v$, one has $\mathbb{E}_p(\hat{e}_k-\tilde{e}_k)^2 = o(1)$ uniformly in k, which we show next. We have

$$\mathbb{E}_p(\widetilde{m}^*_{rf}(\mathbf{x}_k) - \widehat{m}^*_{rf}(\mathbf{x}_k))^2 \leqslant \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}_p(\widetilde{m}^{*(b)}_{tree}(\mathbf{x}_k) - \widehat{m}^{*(b)}_{tree}(\mathbf{x}_k))^2.$$

We can write by using relations (3.35) and (3.36):

$$\widehat{m}^{*(b)}_{tree}(\mathbf{x}_k) - \widetilde{m}^{*(b)}_{tree}(\mathbf{x}_k) \;=\; (\mathbf{z}^{*(b)}_k)^\top(\widehat{\boldsymbol{\beta}}^{*(b)} - \widetilde{\boldsymbol{\beta}}^{*(b)})$$

and then, by using lemma (2),

$$\begin{aligned} \mathbb{E}_p(\widetilde{m}^*_{rf}(\mathbf{x}_k) - \widehat{m}^*_{rf}(\mathbf{x}_k))^2 &\leqslant \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}_p\left(||\mathbf{z}^{*(b)}_k||_2^2||\widehat{\boldsymbol{\beta}}^{*(b)} - \widetilde{\boldsymbol{\beta}}^{*(b)}||_2^2\right) \\ &\leqslant \frac{\tilde{c}_2 N_v}{N_{0v}^2} \end{aligned}$$

quantity going to zero provided that $\frac{N_v^a}{N_{0v}} = O(1)$ with $1/2 < a \leqslant 1$.

Using the same arguments, we obtain that $\mathbb{E}_p|A_2| = o(1)$ and $\mathbb{E}_p|A_3| = o(1)$. We get then

$$n_v \mathbb{E}_p | \widehat{\mathbb{V}}_p \left(N_v^{-1} \widehat{t}_{rf}^* \right) - \widehat{\mathbb{V}}_p \left(N_v^{-1} \widehat{t}_{pgd} \right) | = o(1).$$

The second term from right of (3.46) concerns the consistency of the estimator of the Horvitz-Thompson variance computed for the population residuals $y_k - \widetilde{m}_{rf}^*(\mathbf{x}_k), k \in U_v$. The proof of this consistency (Breidt and Opsomer, 2000) requires assumptions only on the higher order inclusion probabilities (H18) as well as finite forth moment of $y_k - \widetilde{m}_{rf}^*(\mathbf{x}_k)$:

$$\frac{1}{N_v} \sum_{k \in U_v} (y_k - \widetilde{m}_{rf}^*(\mathbf{x}_k))^4 \leq \frac{4}{N_v} \sum_{k \in U_v} (y_k^4 + (\widetilde{m}_{rf}^*(\mathbf{x}_k))^4) < \infty.$$

So,

$$n_v \mathbb{E}_p | \widehat{\mathbb{V}}_p \left(N_v^{-1} \widehat{t}_{pgd} \right) - \mathbb{AV}_p \left(N_v^{-1} \widehat{t}_{rf}^* \right) | = o(1)$$

and the result follows.

3.9.2 Asymptotic results: the sample RF model-assisted estimator $\widehat{t}_{rf}$

The sample RF model-assisted estimator is given by

$$\widehat{t}_{rf} = \sum_{k \in U_v} \widehat{m}_{rf}(\mathbf{x}_k) + \sum_{k \in S_v} \frac{y_k - \widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k},$$

where $\widehat{m}_{rf}$ is the estimator of m built at the sample level and by using RF based on partition built at the sample level (for more details, see relation (17) from the main paper):

$$\widehat{m}_{rf}(\mathbf{x}_k) = \sum_{\ell \in S_v} \frac{1}{\pi_\ell} \widehat{W}_\ell(\mathbf{x}_k) y_\ell,$$

where

$$\widehat{W}_\ell(\mathbf{x}_k) = \frac{1}{B} \sum_{b=1}^{B} \frac{\psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)}}{\widehat{N}(\mathbf{x}_k, \theta_b^{(S)})}$$

and $\widehat{N}(\mathbf{x}_k, \theta_b^{(S)}) = \sum_{\ell \in S_v} \pi_\ell^{-1} \psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)}$ is the estimated number of units falling in the terminal node $A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)$ containing $\mathbf{x}_k$. As in Section 3.9.1, the estimator $\widehat{m}_{rf}(\mathbf{x}_k)$ can be written as a bagged estimator of m as follows:

$$\widehat{m}_{rf}(\mathbf{x}_k) = \frac{1}{B} \sum_{b=1}^{B} \widehat{m}_{tree}^{(b)}(\mathbf{x}_k)$$

and $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)$ is the estimation of m based on the b-th stochastic tree:

$$\widehat{m}_{tree}^{(b)}(\mathbf{x}_k) = \sum_{\ell \in S_v} \frac{1}{\pi_\ell} \frac{\psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)}}{\widehat{N}(\mathbf{x}_k, \theta_b^{(S)})} y_\ell \tag{3.46}$$

As in Section 3.9.1, for more readability, we note in the sequel $\widehat{m}^{(b)}_{tree}(\mathbf{x}_k)$ instead of $\widehat{m}^{(b)}_{tree}(\mathbf{x}_k, \theta^{(S)}_b)$. Consider the pseudo-generalized difference estimator:

$$\hat{t}_{pgd} = \sum_{k \in U_v} \widetilde{m}_{rf}(\mathbf{x}_k) + \sum_{k \in S_v} \frac{y_k - \widetilde{m}_{rf}(\mathbf{x}_k)}{\pi_k}$$

where $\widetilde{m}_{rf}$ is the estimation of m built at the population level by using RF based on partition built also at the population level (relation (9) from the main paper):

$$\widetilde{m}_{rf}(\mathbf{x}_k) = \sum_{\ell \in U_v} \widetilde{W}_\ell(\mathbf{x}_k) y_\ell,$$

where

$$\widetilde{W}_\ell(\mathbf{x}_k) = \frac{1}{B} \sum_{b=1}^{B} \frac{\psi^{(b,U)}_\ell \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}\left(\mathbf{x_k}, \theta^{(U)}_b\right)}}{\widetilde{N}(\mathbf{x}_k, \theta^{(U)}_b)}$$

with $\widetilde{N}(\mathbf{x}_k, \theta^{(U)}_b) = \sum_{\ell \in U_v} \psi^{(b,U)}_\ell \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}\left(\mathbf{x_k}, \theta^{(U)}_b\right)}$ is the number of units falling in the terminal node $A^{(U)}\left(\mathbf{x_k}, \theta^{(U)}_b\right)$ containing $\mathbf{x}_k$. The estimator $\widehat{m}_{rf}$ can be also written as a bagged estimator as follows:

$$\widetilde{m}_{rf}(\mathbf{x}_k) = \frac{1}{B} \sum_{b=1}^{B} \widetilde{m}^{(b)}_{tree}(\mathbf{x}_k)$$

and

$$\widetilde{m}^{(b)}_{tree}(\mathbf{x}_k) = \sum_{\ell \in U_v} \frac{\psi^{(b,U)}_\ell \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}\left(\mathbf{x_k}, \theta^{(U)}_b\right)}}{\widetilde{N}(\mathbf{x}_k, \theta^{(U)}_b)} y_\ell.$$

As in the previous section, we will write $\widetilde{m}_{rf}$ and $\widehat{m}_{rf}$ in equivalent forms. Consider for that the B partitions build at the population level $\widetilde{\mathcal{P}}_U = \{\widetilde{\mathcal{P}}^{(b)}_U\}_{b=1}^{B}$. For a given $b = 1, \ldots, B$, the partition $\widetilde{\mathcal{P}}^{(b)}_U$ is composed by the disjointed regions $\widetilde{\mathcal{P}}^{(b)}_U = \{A^{(bU)}_j\}_{j=1}^{J_{bU}}$. Consider $\mathbf{z}^{(b)}_k = (\mathbb{1}_{\mathbf{x}_k \in A^{(bU)}_1}, \ldots, \mathbb{1}_{\mathbf{x}_k \in A^{(bU)}_{J_{bU}}})^\top$ where $\mathbb{1}_{\mathbf{x}_k \in A^{(bU)}_j} = 1$ if $\mathbf{x}_k$ belongs to the region $A^{(bU)}_j$ and zero otherwise for all $j = 1, \ldots, J_{bU}$. Since $\widetilde{\mathcal{P}}^{(b)}_U$ is a partition, then $\mathbf{x}_k$ belongs to only one region at the b-th step. Suppose for example that $\mathbf{x}_k \in A^{(bU)}_j$, then $\widetilde{m}^{(b)}_{tree}(\mathbf{x}_k)$ is the mean of y-values for individuals ℓ for which $\mathbf{x}_\ell \in A^{(bU)}_j$:

$$\widetilde{m}^{(b)}_{tree}(\mathbf{x}_k) = \sum_{\ell \in U_v} \frac{\psi^{(b,U)}_\ell \mathbb{1}_{\mathbf{x}_\ell \in A^{(bU)}_j}}{\widetilde{N}^{(b)}_j} y_\ell, \quad \text{for} \quad \mathbf{x}_k \in A^{(bU)}_j,$$

where $\widetilde{N}^{(b)}_j$ is the number of units belonging to the region $A^{(bU)}_j$:

$$\widetilde{N}^{(b)}_j = \sum_{\ell \in U_v} \psi^{(b,U)}_\ell \mathbb{1}_{\mathbf{x}_\ell \in A^{(bU)}_j}, \quad j = 1, \ldots, J_{bU}. \tag{3.47}$$

Then, $\widetilde{m}^{(b)}_{tree}(\mathbf{x}_k)$ can be written as a regression-type estimator with $\mathbf{z}^{(b)}_k$ as explanatory variables:

$$\widetilde{m}^{(b)}_{tree}(\mathbf{x}_k) = (\mathbf{z}^{(b)}_k)^\top \widetilde{\boldsymbol{\beta}}^{(b)}, \quad k \in U_v \tag{3.48}$$

where

$$\widetilde{\boldsymbol{\beta}}^{(b)} = \left(\sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{(b)} (\mathbf{z}_\ell^{(b)})^\top \right)^{-1} \sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{(b)} y_\ell.$$

Based on the same arguments as in Section 3.9.1, the matrix $\sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{(b)} (\mathbf{z}_\ell^{(b)})^\top$ is diagonal with diagonal elements equal to $\widetilde{N}_j^{(b)}, j = 1, \ldots, J_{bU}$. By the stopping criterion, we have that all $\widetilde{N}_j^{(b)} \geq N_0 > 0$, so the matrix $\sum_{\ell \in U_v} \psi_\ell^{(b,U)} \mathbf{z}_\ell^{(b)} (\mathbf{z}_\ell^{(b)})^\top$ is invertible and $\widetilde{\boldsymbol{\beta}}^{(b)}$ is well-defined.
Consider now the B partitions build at the sample level $\widehat{\mathcal{P}}_S = \{\widehat{\mathcal{P}}_S^{(b)}\}_{b=1}^B$. For a given $b = 1, \ldots, B$, the partition $\widehat{\mathcal{P}}_S^{(b)}$ is composed by the disjointed regions $\widehat{\mathcal{P}}_S^{(b)} = \{A_j^{(bS)}\}_{j=1}^{J_{bS}}$. Consider $\hat{\mathbf{z}}_k^{(b)} = (\mathbb{1}_{\mathbf{x}_k \in A_{1S}^{(b)}}, \ldots, \mathbb{1}_{\{\mathbf{x}_k \in A_{J_{bS}}^{(b)}\}})^\top$ where $\mathbb{1}_{\mathbf{x}_k \in A_j^{(bS)}} = 1$ if $\mathbf{x}_k$ belongs to the region $A_j^{(bS)}$ and zero otherwise for all $j = 1, \ldots, J_{bS}$. Here, the hat notation is to design the fact that the vector $\hat{\mathbf{z}}_k^{(b)}$ depends on random dummy variables $\mathbb{1}_{\mathbf{x}_k \in A_j^{(bS)}}$. Since $\{A_j^{(bS)}\}_{j=1}^{J_{bS}}$ form a partition, then $\mathbf{x}_k$ belongs to only one terminal node. Suppose for example that $\mathbf{x}_k \in A_j^{(bS)}$, then $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)$ is a Hajek-type estimator:

$$\widehat{m}_{tree}^{(b)}(\mathbf{x}_k) = \sum_{\ell \in S_v} \frac{1}{\pi_\ell} \frac{\psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A_j^{(bS)}} y_\ell}{\widehat{N}_j^{(b)}}, \quad \text{for} \quad \mathbf{x}_k \in A_j^{(bS)},$$

where $\widehat{N}_j^{(b)}$ is the estimated number of units falling in the terminal node $A_j^{(bS)}$:

$$\widehat{N}_j^{(b)} = \sum_{\ell \in S_v} \frac{1}{\pi_\ell} \psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A_j^{(bS)}}, \quad j = 1, \ldots, J_{bS}.$$

Then, $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)$ can be written also as a regression-type estimator with $\hat{\mathbf{z}}_k^{(b)}$ as explanatory variables:

$$\widehat{m}_{tree}^{(b)}(\mathbf{x}_k) = (\hat{\mathbf{z}}_k^{(b)})^\top \widehat{\boldsymbol{\beta}}^{(b)}, k \in U_v, \tag{3.49}$$

where

$$\widehat{\boldsymbol{\beta}}^{(b)} = \left(\sum_{\ell \in S_v} \frac{1}{\pi_\ell} \psi_\ell^{(b,S)} \hat{\mathbf{z}}_\ell^{(b)} (\hat{\mathbf{z}}_\ell^{(b)})^\top \right)^{-1} \sum_{\ell \in S_v} \frac{1}{\pi_\ell} \psi_\ell^{(b,S)} \hat{\mathbf{z}}_\ell^{(b)} y_\ell.$$

As in Section 3.9.1, remark that $\widehat{\boldsymbol{\beta}}^{(b)}$ may be obtained as solution of the following weighted estimating equation:

$$\sum_{\ell \in S_v} \frac{1}{\pi_\ell} \psi_\ell^{(b,S)} \hat{\mathbf{z}}_\ell^{(b)} (y_\ell - (\hat{\mathbf{z}}_\ell^{(b)})^\top \boldsymbol{\beta}^{(b)}) = 0.$$

Since $\{A_j^{(bS)}\}_{j=1}^{J_{bS}}$ is a partition, then the matrix $\sum_{\ell \in S_v} \frac{1}{\pi_\ell} \psi_\ell^{(b,S)} \hat{\mathbf{z}}_\ell^{(b)} (\hat{\mathbf{z}}_\ell^{(b)})^\top$ is diagonal with diagonal elements equal to $\widehat{N}_j^{(b)}, j = 1, \ldots, J_{bS}$. By the stopping criterion and assumption (H15), we have that $\sum_{\ell \in S_v} \frac{1}{\pi_\ell} \psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A_{jS}^{(b)}} \geq n_{0v} > 0$, so $\sum_{\ell \in S_v} \frac{1}{\pi_\ell} \psi_\ell^{(b,S)} \hat{\mathbf{z}}_\ell^{(b)} (\hat{\mathbf{z}}_\ell^{(b)})^\top$ is always invertible is and $\widehat{\boldsymbol{\beta}}^{(b)}$ is well-defined whatever the sample S is.

We need to consider also a second pseudo-generalized difference estimator:

$$\widehat{\widetilde{t}}_{pgd} = \sum_{k \in U_v} \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k) + \sum_{k \in S_v} \frac{y_k - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k)}{\pi_k}$$

where

$$\begin{aligned}\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k) &= \sum_{\ell\in U_v}\left(\frac{1}{B}\sum_{b=1}^{B}\frac{\psi_\ell^{(b,S)}\mathbb{1}_{\mathbf{x}_\ell\in A^{(S)}(\mathbf{x}_k,\theta_b^{(S)})}}{\widehat{\widetilde{N}}(\mathbf{x}_k,\theta_b^{(S)})}\right)y_\ell \\ &= \frac{1}{B}\sum_{b=1}^{B}\widehat{\widetilde{m}}_{tree}^{(b)}(\mathbf{x}_k)\end{aligned}$$

with $\widehat{\widetilde{N}}(\mathbf{x}_k,\theta_b^{(S)}) = \sum_{\ell\in U_v}\psi_\ell^{(b,S)}\mathbb{1}_{\mathbf{x}_\ell\in A(\mathbf{x}_k,\theta_b^{(S)})}$ and

$$\widehat{\widetilde{m}}_{tree}^{(b)}(\mathbf{x}_k) = \sum_{\ell\in U_v}\frac{\psi_\ell^{(b,S)}\mathbb{1}_{\mathbf{x}_\ell\in A^{(S)}(\mathbf{x}_k,\theta_b^{(S)})}y_\ell}{\widehat{\widetilde{N}}(\mathbf{x}_k,\theta_b^{(S)})} = (\hat{\mathbf{z}}_k^{(b)})^\top\widehat{\widetilde{\boldsymbol{\beta}}}^{(b)}, \quad k\in U_v \tag{3.50}$$

for

$$\widehat{\widetilde{\boldsymbol{\beta}}}^{(b)} = \left(\sum_{\ell\in U_v}\psi_\ell^{(b,S)}\hat{\mathbf{z}}_\ell^{(b)}(\hat{\mathbf{z}}_\ell^{(b)})^\top\right)^{-1}\sum_{\ell\in U_v}\psi_\ell^{(b,S)}\hat{\mathbf{z}}_\ell^{(b)}y_\ell.$$

The matrix $\sum_{\ell\in U_v}\psi_\ell^{(b,S)}\hat{\mathbf{z}}_\ell^{(b)}(\hat{\mathbf{z}}_\ell^{(b)})^\top$ is also diagonal with diagonal elements equal to $\sum_{\ell\in U_v}\psi_\ell^{(b,S)}\mathbb{1}_{\mathbf{x}_\ell\in A_j^{(bS)}} \geq n_{0v} > 0, j = 1,\ldots,J_{bS}$ so $\widehat{\widetilde{\boldsymbol{\beta}}}^{(b)}$ is also well-defined whatever the sample S is. In order to prove the consistency of the sample-based RF estimator $\widehat{t}_{rf}$, we use the following decomposition:

$$\frac{1}{N_v}(\hat{t}_{rf}-t_y) = \frac{1}{N_v}(\widehat{\widetilde{t}}_{pgd}-t_y) - \frac{1}{N_v}\sum_{k\in U_v}\alpha_k(\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k)-\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k)). \tag{3.51}$$

We will give first several useful lemmas. The constants used in the following results may not be the same as the ones from Section 3.9.1 even if they are denoted in the same way for simplicity.

Lemma 3. *There exists a positive constant $\tilde{c}_1$ such that:*

$$\frac{n_v}{N_v^2}\mathbb{E}_p(\widehat{\widetilde{t}}_{pgd}-t_y)^2 \leqslant \tilde{c}_1.$$

Proof. The proof is similar to that of lemma 1. We also have that $\sup_{k\in U_v}|\widetilde{m}_{rf}(\mathbf{x}_k)| \leqslant C$ by using assumption (H13). Further,

$$\begin{aligned}n_v\mathbb{E}_p\left(\frac{\widehat{\widetilde{t}}_{pgd}-t_y}{N_v}\right)^2 &\leqslant \left(\frac{n_v}{N_v}\cdot\frac{1}{\lambda}+\frac{n_v\max_{k\neq\ell\in U_v}|\pi_{k\ell}-\pi_k\pi_\ell|}{\lambda^2}\right)\cdot\frac{2}{N_v}\sum_{k\in U_v}\left(y_k^2+(\widetilde{m}_{rf}(\mathbf{x}_k))^2\right) \\ &\leqslant \tilde{c}_1\end{aligned}$$

by assumptions (H13)-(H15). ∎

Lemma 4. *There exists a positive constant $\tilde{c}_2$ such that:*

$$\frac{n_v}{N_v^2}\mathbb{E}_p(\widehat{\widetilde{t}}_{pgd}-t_y)^2 \leqslant \tilde{c}_2.$$

Proof. Using (3.50), we get that $\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k)$ can be written as a weighted sum of y-values with positive weights summing to unity, so $\sup_{k\in U_v}|\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k)| \leqslant C$ by using also assumption (H13). Now,

$$\widehat{\widetilde{t}}_{pgd} - t_y = \sum_{k\in U_v} \alpha_k (y_k - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k))$$

and

$$\begin{aligned}
\frac{n_v}{N_v^2}\mathbb{E}_p(\widehat{\widetilde{t}}_{pgd} - t_y) &= \frac{n_v}{N_v^2}\sum_{k\in U_v}\mathbb{E}_p\left[\alpha_k^2(y_k - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k))^2\right] \\
&\quad + \frac{n_v}{N_v^2}\sum_{k\in U_v}\sum_{\ell\neq k,\ell\in U_v}\mathbb{E}_p\left[(y_k - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k))(y_\ell - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_\ell))\mathbb{E}_p(\alpha_k\alpha_\ell|\hat{\mathcal{P}}_S)\right] \\
&\leqslant \frac{2n_vC^2}{\lambda N_v} + \frac{n_v}{N_v^2}\sum_{k\in U_v}\sum_{\ell\neq k,\ell\in U_v}\mathbb{E}_p\left[|y_k - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k)||y_\ell - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_\ell)|\max_{\ell\neq k\in U_v}|\mathbb{E}_p(\alpha_k\alpha_\ell|\hat{\mathcal{P}}_S)|\right] \\
&\leqslant \tilde{c}_2,
\end{aligned}$$

by assumptions (H14) and (H16). ∎

Lemma 5. *There exists a positive constant $\tilde{c}_3$ not depending on b such that:*

$$\mathbb{E}_p\left\|\widehat{\boldsymbol{\beta}}^{(b)} - \widehat{\widetilde{\boldsymbol{\beta}}}^{(b)}\right\|_2^2 \leqslant \frac{\tilde{c}_3 n_v}{n_{0v}^2},$$

for all $b = 1,\ldots,B$.

Proof. Let denote by $\widehat{\mathbf{T}}^{(b)} = \sum_{\ell\in S_v}\frac{1}{\pi_\ell}\psi_\ell^{(b,S)}\hat{\mathbf{z}}_\ell^{(b)}(\hat{\mathbf{z}}_\ell^{(b)})^\top$. As already mentioned, the $J_{bS}\times J_{bS}$ dimensional matrix $\widehat{\mathbf{T}}^{(b)}$ is diagonal with diagonal elements given by $\widehat{N}_j^{(b)} = \sum_{\ell\in S_v}\frac{1}{\pi_\ell}\psi_\ell^{(b,S)}\mathbb{1}_{\mathbf{x}_\ell\in A_{jS}^{(b)}}$ the weighted somme of units falling in the region $A_{jS}^{(b)}$ for $j = 1,\ldots,J_{bS}$ and by the stopping criterion, we have that $\widehat{N}_j^{(b)} \geq n_{0v} > 0$. The matrix $\widehat{\mathbf{T}}^{(b)}$ is then always invertible with

$$||(\widehat{\mathbf{T}}^{(b)})^{-1}||_2 \leq n_{0v}^{-1} \quad \text{for all} \quad b = 1,\ldots B. \tag{3.52}$$

Now, write

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(b)} - \widehat{\widetilde{\boldsymbol{\beta}}}^{(b)} &= (\widehat{\mathbf{T}}^{(b)})^{-1}\left(\sum_{\ell\in S_v}\frac{1}{\pi_\ell}\psi_\ell^{(b,S)}\hat{\mathbf{z}}_\ell^{(b)}y_\ell - \widehat{\mathbf{T}}^{(b)}\widehat{\widetilde{\boldsymbol{\beta}}}^{(b)}\right) \\
&= (\widehat{\mathbf{T}}^{(b)})^{-1}\sum_{\ell\in S_v}\frac{1}{\pi_\ell}\psi_\ell^{(b,S)}\hat{\mathbf{z}}_\ell^{(b)}\left(y_\ell - \widehat{\widetilde{m}}_{tree}^{(b)}(\mathbf{x}_\ell)\right) \\
&= (\widehat{\mathbf{T}}^{(b)})^{-1}\sum_{\ell\in U_v}\alpha_\ell\widehat{E}_\ell^{(b)}
\end{aligned} \tag{3.53}$$

where $\widehat{E}_\ell^{(b)} = \psi_\ell^{(b,S)}\hat{\mathbf{z}}_\ell^{(b)}(y_\ell - \widehat{\widetilde{m}}_{tree}^{(b)}(\mathbf{x}_\ell))$ with $\sum_{\ell\in U_v}\widehat{E}_\ell^{(b)} = 0$. We have that $||\hat{\mathbf{z}}_\ell^{(b)}||_2 = 1$ and $\sup_{\ell\in U_v}|\widehat{\widetilde{m}}_{tree}^{(b)}(\mathbf{x}_\ell))| \leq C$ for all $\ell\in U_v$ and $b = 1,\ldots,B$, then:

$$||\widehat{E}_\ell^{(b)}||_2^2 \leqslant 2C^2.$$

Following the same lines as in lemma 4, we get that it exists a positive constant $\tilde{C}_0$ not depending on b such that

$$\frac{1}{N_v^2}\mathbb{E}_p\left\|\sum_{\ell\in U_v}\alpha_\ell \widehat{E}_\ell^{(b)}\right\|_2^2 \leqslant \frac{\tilde{C}_0}{n_v}, \quad \text{for all} \quad b=1,\ldots B. \tag{3.54}$$

We obtain then from relations (3.52) and (3.53) that:

$$\begin{aligned}
\mathbb{E}_p\left\|\widehat{\boldsymbol{\beta}}^{(b)}-\widehat{\widetilde{\boldsymbol{\beta}}}^{(b)}\right\|_2^2 &\leqslant \mathbb{E}_p\left(N_v^2||(\widehat{\mathbf{T}}^{(b)})^{-1}||_2^2\frac{1}{N_v^2}\left\|\sum_{\ell\in U_v}\alpha_\ell\widehat{E}_\ell^{(b)}\right\|_2^2\right)\\
&\leqslant \frac{N_v^2}{n_{0v}^2}\frac{1}{N_v^2}\mathbb{E}_p\left\|\sum_{\ell\in U_v}\alpha_\ell\widehat{E}_\ell^{(b)}\right\|_2^2\\
&\leqslant \frac{N_v^2}{n_{0v}^2}\frac{\tilde{C}_0}{n_v}\\
&\leqslant \frac{\tilde{c}_3 n_v}{n_{0v}^2}
\end{aligned} \tag{3.55}$$

by assumption (H14). ∎

Result 3.9.4. *Consider a sequence of sample RF estimators $\{\widehat{t}_{rf}\}$. Then, there exist positive constants $\tilde{C}_1, \tilde{C}_2$ such that*

$$\frac{1}{N_v}\mathbb{E}_p|\hat{t}_{rf}-t_y| \leqslant \frac{\tilde{C}_1}{\sqrt{n_v}}+\frac{\tilde{C}_2}{n_{0v}}.$$

If $\dfrac{n_v^u}{n_{0v}} = O(1)$ with $1/2 \leqslant u \leqslant 1$, then

$$\mathbb{E}_p\left|\frac{1}{N_v}\left(\widehat{t}_{rf}-t_y\right)\right| \leqslant \frac{\tilde{C}}{\sqrt{n_v}}, \quad \textit{with } \xi\textit{-probability one.}$$

Proof. We use the decomposition given in relation (3.51):

$$\frac{1}{N_v}(\hat{t}_{rf}-t_y) = \frac{1}{N_v}(\widehat{\widetilde{t}}_{pgd}-t_y) - \frac{1}{N_v}\sum_{k\in U_v}\alpha_k(\widehat{m}_{rf}(\mathbf{x}_k)-\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k)).$$

Now,

$$\mathbb{E}_p\left|\frac{1}{N_v}\sum_{k\in U_v}\alpha_k(\widehat{m}_{rf}(\mathbf{x}_k)-\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k))\right| \leqslant \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_v}\mathbb{E}_p\left|\sum_{k\in U_v}\alpha_k(\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)-\widehat{\widetilde{m}}_{tree}^{(b)}(\mathbf{x}_k))\right|$$

and using relations (3.49) and (3.50), we get:

$$\begin{aligned}
\frac{1}{N_v}\mathbb{E}_p\left|\sum_{k\in U_v}\alpha_k(\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)-\widehat{\widetilde{m}}_{tree}^{(b)}(\mathbf{x}_k))\right| &\leqslant \mathbb{E}_p\left(\left\|\frac{1}{N_v}\sum_{k\in U_v}\alpha_k\hat{\mathbf{z}}_k^{(b)}\right\|_2\left\|\widehat{\boldsymbol{\beta}}^{(b)}-\widehat{\widetilde{\boldsymbol{\beta}}}^{(b)}\right\|_2\right)\\
&\leqslant \sqrt{\mathbb{E}_p\left\|\frac{1}{N_v}\sum_{k\in U_v}\alpha_k\hat{\mathbf{z}}_k^{(b)}\right\|_2^2\mathbb{E}_p\left\|\widehat{\boldsymbol{\beta}}^{(b)}-\widehat{\widetilde{\boldsymbol{\beta}}}^{(b)}\right\|_2^2}.
\end{aligned}$$

We have that $||\hat{\mathbf{z}}_k^{(b)}||_2 = 1$ for all $k \in U_v$ and $b = 1, \ldots, B$. We can show then by using the same arguments as in the proof of lemma 4, that there exists a positive constant $\tilde{C}_0'$ such that

$$E_p \left\| \frac{1}{N_v} \sum_{k \in U_v} \alpha_k \hat{\mathbf{z}}_k^{(b)} \right\|_2^2 \leqslant \frac{\tilde{C}_0'}{n_v}$$

which together with lemma 5 gives us that there exists a positive constant $\tilde{C}_2$ such that

$$\frac{1}{N_v} \mathbb{E}_p \left| \sum_{k \in U_v} \alpha_k (\widehat{m}_{rf}(\mathbf{x}_k) - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k)) \right| \leqslant \frac{\tilde{C}_2}{n_{0v}}. \tag{3.56}$$

Now,

$$\begin{aligned} \frac{1}{N_v} \mathbb{E}_p \left| \hat{t}_{rf} - t_y \right| &\leqslant \frac{1}{N_v} \mathbb{E}_p \left| \widehat{\widetilde{t}}_{pgd} - t_y \right| + \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_v} \mathbb{E}_p \left| \sum_{k \in U_v} \alpha_k (\widehat{m}_{tree}^{(b)}(\mathbf{x}_k) - \widehat{\widetilde{m}}_{tree}^{(b)}(\mathbf{x}_k)) \right| \\ &\leqslant \frac{\tilde{C}_1}{\sqrt{n_v}} + \frac{\tilde{C}_2}{n_{0v}} \end{aligned}$$

by using lemma 4 and relation (3.56). ■

Result 3.9.5. *Consider a sequence of RF estimators* $\{\widehat{t}_{rf}\}$. *Assume that* $\frac{n_v^u}{n_{0v}} = O(1)$ *with* $1/2 < u \leqslant 1$. *Then,*

$$\frac{\sqrt{n_v}}{N_v} \left(\widehat{t}_{rf} - t_y \right) = \frac{\sqrt{n_v}}{N_v} \left(\widehat{t}_{pgd} - t_y \right) + o_{\mathbb{P}}(1).$$

Proof. We have

$$\frac{\sqrt{n_v}}{N_v} \left(\widehat{t}_{rf} - t_y \right) = \frac{\sqrt{n_v}}{N_v} \left(\widehat{t}_{pgd} - t_y \right) + \frac{\sqrt{n_v}}{N_v} \sum_{k \in U_v} \alpha_k (\widehat{m}_{rf}(\mathbf{x}_k) - \widetilde{m}_{N,rf}(\mathbf{x}_k)). \tag{3.57}$$

Now,

$$\begin{aligned} &\frac{\sqrt{n_v}}{N_v} \sum_{k \in U_v} \alpha_k (\widehat{m}_{rf}(\mathbf{x}_k) - \widetilde{m}_{rf}(\mathbf{x}_k)) \\ = &\frac{\sqrt{n_v}}{N_v} \sum_{k \in U_v} \alpha_k (\widehat{m}_{rf}(\mathbf{x}_k) - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k)) + \frac{\sqrt{n_v}}{N_v} \sum_{k \in U_v} \alpha_k (\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k) - \widetilde{m}_{rf}(\mathbf{x}_k)). \end{aligned} \tag{3.58}$$

Relation (3.56) gives us that

$$\frac{\sqrt{n_v}}{N_v} \sum_{k \in U_v} \alpha_k (\widehat{m}_{rf}(\mathbf{x}_k) - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k)) = O_{\mathbb{P}} \left(\frac{\sqrt{n_v}}{n_{0v}} \right) = o_{\mathbb{P}}(1) \tag{3.59}$$

provided that $\frac{n_v^u}{n_{0v}} = O(1)$ with $1/2 < u \leqslant 1$. Consider now the second term from the right-side of relation (3.58). We have:

$$\begin{aligned} &\frac{n_v}{N_v^2} \mathbb{E}_p \left(\sum_{k \in U_v} \alpha_k (\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k) - \widetilde{m}_{rf}(\mathbf{x}_k)) \right)^2 \\ \leqslant &\frac{n_v}{N_v^2} \frac{(1+\lambda)^2}{\lambda^2} \sum_{k \in U_v} \mathbb{E}_p \left(\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k) - \widetilde{m}_{rf}(\mathbf{x}_k) \right)^2 \end{aligned}$$

$$
\begin{aligned}
&+ \quad \frac{n_v}{N_v^2} \sum_{k \in U_v} \sum_{\ell \neq k, \ell \in U_v} \mathbb{E}_p \left[|\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k) - \widetilde{m}_{rf}(\mathbf{x}_k)| |\widehat{\widetilde{m}}_{rf}(\mathbf{x}_\ell) - \widetilde{m}_{rf}(\mathbf{x}_\ell)| \max_{\ell \neq k \in U_v} |\mathbb{E}_p(\alpha_k \alpha_\ell | \hat{\mathcal{P}}_S)| \right] \\
&\leqslant \quad \left(\frac{n_v}{N_v} \frac{(1+\lambda)^2}{\lambda^2} + \frac{C_1}{\lambda^2} \right) \frac{1}{N_v} \sum_{k \in U_v} \mathbb{E}_p \left(\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k) - \widetilde{m}_{rf}(\mathbf{x}_k) \right)^2 = o(1),
\end{aligned}
$$

by assumptions (H14), (H15), (H16) and (H17). It follows then that

$$
\frac{\sqrt{n_v}}{N_v} \sum_{k \in U_v} \alpha_k (\widehat{\widetilde{m}}_{rf}(\mathbf{x}_k) - \widetilde{m}_{rf}(\mathbf{x}_k)) = o_{\mathbb{P}}(1). \tag{3.60}
$$

Relations (3.57), (3.58), (3.59) and (3.60) give then the result. ■

Result 3.9.6. *Consider a sequence of population RF estimators* $\{\widehat{t}_{rf}\}$. *Assume also that* $\frac{n_v^u}{n_{0v}} = O(1)$ *with* $1/2 < u \leqslant 1$. *Then, the variance estimator* $\widehat{\mathbb{V}}_{rf}(\widehat{t}_{rf})$ *is asymptotically design-consistent for the asymptotic variance* $\mathbb{AV}_p\left(\widehat{t}_{rf}\right)$. *That is,*

$$
\lim_{v \to \infty} \mathbb{E}_p \left(\frac{n_v}{N_v^2} \left| \widehat{\mathbb{V}}_{rf}(\widehat{t}_{rf}) - \mathbb{AV}_p(\widehat{t}_{rf}) \right| \right) = 0. \tag{3.61}
$$

Proof. The proof follows the same steps as those of result (3.9.3). We need to show that

$$
\mathbb{E}_p \left[\left(\widehat{m}_{rf}(\mathbf{x}_k) - \widetilde{m}_{rf}(\mathbf{x}_k) \right)^2 \right] = o(1), \tag{3.62}
$$

uniformly in $k \in U_v$. We have $\widehat{m}_{rf}(\mathbf{x}_k) - \widetilde{m}_{rf}(\mathbf{x}_k) = \widehat{m}_{rf}(\mathbf{x}_k) - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k) + \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k) - \widetilde{m}_{rf}(\mathbf{x}_k)$ and

$$
\begin{aligned}
\mathbb{E}_p(\widehat{m}_{rf}(\mathbf{x}_k) - \widehat{\widetilde{m}}_{rf}(\mathbf{x}_k))^2 &\leqslant \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}_p(\widehat{m}_{tree}^{(b)}(\mathbf{x}_k) - \widehat{\widetilde{m}}_{tree}^{(b)}(\mathbf{x}_k))^2 \\
&\leqslant \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}_p \left(||\hat{\mathbf{z}}_k^{(b)}||_2^2 \left\| \widehat{\boldsymbol{\beta}}^{(b)} - \widehat{\widetilde{\boldsymbol{\beta}}}^{(b)} \right\|_2^2 \right) \\
&\leqslant \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}_p \left(\left\| \widehat{\boldsymbol{\beta}}^{(b)} - \widehat{\widetilde{\boldsymbol{\beta}}}^{(b)} \right\|_2^2 \right) \\
&\leqslant \frac{\tilde{c}_3 n_v}{n_{0v}^2} = o(1)
\end{aligned}
$$

by lemma 5 and provided that $\frac{n_v^u}{n_{0v}} = O(1)$ with $1/2 < u < 1$. The result (3.62) follows then by using also assumption (H17). ■

4 IMPUTATION PROCEDURES IN SURVEYS USING NONPARAMETRIC AND MACHINE LEARNING METHODS: AN EMPIRICAL COMPARISON

Abstract.[1] Nonparametric and machine learning methods are flexible methods for obtaining accurate predictions. Nowadays, data sets with a large number of predictors and complex structures are fairly common. In the presence of item nonresponse, nonparametric and machine learning procedures may thus provide a useful alternative to traditional imputation procedures for deriving a set of imputed values used next for the estimation of study parameters defined as solution of population estimating equation. In this paper, we conduct an extensive empirical investigation that compares a number of imputation procedures in terms of bias and efficiency in a wide variety of settings, including high-dimensional data sets. The results suggest that a number of machine learning procedures perform very well in terms of bias and efficiency.

Keywords: Additive models; Bayesian additive regression trees (BART); CART; Cubist algorithm; Ensemble Methods; Nearest Neighbor; Item nonresponse; Random forest; Support vector regression (SVR); Survey data; Statistical learning; Tree boosting.

4.1 Introduction

In the last decade, the interest in machine learning methods has been growing in national statistical offices (NSO). These data-driven methods provide flexible tools for obtaining accurate predictions. The increasing availability of data sources (e.g., big data sources and satellite information) provides a rich pool of potential predictors that may be used to obtain predictions at different stages of a survey. These stages include the nonresponse treatment stage (e.g., propensity score weighting and imputation) and the estimation stage (e.g., model-assisted estimation and small area estimation). The imputation stage is the focus of the current paper.

Item nonresponse refers to the presence of missing values for some, but not all, survey variables. Frequent causes of item nonresponse include refusal to answer a sensitive question (e.g., income) and edit failures. The most common way of treating item nonresponse in NSOs is to replace a missing value with a single imputed value, constructed on the basis of a set of p explanatory variables, $\mathbf{X} = (X_1, \ldots, X_p)$, available for both respondents and nonrespondents. A variety of imputation procedures are available, ranging from simple (e.g., mean, historical and ratio imputation) to more complex (e.g., nonparametric procedures); e.g., see Chen and Haziza (2019) for an overview of imputation procedures in surveys. Every imputation procedure makes some (implicit of explicit) assumptions about the distribution of the variable Y requiring imputation. This set of assumptions is often referred to as an imputation model. At the imputation stage, it is therefore important to identify and include in the model all the appropriate explanatory variables that are predictive of the variable requiring imputation and determine a suitable model describing the relationship between Y and the set of explanatory variables $\mathbf{X}$.

We distinguish parametric imputation procedures from nonparametric imputation procedures. In parametric imputation, the shape of the relationship between Y and $\mathbf{X}$ is predetermined; e.g., linear and generalized linear regression models. However, point estimators based on parametric imputation procedures may suffer from bias if the functional form is misspecified or if the vector $\mathbf{X}$ fails to include interactions or predictors accounting for curvature. In contrast, with nonparametric methods, the shape

[1] The article is accepted for publication in Journal of Survey Statistics and Methodology.

of the relationship between Y and $\mathbf{X}$ is left unspecified. These methods have the ability to capture nonlinear trends in the data and tend to be robust to the non-inclusion of interactions or predictors accounting for curvature.

Commonly used nonparametric methods include kernel smoothing, local polynomial regression and spline-based regression models. While these methods provide some robustness against model misspecification, they tend to breakdown when the number predictors is large, a problem known as the curse of dimensionality. To mitigate this problem, one may employ additive models (Hastie and Tibshirani, 1986). However, when the dimension of $\mathbf{X}$ is very large, these models tend to fail and machine learning methods may provide an interesting alternative. The class of machine learning methods, that includes tree-based models such as random forests and boosting methods, provide more flexible approaches able to adapt to complex non-linear and non-additive relationships between the survey variable requiring imputation and a set of predictors. These methods may also prove useful in the case of large data sets exhibiting a large number of observations on a large number of variables. Many machine learning procedures are relatively computationally efficient and can produce accurate predictions by offering the user a kind of automatic variable selection that may prove useful in a high-dimensional setting.

However, both a theoretical treatment and an empirical comparison of machine learning imputation procedures in the context of missing survey data are currently lacking. In this paper, we aim to fill the latter gap by conducting an extensive simulation study that investigates the performance of several nonparametric and machine learning procedures in terms of bias and efficiency. To that end, we generated several finite populations with relationships between Y and $\mathbf{X}$, ranging from simple to complex and generated the missing values according to several nonresponse mechanisms. We also considered both a low-dimensional and high dimensional settings. The simulation setup and the models are described in Section 4.4. We restricted our attention to population totals (Section 4.4) and population quantiles (Section 4.5) as the target parameters. The following procedures were included in our comparisons: the score method (Haziza and Beaumont, 2007, Little, 1986), K nearest-neighbour (Chen and Shao, 2000), additive models based on B-spline regression, regression trees (Breiman, 1984), random forests (Breiman, 2001), tree-based boosting methods (Friedman, 2001) including XGBoost (Chen and Guestrin, 2016) and Bayesian additive regression trees (Chipman et al., 2010), the cubist algorithm (Quinlan, 1993, Quinlan et al., 1992) and support vector regression (Vapnik, 1998, 2000). In Section 4.3, we describe these models and the corresponding imputation procedures.

In recent years, machine learning procedures have received some attention in a survey sampling context. In the ideal situation of 100% response, the theoretical properties of model-assisted estimation procedures based on regression trees (McConville and Toth, 2019) and random forests (Dagdoug et al., 2021b) have been recently established. Dagdoug et al. (2022b) studied the theoretical properties of point and variance estimators based on random forests in a context of imputation for item nonresponse and data integration; see also De Moliner and Goga (2018), Tipton et al. (2013) for applications of random forests in surveys. A number of empirical investigations have been conducted to assess the performance of machine learning procedures in a context of propensity score estimation for unit nonresponse; e.g., Lohr et al. (2015), Gelein (2017) and Kern et al. (2019).

The machine learning procedures described in Section 4.3 slightly differ from their traditional implementation because of the inclusion of the sampling weights in the construction of imputed values. However, it should be noted that most of the machine learning software packages for obtaining predicted values assume simple random sampling and cannot handle unequal weights. Modifying machine learning algorithms to account for unequal weights may prove challenging.

When the design features (e.g., sampling weights, stratum indicators, etc.) are related to the survey variable requiring imputation, failing to incorporate them in the models may lead to biased estimators. To cope with this issue, we suggest to include all the appropriate design variables in the specification of the model. Standard machine learning software packages may then be safely used for creating a set of

imputed values. In Section 4.4, we use Poisson sampling with inclusion probabilities proportional to a size variable X to select repeated samples from the finite population. The size variable X being related to the variable requiring imputation, including the X-variable in the specified models led to satisfactory results.

4.2 Preliminaries

Consider a finite population $U = \{1, 2, ..., N\}$ of size N. Let Y denote a survey variable and y_i be the y-values attached to unit i, $i = 1, \cdots, N$. We are interested in estimating (i) the finite population total of the y-values, $t_y = \sum_{i \in U} y_i$ and (ii) the finite population quantile of order γ defined as $Q_\gamma := \inf \{t \in \mathbb{R}; F_N(t) \geqslant \gamma\}$, where

$$F_N(t) = \sum_{i \in U} \mathbb{1}(y_i \leqslant t)/N$$

denotes the finite population distribution function.

From U, we select a sample S, of size n, according to a sampling design $\mathcal{P}(S = s)$ with first-order inclusion probabilities $\pi_i = Pr(i \in S)$.

A complete data estimator of t_y is the well-known Horvitz-Thompson estimator

$$\widehat{t}_\pi = \sum_{i \in S} \frac{y_i}{\pi_i}, \tag{4.1}$$

which is design-unbiased for t_y provided that $\pi_i > 0$ for all $i \in U$. A complete data estimator of the finite population quantile Q_γ is given by

$$\widehat{Q}_\gamma := \inf \left\{ t \in \mathbb{R}; \widehat{F}(t) \geqslant \gamma \right\}, \tag{4.2}$$

where

$$\widehat{F}(t) = \frac{1}{\widehat{N}} \sum_{i \in S} \frac{\mathbb{1}(y_i \leqslant t)}{\pi_i} \tag{4.3}$$

with $\widehat{N} = \sum_{i \in S} 1/\pi_i$ denoting the Horvitz-Thompson estimator of the population size N. Under mild regularity conditions (Wang and Opsomer, 2011), the complete data estimator $\widehat{Q}_\gamma$ is design-consistent for Q_γ.

In practice, the Y-variable may be prone to missing values. Let r_i be a response indicator such that $r_i = 1$ if y_i is observed and $r_i = 0$, otherwise. Let $S_r = \{i \in S;\ r_i = 1\}$ denote the set of respondents, of size n_r, and $S_m = \{i \in S;\ r_i = 0\}$ the set of nonrespondents, of size n_m, such that $S_r \cup S_m = S$ and $n_r + n_m = n$. Available to the imputer is the data $(y_i, \mathbf{x}_i)$ for $i \in S_r$ as well as the values of the vector $\mathbf{x}_i$ for $i \in S_m$.

Let $\widehat{y}_i$ be the imputed value used to replace the missing value y_i and

$$\widetilde{y}_i = r_i y_i + (1 - r_i)\widehat{y}_i$$

be the ith value of the Y-variable after imputation. Point estimators of t_y and Q_γ after imputation, often referred to as imputed estimators, are readily obtained from the complete data estimators (4.1) and (4.2) by replacing y_i with $\widetilde{y}_i$. This leads to

$$\widehat{t}_{imp} = \sum_{i \in S} \frac{\widetilde{y}_i}{\pi_i} \tag{4.4}$$

and

$$\widehat{Q}_{\gamma,imp} = \inf \left\{ t \in \mathbb{R}; \widehat{F}_{imp}(t) \geqslant \gamma \right\}, \tag{4.5}$$

where

$$\widehat{F}_{imp}(t) = \frac{1}{\widehat{N}} \sum_{i \in S} \frac{\mathbb{1}\left(\widetilde{y}_i \leqslant t\right)}{\pi_i} \tag{4.6}$$

denotes the imputed estimator of $F_N(t)$.

Remark 4.2.1. *The population total t_y, the distribution function $F_N(t)$ and the quantile of order γ, Q_γ, may all be obtained as the solution of the following census estimating equation (Binder, 1983, Chen and Haziza, 2019):*

$$U_N(\theta_N) = \sum_{i \in U} u(y_i; \theta_N) = 0, \tag{4.7}$$

where θ_N is a generic notation denoting a finite population parameter and $u(y_i;\theta)$ is a function of θ_N. We assume that a solution to (4.7) exists and is unique. For instance, the population total t_y can be obtained as a solution of (4.7) with $u(y_i;\theta_N) = y_i - n^{-1}\pi_i\theta_N$; the finite population distribution function $F_N(t)$ can be obtained as a solution of (4.7) with $u(y_i;\theta_N) = \mathbb{1}\left(y_i \leqslant t\right) - \theta_N$. Finally, the quantile Q_γ of order γ can be obtained as a solution of (4.7) with $u(y_i;\theta_N) = \mathbb{1}\left(y_i \leqslant \theta_N\right) - \gamma$. Other finite population parameters can be obtained as a solution of (4.7); e.g., see Chen and Haziza (2019). The imputed estimators $\widehat{t}_{imp}$, $\widehat{Q}_{\gamma,imp}$ and $\widehat{F}_{imp}(t)$ given respectively by (4.4)-(4.6) can be obtained by solving the following sample estimating equation:

$$\widehat{U}_{imp}(\widehat{\theta}_{imp}) = \sum_{i \in S} \frac{1}{\pi_i} u(\widetilde{y}_i; \widehat{\theta}_{imp}) = 0,$$

where $\widehat{\theta}_{imp}$ denotes an imputed estimator of θ_N.

To construct the imputed values $\widetilde{y}_i$, we postulate the following imputation model ξ:

$$\begin{aligned} &\mathbb{E}_\xi(y_i|\mathbf{x}_i) = f(\mathbf{x}_i), \\ &\mathbb{V}_\xi\left(y_i|\mathbf{x}_i\right) = \sigma_i^2, \\ &\mathbb{C}ov_\xi\left(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j\right) = 0 \quad \text{for} \qquad i \neq j, \end{aligned} \tag{4.8}$$

where f is an unknown function. Often, the variance structure σ_i^2 is assumed to have the form $\sigma_i^2 = \sigma^2 a_i$, where $a_i > 0$ is a known coefficient attached to unit i and σ^2 is an unknown parameter.

We assume that the data are Missing At Random (Rubin, 1976):

$$f(y_i|\mathbf{x}_i, r_i = 1) = f(y_i|\mathbf{x}_i, r_i = 0). \tag{4.9}$$

That is, we assume that the distribution of Y given $\mathbf{x}$ is the same for both respondents and nonrespondents. If Condition (4.9) holds, the imputed values can be safely generated from $f(y_i|\mathbf{x}_i, r_i = 1)$, which can be estimated from the observed data. In the context of imputation, the properties of point estimators are evaluated with respect to the joint distribution induced by the imputation, the sampling design and the unknown nonresponse mechanism. This framework is often referred to as the ξpq-framework (Chen and Haziza, 2019). Note that our simulation setup in Section 4.4 is consistent with the ξpq-framework as the simulation process involves (i) generating repeated finite populations; (ii) selecting a sample from each of population and (iii) generating a set of response indicators in each sample.

Deterministic imputation consists of replacing the missing y_i by $\widetilde{y}_i = \widehat{f}(\mathbf{x}_i)$, where $\widehat{f}$ is an estimator of the unknown regression function f based on the responding units $i \in S_r$. However, deterministic imputation methods tend to distort the distribution of the survey variable Y requiring imputation, potentially leading to biased estimators of quantiles (Chen and Haziza, 2019, Haziza, 2009). To cope with this issue, one can recourse to random imputation that consists of adding an appropriate amount of

random noise to the deterministic value $\widehat{f}(\mathbf{x}_i)$. More specifically, let $e_j := \widehat{\sigma}_j^{-1}\{y_j - \widehat{f}(\mathbf{x}_j)\}$ for $j \in S_r$, where $\widehat{\sigma}_j$ of an estimator of σ_j (see Remark 4.2.2 below). We define the standardized residual

$$\widetilde{e}_j = e_j - \frac{\sum_{\ell \in S_r} w_\ell e_\ell}{\sum_{\ell \in S_r} w_\ell}, \quad j \in S_r.$$

In the case of random imputation, the missing y_i is replaced by

$$\widehat{y}_i = \widehat{f}(\mathbf{x}_i) + \widehat{\sigma}_i \widehat{e}_i, \tag{4.10}$$

where $\widehat{e}_i$ is selected at random from the set of standardized residuals $\{\widetilde{e}_j\}_{j \in S_r}$ with probability $w_j / \sum_{\ell \in S_r} w_\ell$.

Remark 4.2.2. *To obtain an estimator $\widehat{\sigma}_i$ of σ_i, one can postulate a model $\mathbb{E}(\epsilon_i^2 \mid \mathbf{x}_i) = m(\mathbf{x}_i)$, where m is an unknown function. An estimator $\widehat{\sigma}_i^2$ of σ_i^2 is obtained by fitting a parametric or a nonparametric procedure with the square residuals e_i^2 as the response and $\mathbf{x}_i$ as the set of predictors.*

In Section 4.3, except for the parametric imputation procedure discussed in Section 4.3.1, all the other procedures (Section 4.3.2-4.3.9) are nonparametric. In Section 4.4, these procedures are compared empirically in terms of bias and efficiency under a variety of settings.

4.3 A description of imputation methods

4.3.1 Parametric regression imputation

Parametric regression assumes that the first moment (4.8) is given by

$$\mathbb{E}_\xi(y_i|\mathbf{x}_i) = f(\mathbf{x}_i, \boldsymbol{\beta}), \tag{4.11}$$

where $\boldsymbol{\beta}$ is a vector of coefficients to be estimated and $f(\cdot)$ is a predetermined function. An estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by solving the following estimating equations based on the responding units:

$$\sum_{i \in S_r} \frac{w_i}{\sigma_i^2} \{y_i - f(\mathbf{x}_i, \boldsymbol{\beta})\} \frac{\partial f(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0, \tag{4.12}$$

where $w_i > 0$ is a weight attached to element i. Common choices for w_i include $w_i = 1$ and $w_i = \pi_i^{-1}$ (Chen and Haziza, 2019). The imputed value $\widehat{y}_i$ under deterministic parametric regression imputation is given by

$$\widehat{y}_i = f(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}), \quad i \in S_m. \tag{4.13}$$

A special case of (4.13) is $f(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i^\top \boldsymbol{\beta}$, which corresponds to the customary linear regression model. In this case, the imputed value (4.13) reduces to

$$\widehat{y}_i = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}, \quad i \in S_m, \tag{4.14}$$

where

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{j \in S_r} w_j \sigma_j^{-2} \mathbf{x}_j \mathbf{x}_j^\top \right)^{-1} \sum_{j \in S_r} w_j \sigma_j^{-2} \mathbf{x}_j y_j. \tag{4.15}$$

The imputed value $\widehat{y}_i$ given by (4.14) can be written as a weighted sum of the respondent y-values:

$$\widehat{y}_i = \sum_{j \in S_r} w'_{ij} y_j, \quad i \in S_m, \tag{4.16}$$

where $w'_{ij} = \mathbf{x}_i^\top \left(\sum_{j' \in S_r} w_{j'} \sigma_{j'}^{-2} \mathbf{x}_{j'} \mathbf{x}_{j'}^\top\right)^{-1} w_j \sigma_j^{-2} \mathbf{x}_j$. If the intercept is among the X-variables, then $\sum_{j \in S_r} w'_{ij} = 1$ for all $i \in S_m$. A random counterpart of (4.13) is given by (4.10).

Another important special case of (4.13) is the logistic regression model,

$$f(\mathbf{x}_i, \boldsymbol{\beta}) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})/(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})),$$

which can be used for modeling binary variables. An estimator of $\boldsymbol{\beta}$ is obtained by solving (4.12), which requires a numerical algorithm such as the Newton-Raphson procedure. To eliminate the possibility of an impossible imputed value, a missing value to a $0-1$ variable is typically imputed by $\widehat{y}_i$, where $\widehat{y}_i$ is a realization of a Bernoulli variable with parameter $f(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})$.

Under deterministic or random parametric regression imputation, the imputed estimator $\widehat{t}_{imp}$ is consistent for t_y provided that the first moment of the imputation model (4.8) is correctly specified. However, this type of imputation may lead to biased estimators of quantiles. In contrast, the use of a random parametric regression imputation procedure tend to preserve the distribution of the variable requiring imputation, leading to valid estimators; see Chen and Haziza (2019) for a discussion.

4.3.2 Imputation classes : the score method

The score method (Haziza and Beaumont, 2007, Little, 1986) consists of partitioning the sample S into H (say) imputation classes and imputing the missing values within each class independently from one class to another. It can be implemented as follows:

Step 1: For all $i \in S$, compute the preliminary values $\widehat{y}_i^{LR} = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is given by (4.15).

Step 2: Compute the empirical quantiles $q_1, q_2, \ldots, q_{H-1}$ of order $1/H, 2/H, \ldots, (H-1)/H$ of the $\widehat{y}^{LR}$-values.

Step 3: Split the sample S into H classes, $C_1, \ldots, C_h, \ldots, C_H$, such that

$$C_h = \left\{i \in S : \widehat{y}_i^{LR} \in [q_{h-1}; q_h)\right\}, \quad h = 1, \ldots, H,$$

with $q_0 = -\infty$ and $q_H = +\infty$.

It is common practice to use either mean imputation or random hot-deck imputation within classes. For mean imputation, the imputed value for missing y_i in the hth imputation class is given by

$$\widehat{y}_i = \frac{\sum_{j \in S_r \cap C_h} w_j y_j}{\sum_{j \in S_r \cap C_h} w_j} = \sum_{j \in S_r \cap C_h} w'_{ij} y_j, \quad i \in S_m \cap C_h,$$

where $w'_{ij} = w_j / \sum_{j' \in S_r \cap C_h} w_{j'}$ are the same for all $i \in S_m \cap C_h$ and $\sum_{j \in S_r \cap C_h} w'_{ij} = 1$ for all $i \in S_m \cap C_h$. For random hot-deck imputation, the imputed value is given by $\widehat{y}_i = y_j$, where the donor $j \in S_r \cap C_h$ is selected at random from the set of donors belonging to the hth imputation class with probability $w_j / \sum_{j' \in S_r \cap C_h} w_{j'}$. Note that random hot-deck imputation within classes can be viewed as mean imputation within classes with added residuals.

4.3.3 K-nearest neighbours imputation

K-nearest neighbour (KNN) imputation is one of the simplest and widely used nonparametric imputation procedures. No explicit assumption is made about the regression function f relating Y and $\mathbf{X}$. KNN imputation consists of replacing the missing value of a recipient by the weighted average of the y-values of its K closest respondents in terms of the X-variables.

Nearest-neighbour (NN) imputation corresponds to the limiting case of KNN obtained with $K = 1$. NN is a donor imputation belonging to the class of hot-deck procedures (Chen and Shao, 2000) since a missing value is replaced by an actual respondent y-value from the same file. NN imputation is especially useful for imputing categorical or discrete Y-variables; e.g., see Chen and Shao (2000), Beaumont and Bocci (2009) and Yang and Kim (2019).

Let $\mathcal{N}_K(i)$ be the set of K responding units closest to $\mathbf{x}_i$. Any distance function in $\mathbb{R}^p$ may be used to measure the closeness between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. In the simulation study presented in Section 4.4, we used the customary Euclidean distance. The KNN imputed value for missing y_i is given by

$$\widehat{y}_i = \frac{\sum_{j\in\mathcal{N}_K(i)\cap S_r} w_j y_j}{\sum_{j\in\mathcal{N}_K(i)\cap S_r} w_j}, \quad i \in S_m.$$

The imputed value $\widehat{y}_i$ obtained with KNN can be written as a weighted sum of the respondent y-values:

$$\widehat{y}_i = \sum_{j\in S_r} w'_{ij} y_j, \quad i \in S_m,$$

where $w'_{ij} = w_j \mathbb{1}(j \in \mathcal{N}_K(i))/\sum_{j'\in\mathcal{N}_K(i)\cap S_r} w_{j'}$ for $j \in S_r$ with $\sum_{j\in S_r} w'_{ij} = 1$. KNN imputation is a locally weighted procedure since the respondents j lying not close enough to unit i with respect to the X-variables are assigned a weight equal to 0; i.e., $w'_{ij} = 0$. The indicator function in the expression of w'_{ij} can be replaced by a one-dimensional continuous kernel smoother $\mathcal{K}_h$, whose role is to control the size of the weight through a tuning parameter h : the units j lying farther from unit i will be assigned a smaller weight than units lying close to it (Hastie et al., 2011).

The imputed estimator under KNN imputation tends to be inefficient when the dimension p of $\mathbf{x}$ is large. Indeed, as p increases, it becomes more difficult to find enough respondents around the point at which we aim to make a prediction. This phenomenon is known as the curse of dimensionality (Hastie et al., 2011, Chap. 1) for a more in-depth discussion ok the KNN procedure. Also, it suffers from a model bias which is of order $(K/n)^{1/p}$. Nearest-neigbour imputation for missing survey data has been considered in Chen and Shao (2000), Beaumont and Bocci (2009) and Yang and Kim (2019).

4.3.4 B-splines and additive model nonparametric regression

Spline regression is a flexible nonparametric method for fitting non-linear functions $f(\cdot)$. It can be viewed as a simple extension of linear models. For simplicity, we start with a univariate X-variable supported on the interval $[0; 1]$. A spline function of order v with κ equidistant interior knots, $0 = \xi_0 < \xi_1 < ... < \xi_\kappa < \xi_{\kappa+1} = 1$, is a piecewise polynomial of degree $v - 1$ between knots and smoothly connected at the knots. These spline functions span a linear space of dimension of $q = v + \kappa$ with a basis function given by the B-splines functions:

$$B_\ell(x) = (\xi_\ell - \xi_{\ell-v}) \sum_{l=0}^{v} (\xi_{\ell-l} - x)_+^{v-1} / \Pi_{r=0, r\neq l}^{v} (\xi_{\ell-l} - \xi_{\ell-r}), \quad \ell = 1, \ldots, q,$$

where $(\xi_{\ell-l} - x)_+^{v-1} = (\xi_{\ell-l} - x)^{v-1}$ if $\xi_{\ell-l} \geq x$ and equal to zero, otherwise; see (Dierckx, 1993, Schumaker, 1981). The B-spline basis is appealing because the basis functions are strictly local: each function $B_\ell(\cdot)$ has the knots $\xi_{\ell-v}, \ldots, \xi_\ell$ with $\xi_r = \xi_{\min(\max(r,0),\kappa+1)}$ for $r = \ell - v, \ldots, \ell$ (Zhou et al., 1998), which means that its support consists of a small, fixed, finite number of intervals between knots. The unknown function $f(\cdot)$ is then approximated by $\widehat{f}(\cdot)$, a linear combination of basis functions $\{B_\ell\}_{\ell=1}^{q}$

with coefficients determined by a least squares criterion computed on the data $(y_i, x_i)_{i \in S_r}$ (Goga et al., 2019). The missing value y_i is then imputed by $\widehat{y}_i = \widehat{f}(x_i)$, where

$$\widehat{f}(x_i) = \sum_{\ell=1}^{q} \widehat{\beta}_\ell B_\ell(x_i) = \mathbf{b}_i^\top \widehat{\boldsymbol{\beta}}, \qquad x_i \in [0;1], \tag{4.17}$$

with $\mathbf{b}_i = (B_\ell(x_i))_{\ell=1}^q$ denoting the vector of B-spline basis functions, and $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_\ell)_{\ell=1}^q$ minimizes

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^q} \sum_{j \in S_r} w_j \left(y_j - \sum_{\ell=1}^{q} \beta_\ell B_\ell(x_j) \right)^2 = \left(\sum_{j \in S_r} w_j \mathbf{b}_j \mathbf{b}_j^\top \right)^{-1} \sum_{j \in S_r} w_j \mathbf{b}_j y_j; \tag{4.18}$$

see Goga et al. (2019). The expression of $\widehat{\boldsymbol{\beta}}$ is similar to that obtained with linear regression imputation given by (4.15) but unlike (4.15), the estimator (4.18) uses the B-spline functions $B_1, \ldots, B_q$, whose number can vary as a function of the number of knots κ and the order v of the B-spline functions. The degree v of the piecewise polynomial does not seem to have a great impact on the model fits if a large enough number of interior knots is used (Ruppert et al., 2003). This is why quadratic or cubic splines are mostly used in practice and an adequate number of interior knots will allow to obtain flexible fits that capture local non-linear trends in the data. Knots are usually placed at the X-quantiles and their number may have a great effect on the model fits: a large value of κ will lead to overfitting, in which case a penalization criterion may be used in (4.18), while a small value of κ may lead to underfitting. Ruppert et al. (2003) give a practical rule for choosing the number κ of interior knots :

$$\kappa = \min\left(\frac{1}{4} \times \text{number of unique } x_i, 35 \right).$$

The imputed value (4.17) with B-spline regression can be also written as a weighted sum of the respondent y-values similar to (4.16), $\widehat{y}_i = \sum_{j \in S_r} w'_{ij} y_j$ for all $i \in S_m$ with weights now given by $w'_{ij} = \mathbf{b}_i^\top \left(\sum_{j' \in S_r} w_{j'} \mathbf{b}_{j'} \mathbf{b}_{j'}^\top \right)^{-1} w_j \mathbf{b}_j$. These weights do not depend on the y-values as in linear regression imputation and $\sum_{j \in S_r} w'_{ij} = 1$ since $\sum_{j=1}^q B_j(x) = 1$ for all $x \in [0;1]$. Unlike linear regression imputation, the weights w'_{ij} are now local due to the B-spline functions ensuring more flexibility to model local nonlinear trends in the data.

We now turn to the multivariate case. For ease of presentation, we confine to the case of two predictors, X_1 and X_2. Additive models provide a simple way to model nonlinear trend in the data (Hastie and Tibshirani, 1986) and extend the standard linear model by allowing non-linear functions between the response variable Y and each of the explanatory variables, while maintaining additivity. In the case of two predictors, the relationship between Y and X_1, X_2 is expressed as a linear combination of unknown smooth functions f_1 and f_2:

$$y_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + \epsilon_i, \tag{4.19}$$

where the ϵ_i's are independent errors with mean equal to zero. The model (4.19) is restricted to be additive and does not account for the potential interactions among the predictors. Accounting for interactions between X_1 and X_2 would require the additional predictor $X_1 X_2$ to be included in the model, leading to

$$y = f_1(x_1) + f_2(x_2) + f_3(x_1, x_2) + \xi,$$

where f_3 is a low-dimensional interaction function fitted by using two-dimensional smoothers, such as local regression or two-dimensional splines. This is beyond the scope of this article. When the number of

predictors is large, the number of potential interactions may be considerable, making the implementation of this procedure challenging. In such situations, random forests and boosting, discussed in sections 4.3.6 and 4.3.7, provide more flexible approaches. But, as pointed out by James et al. (2015), additive models provide a useful compromise between linear and fully nonparametric models.

The unknown functions f_1 and f_2 in (4.19) can be estimated by using two B-spline basis $\mathcal{B}_1 = \{B_{11}, \ldots, B_{1q_1}\}$ and $\mathcal{B}_2 = \{B_{21}, \ldots, B_{2q_2}\}$, which leads to $\widehat{f_1}(x_{i1}) = \sum_{\ell=1}^{q_1} \widehat{\beta}_{1\ell} B_{1\ell}(x_{i1})$ and $\widehat{f_2}(x_{i2}) = \sum_{\ell=1}^{q_2} \widehat{\beta}_{2\ell} B_{2\ell}(x_{i2})$, where $\widehat{\beta}_{1\ell}$ and $\widehat{\beta}_{2\ell}$ are determined, as before, by a least square criterion. To ensure the identifiability of α, additional constraints such as $\sum_{i=1}^{n_r} \widehat{f_1}(x_{i1}) = \sum_{i=1}^{n_r} \widehat{f_2}(x_{i2}) = 0$ are usually imposed. With these constraints, the estimators $(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$ are simply obtained as a regression coefficient estimator, for $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\beta}_{1\ell})_{\ell=1}^{q_1}$ and $\widehat{\boldsymbol{\beta}}_2 = (\widehat{\beta}_{2\ell})_{\ell=1}^{q_2}$. The imputed value for missing y_i is given by

$$\widehat{y}_i = \widehat{\alpha} + \widehat{f_1}(x_{i1}) + \widehat{f_2}(x_{i2}), \quad i \in S_m. \tag{4.20}$$

In practice, a backfitting algorithm is used to compute $f_1(\cdot)$ and $f_2(\cdot)$ iteratively (Hastie et al., 2011). However, when the number p of explanatory variables is large, the algorithm may not converge and additive models tend to breakdown. Finally, random versions of (4.17) and (4.20) are obtained by adding random residuals as in (4.10).

4.3.5 Regression trees

Regression trees through the CART algorithm have been initially suggested by Breiman (1984). Tree-based methods are simple to use in practice for both continuous and categorical variables and useful for interpretation. They form a class of algorithms which recursively split the p-dimensional predictor space, the set of possible values for the X-variables, into distinct and non-overlapping regions of $\mathbb{R}^p$. The prediction $\widehat{f_{tree}}(\mathbf{x}_i)$ at point $\mathbf{x}_i$ corresponds to the average of the respondent y-values falling in the same region as unit i. When the number of X-variables is not too large, the splitting algorithm is quite fast, otherwise it may be time-consuming.

Following Creel and Krotki (2006), we slightly adapt the original CART algorithm as well as the estimation procedure of $f(\cdot)$. The CART algorithm recursively searches for the splitting variable and the splitting position (i.e., the coordinates on the predictor space where to split) leading to the greatest possible reduction in the residual mean of squares before and after splitting.

More specifically, let A be a region or node and let $\#(A)$ the number of units belonging to A. A split in A consists of finding a pair (ℓ, z), where ℓ is the variable coordinates taking value between 1 and p, and z is the position of the split along the ℓth coordinate, within the limits of A. Let C_A be the set of all possible pairs (ℓ, z) in A. The splitting process is performed by searching for the best split (ℓ^*, z^*) in the sense that

$$(\ell^*, z^*) = \underset{(\ell,z)\in C_A}{\arg\max}\, L(\ell, z) \tag{4.21}$$

with

$$L(\ell, z) = \frac{1}{\#(A)} \sum_{i \in S_r} \mathbb{1}(\mathbf{x}_i \in A) \left\{ (y_i - \bar{y}_A)^2 - \left(y_i - \bar{y}_{A_L} \mathbb{1}(X_{i\ell} < z) - \bar{y}_{A_R} \mathbb{1}(X_{i\ell} \geqslant z)\right)^2 \right\}, \tag{4.22}$$

where X_{ij} is the measure of jth variable X_j for the ith individual, $A_L = \{\mathbf{X} \in A; \mathbf{X}_\ell < z\}$, $A_R = \{\mathbf{X} \in A; \mathbf{X}_\ell \geqslant z\}$ and $\mathbf{X}_\ell$ the ℓth coordinate of X; $\bar{y}_A$ is the average of y_i for those units i such that $\mathbf{x}_i \in A$. In (4.21), $\mathbb{1}(\mathbf{x}_i \in A) = 1$ if $\mathbf{x}_i \in A$, and $\mathbb{1}(\mathbf{x}_i \in A) = 0$, otherwise. From (4.21), the best split (ℓ^*, z^*)

is the one that produces a tree with the smallest residuals sum of squares (James et al., 2015, Chap. 8); that is, we seek (ℓ^*, z^*) that minimizes

$$(\ell^*, z^*) = \arg \min_{(\ell,z)\in C_A} \left\{ \sum_{i\in S_r : \mathbf{x}_i \in A} \left(y_i - \bar{y}_{A_L}\right)^2 \mathbb{1}(X_{i\ell} < z) + \sum_{i\in S_r : \mathbf{x}_i \in A} \left(y_i - \bar{y}_{A_R}\right)^2 \mathbb{1}(X_{i\ell} \geqslant z) \right\}.$$

The missing y_i is replaced by $\widehat{y}_i = \widehat{f}_{tree}(\mathbf{x}_i)$, which corresponds to the weighted average of the respondent y-values falling into the same region as $i \in S_m$:

$$\widehat{y}_i = \sum_{j\in S_r} \frac{w_j \mathbb{1}(\mathbf{x}_j \in A(\mathbf{x}_i))}{\sum_{j'\in S_r} w_{j'} \mathbb{1}(\mathbf{x}_{j'} \in A(\mathbf{x}_i))} y_j, \quad i \in S_m, \tag{4.23}$$

where $A(\mathbf{x}_i)$ is the region from $\mathbb{R}^p$ containing the point $\mathbf{x}_i$. With tree-based methods, the imputed value $\widehat{y}_i$ can also be expressed as

$$\widehat{y}_i = \sum_{j\in S_r} w'_{ij} y_j, \quad i \in S_m, \tag{4.24}$$

where $w'_{ij} = w_j \mathbb{1}(\mathbf{x}_j \in A(\mathbf{x}_i))/\sum_{j'\in S_r} w_{j'} \mathbb{1}(\mathbf{x}_{j'} \in A(\mathbf{x}_i))$ with $\sum_{j\in S_r} w'_{ij} = 1$. With regression trees and tree-based methods in general, the non-overlapping A-regions obtained by means of the CART algorithm depend on the respondent data $\{(y_i, \mathbf{x}_i)\}_{i\in S_r}$; i.e., the same set of X-variables with a different set of respondents will lead to different non-overlapping A-regions. The resulting imputed estimator is similar to a post-stratified estimator based on adaptative post-strata.

Regression trees are simple to interpret and often exhibit a small model bias. However, they tend to overfit the data if each A-region contains too few elements. To cope with this issue, regression trees may be pruned, meaning that superfluous splits (with respect to a penalized version of (4.21)) are removed from the tree. Pruning a regression tree tends to reduce its model variance at the expense of increasing the model bias; see Hastie et al. (2011). A random version of (4.24) is obtained by adding random residuals as in (4.10). Bagging and boosting methods may be used to improve the efficiency of tree-based procedures. This is discussed next.

4.3.6 Random forests

Random forest (Breiman, 2001) is an ensemble method which achieves better accuracy than tree-regression methods by creating a large number of different regression trees and combining them to produce more accurate predictions than a single model would. Random forests are especially efficient in complex settings such as small sample sizes, high-dimensional predictor space and complex relationships (Hamza and Larocque (2005), Díaz-Uriarte and de Andrés (2006), among others). Since the article of Breiman (2001), random forests have been extensively used in various fields such as medicine (Fraiwan et al., 2012), time series analysis (Kane et al., 2014), agriculture (Grimm et al., 2008), to cite just a few. Recently, their theoretical properties have been established by Scornet et al. (2015).

There exist a number of random forest algorithms (see Biau and Scornet (2016) for discussion). A widely used algorithm proceeds as follows (Dagdoug et al., 2022b):

Step 1: Consider B bootstrap data sets $D_1, D_2, ..., D_B$, obtained by selecting with replacement n_r pairs $(y_i, \mathbf{x}_i)$ from $D = \{(y_i, \mathbf{x}_i)\}_{i\in S_r}$.

Step 2: In each bootstrap data set D_b for $b = 1, \ldots, B$, fit a regression tree and determine the prediction $\widehat{f}^{(b)}_{tree}$ for the unknown f in (4.8) as described in section 4.3.5. For each regression tree, only p' variables randomly chosen among the p variables are considered in the search for the best split in (4.21).

Step 3: The imputed value for missing y_i is obtained by averaging the predictions at the point $\mathbf{x}_i$ of the B regression tree predictions:

$$\widehat{y}_i = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}_{tree}^{(b)}(\mathbf{x}_i), \quad i \in S_m, \tag{4.25}$$

where $\widehat{f}_{tree}^{(b)}(\mathbf{x}_i)$ is the prediction for the unknown f in (4.8) computed at $\mathbf{x}_i$ and obtained with the bth regression tree as described in Section 4.3.5. More specifically, from (4.23), the prediction $\widehat{f}_{tree}^{(b)}(\mathbf{x}_i)$ corresponds to the weighted average of y-values for $j \in S_r$ falling in the same region $A^{(b)}(\mathbf{x}_i)$ containing $i \in S_m$.

A random version of (4.25) is obtained by adding random residuals as in (4.10). Although random forests are based on fully-grown trees, the accuracy of the predictions is improved by considering bootstrap of units and model aggregation, a procedure called *bagging* and used in statistical learning for reducing the variability. The number B of regression trees should be large enough to ensure a good performance without harming the processing time; see Scornet (2017). The second improvement brought by random forest is the random selection at each split of p' predictors, achieving decorrelated trees. The value of p' is typically chosen as $p' \simeq \sqrt{p}$ (Hastie et al., 2011). In random forest algorithms, a stopping criterion is usually specified so that the algorithm stops once a certain condition (e.g., on the minimum number of units in each final nodes) is met.

4.3.7 Least square tree-boosting and other tree-boosting methods

As in bagging, boosting (Friedman, 2001) is a procedure that can be applied to any statistical learning methods for improving the accuracy of model predictions and is typically used with tree-based methods. While bagging involves the selection of bootstrap samples to create many different predictions, boosting is an iterative method that starts with a weak fit (or learner) and improves it at each step of the algorithm by predicting the residuals of prior models and adding them together to make the final prediction.

To understand how boosting works, consider a regression tree with non-overlapping regions $A_1, \ldots, A_J$, expressed as

$$T(x, \Theta) = \sum_{j=1}^{J} \gamma_j \mathbb{1}(\mathbf{x}_i \in A_j). \tag{4.26}$$

The parameter $\Theta = \{\gamma_j, A_j\}_{j=1}^{J}$ is obtained by minimizing

$$\widehat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{i:\mathbf{x}_i \in A_j} \mathcal{L}(y_i, \gamma_j) = \arg\min_{\Theta} \sum_{i \in S_r} \mathcal{L}(y_i, T(\mathbf{x}_i, \Theta)), \tag{4.27}$$

where $\mathcal{L}$ denotes a loss function; e.g., the quadratic loss function. With the latter, given a region A_j, estimating the constant γ_j is usually straightforward as $\widehat{\gamma}_j = \overline{y}_j$ the average the y-values belonging to A_j. However, finding the regions $\{A_j\}_{j=1}^{J}$ and solving (4.27) in a traditional way may prove challenging and computationally intensive as it requires optimizing over all the parameters jointly. To overcome this difficulty, one may use a greedy top-down recursive partitioning algorithm to find $\{A_j\}_{j=1}^{J}$ as described in Section 4.3.5. Alternatively, one may split the optimization problem (4.27) into many simple

subproblems that can be solved rapidly. Boosting uses the latter and considers that the unknown f has the following additive form:

$$f(x) = \sum_{m=1}^{M} T(x, \Theta_m), \tag{4.28}$$

where $T(x, \Theta_m)$ for $m = 1, \ldots, M$ are trees determined iteratively by using a forward stagewise procedure (Hastie et al., 2011): at each step, a new tree is added to the expansion without modifying the coefficients and parameters of trees already added. Each added tree, usually referred to as a weak-learner, has a small size and slowly improves the estimation of f in areas where it does not perform well. For the quadratic loss function, after accounting for the survey weights, the algorithm becomes:
Step 1: Initialize the algorithm with a constant value: $\widehat{f}_0(\mathbf{x}_i) = 0$ and

$$\widehat{\gamma}_0 = \arg\min_{\gamma \in \mathbb{R}} \sum_{i \in S_r} w_i (y_i - \gamma)^2 = \frac{1}{\sum_{i \in S_r} w_i} \sum_{i \in S_r} w_i y_i.$$

Step 2: For $m = 1$ to M:

(a) Given the current model $\widehat{f}_{m-1}$, fit the regression tree that best predicts the residuals values $y_i - \widehat{f}_{m-1}(\mathbf{x}_i), i \in S_r$ and get the terminal regions $(A_{jm})_{j=1}^{J_m}$.

(b) Given the terminal regions A_{jm}, the optimal constants $\widehat{\gamma}_{jm}$ are found as follows:

$$\widehat{\gamma}_{jm} = \arg\min_{\gamma_{jm}} \sum_{i \in S_r : \mathbf{x}_i \in A_{jm}} w_i \mathcal{L}(y_i, \widehat{f}_{m-1}(\mathbf{x}_i) + \gamma_{jm}) = \arg\min_{\gamma_{jm}} \sum_{i \in S_r : \mathbf{x}_i \in A_{jm}} w_i (y_i - \widehat{f}_{m-1}(\mathbf{x}_i) - \gamma_{jm})^2$$

for $j = 1, \ldots, J_m$.

(c) Update $\widehat{f}_m(\mathbf{x}_i) = \widehat{f}_{m-1}(\mathbf{x}_i) + T(\mathbf{x}_i, \widehat{\Theta}_m)$ where $\widehat{\Theta}_m = \{A_{jm}, \widehat{\gamma}_{jm}\}_{j=1}^{J_m}$ and $T(\mathbf{x}_i, \widehat{\Theta}_m) = \sum_{j=1}^{J_m} \widehat{\gamma}_{jm} \mathbb{1}(\mathbf{x}_i \in A_{jm})$.

Step 3: Output $\widehat{f}_M(\mathbf{x}_i)$ and get the imputed value

$$\widehat{y}_i = \widehat{f}_M(\mathbf{x}_i). \tag{4.29}$$

A random version of (4.29) is obtained by adding random residuals as in (4.10). The number M of trees should not be too large and, for better performances, Hastie et al. (2011) recommend to consider the same number of splits $J_m = J$ at each iteration. The value of J reflects the level of dominant interactions between the X-variables. The value $J = 2$ (one split) produces boosted models with only main effects without interactions, whereas the value $J = 3$ allows for two-variable interactions. Empirical studies suggest that $J = 6$ generally leads to good results. As in ridge regression, shrinkage is used with tree boosting. In this case, Step 2. (c) of the above algorithm is replaced by a penalized version:

$$\widehat{f}_m(\mathbf{x}_i) = \widehat{f}_{m-1}(\mathbf{x}_i) + \nu T(\mathbf{x}_i, \widehat{\Theta}_m),$$

where the parameter $\nu \in (0, 1)$, called learning rate, is used to penalized large trees; usually $\nu = 0.1$ or 0.01. Both M and ν control the performance of the model prediction.

XGBoost

Chen and Guestrin (2016) suggested a scalable end-to-end tree boosting system called XGBoost which is extremely fast. Here, we adapt the algorithm in order to account for the survey weights. Consider again

a tree with formal expression given in (4.26). This tree learning algorithm consists of minimizing the following objective function at the m-th iteration:

$$\widehat{\Theta}_m = \arg\min_{\Theta_m}\{\sum_{i\in S_r} w_i \mathcal{L}(y_i, \widehat{f}_{m-1}(\mathbf{x}_i) + T(\mathbf{x}_i, \Theta_m))\} + \Omega(T(x, \Theta_m)), \tag{4.30}$$

where the penalty function $\Omega(T(x,\Theta_m)) = \gamma J + \frac{\lambda}{2}\sum_{j=1}^{J}\gamma_j^2$ penalizes large trees in order to avoid overfitting. The search problem is optimized by using a second-order Taylor approximation of $\mathcal{L}$, and ignoring the constant term, the new optimization problem reduces to:

$$\widehat{\Theta}_m = \arg\min_{\Theta_m} \sum_{j=1}^{J}\left[\gamma_j \sum_{i\in S_r:\mathbf{x}_i\in A_j} w_i g_i + \frac{1}{2}\gamma_j^2\Big(\sum_{i\in S_r:\mathbf{x}_i\in A_j} w_i h_i + \lambda\Big)\right] + \gamma J, \tag{4.31}$$

where g_i and h_i are the first and second-order derivatives of the loss function computed at $\widehat{f}_{m-1}(\mathbf{x}_i)$. With the quadratic loss function, $g_i = 2(\widehat{f}_{m-1}(\mathbf{x}_i) - y_i)$ and $h_i = 2$. The new objective function from (4.31) is a second-order polynomial with respect to γ_j, so the optimal γ_j is easily obtained as $\gamma_j^* = -(\sum_{i\in S_r:\mathbf{x}_i\in A_j} w_i g_i)/(\sum_{i\in S_r:\mathbf{x}_i\in A_j} w_i h_i + \lambda)$, leading to the optimal value of the objective function as $-(1/2)\sum_{j=1}^{J}(\sum_{i\in S_r:\mathbf{x}_i\in A_j} w_i g_i)^2/(\sum_{i\in S_r:\mathbf{x}_i\in A_j} w_i h_i + \lambda) + \gamma J$. This value is then used next as a decision criterion in a greedy top-down recursive algorithm to find the optimal regions A_j of the m-th tree to be added.

Bayesian additive regression trees (BART)

Bayesian additive regression trees (Chipman et al., 2010, BART) is similar to boosting in the sense that the unknown regression function f has an additive form as in (4.28). While boosting is completely nonparametric, BART makes a Gaussian assumption on the model errors:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right),$$

where $f(x) = \sum_{m=1}^{M} T(x,\Theta_m) = \sum_{m=1}^{M} T_m(x,\Gamma_m)$ is assumed to be a sum of tree functions and $\Gamma_m = \left\{\gamma_j, \gamma_2, \ldots, \gamma_{Jm}\right\}$ is the set of parameter values associated with the J_m terminal nodes in each tree $T(x,\Theta_m)$.

As stated in Chipman et al. (2010), although similar in spirit to gradient boosting, BART differs from boosting algorithms both by the way it weakens the individual trees by relying on a Bayesian framework, but also on how it performs the iterative fitting. More specifically, a prior is specified for the parameters of the model $(T_1,\Gamma_1),(T_2,\Gamma_2),\ldots,(T_m,\Gamma_m)$ and σ^2. The prior of T_m can be decomposed into three components :

1. The probability that a node at depth J is a terminal node is given by $\alpha\,(1+J)^{-\beta}$ for $\alpha \in (0;1)$, $\beta \geq 0$.

2. The distribution on the splitting variable assignments in each interior node is uniform.

3. The distribution of the splitting value conditional on the chosen splitting variable is also uniform.

Borrowing the illustrative example of Chipman et al. (2010), with the parameters $\alpha = 0.95$ and $\beta = 2$, trees with $1, 2, 3, 4, 5$ terminal nodes receive prior probabilities of $0.05, 0.55, 0.28, 0.09$ and 0.03, respectively. Therefore, as in boosting, the BART model tends to favor trees with a small number of terminal nodes. However, the process of restricting the depth of regression trees (or equivalently the number of terminal nodes) in BART is different from the one used in boosting. For boosting, the depth of the trees is fixed by the user and is similar for all trees used in the forest. For BART, the user specifies a

probability for the trees to have a certain number of terminal nodes. As a result, the number of terminal nodes is random rather tan fixed. Therefore, it is likely that trees have only a small number of terminal nodes with the BART model, but this number can vary depending on the data at hand. For γ_j, a conjugate prior is chosen to make computations simpler; e.g., $p(\gamma_{jm}|T_m)$ is assumed to be $\mathcal{N}(\gamma_\gamma, \sigma^2_\gamma)$. Similarly, a conjugate prior is chosen for σ^2, e.g., the inverse chi-square distribution. To generate the posterior distribution, the authors suggest the use of a Gibbs sampler. For general guidelines about the choices of these parameters, see Chipman et al. (2010). The imputed value for missing y_i is obtained as with the general boosting algorithm given in Section 4.3.7, where the prediction of each regression tree is the weighted average of the values in the terminal node containing $\mathbf{x}_i$.

4.3.8 Cubist algorithm

Cubist is an updated implementation of the M5 algorithm introduced by Quinlan et al. (1992) and Quinlan (1993). It is an algorithm based on regression trees and linear models, among other ingredients. Initially, Cubist was only available under a commercial license. In 2011, the code was released as open-source. The algorithm proceeds as follows (Kuhn and Johnson, 2013, Chap. 8):

Step 1: Create a partition $\mathcal{P} = \{A_1, A_2, ..., A_T\}$ of $\mathbb{R}^p$. To do so, let C_A be the set of all possible splits in a node A of cardinality ℓ, that is, the set of all possible pairs (position, variable). Then, the split is performed using the following criterion:

$$L'(z,j) = \underset{(z,j)\in C_A}{\arg\max} \sqrt{\sum_{i\in S_r}\left(y_i - \left(\frac{1}{n_r}\sum_{j'\in S_r} y_{j'}\right)\right)^2} - \sum_{h=1}^{\ell}\frac{n_h}{n_r}\sqrt{\sum_{i:\mathbf{x}_i\in D_h}\left(y_i - \left(\frac{1}{n_r}\sum_{j':\mathbf{x}_i\in D_h} y_{j'}\right)\right)^2},$$

where $D_1, \ldots, D_\ell$ denote the ℓ non-terminal nodes after each of the $\ell - 1$ previous splits and n_h denotes the cardinal of elements in the node D_h.

Step 2: In each node, a linear model is fitted between the survey variable Y and the auxiliary variables that have been used to split the tree. More specifically, consider the jth terminal node A_j. Then, there exists a path from the first node to the current node A_j in the graph formed by the tree. This path uses p'_j variables among the set $\{X_1, X_2, ..., X_p\}$. For instance, assume that a partition of 5 elements is created by the tree shown in Figure 22. Then, the linear model in the node A_1 is fitted using the variables that created the path in red, that is, X_1, X_4 and X_6, and so $p'_1 = 3$ for this node. The linear model fitted in the node A_4 uses only one variable, X_1, (the green path), so $p'_4 = 1$. The coefficients $\boldsymbol{\beta}_j \in \mathbb{R}^{p'_j}$ of the linear model in the node A_j are estimated using the customary weighted least squares criterion:

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}_j\in\mathbb{R}^{p'_j}}{\arg\min} \sum_{i\in S_r} w_i \left\{y_i - \boldsymbol{\beta}_j^\top \mathbf{x}_i^{(j)}\right\}^2 \mathbb{1}\left(\mathbf{x}_i \in A_j\right),$$

where $\mathbf{x}_i^{(j)}$ is the vector containing the measurements of the p'_j variables for unit i.

Step 3: In each node, a backward elimination procedure is performed using the adjusted error rate (AER) criterion. For instance, in the jth terminal node, we have

$$AER(A_j) = \frac{\#(A_j) + p^*}{\#(A_j) - p^*} \sum_{i\in S_r:\mathbf{x}_i\in A_j} |y_i - \widehat{y}_i|,$$

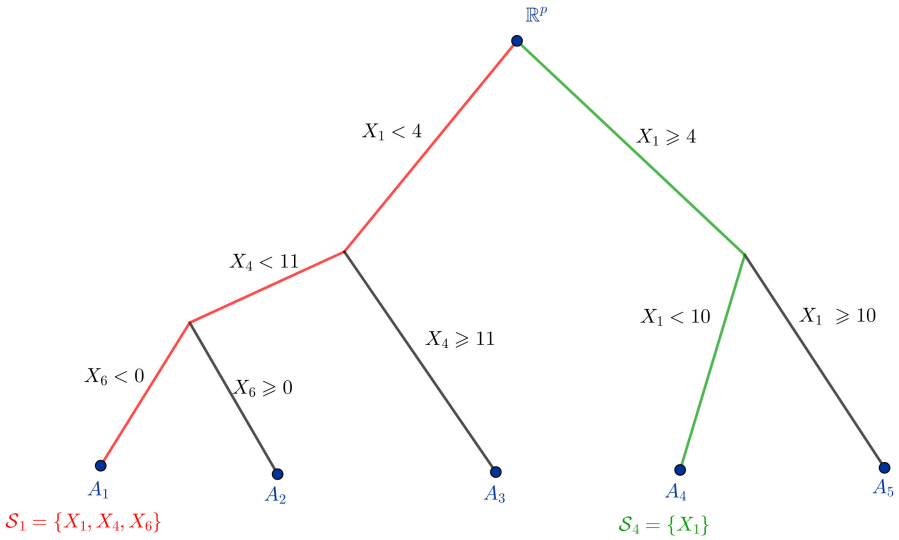


Figure 22: Example of a graph induced by a tree algorithm.

where p^* denotes the number of variables used in the current model which predicts $\widehat{y}_i$ for a prediction at the point $\mathbf{x}_i$. Each variable in the initial model is dropped and the AER is recomputed. Terms are dropped from the model as long as the AER decreases.

Step 4: Once the tree is fully grown, it is pruned by removing unnecessary splits. Starting at the terminal nodes, the AER is computed with and without the node. Whenever the node does not result in a decrease of the AER, it is pruned. This process is performed until no more node can be removed.

Step 5: To avoid over-fitting, a smoothing procedure is performed. Let $\widehat{y}_{i(j)}$ be the predicted value obtained by fitting the linear model in the jth child node and $\widehat{y}_{i(p)}$ be the predicted value obtained from the direct parent node. These predictions are combined as

$$\widehat{y}_i = a y_{i(j)} + (1 - a)\widehat{y}_{i(p)},$$

where

$$a = \frac{\widehat{V}(\mathbf{e}_{(p)}) - \widehat{Cov}(\mathbf{e}_{(j)}, \mathbf{e}_{(p)})}{\widehat{V}(\mathbf{e}_{(j)} - \mathbf{e}_{(p)})}$$

with $e_{i(j)} = y_i - \widehat{y}_{i(j)}$ denoting the ith coordinate of the vector $\mathbf{e}_{(j)}$, $e_{i(p)} = y_i - \widehat{y}_{i(p)}$ denoting the ith coordinate of the vector $\mathbf{e}_{(p)}$ and $\widehat{V}(\cdot)$ and $\widehat{Cov}(\cdot, \cdot)$ denoting the empirical model variance and covariance, respectively.

Step 6: Cubist can be used as an ensemble model. Once the Cubist algorithm is fitted, the subsequent iterations of the algorithm use the previously trained algorithm to define an adjusted response $y_i^{(m)}$ so that the next iteration of the algorithm uses

$$y_i^{(m)} = y_i - (y_i^{(m-1)} - y_i),$$

where $y_i^{(m)}$ is the value of the adjusted response y_i for the mth iteration of the Cubist algorithm.

Step 7: The final imputed value for missing y_i is derived using a K nearest-neighbour rule:

$$\widehat{y}_i = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{0.5+d_k}(t_k + \widehat{y}^{(k)} - \widehat{t}_k), \tag{4.32}$$

where d_k denotes the distance between $\mathbf{x}_i$ and the kth neighbor, t_k denotes the outcome of the kth neighbor and $\widehat{t}_k$ its predicted value.

A random version version of (4.32) is obtained by adding random residuals as in (4.10).

4.3.9 Support vector regression

Support vector machines (Cortes and Vapnik, 1995, Smola and Schölkopf, 2004, Vapnik, 1998, 2000) belong to the class of supervised learning algorithms and may be used for regression analysis. We start by considering the linear regression model

$$f(\mathbf{x}_i) = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}, \quad \beta_0 \in \mathbb{R}, \quad \boldsymbol{\beta} \in \mathbb{R}^p,$$

before discussing the case of nonlinear relationships. In the customary regression framework, the goal is to minimize the residuals sum of squares. In Support Vector Regression (SVR), the goal is to minimize a function of the residuals plus a L^2-penalization on the regression coefficient:

$$\mathcal{S} = \sum_{i\in S_r} V_\epsilon(y_i - f(\mathbf{x}_i)) + \frac{\lambda}{2}||\boldsymbol{\beta}||^2, \tag{4.33}$$

where V_ϵ is the so-called ϵ-insensitive error measure defined as $V_\epsilon(x) = 0$ if $|x| < \epsilon$ and $|x| - \epsilon$ otherwise (Vapnik, 2000) for $\epsilon > 0$; ε can be viewed as the allowed tolerance for fitting; see Figure 1 in Smola and Schölkopf (2004). The optimization problem (4.33) may not have solution and supplementary tolerances ξ_i, ξ_i^* (called also "the slack variables") on the individual fitted errors are considered (Smola and Schölkopf, 2004). There exist several ways for incorporating weights in the optimization problem, leading to different weighted support vector regression solutions. We consider the method suggested by Lee et al. (2005) and Han and Clemmensen (2014):

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{2}||\boldsymbol{\beta}||^2 + C\sum_{i\in S_r}\widetilde{w}_i\left(\xi_i + \xi_i^*\right) \tag{4.34}$$

and

$$\begin{aligned} \text{subject to} \quad & y_i - \beta_0 - \mathbf{x}_i^T\boldsymbol{\beta} \leqslant \epsilon + \xi_i, \\ & \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} - y_i \leqslant \epsilon + \xi_i^*. \\ & \xi_i, \xi_i^* > 0, \end{aligned} \tag{4.35}$$

where $C > 0$ is the tuning parameter that provides a trade-off between the smoothness of the fitted function and the deviation from the training data and $\widetilde{w}_i = w_i/\sum_{j\in S_r} w_j \in (0,1)$ denotes the normalized sampling weight associated with unit i. As a result, the $\widetilde{w}_i$'s are all smaller than one. As argued by Han and Clemmensen (2014), incorporating weights in the objective function as in (4.34) has the effect of shrinking the estimators $\widehat{\beta}_j$ to different extents. The solution of (4.33) and (4.35) is given by $\widehat{\boldsymbol{\beta}} = \sum_{i\in S_r}\left(\widehat{\alpha}_i - \widehat{\alpha}_i^*\right)\mathbf{x}_i$, which leads to

$$\widehat{f}(\mathbf{x}) = \sum_{i\in S_r}\left(\widehat{\alpha}_i - \widehat{\alpha}_i^*\right) < \mathbf{x}_i, \mathbf{x} > + \beta_0, \tag{4.36}$$

where $< \cdot, \cdot >$ is an inner product and $\widehat{\alpha}_i > 0$ and $\widehat{\alpha}_i^* > 0$ denote the Lagrange multipliers verifying the quadratic programming problem:

$$\min_{\alpha_i, \alpha_i^*} \epsilon \sum_{i \in S_r} (\alpha_i + \alpha_i^*) - \sum_{i \in S_r} y_i (\alpha_i - \alpha_i^*) + \frac{1}{2} \sum_{i,j \in S_r} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) < \mathbf{x}_i, \mathbf{x}_j >$$

subject to $0 \leq \alpha_i, \alpha_i^* \leq C_i := C \times \widetilde{w}_i$, $\sum_{i \in S_r} (\alpha_i - \alpha_i^*) = 0$ and $\alpha_i \alpha_i^* = 0$. As a result, only a subset of the solution values $(\widehat{\alpha}_i - \widehat{\alpha}_i^*)$ are nonzero and the associated data values are called the support vectors. The solution $\widehat{\boldsymbol{\beta}}$ is written as a linear combination of these support vectors. Moreover, the prediction $\widehat{f}(\mathbf{x})$ uses only the support vectors and the inner products between $\mathbf{x}$ and $\mathbf{x}_i$ without requiring the computation of $\widehat{\boldsymbol{\beta}}$. This property is useful for extending the method to handle nonlinear relationships.

We now consider the case of a nonlinear and unknown function f. We approximate f in a basis of functions $\{\phi_m\}_{m=1}^M$ as follows:

$$f(x) = \sum_{m=1}^{M} \beta_m \phi_m(x) + \beta_0$$

and β_0 and $\boldsymbol{\beta} = (\beta_m)_{m=1}^M$ minimize (4.34) and

$$\begin{aligned} \text{subject to} \quad & y_i - \beta_0 - \sum_{m=1}^{M} \beta_m \phi_m(x_i) \ \leqslant \epsilon + \xi_i, \\ & \beta_0 + \sum_{m=1}^{M} \beta_m \phi_m(x_i) - y_i \ \leqslant \epsilon + \xi_i^*. \\ & \xi_i, \xi_i^* > 0. \end{aligned} \tag{4.37}$$

A similar derivation as before leads to $\widehat{\boldsymbol{\beta}} = \sum_{i \in S_r} \left(\widehat{\alpha}_i - \widehat{\alpha}_i^*\right) \phi(\mathbf{x}_i)$ for $\phi(\mathbf{x}_i) = (\phi_m(\mathbf{x}_i))_{m=1}^M$ and

$$\widehat{f}(\mathbf{x}) = \sum_{i \in S_r} \left(\widehat{\alpha}_i - \widehat{\alpha}_i^*\right) \mathcal{K}(\mathbf{x}_i, \mathbf{x}) + \beta_0,$$

where $\mathcal{K}(\mathbf{x}_i, \mathbf{x}) =< \phi(\mathbf{x}_i), \phi(\mathbf{x}) >= \sum_{m=1}^M \phi_m(\mathbf{x}_i)\phi_m(\mathbf{x})$ is a positive definite kernel (Smola and Schölkopf, 2004). The computation of $\widehat{f}(\mathbf{x})$ involves $\phi(\mathbf{x})$ only through inner products and using a kernel function makes the computation of $\widehat{f}(\mathbf{x})$ possible without requiring $\phi(\mathbf{x})$. All is needed is the knowledge of $\mathcal{K}$. Using $\mathcal{K}$, it is possible to solve the optimization problem in a higher-dimensional space without having to compute any product in this space. Common choices of $\mathcal{K}(\cdot, \cdot)$ include the Gaussian kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-||\mathbf{x}_i - \mathbf{x}_j||^2\right)$ and the polynomial kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \left(1 + \mathbf{x}_i^\top \mathbf{x}_j\right)^q$, $\quad q = 2, 3, \ldots$. The imputed value for the missing y_i is given by

$$\widehat{y}_i = \sum_{j \in S_r} \left(\widehat{\alpha}_j - \widehat{\alpha}_j^*\right) \mathcal{K}(\mathbf{x}_j, \mathbf{x}_i) + \widehat{\beta}_0. \tag{4.38}$$

A random version version of (4.38) is obtained by adding random residuals as in (4.10). The reader is referred to Smola and Schölkopf (2004) for a discussion on how to estimate β_0.

4.4 Simulation study: the case of population totals

We conducted an extensive simulation study to investigate the performance of the imputation procedures described in Section 4.3 in terms of bias and efficiency.

4.4.1 The setup

For each scenario, we repeated $R = 5,000$ iterations of the following process:

(i) A finite population of size $N = 10,000$ was generated. The population consisted of a survey variable Y and a set of predictors $X_1, \ldots, X_p$.

(ii) From the finite population generated in Step (i), a sample, of size n, was selected according to a given probability sampling design.

(iii) In each sample, nonresponse to item Y was generated according to a given nonresponse mechanism.

(iv) The missing values in each sample were imputed using several imputation procedures.

We now give a more in-depth discussion of each of the steps (i)-(iv).

We first generated five predictors $X_1, \ldots, X_5$, according to the following distributions: X_1 followed a normal distribution, $X_1 \sim \mathcal{N}(0,1)$; X_2 followed a Beta distribution, $X_2 \sim \text{Beta}(3,1)$; X_3 followed a Gamma distribution, $X_3 \sim 2 \times \text{Gamma}(3,2)$; X_4 followed a Bernoulli distribution, $X_4 \sim \mathcal{B}(0.7)$; and X_5 followed a multinomial distribution, $X_5 \sim \text{Mult}(0.4, 0.3, 0.3)$. The predictors X_1-X_3 were continuous, whereas the predictors X_4 and X_5 were discrete. The predictors X_1-X_3 were standardized so as to have a zero mean and a variance equal to one. Given the predictors X_1-X_5, we generated the continuous survey variables $Y_1, \ldots, Y_8$, according to the following models:

- $Y_1 = 2 + 2X_1 + X_2 + 2X_3 + \mathcal{N}(0,1)$;
- $Y_2 = 2 + 2X_1 + X_2 + 2X_3 + \text{Pareto}(1,4)$;
- $Y_3 = 2 + X_1 + X_2^2 + X_3 + \mathcal{N}(0,1)$;
- $Y_4 = 2 + 2X_1 + X_2 + 3X_3X_4 + 1.5\mathbb{1}(X_5 = 1) - 2\mathbb{1}(X_5 = 2) + \mathcal{N}(0,1)$;
- $Y_5 = 2 + 5X_1^3 + 4X_2^2 + X_3X_4 + 1.5\mathbb{1}(X_5 = 1) - 2\mathbb{1}(X_5 = 2) + \mathcal{N}(0,1)$;
- $Y_6 = 2 + (2X_1 + X_2 + 2X_3)^2 + \mathcal{N}(0,1) + \text{Beta}(3,1)$;
- $Y_7 = 2 + (2X_1 + X_2 + 3X_3X_4 + 1.5\mathbb{1}(X_5 = 1) - 2\mathbb{1}(X_5 = 2))^2 + \mathcal{N}(0,1)$;
- $Y_8 = 4\cos(X_1) + \mathcal{N}(0,1)$;

and the binary survey variables as follows:

- $Y_9 = \mathbb{1}(S_1 > 1/2)$, where

$$S_1 = 0.1 + 0.79 \exp\{1 + 0.5\,(0.75 + 2X1 + 2X_2 + 2X_3 - X_4 - X_3X_4 + 1.5\mathbb{1}(X_5 = 1) - 2\mathbb{1}(X_5 = 2))\}^{-1};$$

- $Y_{10} = \mathbb{1}(S_2 > 1/2)$, where

$$S_2 = 0.55 \times Q + 0.02 - 0.01X_2^3$$

with

$$Q = \exp\{1 + 0.4 \times (6.5 + 2X_1 + 2X_2 + 2X_3 - X_4 - X_3X_4 + 1.5\mathbb{1}(X_5 = 1) - 2\mathbb{1}(X_5 = 2))\}^{-1}. \quad (4.39)$$

For the survey variables Y_2 and Y_6, note that we have generated errors for non-normal distribution to assess the robustness of the BART procedure that assumes a Gaussian distribution for the errors.

From each population, we selected samples, of (expected) size $n = 1,000$, according to two sampling designs: (a) simple random sampling without replacement and (b) Poisson sampling with probability proportional to the values of the variable X_5; i.e., $\pi_i = 1,000 \times (x_{5i} / \sum_{i \in U} x_{5i})$ for all $i \in U$. Simple random sampling without replacement was used for estimating the finite population total of the continuous survey variables Y_1-Y_6 and Y_8 and the binary variables Y_9 and Y_{10}, whereas Poisson sampling was used for estimating the totals of the survey variables Y_4 and Y_7.

In each sample, nonresponse to the survey variable Y_ℓ, $\ell = 1, \ldots, 10$, was generated according to four nonresponse mechanisms. That is, the response indicators r_i were generated from a Bernoulli distribution with probability p_{gi}, $g = 1, \ldots, 4$, where

$$
\begin{aligned}
\text{(NR1): } p_{1i} &= 0.1 + 0.79 \exp\{1 + 0.5\,(0.75 + 2x_{i1} + 2x_{i2} \\
&\qquad + 2x_{i3} - x_{i4} - x_{i3}x_{i4} + 1.5\mathbb{1}(x_{i5} = 1) - 2\mathbb{1}(x_{i5} = 2))\}^{-1}; \\
\text{(NR2): } p_{2i} &= 0.5; \\
\text{(NR3): } p_{3i} &= 0.55 \times q_i + 0.02 - 0.01 x_{i2}^3; \\
\text{(NR4): } p_{4i} &= 0.5 \times q_i + 0.13 - 0.1\,(\sin(x_{i1}) + \cos(x_{i2}));
\end{aligned}
$$

where q_i is the ith value of Q given by (4.39). In (NR1)-(NR4), the model parameters were set so as to obtain a response rate of about 50% in each sample.

In each sample, the missing values were imputed according to eleven imputation procedures described in section 4.3. Some of the imputation procedures required the specification of some parameters (e.g., regularization parameter, depth of a regression tree, choice of a kernel, etc.). We have included several configurations to assess the impact of these parameters on the performance of these procedures. Based on the different configurations, we ended up with twenty-seven imputation procedures. More specifically, we included the following procedures:

Procedure 1: "LR" : Deterministic linear regression imputation; see Section 4.3.1.

Procedure 2: "MWCα" : Mean imputation within classes, where the number of units in each class was set to $\alpha \in \{50, 100, 250, 500\}$; see Section 4.3.2.

Procedure 3: "HDWCα" : Random hot-deck imputation within classes, where the number of units in each class was set to $\alpha \in \{50, 100, 250\}$; see Section 4.3.2.

Procedure 4: "KNN" : K-Nearest-Neighbours imputation with $K = 1$ and $K = 5$ nearest neighbours and the euclidian distance and implemented with the R-package `caret`; see Section 4.3.3.

Procedure 5: "AMSα" : Additive models based on cubic B-splines with α equidistant interiors knots placed at the x-quantiles, where $\alpha \in \{5, 10\}$ and implemented with the R-package `mgcv`; see Section 4.3.4.

Procedure 6: "CART" : Imputation through regression trees with the CART algorithm and implemented with the R-package `rpart`; see Section 4.3.5.

Procedure 7: "RF1" : Imputation through random forest with $B = 1000$ trees, one observation per terminal node and 1 predictor considered for the search in each split. "RF2": Random forest with $B = 1000$ trees, 5 observations per terminal node and $\sqrt{p}$ predictors considered for each split, where p is the number of X-variables used in the imputation model, in our case $p = 5$. "RF3" : Random forest with $B = 1000$ trees, 10 observations per terminal node and $\sqrt{p}$ predictors considered for each split. Simulations were implemented with the R-package `ranger`; see Section 4.3.6.

Procedure 8: "XGB1": XGBoost algorithm with $M = 50$ trees each one with $J = 3$ final splits and a learning rate of 0.1. "XGB2": XGBoost algorithm with $M = 100$ trees with $J = 6$ and a learning rate of 0.05. "XGB3": XGBoost algorithm with $M = 250$ trees with $J = 10$ and a learning rate of 0.01. Simulations were implemented with the *R*-package `xgboost`; see Section 4.3.7.

Procedure 9: "BART" : Imputation through Bayesian additive regression trees. Simulations were implemented with the *R*-package `bartMachine`; see Section 4.3.7.

Procedure 10: "CUBIST1": Cubist with one model. "CUBIST2" : Cubist with five models. "CUBIST3" : Cubist with 5 models and unbiased estimation. Simulations were implemented with the *R*-package `Cubist`; see Section 4.3.8.

Procedure 11: "SVR1": Support vector regression imputation with a Gaussian kernel and the ν objective function. "SVR2": Support vector regression imputation with a polynomial kernel of degree 3 and the ϵ-insensitive objective function. "SVR3": Support vector regression imputation with a Gaussian kernel and the ϵ-insensitive objective function. "SVR4": Support vector regression imputation with a linear kernel and the ϵ-insensitive objective function. Simulations were implemented with the *R*-package `e1071`; see Section 4.3.9.

The imputation procedures used in our simulations were based on an imputation model that included the predictors $X_1, \ldots, X_5$, without any interaction terms. Except for random hot-deck imputation (Procedure 3) and nearest-neighbour imputation (Procedure 4 with $K = 1$), for the binary variables Y_9 and Y_{10}, note that we have generated zeroes and ones from independent Bernoulli distributions with parameter $\widehat{y}_i$, where $\widehat{y}_i$ denotes the predicted value associated with unit i. Whenever $\widehat{y}_i < 0$, we set it to $\widehat{y}_i = 0$. Similarly, when $\widehat{y}_i > 1$, we set it to $\widehat{y}_i = 1$.

As a measure of bias of the imputed estimator $\widehat{t}_{imp}$ given by (4.4), we computed the Monte Carlo percent relative bias defined as

$$RB_{MC}(\widehat{t}_{imp}) = 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{(\widehat{t}_{imp}^{(r)} - t_y)}{t_y}, \tag{4.40}$$

where $\widehat{t}_{imp}^{(r)}$ denotes the imputed estimator $\widehat{t}_{imp}$ at the rth iteration, $r = 1, \ldots, 5,000$.

As a measure of efficiency, we computed the relative of efficiency, using the complete data estimator $\widehat{t}_{\pi}$ given by (4.1), as the reference. That is,

$$RE_{MC}(\widehat{t}_{imp}) = 100 \times \frac{MSE_{MC}(\widehat{t}_{imp})}{MSE_{MC}(\widehat{t}_{\pi})}, \tag{4.41}$$

where $MSE_{MC}(\widehat{t}_{imp}) = R^{-1} \sum_{r=1}^{R} (\widehat{t}_{imp}^{(r)} - t_y)^2$ and $MSE_{MC}(\widehat{t}_{\pi})$ is defined similarly.

4.4.2 Simulation results

In Section 4.4.2, we discuss the simulation results pertaining to the continuous survey variables $Y_1, \ldots, Y_6$ and Y_8, with simple random sampling without replacement. The results for Poisson sampling used in the case of Y_4 and Y_7 are discussed in Section 4.4.2. Finally, the case of the binary variables Y_9 and Y_{10}, whose totals were estimated with simple random sampling without replacement, is discussed in Section 4.4.2.

Continuous survey variables and simple random sampling without replacement

For simple random sampling without replacement, for each of the twenty-seven imputation procedures, we had seven survey variables and four nonresponse mechanisms, leading to $27 \times 4 \times 27 = 756$ sets of simulation results. For ease of presentation, we present the results in tabular and graphic forms. The displayed statistical analyses were obtained from $4 \times 7 = 28$ scenarios obtained by crossing all the nonresponse models and the survey variables.

For each imputation procedure, Table 6 and Table 7 display, respectively, some descriptive statistics regarding the Monte Carlo absolute percent relative bias (absolute value of RB) and the Monte Carlo relative efficiency (RE) of $\widehat{t}_{imp}$ calculated across the twenty-eight scenarios. The corresponding side-by-side boxplots obtained from the twenty-eight scenarios are given in Figures 23 and 24. In Tables 6 and 7, the imputation procedures are ordered from the best to the worst with respect to the median absolute percent RB (the median of the twenty-eight values of absolute RB) and the median percent RE (the median of the twenty-eight values of RE), respectively. Figure 25 shows the distribution of the imputed estimator for the best ten imputation procedures in terms of RE. Finally, Table 8 displays the best five imputation procedures for each Y-variable.

From Table 6 and Table 7, among the twenty-seven imputation procedures, the best methods were: CUBIST, XGboost, AMS and BART. The performance of CUBIST3 was especially impressive with a median RE of 115%, a value of Q_{95} equal to 158% and a maximum value of 211%. The methods XGboost, AMS and BART exhibited similar performances with values of median RE ranging from 122% and 129%. However, for some scenarios, these methods did not perform well. For instance, the procedure XGB2 showed a value of max RE of about 438%, whereas it was equal to 1728% for AM5. Results suggest that additive models with 5 interiors knots perform better than those with 10 interior knots. The next group of imputation procedures includes SVR and RF, with values of median RE ranging from 141% and 151%. Again, for some scenarios, both methods displayed poor performances with values of max RE ranging from 322% to 1138%. The procedure CART was less efficient than RF2 and RF3. The procedure 1-NN did relatively well with a median RE equal to 194%. On the other hand, the procedure 5-NN was rather inefficient with a median RE of 229%, which suggests that KNN with survey data works well only with a small number of neighbour. Turning to mean and random hot-deck imputation within classes, the score method was outperformed by the aforementioned procedures. Among the different versions of MCW and HDWC, the procedure MWC50 (which corresponds to 20 classes) led to the best results. This is consistent with the results of Haziza and Beaumont (2007). As expected, the procedure HDWC50 was less efficient than MWC50 as random hot-deck imputation suffers from the imputation variance, arising from the random selection of donors within classes. Finally, for some scenarios, it is worth noting that some of the procedures were better than the complete data estimator. For instance, for SVR4, the minimum value of RE and the value of $Q_{0.05}$ were respectively equal to 82% and 89%, respectively (see Table 7). Finally, the results in Table 5 suggest that the best methods were CUBIST, XGBoost, additive models and BART, which is consistent with the discussion above.

For each of the best ten imputation procedures displayed Table 7, Figure 26 displays the distribution of $\widehat{t}_{imp}$ for each nonresponse mechanism. Figure 26 suggests that the nonresponse mechanism may have a considerable impact on the behavior of the imputed estimator. For instance, in our experiments, we note that most of the imputation procedures performed poorly in the case of the nonresponse mechanism (NR1). Notable exceptions were AMS5, BART and Cubist3. In particular, Cubist3 seemed to be insensitive to the nonresponse mechanism, which is a desirable feature.

Ranking	Model	Min	$Q_{0.05}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.95}$	Max
1	CUBIST3	0.0	0.0	0.0	0.1	0.9	2.8	3.5
2	AMS5	0.0	0.0	0.0	0.1	1.8	7.7	13.8
3	AMS10	0.0	0.0	0.0	0.1	1.8	7.6	13.5
4	CUBIST1	0.0	0.0	0.1	0.5	3.4	7.5	7.5
5	XGB1	0.0	0.0	0.2	0.6	1.8	4.2	5.4
6	MWC50	0.0	0.0	0.1	0.6	2.7	8.3	11.7
7	HDWC50	0.0	0.0	0.1	0.6	2.7	8.3	11.8
8	CUBIST2	0.0	0.0	0.1	0.6	3.6	7.5	7.5
9	BART	0.0	0.1	0.4	0.8	2.2	4.0	4.6
10	XGB2	0.1	0.2	0.4	0.9	2.8	5.4	10.1
11	LR	0.0	0.0	0.1	0.9	3.8	12.8	20.4
12	SVR3	0.1	0.1	0.4	1.0	3.2	7.1	13.5
13	MWC100	0.0	0.0	0.3	1.0	3.6	10.1	12.9
14	HDWC100	0.0	0.0	0.3	1.0	3.6	10.1	12.9
15	SVR1	0.0	0.1	0.4	1.2	3.4	7.4	14.0
16	RF3	0.0	0.2	0.5	1.3	3.8	16.6	20.7
17	RF2	0.0	0.1	0.4	1.4	4	15.6	18.6
18	MWC250	0.0	0.0	0.7	1.7	4.9	14.6	18.1
19	HDWC250	0.0	0.0	0.6	1.7	4.9	14.6	18.1
20	RF1	0.1	0.2	0.9	1.7	7.7	32.1	39.5
21	NN	0.0	0.1	1.0	2.1	5.2	8.0	9.4
22	MWC500	0.0	0.0	0.7	2.2	7.2	25.5	30.6
23	CART	0.0	0.1	0.1	2.4	4.9	17.4	28.0
24	X5NN	0.0	0.2	1.5	3	7.3	12.0	13.7
25	SVR2	0.1	0.2	1.0	3.7	11.7	19.9	27.0
26	XGB3	0.6	1.5	3.1	4.3	5.0	9.5	10.3
27	SVR4	0.0	0.0	2.4	5.3	7.8	22.2	33.3

Table 6: Monte Carlo percent absolute relative bias of the imputed estimator: Descriptive statistics over all the scenarios

Ranking	Model	Min	$Q_{0.05}$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$Q_{0.95}$	Max
1	CUBIST3	102	102	111	115	125	158	211
2	BART	113	113	116	122	131	154	204
3	AMS5	100	101	111	123	147	378	1728
4	AMS10	100	101	112	123	167	1195	1749
5	XGB1	101	103	115	129	153	203	288
6	CUBIST2	102	103	119	133	187	360	365
7	XGB2	102	102	117	133	166	316	438
8	CUBIST1	103	105	120	136	182	360	365
9	SVR1	94	103	122	141	180	284	322
10	SVR3	95	106	122	143	181	269	299
11	RF3	115	118	131	149	192	919	1138
12	RF2	113	118	130	151	202	824	1025
13	CART	125	134	143	168	248	1498	2683
14	LR	110	111	114	169	315	823	3494
15	MWC50	113	114	122	171	205	308	583
16	HDWC50	120	120	128	189	240	332	600
17	MWC100	116	116	136	191	217	296	670
18	NN	101	111	125	194	378	486	526
19	XGB3	92	100	128	194	663	1082	1104
20	HDWC100	123	125	142	213	246	322	686
21	RF1	136	137	149	223	375	3656	3916
22	MWC250	128	130	159	229	279	383	1162
23	5NN	94	108	123	229	659	775	855
24	SVR2	97	102	151	242	1616	3849	6355
25	SVR4	82	89	117	258	1439	4301	8675
26	HDWC250	141	143	185	265	325	411	1184
27	MWC500	151	155	202	269	336	1783	3021

Table 7: Monte Carlo percent absolute relative efficiency of the imputed estimator: Descriptive statistics over all the scenarios

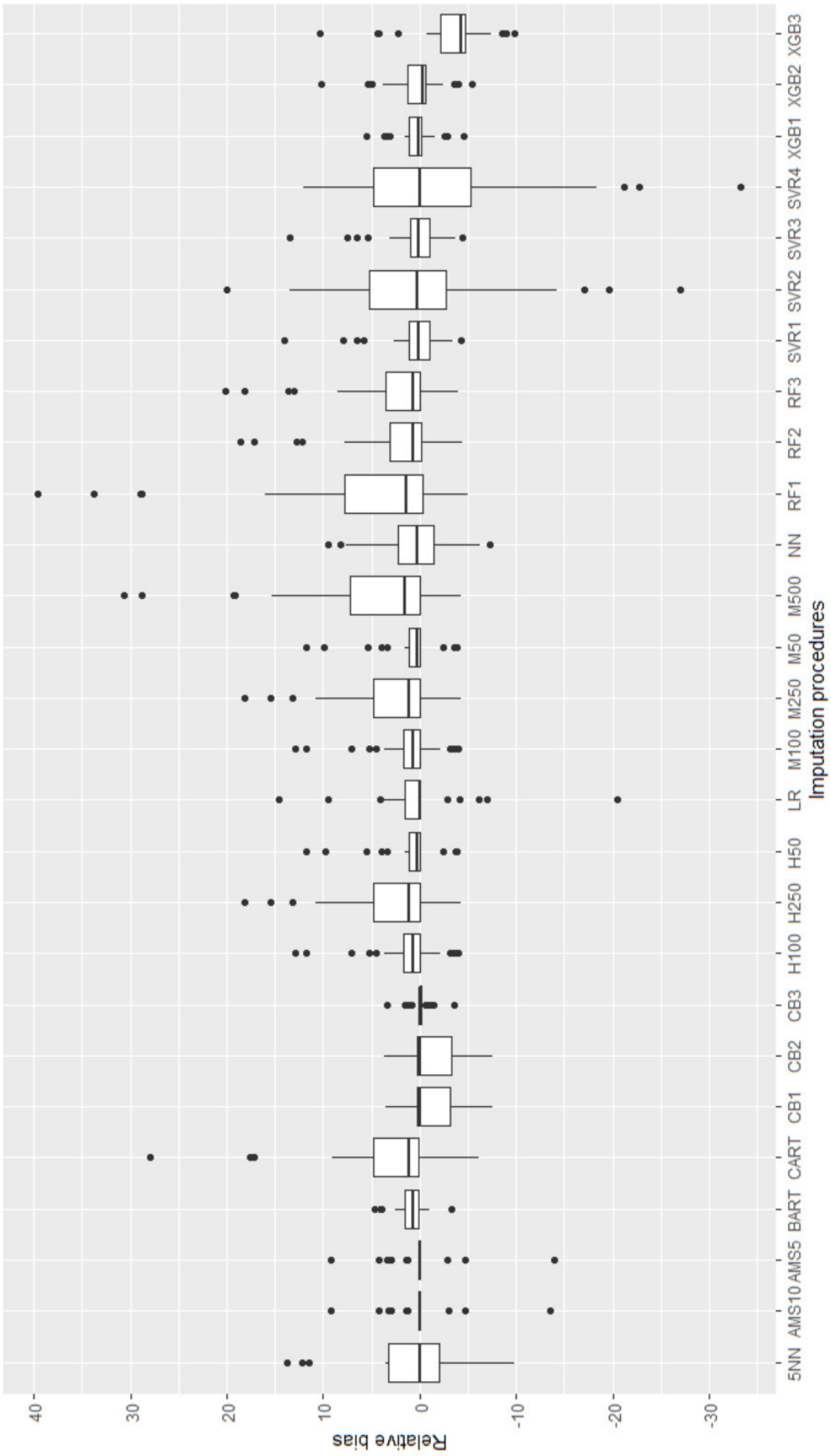


Figure 23: Monte Carlo percent relative bias across the scenarios.

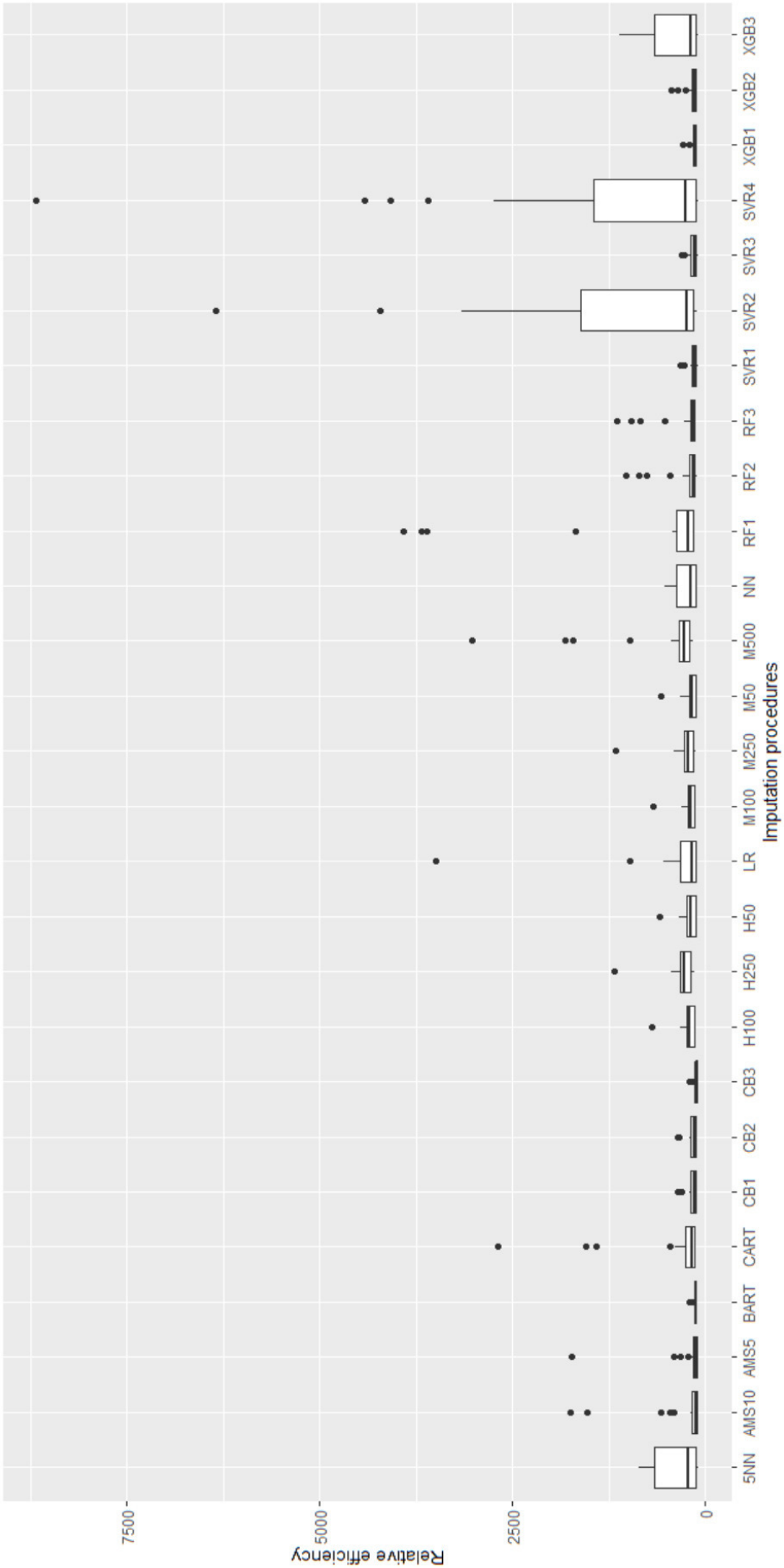


Figure 24: Monte Carlo percent relative efficiency across the scenarios.

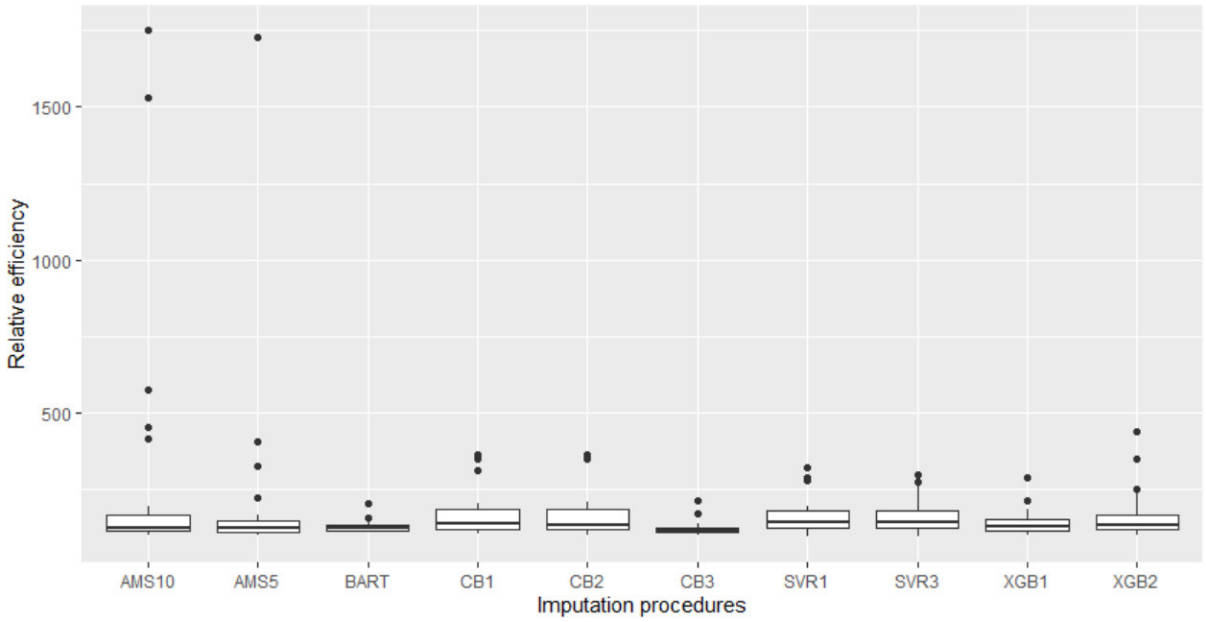


Figure 25: Monte Carlo percent relative efficiency across the scenarios: the best 10 procedures.

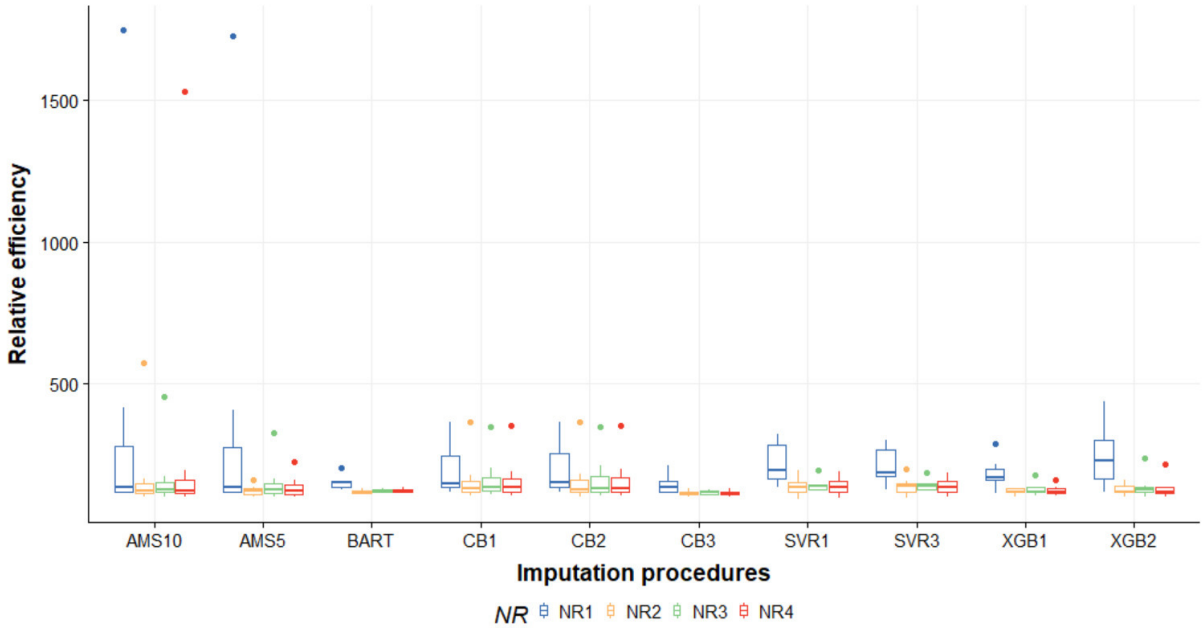


Figure 26: The effects of the nonresponse mechanism on the performance of the 10 best imputation procedures.

Continuous survey variables with Poisson sampling

Recall that Poisson sampling was used for estimating the population total of the survey variables Y_4 and Y_7. This led to $2 \times 4 \times 27 = 216$ sets of results. Due to the small number of scenarios ($2 \times 4 = 8$) for each of the survey variables Y_4 and Y_7, Tables 9 and 10 show the minimum, the median and the maximum Monte Carlo percent absolute RB and Monte Carlo percent RE only. The size variable X_5 used to obtain the first-order inclusion probabilities was included as a predictor in the imputation models. The results

Ranking	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
1	LR	CUBIST3	AMS5	BART	XGB3	CUBIST3	CUBIST3	CUBIST3
2	CUBIST3	LR	AMS10	CUBIST3	AMS5	BART	AMS5	AMS5
3	MW50	AMS5	BART	CUBIST1	AMS10	SVR3	AMS10	AMS10
4	AMS5	MWC50	CUBIST3	CUBIST2	XGB1	SVR1	MWC50	XGB1
5	AMS10	AMS10	CUBIST2	XGB1	XGB2	XGB1	BART	BART

Table 8: Best 5 imputation procedures for each survey variable.

in Tables 9 and 10 were consistent with those obtained for simple random sampling without replacement. Again, the best methods were CUBIST3, BART and XGB1 in terms of either bias or efficiency.

Binary survey variables

In this section, we present the results pertaining to the binary variables Y_9 and Y_{10}. Again, for each imputation procedure, we obtained $2 \times 4 = 8$ sets of results. Tables 11 and 12 show the minimum, the median and the maximum Monte Carlo percent absolute RB and Monte Carlo percent RE, respectively.

The ranking for binary survey variables was slightly different from that obtained for the continuous survey variables. Nearest-neighbor (NN) imputation procedure was the best in terms of bias and efficiency. Recall that NN imputation did not rank among the best procedures for the continuous variables. NN imputation was followed by CUBIST, XGBOOST and BART.

4.4.3 High-dimensional setting

In this section, we investigate the performance of a subset of the imputation procedures considered in Section 4.4.1 in a high-dimensional setting. To that end, we used data from the Irish Commission for Energy Regulation (CER) Smart Metering Project conducted in 2009-2010 (CER, 2011) that focused on energy consumption and energy regulation[2]. About 6000 smart meters were installed in Irish residences and businesses. The customer's electrical consumption was collected every half an hour over a period of about two years.

We considered a subset of the original data set. We ended up with a population of $N = 6291$ smart meters (households and businesses) for a period of 14 consecutive days. For each population unit i (household or business), we had $2 \times 7 \times 48 = 672$ measurements denoted by $X_j = X(t_j), j = 1, \ldots 672$. Each of these 672 measurements represents the electricity consumption (in kW) at instant t_j. We denote by x_{ij} the value of X_j recorded by the smart meter i for $i = 1, \ldots, N$ at instant t_j. It should be noted that these variables were highly correlated among themselves with a condition number of the matrix $N^{-1}\mathbf{X}^T\mathbf{X}$ computed using all the data, of about 60.000.

We created four survey variables based on a subset of the auxiliary variables $X_1, \ldots, X_{672}$:

$$
\begin{aligned}
Y_1 &= 400 + 2X_1 + X_2 + 2X_3 + \mathcal{N}(0, 1500); \\
Y_2 &= 400 + X_1X_2 + 2X_3 + \mathcal{N}(0, 1500); \\
Y_3 &= 500 + 2X_4 + 400\mathbb{1}_{\{X_5>156\}} - 400\mathbb{1}\,(X_5 \leqslant 156) + 1000\mathbb{1}\,(X_2 > 190) \\
&\quad + 300\mathbb{1}\,(X_5 > 200) + \mathcal{N}(0, 1500);
\end{aligned}
$$

[2] The data are available on request at: `https://www.ucd.ie/issda/data/commissionforenergyregulationcer/`.

Ranking	Model	Min	$Q_{0.5}$	Max
1	BART	0.1	0.9	3.0
2	CUBIST3	0.0	1	6.5
3	XGB1	0.0	2.4	5.2
4	CUBIST1	0.0	3.4	10.9
5	RF2	0.3	3.5	15.8
6	RF3	0.5	3.5	16.8
7	XGB2	0.4	3.9	8.6
8	AMS5	0.2	4.3	11.1
9	AMS10	0.2	4.3	10.7
10	CUBIST2	0.0	4.3	12.6
11	RF1	0.8	4.4	31.4
12	SVR3	0.1	4.4	6.7
13	LR	0.2	4.9	16.8
14	SVR1	0.1	4.9	7.1
15	MWC500	0.0	5.0	26.1
16	NN	0.0	5.0	7.3
17	MWC250	0.0	5.1	14.7
18	HDWC50	0.8	5.1	9.9
19	MWC50	0.0	5.2	9.9
20	MWC100	0.0	5.2	10.1
21	HDWC100	0.1	5.2	10.0
22	HDWC250	0.0	5.2	14.7
23	CART	0.2	5.6	24.6
24	5NN	1.3	7.1	11.7
25	XGB3	2.5	8.8	11.1
26	SVR2	1.0	11.7	22.6
27	SVR4	0.2	15.4	27.5

Table 9: Monte Carlo percent absolute relative bias of the imputed estimator: Descriptive statistics for Poisson sampling.

$$Y_4 = 1 + \cos(2X_1 + X_2 + 2X_3)^2 + \epsilon_1,$$

where $\epsilon_1 \sim \mathcal{E}(2)$ and these error terms were centered so as to have a mean equal to zero. We were interested in estimating the population total of the survey variables Y_1-Y_4. Again, the simulation was based of $R = 5,000$ iterations of the process described in Section 4.4. Samples of size $n = 1000$ were selected according to simple random sampling without replacement. The missing values to the survey variables Y_1-Y_4 were generated according to

$$p_i = 0.1 + 0.89 \times \text{sigmoid}\left\{-0.83 + 0.001 \times (2x_{i1} + 2x_{i2} - 2.5x_{i3})\right\},$$

leading to an average response rate of about 50%.

Ranking	Model	Min	$Q_{0.5}$	Max
1	BART	106	117	139
2	CUBIST3	111	118	239
3	XGB1	108	133	207
4	RF2	114	144	565
5	RF3	114	145	621
6	XGB2	110	156	246
7	SVR3	109	165	198
8	AMS5	124	168	486
9	SVR1	109	175	209
10	CUBIST1	114	175	469
11	NN	117	178	234
12	MWC50	125	188	396
13	MWC100	125	188	363
14	RF1	122	188	1868
15	LR	123	189	923
16	MWC250	128	190	525
17	CUBIST2	111	193	548
18	CART	133	198	1224
19	MWC500	133	198	1346
20	HDWC50	135	210	409
21	HDWC100	139	213	381
22	HDWC250	145	217	539
23	5NN	120	241	370
24	XGB3	116	272	441
25	AMS10	130	313	592
26	SVR2	142	493	1619
27	SVR4	141	769	2119

Table 10: Monte Carlo percent relative efficiency of the imputed estimator: Descriptive statistics for Poisson sampling.

Three high and very high dimensional settings were considered: in the first setting, the imputation models used the first 15 auxiliary variables $X_1, ..., X_{15}$, in the data set. In the second and third settings, the imputation models were based on the first 100 and 300 auxiliary variables $X_1, ..., X_{100}$, and $X_1, ..., X_{300}$, respectively.

To impute the missing values, we confined to a subset of the imputation procedures considered in Section 4.4.1: additive models, BART, CUBIST, XGBoost, random forests, nearest-neighbour imputation and support vector regression. Linear regression imputation and mean imputation within 20 classes were also considered. It is well known that the quality of predictions based on linear models tend to deteriorate substantially in the presence of a very large number of auxiliary variables. To cope with this issue, we also considered principal components analysis as a reduction-dimension method; see Cardot et al. (2017).

Ranking	Model	Min	$Q_{0.5}$	Max
1	NN	136	144	428
2	XGB3	153	165	860
3	XGB2	156	167	827
4	CUBIST3	156	167	841
5	XGB1	156	171	932
6	BART	156	173	1052
7	5NN	152	174	1191
8	CUBIST2	163	179	873
9	CUBIST1	169	191	904
10	RF2	158	192	1572
11	RF3	162	198	1769
12	AMS5	169	219	2453
13	MWC100	160	221	1120
14	MWC50	159	222	1067
15	SVR1	171	222	3196
16	AMS10	165	223	2472
17	MWC50	159	223	1061
8	M100	159	225	1116
19	CART	176	229	1882
20	LR	164	230	2707
21	MWC250	172	244	1460
22	MWC250	173	246	1471
23	SVR3	191	280	2899
24	RF1	190	305	4666
25	M500	186	365	4977
26	SVR4	219	409	26429
27	SVR2	413	1839	17279

Table 11: Monte Carlo percent relative efficiency of the imputed estimator: Descriptive statistics for the binary survey variables.

Table 13 shows the Monte Carlo percent relative bias (RB) and relative efficiency (RE) for $p = 15$ predictors. Table 14 shows the results for $p = 100$ and $p = 300$ predictors. For each scenario, the best imputation procedures are highlighted in bold. Note that the relative efficiency is now computed with respect to the mean square error of the imputed estimator based on the true imputation model. The additive models were considered in the first setting only ($p = 15$ variables) because their performance deteriorated rapidly as the number p of variables increased. For $p = 100$ and $p = 300$ the backfitting algorithm did not reach convergence in most scenarios.

From Tables 13 and 14, we note that CUBIST and XGBoost were the best method in the vast majority of the scenarios. These methods were followed by BART and random forests. As expected, additive models performed poorly, which illustrates the curse of dimensionality. It is worth pointing out that random forests performed better in the high-dimensional setting than they did in the low-dimension

Ranking	Model	Min	$Q_{0.5}$	Max
1	NN	0.0	0.5	3.6
2	CUBIST3	0.02	0.7	6.7
3	XGB3	0.03	0.8	7.7
4	BART	0.1	0.8	8.8
5	XGB1	0.14	0.9	7.9
6	XGB2	0.0	0.9	6.9
7	5NN	0.0	1.0	7.3
8	CUBIST2	0.2	1.0	7.0
9	CUBIST1	0.0	1.1	6.8
10	RF2	0.12	1.5	10.3
11	RF3	0.13	1.6	11.0
12	AMS5	0.04	1.6	11.9
13	AMS10	0.1	1.6	11.9
14	SVR1	0.3	1.7	12.0
15	LR	0.19	1.8	12.3
16	CART	0.18	1.8	11.4
17	MWC50	0.0	1.8	7.5
18	MWC100	0.0	1.8	7.7
19	HDWC50	0.03	1.8	7.5
20	HDWC100	0.01	1.8	7.7
21	MWC250	0.0	2.0	9.4
22	HDWC250	0.0	2.0	9.4
23	SVR3	0.43	2.3	11.5
24	RF1	0.08	2.7	19.0
25	SVR4	0.17	3.0	36.5
26	MWC500	0.0	3.2	16.4
27	SVR2	1.9	9.5	33.9

Table 12: Monte Carlo percent absolute relative bias of the imputed estimator: Descriptive statistics for the binary survey variables.

setting considered in section 4.4.1. Finally, the strategy based on principal components analysis did relatively well in most scenarios.

4.5 Simulation study: the case of population quantiles

In this section, we turn our attention to population quantiles. Except for nearest-neighbour imputation, we confined to the random versions of the imputation procedures described in Section 4.3. The target parameters were the quantiles of order $\gamma_1 = 0.25$, $\gamma_2 = 0.5$ and $\gamma_3 = 0.75$ that correspond to the first quartile, the median and the third quartile, respectively. We considered a subset of the scenarios described in Section 4.4.1. First, we confined to the case of the survey variables Y_3 and Y_6 and the nonresponse mechanisms (NR1) and (NR3) described in Section 4.4.1, leading to $2 \times 2 = 4$ scenarios. Also, samples

Variable	Criterion	LR	MWC50	RF2	XGB1	NN	SVR3	AMS5	CB3	PCR1	PCR2	PCR3	BART
Y_1	RE	**100**	117	110	**103**	111	124	**101**	**100**	160	113	**100**	**101**
	RB	-0,18	1,7	1,7	0	-0,1	2,6	-0,0	-0,1	4,0	0,6	-0,5	0,3
Y_2	RE	184	176	**103**	**100**	**100**	295	7041	**101**	159	213	207	106
	RB	-44,3	15,7	3,8	0	0,7	19,2	9,5	-0,0	-47,0	-53,1	-48,5	2,1
Y_3	RE	190	135	**102**	108	128	134	403	109	188	178	210	**105**
	RB	4,6	2,1	0,1	-0,2	0,1	2,08	-0,0	1,2	4,6	4,3	5,2	0,0
Y_4	RE	125	126	143	147	188	195	130	**118**	**119**	121	123	131
	RB	-0,0	-0,0	0,5	0,2	-0,1	-1,3	0,0	-0,0	-0,11	-0,1	-0,0	0,0

Table 13: Relative biais (RB) and relative efficiency (RE) of imputation procedures with $p = 15$ auxiliary variables.

Variable	Dim	Criterion	LR	MWC50	RF2	XGB1	NN	SVR3	CB3	PCR1	PCR2	PCR3	BART
Y_1	p=100	RE	**102**	122	149	**103**	216	187	**100**	269	226	151	**105**
		RB	0,14	2,1	4,2	0,3	6,2	5,1	0	7,8	6,6	4,0	0,6
Y_2	p=100	RE	115	287	**109**	**100**	**100**	340	**100**	**100**	**108**	140	127
		RB	-23,8	34,3	7,5	0,1	3,3	26,1	-0,0	-31,0	-28,9	-32,5	5,8
Y_3	p=100	RE	158	185	**107**	**107**	354	162	**108**	236	224	196	129
		RB	3,2	3,9	1,1	-0,0	7,0	3,4	0,9	5,9	5,5	4,8	7,7
Y_4	p=100	RE	140	141	151	146	243	217	**122**	**120**	**120**	**121**	135
		RB	0,0	0,1	0,7	0,28	0,4	-1,5	-0,0	-0,0	-0,1	-0,1	-0,0
Y_1	p=300	RE	120	215	190	103	286	237	**100**	290	262	189	**110**
		RB	-0,2	1	5,7	0,6	7,05	6,7	0,06	8,3	7,7	5,7	1,3
Y_2	p=300	RE	**102**	1106	112	**100**	**100**	405	**100**	**91**	**85**	109	243
		RB	-6,3	89,1	9,5	0,1	4,01	35,	-0,0	-28,4	-25,3	-26,9	4,6
Y_3	p=300	RE	197	378	118	107	630	180	**108**	350	245	224	242
		RB	1,0	6,7	2,0	0,0	9,1	4,1	0,8	6,2	6,1	5,6	6,4
Y_4	p=300	RE	276	584	155	143	443	214	**124**	**120**	**120**	**121**	131
		RB	0,1	2,4	0,7	0,3	0,6	-1,5	0,06	-0,0	-0,1	-0,1	-0,0

Table 14: Relative biais (RB) and relative efficiency (RE) of imputation procedures with $p = 100$ and respectively, $p = 300$ auxiliary variables.

were selected according to simple random sampling without replacement only. In each sample, we computed the imputed estimator $\widehat{Q}_{\gamma,imp}$ given by (4.5) for $\gamma_1 = 0.25$, $\gamma_2 = 0.5$ and $\gamma_3 = 0.75$. As in Section 4.4, we computed the Monte Carlo percent relative bias of $\widehat{Q}_{\gamma,imp}$ and the relative efficiency,

given respectively by (4.40) and (4.41) with $\widehat{t}_{imp}$ replaced with $\widehat{Q}_{\gamma,imp}$, $\widehat{t}_{\pi}$ replaced with $\widehat{Q}_{\gamma}$ and t_y replaced with Q_γ.

The results are presented in Figures 27-29. In each figure, the x-axis corresponds to the median of the Monte Carlo percent relative bias of $\widehat{Q}_{\gamma,imp}$ computed across the 4 scenarios, whereas the y-axis corresponds to the median of the Monte Carlo relative efficiency. For the purpose of clarity, we have excluded from Figures 27-29 any imputation procedure whose median of the Monte Carlo percent relative bias lied outside the interval $[-20; 20]$ or whose median of the Monte Carlo relative efficiency was above 500.

From Figures 27-29, Cubist displayed a very good performance in terms of bias and efficiency for the three quantiles. The procedure XGBoost led to good results for $Q_{0.25}$ and $Q_{0.75}$ but performed poorly for $Q_{0.5}$. Similarly, BART performed very well for both $Q_{0.5}$ and $Q_{0.75}$ but exhibited a poor performance for $Q_{0.25}$. Support vector machine (SVR3) did relatively well for both $Q_{0.5}$ and $Q_{0.75}$ but was outperformed by Cubist and XGBoost for $Q_{0.25}$. Again, the Cubist algorithm seemed to be insensitive to the target parameter, the model that has generated the Y-variable and the nonresponse mechanism, at least in our experiments.

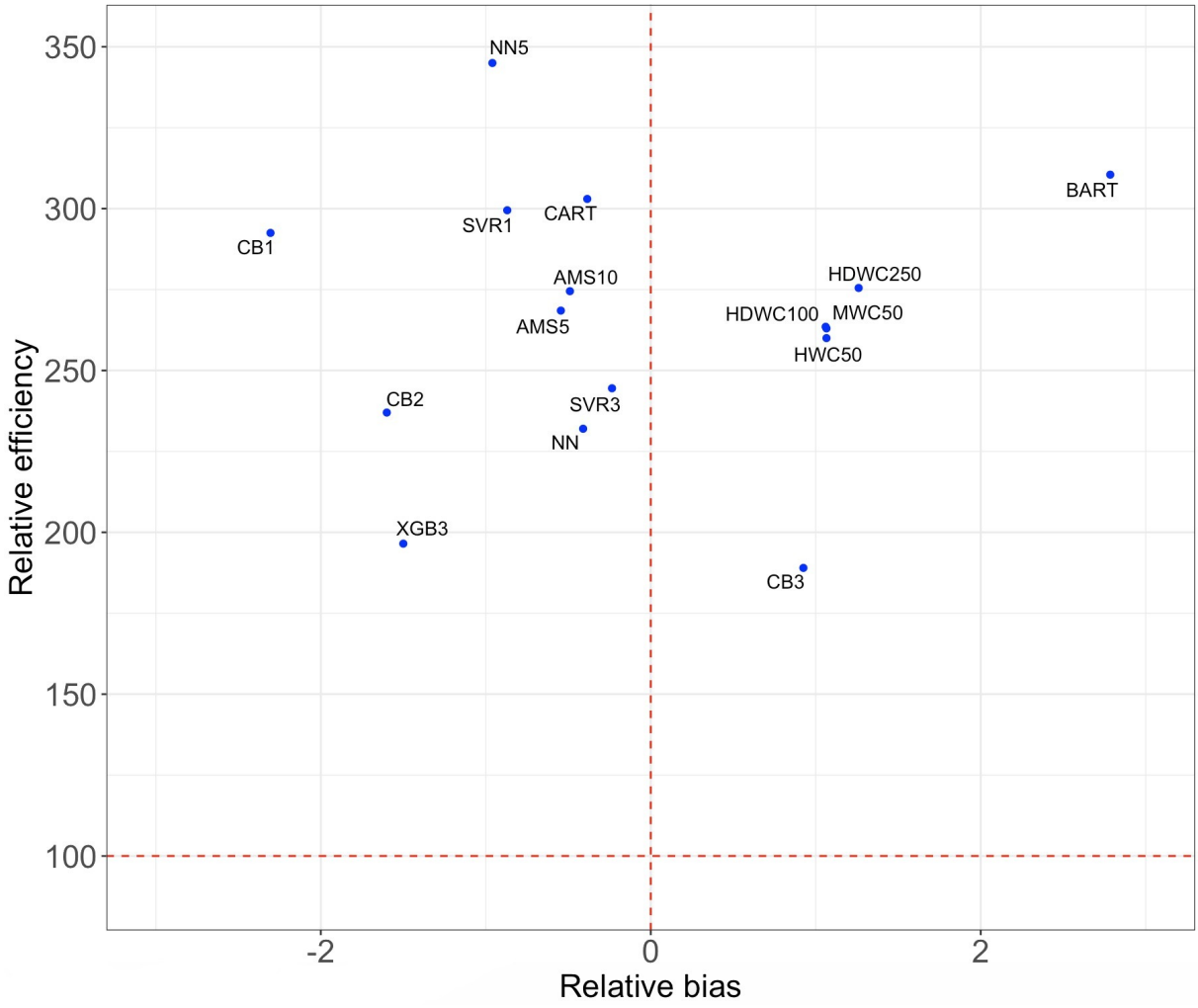


Figure 27: Median performances of the best imputed estimators for the estimation of $Q_{0.25}$.

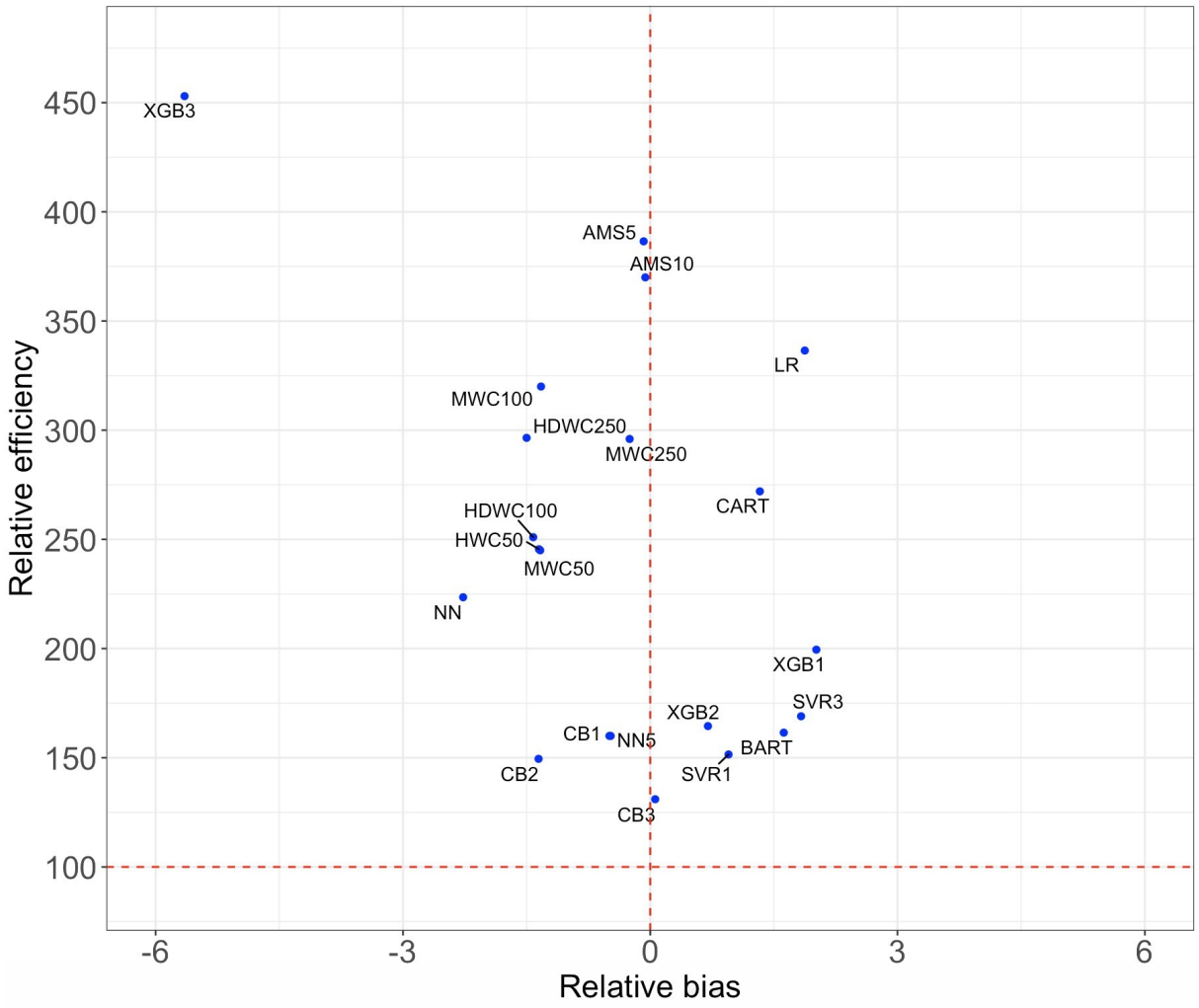


Figure 28: Median performances of the best imputed estimators for the estimation of $Q_{0.5}$.

4.6 Final remarks

In this paper, we have conducted an extensive simulation study to compare several nonparametric and machine learning imputation procedures in terms of bias and efficiency. The imputation procedures were evaluated in the case of finite population totals of continuous and binary variables and for population quantiles under both simple random sampling without replacement and proportional-to-size Poisson sampling. The Cubist algorithm, BART and XGBoost performed very well in a wide variety of settings. In general, these methods seem to be highly robust to model misspecification and seem to have the ability to capture nonlinear trends in the data. Additive models based on B-splines performed well in the case of population totals when the number of explanatory variables was small but broke down for large values of p. Finally, random forests performed relatively well in a high-dimensional setting. In practice, the choice of an imputation procedure is not clear-cut and depends on the data at hand. If one is reasonably confident about the correct specification of the first moment of the imputation model (that includes the correct specification of the functional form and the correct specification of the vector of explanatory variables), parametric imputation procedures are expected to do well in terms of bias and efficiency. In addition, parametric imputation is simpler to understand and the results are easier to interpret, in general. In the case of complex/nonlinear relationships and/or in a high-dimensional setting, our empirical investigations suggest that machine learning procedures outperform traditional imputation procedures as they tend to be robust against model misspecification. However, these procedures require

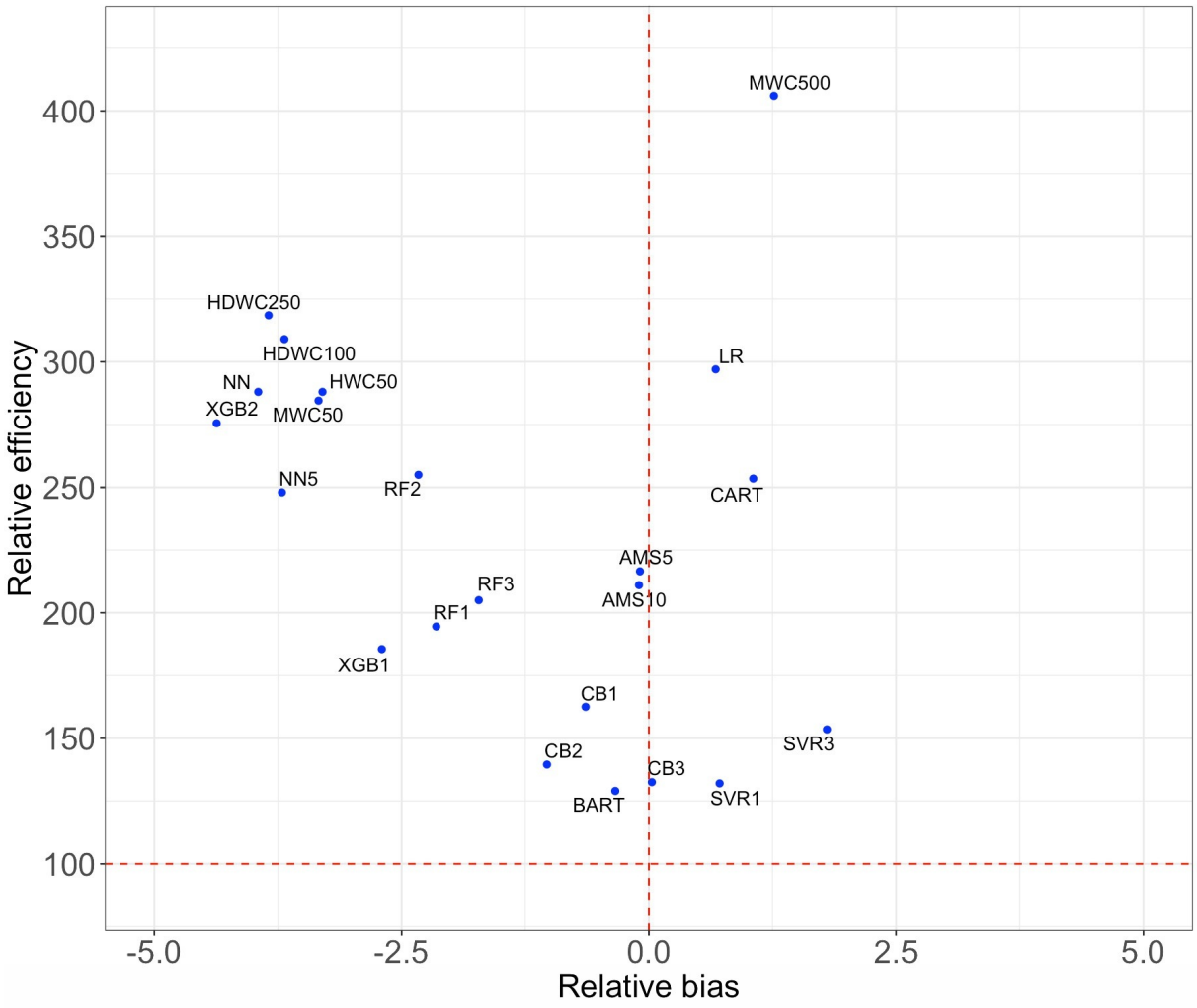


Figure 29: Median performances of the best imputed estimators for the estimation of $Q_{0.75}$.

the specification of some regularization parameters. For instance, for XGBoost, one must specify the learning rate, the maximal depth and the coefficient of penalization. In support vector regression, the cost function and the kernel function must be selected, among others. In practice, the value for some of these parameters are determined through a cross-validation procedure. To keep the processing time at a reasonable level, all the regularization parameters were predetermined in our experiments. Overall, it seems that Cubist is an excellent choice as it performed well in all the scenarios, unlike its main competitors (e.g., XGBoost, BART, random forest, etc.) whose performance varied from one scenario to another. From a computational point of view, most procedures were efficient. One notable exception is BART that proved to be highly computer intensive with an average processing time approximately twenty times larger than what was required for the other procedures.

Drawing inferences from survey data requires a variance estimate. It is well known that imputed values should not be treated as observed values. Otherwise, the resulting variance estimates tend to be much smaller, on average, than the true variance, especially if the nonresponse rates are appreciable. In the last three decades, a number of variance estimation procedures have been proposed for obtaining variance estimates that account for sampling, nonresponse and imputation. The reader is referred to Haziza and Vallée (2020) for a comprehensive overview of variance estimation procedures in the presence of singly imputed data sets. Estimating the variance of imputed estimators obtained through machine learning procedures is challenging and requires further research. If the sampling fraction is negligible, one can recourse to the bootstrap procedure of Shao and Sitter (1996) that consists of selecting

bootstrap samples according to a complete data bootstrap procedure and reimputing the missing values within each bootstrap sample using the same imputation method that was used on the original data. If a machine learning procedure is used to impute the missing data, the Shao-Sitter procedure may be highly computer intensive. When the sampling fraction is not negligible, the problem of bootstrap variance estimation is more intricate (Chen et al., 2019). To make the variance estimation process simpler for survey practitioners, it would be desirable to derive a "universal" variance estimator based on Taylor expansion procedures that could be applicable to a wide class of machine learning imputation procedures, at least in the case of negligible sampling fractions. This is currently under investigation.

Investigating the performance of deep learning methods in the context of imputation for missing survey data would constitute a promising direction for future research. There exist a wide class of deep learning procedures based on relatively sophisticated algorithms that proved to be extremely efficient in the context of unstructured data such as signal processing or text analysis. However, for deep learning procedures to "shine" in terms of efficiency typically requires a huge volume of unstructured data, which is seldom the case in surveys. In practice, most data sets in surveys consist of structured data and contains, at most, a few millions observations and a few hundred survey variables. As noted by Choley (2018):

“(...) *gradient boosting (such as XGBoost) is used for problems where structure data is available, whereas deep learning is used for perceptual problems such as image classification*”.

We believe that the class of imputation procedures considered in this article, that includes bagging and boosting among others, offers a number of very good options that may be applicable to virtually all the surveys conducted by NSOs.

5 REGRESSION TREE AND RANDOM FOREST IMPUTATION IN SURVEYS WITH APPLICATION TO DATA INTEGRATION

Abstract. Item nonresponse in surveys is usually handled through some form of imputation. Regression trees and random forests provide flexible tools for obtaining a set of imputed values. We lay out a set of conditions on the imputation model sufficient for establishing the L^2-consistency of an imputed estimator. We consider several regression trees and random forest algorithms for which we establish the L^2-consistency. We derive the asymptotic variance of imputed estimators and propose corresponding variance estimators. We apply our results to the particular case of mass imputation for data integration. We present the results from a simulation study that investigates the performance of point and variance estimators based on random forest imputation in terms of bias, efficiency and coverage rates.

Keywords: Missing data; Imputation; Survey sampling; Regression trees; Random forests.

5.1 Introduction

Since the seminal paper of Breiman (2001), random forests have been used in a variety of applications including medicine (Fraiwan et al., 2012), time series analysis (Kane et al., 2014), agriculture (Grimm et al., 2008), missing data (Stekhoven and Buhlmann, 2011), genomics (Qi, 2012) and pattern recognition (Rogez et al., 2008). Random forests belong to the class of ensemble models, whereby a collection of B regression trees are constructed and a prediction is generated from each of the B trees. Unlike many nonparametric statistical procedures (e.g., kernel predictors, k-nearest neighbors, splines), random forests perform relatively well with high-dimensional data; see e.g., Hamza and Larocque (2005) and Díaz-Uriarte and de Andrés (2006). Some recent theoretical investigations (Biau, 2012, Klusowski, 2021, Scornet et al., 2015) also suggest that random forests adapt well to sparse situations.

In surveys, the problem of missing data is ubiquitous. Estimators of population totals based on complete cases only, often referred to as unadjusted estimators, tend to exhibit large biases when the proportion of missing data is appreciable and the behavior of the responding units is different from that of the nonresponding units. In this article, we focus on the problem of item nonresponse, a term used to describe the absence of information on some, but not all, survey variables for a sample unit. The missing values are replaced by a plausible value constructed on the basis of auxiliary variables available for both respondents and nonrespondents, a process known as imputation. A large number of imputation procedures have been developed to compensate for missing values and to reduce the nonresponse bias to the best possible extent. The reader is referred to Haziza (2009) and Chen and Haziza (2019) for comprehensive discussions of imputation procedures in a survey sampling setting. Every imputation procedure relies on some implicit or explicit assumptions about the distribution of the survey variable requiring imputation. This set of assumptions is called an imputation model. In this context, tree-based methods such as random forests may prove useful for obtaining a set of imputed values. Because they are nonparametric in nature, random forests tend to be robust against model misspecification. Also, with the emergence of large data sets in National Statistical Offices (NSO), random forests have attracted a lot of attention in NSOs in recent years and are being scrutinized as an alternative to traditional imputation procedures. However, to the best of our knowledge, a theoretical investigation on the properties of random forests in the context of imputation for missing survey data is currently lacking.

In this paper, the aim is to study a number of random forest algorithms that have been suggested in the literature. In Section 5.2, we begin by introducing a set of two sufficient conditions on an imputation model, so that, whenever satisfied, leads to the L^2-consistency of the resulting imputed estimator. In Section 5.3, we provide an analysis of trees imputed estimators. Finite sample properties are derived through the analysis of the corresponding weighting system. The L^2-consistency of the tree imputed estimator based on the CART algorithm (Breiman, 1984) is established. In Section 5.4, we focus on random forest imputed estimators. We begin by establishing the connection between tree imputed estimators and random forests imputed estimators. As such, random forest estimators inherit many of the properties of tree estimators, minor a few differences that are highlighted. The L^2-consistency of forest imputed estimators based on uniform random forests (Biau et al., 2008, Scornet, 2016a) and Breiman's original algorithm (Breiman, 2001) is established. In Section 5.5, using the reverse approach of Shao and Steel (1999) and the approach of Särndal (1992), we suggest two variance estimators that account for the sampling, nonresponse and imputation variability. In Section 5.6, we apply random forest imputation to the case of data integration. Before concluding, we investigate the empirical properties through a simulation study presented in Section 5.7. All proofs and further technical details are relegated to the Appendix.

5.2 Mean square consistency of imputed estimators

Consider a finite population $U = \{1, 2, ..., N\}$ of size N. We are interested in estimating the population total, $t_y = \sum_{k \in U} y_k$, of a survey variable Y. We select a sample S, of size n, according to a sampling design $\mathcal{P}(S)$ with first-order inclusion probabilities $\{\pi_k\}_{k \in U}$ and second-order inclusion probabilities $\{\pi_{k\ell}\}_{k \neq \ell \in U}$; we shall denote by $\Delta_{k\ell} := \pi_k \pi_\ell - \pi_{k\ell}$ the sampling covariances, for elements $k, \ell \in U$. The sample S is completely characterized by the vector of sample selection indicators $\boldsymbol{I} = (I_1, \ldots, I_k, \ldots, I_N)^\top$, where $I_k = 1$ if $k \in S$ and $I_k = 0$, otherwise. A complete data estimator of t_y is the well-known Horvitz-Thompson (HT) estimator:

$$\widehat{t}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} d_k y_k, \tag{5.1}$$

with $d_k := 1/\pi_k$, the sampling weight attached to element $k \in S$. Provided that $\pi_k > 0$, for all $k \in U$, the estimator (5.1) is design-unbiased for t_y.

In practice, the Y variable is subject to missingness. Let $\boldsymbol{r} = (r_1, \ldots, r_k, \ldots, r_N)^\top$ denote the vector of response indicators such that $r_k = 1$ if y_k is observed and $r_k = 0$, otherwise. Let $S_r = \{k \in S;\ r_k = 1\}$ denote the set of respondents, of size n_r, and $S_m = \{k \in S;\ r_k = 0\}$ the set of nonrespondents, of size n_m, such that $S_r \cup S_m = S$ and $n_r + n_m = n$. Let $\mathbf{x}_k = (x_{1k}, x_{2k}, ..., x_{pk})^\top$ be the vector of fully observed auxiliary variables attached to unit k and $\boldsymbol{X} := \{\mathbf{x}_k\}_{k \in U}$. Available to the imputer are the pairs $(y_k, \mathbf{x}_k)$, for $k \in S_r$, and the vector $\mathbf{x}_k$ for $k \in S_m$. In this paper, we restrict our investigations to response mechanisms satisfying the missing at random assumption (MAR, Rubin (1976)), defined below.

(H19) The missing data mechanism is such that:

a) The random vectors $\{[r_k, y_k, \mathbf{x}_k]^\top\}_{k \in U_v}$ are independently and identically distributed (i.i.d.).

b) The nonresponse mechanism is missing at random (MAR), that is, $\mathbb{P}(r_k = 1 | \mathbf{x}_k, y_k) = \mathbb{P}(r_k = 1 | \mathbf{x}_k)$, and, for all $k \in U$, $\mathbb{P}(r_k = 1 | \mathbf{x}_k) > 0$.

Assumption (H19) is common in the missing data literature (Rubin, 1976) and is required in order to estimate the regression function from the observed data. It states that, given the covariates $\mathbf{x}_k$, the survey variable y_k is independent of the response indicators r_k.

We assume that the relationship between the survey variable Y and the set of auxiliary variables $\mathbf{x}$ can be described by the following imputation model:

$$\xi : \qquad y_k = m(\mathbf{x}_k) + \epsilon_k, \qquad k \in S_r, \tag{5.2}$$

where $m(\mathbf{x}) := \mathbb{E}[Y|\boldsymbol{X} = \mathbf{x}]$ denotes the regression function, and $\{\epsilon_k\}_{k\in S_r}$ is a sequence of independent and identically distributed (i.i.d.) white noise. We assume that: i) the regression function m is continuous; ii) the distribution of the covariates $\mathbb{P}_{\mathbf{x}}$ is supported on $Supp(\mathbb{P}_{\mathbf{x}})$, a compact subset the unit cube $[0;1]^p$; iii) the residuals have a compact support; the survey variable Y has a distribution absolutely continuous with respect to the Lebesgue measure. These assumptions imply that the survey variable Y is almost surely bounded.

Let $\widehat{m}$ be a estimator of m fitted on $D_{n_r} := \{(\mathbf{x}_k, y_k); k \in S_r\}$. The imputed estimator $\widehat{t}_{\widehat{m}}$ of t_y based on the imputation procedure $\widehat{m}$ is given by

$$\widehat{t}_{\widehat{m}} := \sum_{k\in S_r} \frac{y_k}{\pi_k} + \sum_{k\in S_m} \frac{\widehat{m}(\mathbf{x}_k)}{\pi_k}, \tag{5.3}$$

where $\widehat{m}(\mathbf{x}_k)$ denotes the imputed value associated with $k \in S_m$.

To establish the asymptotic properties of (5.3), we consider the asymptotic framework of Isaki and Fuller (1982). We consider an increasing sequence of embedded finite populations $\{U_v\}_{v\in\mathbb{N}}$ of size $\{N_v\}_{v\in\mathbb{N}}$. In each finite population U_v, a sample S_v, of size n_v, is selected according to a sampling design $\mathcal{P}_v$ with inclusion probabilities $\pi_{k,v}$ and $\pi_{k\ell,v}$. While the finite populations are assumed to be embedded, we do not require this property to hold for the samples $\{S_v\}_{v\in\mathbb{N}}$. This asymptotic framework assumes that v goes to infinity, so that both the finite population size N_v and the sample size n_v go to infinity. To improve readability, we shall use the subscript v only in the quantities U_v, N_v and n_v; quantities such as $\pi_{k,v}$ and $\pi_{k\ell,v}$ shall simply be denoted as π_k and $\pi_{k\ell}$.

We now describe a set of conditions on $\widehat{m}$ sufficient for establishing the L^2-consistency of $\widehat{t}_{\widehat{m}}$. An imputed estimator $\widehat{t}_{\widehat{m}}$ is said to be mean square consistent or L^2-consistent for t_y if

$$\mathbb{E}\left[\left\{\frac{1}{N_v}\left(\widehat{t}_{\widehat{m}} - t_y\right)\right\}^2\right] \xrightarrow[v\to\infty]{} 0. \tag{5.4}$$

Throughout this chapter, the expectation and variance operators are evaluated with respect to the joint distribution induced by the imputation model, the sampling design and the nonresponse mechanism.

We start with the regularity conditions needed for the L^2-consistency of the complete data estimator (5.1).

(H20) We assume that the sampling design $\mathcal{P}_v(\cdot)$ is non-informative sampling and that

a) The sampling fraction is such that $\lim_{v\to\infty} \frac{n_v}{N_v} = \pi^* \in (0;1)$.

b) There exists positive constants λ and λ^* such that $\min_{k\in U_v} \pi_k \geqslant \lambda > 0$, $\min_{k,\ell\in U_v} \pi_{k\ell} \geqslant \lambda^* > 0$.

c) The sampling covariances are such that $\limsup_{v\to\infty} n_v \max_{k\neq\ell\in U_v} |\pi_{k\ell} - \pi_k\pi_\ell| < \infty$.

Assumption (H20) is commonly used in the literature, see e.g., Robinson and Särndal (1983) and Breidt and Opsomer (2000). It is known to hold for commonly used sampling designs.

Result 5.2.1. *Assume (H19) and (H20). Consider a sequence of predictors* $\{\widehat{m}\}$ *fitted on* D_{n_r} *and its population counterparts* $\{\widetilde{m}\}$ *fitted on* $D_N := \{(\mathbf{x}_k, y_k)\,; k \in U\}$ *. If*

i) The sequence of population predictors $\{\widetilde{m}\}$ *satisfies*

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\widetilde{m}(\mathbf{x}) - m(\mathbf{x})\right)^2\right] = 0,$$

with a convergence rate denoted γ_v*.*

ii) There exists a positive constant C*, independent of* v*, such that*

$$\mathbb{E}\left\{\left(\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right)^2 \,\middle|\, \boldsymbol{r}, \boldsymbol{X}, \boldsymbol{I}\right\} \leqslant C. \qquad a.s.$$

Then, the sequence of imputed estimators $\{\widehat{t}_{\widehat{m}}\}$ *is* L^2*-consistent with rate*

$$\mathbb{E}\left[\left(\frac{1}{N_v}\left\{\widehat{t}_{\widehat{m}} - t_y\right\}\right)^2\right] = O\left(\gamma_v\right). \tag{5.5}$$

Proof. See Appendix 5.2.1. ■

Condition (i) in Result 5.2.1 requires the convergence of the L^2-error of prediction towards zero whereas Condition (ii) requires that the conditional error is almost surely bounded, even for finite samples. Under some regularity assumptions, Condition (i) is satisfied for a large number of statistical procedures including linear regression and nonparametric regression procedures such as k-nearest neighbors and kernel regression under appropriate regularity conditions, see Devroye et al. (2013). Assuming a framework in which condition (ii) is satisfied, this result suggests that, in order to build a consistent imputed estimator, it is enough to use a consistent predictor to produce the imputed values. In that respect, imputation is not more difficult than regression. In a way, this is closely related to Theorem 2.2 of Devroye et al. (2013), which states that it is enough to have a consistent regression estimate to obtain a consistent classification rule.

Remark 5.2.1. *Result 5.2.1 can be used for establishing the* L^2*-consistency of imputed estimators based on a wide class of parametric and nonparametric procedures, without having to impose any assumption on the superpopulation model. However, it may lead to suboptimal consistency rates without additional assumptions. Indeed, if no assumption about the joint distribution of* $(\boldsymbol{X}, y)$ *or about the smoothness of the regression function* $m(\cdot)$ *is made, an approach referred to as fully nonparametric, then there does not exist any guaranteed convergence rate in Result 5.2.1, no matter which imputation procedure is used (see Theorem 3.1 of Györfi et al. (2006)). It follows that if one aims to obtain the convergence rate of* $\widehat{t}_{\widehat{m}}$ *based on Result 5.2.1, one has to consider a restricted class for the distributions of* $(\boldsymbol{X}, Y)$*. By doing so, the rate of convergence of* $\widehat{t}_{\widehat{m}}$ *will be restricted by the minimax rate of convergence over the selected class, see e.g., Györfi et al. (2006).*

5.3 Tree imputation

In this section we begin by describing regression trees and partitioning predictors. We define an imputed estimator based on a regression tree. Its finite sample properties are discussed through the analysis of the underlying weighting system.

5.3.1 Trees and partitioning predictors

Partitioning predictors are algorithms that first create a partition $\mathcal{P} = \{A_1, A_2, ..., A_T\}$ of the predictor space based on D_{n_r}, and use it to make their predictions. The elements of $\mathcal{P}$ are called the (terminal) nodes. For a point $\mathbf{x}$, define

$$\widehat{m}_{tree}(\mathbf{x}, \mathcal{P}) := \sum_{k \in S_r} \frac{\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} \cdot y_k = \sum_{k \in S_r} \widehat{W}_k(\mathbf{x}, \mathcal{P}) y_k, \tag{5.6}$$

where $A(\mathbf{x})$ denotes the node of $\mathcal{P}$ containing $\mathbf{x}$ and

$$\widehat{W}_k(\mathbf{x}, \mathcal{P}) := \frac{\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}}, \qquad k, \ell \in U, \tag{5.7}$$

denotes the prediction weights of $\widehat{m}_{tree}$. In other words, the prediction $\widehat{m}_{tree}(\mathbf{x})$ of $\widehat{m}_{tree}$ is obtained by averaging the observations for the points that fall into the same node as the point $\mathbf{x}$. We use the convention that $\widehat{m}_{tree}(\mathbf{x}, \mathcal{P}) = 0$ if $\mathbf{x}$ is in an empty node. Since $\widehat{m}_{tree}$ can be written as a weighted sum of the observations, the properties of the tree imputed estimators are closely related to the properties of the prediction weights $\{\widehat{W}_k(\mathbf{x}_\ell, \mathcal{P})\}_{k,\ell \in S_r}$. For simplicity, in the sequel, we write $\widehat{W}_k(\mathbf{x})$ for the weights defined in (5.7). Properties of the weights $\{\widehat{W}_k(\mathbf{x}_\ell, \mathcal{P})\}_{k,\ell \in S_r}$ are given in Technical lemma 1.

Given the sample points, changing the partition may lead to different predictions. As such, a partitioning predictor is fully determined by both a set of points D_{n_r} and a partition $\mathcal{P}$. Most often, the partition $\mathcal{P}$ is obtained as the output of a data partitioning algorithm; that is; an algorithm which takes the sample points as input, and outputs a partition of the regressor space. When the partitioning algorithm creates a partition by splitting recursively the regressor space, the algorithms are often called trees. Adopting the terminology of Devroye et al. (2013), when the partitioning algorithm does not make use of the survey variable Y, we say that the partitioning rule, and, by extension, the partitioning predictor, has the X-property.

Example 5.3.1. *CART algorithm (Breiman, 1984).*
In the CART algorithm, splits are created by a greedy algorithm that splits recursively the regressor space. More specifically, let A denote a node containing #(A) respondents, considered for the next split, and C_A the set of possible splits in the node A, which corresponds to the set of all possible pairs (j, z) = (variable, position). Let

$$mse(A) = \frac{1}{\#(A)} \sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A} (y_k - \bar{y}_A)^2$$

and $\bar{y}_A$ be the average of the y-values of units belonging to A. This splitting process is performed by searching for the best split (j^*, z^*), i.e. the split for which the following criterion is maximized:

$$L(j, z) = mse(A) - mse(A_L) - mse(A_R)$$

where $A_L = \{k \in A; x_{kj} < z\}$, $A_R = \{k \in A; x_{kj} \geqslant z\}$. This criterion therefore searches for the split which would generate child nodes as homogeneous as possible, in terms of mean square error. Splits are always performed in the middle of two points. The procedure continues until a stopping criterion is reached. Usual stopping criteria consist of specifying a minimum number of elements (n_0) in the terminal nodes, or a maximum depth (K) for the tree.

For more details about trees and partitioning procedures, the reader is referred to Hastie et al. (2011) or Györfi et al. (2006).

5.3.2 Regression tree imputation

Let $\widehat{m}_{tree}(\cdot, D_{n_r}) := \widehat{m}_{tree}(\cdot)$ be a tree predictor. We define the regression tree imputed estimator $\widehat{t}_{tree}$ of t_y as

$$\widehat{t}_{tree} := \sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_m} \frac{\widehat{m}_{tree}(\mathbf{x}_k)}{\pi_k}. \tag{5.8}$$

The properties of $\widehat{t}_{tree}$ are closely related to the behavior of its underlying weighting system.

Proposition 5.3.1. *The tree estimator $\widehat{t}_{tree}$ defined by (5.8) can be written as*

$$\widehat{t}_{tree} = \sum_{k \in S_r} w_k y_k,$$

where the weights $\{w_k\}_{k \in S_r}$ are given by

$$w_k = \frac{1}{\pi_k} + \sum_{\ell \in S_m} \frac{\widehat{W}_k(\mathbf{x}_\ell)}{\pi_\ell} = \frac{1}{\pi_k} + \frac{\widehat{N}(\mathbf{x}_k, S_m)}{N(\mathbf{x}_k, S_r)}, \qquad k \in S_r, \tag{5.9}$$

with $\widehat{N}(\mathbf{x}_k, S_m) := \sum_{\ell \in S_m} \pi_\ell^{-1} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x}_k)}$ denoting the estimated number of nonrespondents in $A(\mathbf{x}_k)$; similarly, $N(\mathbf{x}_k, S_r) := \sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x}_k)}$ is used to denote the number of respondents that fall in $A(\mathbf{x}_k)$.

To get a better understanding of the weighting system $\{w_k\}_{k \in S_r}$, consider the special case of $\pi_k = \pi$, for all $k \in U$. In that particular case, the estimation weights w_k reduce to

$$w_k = \pi^{-1} \times \left(1 + \frac{N(\mathbf{x}_k, S_m)}{N(\mathbf{x}_k, S_r)}\right) = \pi^{-1} \times \left\{1 + R_{mr}(\mathbf{x}_k)\right\} = \pi^{-1} \cdot g_k,$$

where $g_k := 1 + R_{mr}(\mathbf{x}_k)$ and $R_{mr}(\mathbf{x}_k)$ denotes the ratio between the number of nonrespondents and the number of respondents in the node $A(\mathbf{x}_k)$. We see that a set "g-weights" is induced by the tree estimator in the case of equal inclusion probabilities (and only in that case). The weight w_k will be large when the ratio of nonrespondents over respondents for elements similar to element k is large. If the neighbors of element k are mostly nonrespondents, there are only few such elements in the sample of respondents, so that this "category" would, in some sense, be under-represented. In that case, the ratio $R_{mr}(\mathbf{x}_k)$ is large, so that the correction factor $1 + R_{mr}(\mathbf{x}_k)$ is significantly greater than 1.

Remark 5.3.1. *Although Proposition 5.3.1 states that $\widehat{t}_{tree}$ can be written as a weighted sum of the sequence $\{y_k\}_{k \in S_r}$, it does not imply that $\widehat{t}_{tree}$ is a linear estimator of t_y. Indeed, if the partition $\mathcal{P}$ induced by $\widehat{m}_{tree}$ does not have the X-property, then the estimation weights $\{w_k\}_{k \in S_r}$ also have a Y-dependency, which implies that $\widehat{t}_{tree}$ is a nonlinear function of $\{y_k\}_{k \in S_r}$.*

Proposition 5.3.2. *The weights $\{w_k\}_{k \in S_r}$ in (5.31) have the following properties.*

i) The weights are calibrated to the population size N whenever the original weighting system $\{d_k\}_{k \in U}$ is:

$$\sum_{k \in S_r} w_k = \sum_{k \in S} d_k := \widehat{N}.$$

ii) If there are at least n_0 elements in each node, the weights $\{w_k\}_{k \in S_r}$ are bounded,

$$d_k \leqslant w_k \leqslant d_k \left(1 + \frac{n_m}{n_0}\right), \quad a.s. \qquad k \in S_r. \tag{5.10}$$

The bounds are sharp, i.e., each of the bounds can be attained.

The lower bound in (5.10) given by $w_k = d_k$ is obtained where there is no missing elements in the node containing element k. The upper bound is obtained when all missing elements are in the node of element k, and if this node contains precisely the minimal number of elements possible. Two elements can also be observed from Proposition 5.3.2: 1) more nonrespondents leads to a more conservative inequality; 2) more elements in the nodes implies less volatile weights: the larger n_0 is chosen, the lower is the diameter of the support of $\{w_k\}_{k\in S_r}$, leading to less volatile weights.

Proposition 5.3.3. *If the sampling design is such that $\pi_k = \pi$ for all $k \in U$, then $\widehat{t}_{tree}$ can be written in projection form, that is,*

$$\widehat{t}_{tree} = \sum_{k\in S} \frac{\widehat{m}_{tree}(\mathbf{x}_k)}{\pi_k}.$$

Remark 5.3.2. *When the sampling design induces unequal first order inclusion probabilities, we can find cases where Proposition 5.3.3 does not hold. This is due to the fact that the predictions of the tree predictor $\widehat{m}_{tree}$ are not weighted by the inclusion probabilities. If it was, Proposition 5.3.3 would hold without the equal inclusion probabilities assumption.*

5.3.3 Properties of the tree imputed estimator

As for most imputed estimators, information about the distribution of $\widehat{t}_{tree}$ are difficult to obtain. In general, its bias and variance are unknown. In some cases, however, it is possible to obtain information about these quantities. First, in the particular case where the survey variable is constant, it is possible to fully characterize the distribution of $\widehat{t}_{tree}$, as shown in Example 5.3.2.

Example 5.3.2. *Assume that*

$$\xi: \qquad y_k = C, \qquad k \in U,$$

and that $\sum_{k\in S} d_k = N$. Then, we have $\widehat{t}_{tree} = t_y$, with probability one. In this particular case, t_y is always perfectly estimated by $\widehat{t}_{tree}$. Note that, in this case, we have $\widehat{t}_{\pi} = t_y$, with probability one as well.

Obviously, this scenario is not realistic as it is too simple to represent practical situations; consider the more practical example in which the survey variable is not constant, but the regression function is. In that case, the first two moments of $\widehat{t}_{tree}$ can be obtained, see Example 5.3.3 below.

Example 5.3.3. *Assume that*

$$\xi: \qquad y_k = C + \epsilon_k, \qquad k \in U,$$

with $\{\epsilon_k\}_{k\in U}$ is a sequence of i.i.d. random variables such that $\mathbb{E}[\epsilon_k|\mathbf{x}_k] = 0$ and $\mathbb{E}\left[\epsilon_k^2|\mathbf{x}_k\right] = \sigma^2$. Assume also that $\sum_{k\in S} d_k = N$ and that the tree predictor has the X-property. Then, it can be shown that $\widehat{t}_{tree}$ remains unbiased but its variance is now strictly positive.

In a general setup, Result 5.3.1 shows that the tree imputed estimator based on the CART criterion is L^2 consistent for t_y, under some regularity conditions.

Result 5.3.1. *Assume (H19) and (H20). Consider a sequence of tree imputed estimators $\{\widehat{t}_{tree}\}$ based on the CART criterion described in Example 5.3.1. Assume that:*

1. *No additional split is performed in a node if it contains one element or if the maximal depth K_v is reached.*

2. *The regression function m is additive and bounded, i.e.*

$$m_v \in \mathcal{G}_v := \{g(\mathbf{x}) = \sum_{j=1}^{p_v} g_j(x_j),\ g_j \text{ is bounded and Borel measurable},\ j = 1, 2, \ldots p_v\},$$

and $||m_v||_{l_0} = \#\{j = 1, 2, ..., p_v;\ m_j \text{ non-constant}\} = o(\sqrt{K_v})$.

Then, if $\lim_{v\to\infty} K_v = +\infty$ *and* $\lim_{v\to\infty} 2^{K_v} \log(n_r p_v)/n_r = 0$, *the tree estimator* $\{\widehat{t}_{tree}\}$ *is mean-square consistent for* t_y, *i.e.*

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{tree} - t_y\right)\right)^2\right] = 0.$$

The conditions given in Result 5.3.1 follow from the conditions of results from Klusowski (2021). The conditions on the tree predictor states that the depth of the trees should increase as the sample and population sizes increase, but not too fast with respect to the number of respondents. The assumption that the regression function is additive in its covariates is technical only.

Remark 5.3.3. *Corollary 4.3 of Klusowski (2021) holds in a high-dimensional framework as well, in which the number of covariates is allowed to increase to infinity, with "noise" variables. Interestingly, Result 5.2.1 carries over every property regarding the conditions and the convergence rate of the sequence of predictors, including high-dimensional convergence. As such, Result 5.3.1 also holds if* p_v *diverges. As such, Result 5.3.1 also proves the existence of* L^2*-consistent imputed estimators in a high-dimensional framework.*

5.4 From trees to forest estimators

5.4.1 Randomized predictors and random forests

A random forest predictor is an ensemble method based on a large collection of regression trees. Its predictions are defined as the average of the predictions of each of the trees in the forest. By noting that the prediction rules described in Examples 1.2.1 and 1.2.2 are deterministic, it is clear that, for a fixed set of elements, using the same partitioning rule to construct B trees would simply result in constructing the same tree B times. Breiman suggested (Breiman, 1996, 2001) to introduce an additional randomness in the partitioning algorithm and/or prediction rule. The additional randomness introduced in the predictors can be defined using the concept of *stochastic predictors*. Let Θ be defined in a measurable space $(J, \mathcal{J})$. A stochastic predictor $\widetilde{m}$ is a measurable function such that $\widetilde{m} : \mathbb{R}^p \times J \to \mathbb{R}$. In other words, the predictor $\widetilde{m}$ might use a random variable to make its predictions. It follows that the prediction method $\widetilde{m}$ is random with respect to Θ and, as such, an additional source of randomness is present.

Example 5.4.1. *Let* $q \in]0;1[$ *and* Θ *be a random variable with Bernoulli distribution* $\mathcal{B}(q)$; *define* $\widetilde{m}(\mathbf{x}, \Theta) := \Theta||\mathbf{x}||_2$, *where* $||\cdot||_2$ *denotes the Euclidean norm. Then,* $\widetilde{m}$ *is a stochastic prediction model, meaning that, for two different realizations of* Θ, *the prediction* $\widetilde{m}$ *may generate different values. An additional random source is present, i.e. one can show that* $\mathbb{V}_\Theta(\widetilde{m}(\mathbf{x}, \Theta)) = q(1-q)||\mathbf{x}||_2^2 > 0$.

Two additional examples of how the randomization procedure can be incorporated and used in regression trees are given below.

Example 5.4.2. *Uniform random forest (Biau et al., 2008, Scornet, 2016a).*
All the B *trees of the forest have the same behavior; as such, we describe only the behavior of a generic tree among the* B *belonging in the forest. We begin by considering* $[0;1]^p$ *as the initial leaf. Then, recursively, the algorithm splits as follows:*

1. *A node* G *is selected uniformly at random.*
2. *A splitting variable* X_j *is selected uniformly at random among the* p *auxiliary variables* $X_1, X_2, ..., X_p$.

3. *A split is performed in the node G along the axis induced by X_j with a location chosen uniformly at random.*

The process is repeated K times, with $K \in \mathbb{N}$, a parameter chosen by the user.

Example 5.4.3. *Breiman's original algorithm.*
The algorithm poceeds as follows:

Step 1: *Select B bootstrap samples from S_r denoted $\{S_r(\Theta_b)\}_{b=1}^B$.*

Step 2: *On each bootstrap sample $S_r(\Theta_b)$, fit a tree predictor $\widehat{m}(\cdot, \Theta_b)$ using the CART algorithm on $D_{n_r}(\Theta_b)$, where the CART criterion is optimized on p_0 covariates instead of p. The p_0 covariates are chosen uniformly at random (without replacement) among the p covariates available, at each split, according to Θ_b.*

Uniform random forests are mostly studied in the literature because the partitions of its trees are independent of the observed data, thus making their theoretical analysis simpler. However, because they do not use the data for building the partitions, they are of little practical interest. In practice, Breiman's original algorithm is typically used, but its theoretical analysis is more complicated.

Generally speaking, random forest predictions can be obtained as follows. Let $\{\Theta_b\}_{b=1}^B$ denote a sequence of i.i.d. random variables distributed according to some generic random variable Θ and assumed to be independent of the observed data. Let $\{\widehat{m}_{tree}(\cdot, \Theta_b)\}_{b=1}^B$ be a sequence of randomized tree predictors. Then, the RF prediction at $\mathbf{x}$ is given by

$$\widehat{m}_{rf}\left(\mathbf{x}, \{\Theta_b\}_{b=1}^B\right) = \frac{1}{B}\sum_{b=1}^B \widehat{m}_{tree}(\mathbf{x}, \Theta_b) = \frac{1}{B}\sum_{b=1}^B \sum_{k \in S_r(\Theta_b)} \frac{\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x}, \Theta_b)}}{\sum_{\ell \in S_r(\Theta_b)} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x}, \Theta_b)}} \cdot y_k, \tag{5.11}$$

where $S_r(\Theta_b) = S_r$ if there is no resampling mechanism in the forest. It follows that predictions can be also be written as

$$\widehat{m}_{rf}\left(\mathbf{x}, \{\Theta_b\}_{b=1}^B\right) = \sum_{k \in S_r} \widehat{W}_k^{(B)}(\mathbf{x}, \{\Theta_b\}_{b=1}^B) \cdot y_k,$$

with weights given by

$$\widehat{W}_k^{(B)}(\mathbf{x}, \{\Theta_b\}_{b=1}^B) = \frac{1}{B}\sum_{b=1}^B \frac{\psi_k^{(b)} \mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x}, \Theta_b)}}{\sum_{\ell \in S_r} \psi_\ell^{(b)} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x}, \Theta_b)}}, \tag{5.12}$$

with $\psi_k^{(b)}$ denoting the indicator of selection of element k in $S_r(\Theta_b)$, meaning that $\psi_k^{(b)} = 1$ if $k \in S_r(\Theta_b)$ and $\psi_k^{(b)} = 0$ otherwise. In the sequel, for ease of notation, we suppress the dependence of $\{\Theta_b\}_{b=1}^B$ on the predictor and its weight functions in the notations; we write $\widehat{m}_{rf}^{(B)}$ for the predictor, $\widehat{W}_k^{(B)}$ for the weight functions, and A_b for a node of the b-th tree. Note that $\widehat{m}_{rf}^{(B)}$ is also dependent of D_{n_r}, a dependence which is omitted in the notation for readability.

For more details about random forests and their implementation, the reader is referred to Biau and Scornet (2016) and Genuer and Poggi (2019).

5.4.2 Random forest imputation

Let $\widehat{m}_{rf}^{(B)}$ be a random forest predictor built on B trees. The random forest imputed estimator $\widehat{t}_{rf}^{(B)}$ of t_y is defined as

$$\widehat{t}_{rf}^{(B)} := \sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_m} \frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k)}{\pi_k}, \tag{5.13}$$

where $\widehat{m}_{rf}^{(B)}(x_k)$ denotes the prediction of $\widehat{m}_{rf}^{(B)}$ defined in (5.11) at the point $\mathbf{x}_k$.

We begin our analysis of $\widehat{t}_{rf}^{(B)}$ by establishing the link between forest estimators and tree estimators, described in Proposition 5.4.1.

Proposition 5.4.1. *The forest imputed estimator $\widehat{t}_{rf}^{(B)}$ defined in (5.13) is an average of (randomized) tree imputed estimators:*

$$\widehat{t}_{rf}^{(B)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{t}_{tree}^{(b)},$$

where $\widehat{t}_{tree}^{(b)}$ is the imputed estimator based on the b-th tree of the forest $\widehat{m}_{tree}^{(b)}$, that is,

$$\widehat{t}_{tree}^{(b)} = \sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_m} \frac{\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)}{\pi_k}.$$

Thus a forest estimator is an average of randomized tree estimators. Many of the properties of tree estimators are also shared by randomized tree estimators, and thus with forest estimators as well.

In terms of finite sample properties, we need to distinguish deterministic tree estimators from randomized tree estimators when there is a resampling mechanism involved. If there is not resampling mechanism, then the finite sample properties of both types of estimators are similar. Indeed, if $\psi_k = 1$ for all $k \in S_r$, then every property stated in Section 5.3.1 hold. If not, however, some properties no longer hold. This is due to the fact that the weights of a randomized tree

$$\widehat{W}_k(\mathbf{x}_\ell) = \frac{\psi_k \mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x}_\ell, \Theta)}}{\sum_{i \in S_r} \psi_i \mathbb{1}_{\mathbf{x}_i \in A(\mathbf{x}_\ell, \Theta)}}, \qquad k, \ell \in S,$$

are not always symmetrical in k, ℓ. Specifically, property iv) of Technical lemma 1 no longer holds anymore for such trees (in fact, symmetry holds only for elements $k, \ell \in \cap_{b=1}^{B} S_r(\Theta_b)$). To analyze the properties of forests, two main paths might be followed: first, through Proposition 5.3.1, thus deducing properties of forests through the properties of randomized trees; second, through the fact that a forest predictor can be written in (almost) the same way as a tree predictor, with weights defined in (5.12). As such, most proofs can be reproduced almost identically. For conciseness, we omit proofs that are obtained easily from similar arguments than those given in Section 5.3.

Proposition 5.4.2. *The forest $\widehat{t}_{rf}^{(B)}$ estimator defined in (5.13) can be written as*

$$\widehat{t}_{rf}^{(B)} = \sum_{k \in S_r} w_k^{(B)} y_k,$$

where the weights $\{w_k^{(B)}\}_{k \in S_r}$ are given by

$$w_k^{(B)} = \frac{1}{\pi_k} + \sum_{\ell \in S_m} \frac{\widehat{W}_k^{(B)}(\mathbf{x}_\ell)}{\pi_\ell} = \frac{1}{\pi_k} + \frac{1}{B} \sum_{b=1}^{B} \psi_k^{(b)} \frac{\widehat{N}_b(\mathbf{x}_k, S_m)}{N_b(\mathbf{x}_k, S_r(\Theta_b))}. \tag{5.14}$$

Similarly as for individual tree estimators, forest estimators belong to the class of linear estimators in Y if and only if each tree of the forest has the X-property; see Remark 5.3.1. The weights $\{w_k^{(B))}\}_{k \in S_r}$ share the properties of the weights $\{w_k\}_{k \in S_r}$ detailed in Proposition 5.3.2.

5.4.3 From finite to infinite forests

We now consider forests with a large number of trees. Such forests are more stable, hence easier to analyze. We begin our discussion on large forests by considering the notion of infinite forests predictors, defined as

$$\widehat{m}^{(\infty)} := \mathbb{E}\left[\widehat{m}_{rf}^{(B)} | \boldsymbol{X}, \boldsymbol{I}, \boldsymbol{r}, \boldsymbol{y}\right].$$

We emphasize that, in practice, $\widehat{m}^{(\infty)}$ cannot be computed explicitly (but can be approached, see below). It is called an infinite forest predictor because, by the strong law of large numbers, we have

$$\lim_{B\to\infty} \widehat{m}_{rf}^{(B)} = \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} \widehat{m}_{tree}^{(b)} = \widehat{m}^{(\infty)}. \qquad a.s.$$

Accordingly, define the infinite forest estimator as

$$\widehat{t}_{rf}^{(\infty)} := \sum_{k\in S_r} \frac{y_k}{\pi_k} + \sum_{k\in S_m} \frac{\widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k} = \sum_{k\in S_r} w_k^{(\infty)} y_k. \tag{5.15}$$

Using the fact that an imputed forest estimator is a average of tree imputed estimators and the strong law of large numbers, it follows that

$$\lim_{B\to\infty} \widehat{t}_{rf}^{(B)} \overset{a.s.}{=} \mathbb{E}\left[\widehat{t}_{rf}^{(B)} | \boldsymbol{X}, \boldsymbol{I}, \boldsymbol{r}, \boldsymbol{y}\right] = \widehat{t}_{rf}^{(\infty)}. \tag{5.16}$$

We see therefore that, even though the infinite forest estimator cannot be computed, there is hope to approach it with a finite forest estimator based on large B. In fact, approaching the infinite forest is of particular interest, as reveals next lemma.

Lemma 6. *Consider sequences of finite* $\{\widehat{t}_{rf}^{(B)}\}$ *and infinite* $\{\widehat{t}_{rf}^{(\infty)}\}$ *forest estimators.*

There exists C *such that*

$$0 \leqslant \mathbb{E}\left[\left(\frac{\widehat{t}_{rf}^{(B)} - t_y}{N_v}\right)^2\right] - \mathbb{E}\left[\left(\frac{\widehat{t}_{rf}^{(\infty)} - t_y}{N_v}\right)^2\right] \leqslant \frac{C}{B}.$$

We also obtain that

$$\frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{rf}^{(B)} - t_y\right) = \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{rf}^{(\infty)} - t_y\right) + O_{\mathbb{P}}\left(\sqrt{\frac{n_v}{B}}\right).$$

By Lemma 6, we see that the mean squared error of infinite forest is always lower of equal to the mean squared error of finite forest. As a consequence, it follows that infinite forests are more efficient than finite forests. Lemma 6 also reveals that the difference between the two errors is bounded, and even decreases to 0 if B diverges. The following proposition postulates that, with high probability, the finite forest estimator can be made arbitrarily close to the (unknown) infinite forests. Stability is also recovered.

Proposition 5.4.3. *Fix $B \in \mathbb{N}$ and $\epsilon > 0$. The probability (with respect to $\mathbb{P}_\Theta$) that the finite forest estimator is not in an ϵ-neighborhood of the infinite forest estimator is bounded by*

$$\mathbb{P}_\Theta\left(|\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)}| > \epsilon\right) \leqslant 2\exp\left(\frac{-B\epsilon^2}{2n_m^2\left(\frac{\sup_{\omega\in\Omega_Y} Y(\omega) - \inf_{\omega\in\Omega_Y} Y(\omega)}{\min_{k\in U}\pi_k}\right)^2}\right), \tag{5.17}$$

where Ω_Y denotes the sample space of the random variable Y.

Obviously, when $B \to \infty$, the bound given in Proposition 5.4.3 decreases to 0, which is convergence in probability of $\widehat{t}_{rf}^{(B)}$ towards $\widehat{t}_{rf}^{(\infty)}$. This, naturally, is not surprising as both convergence in L^2 and almost sure hold, as stated for instance in (5.16). However, the bound (5.17) provides the quantitative guarantee that, with large probability, the two estimators are close. In particular, the bound (5.17) can be used to choose the number of trees to be used in practical situations. The bound also illustrates the impact of the number of nonrespondents, with regards to the number of trees. A similar concentration inequality can be obtained for the estimation weights as well: for $\epsilon > 0$, we have

$$\mathbb{P}_\Theta\left(|w_k^{(B)} - w_k^{(\infty)}| > \epsilon\right) \leqslant \exp\left(\frac{-2B\epsilon^2}{d_k^2\frac{n_m^2}{n_0^2}}\right), \qquad k \in S_r.$$

5.4.4 Convergence of random forest imputed estimators

We conclude this section by two examples of L^2-consistent forest estimators. We begin by considering the case of uniform random forests, as described in Example 5.4.2, followed by Breiman's random forests. The proofs will rely on mainly the same ideas, but will require more restrictive assumptions for Breiman's random forests due to a high level of data dependency. An important part of our proof is based on the idea that the forests that we consider are, in some sense, large and stable: we will assume that, without rate requirement, the number of trees is strictly increasing: let $v_1 < v_2$ be positive integers, then the number of trees B_{v_1} in the forest predictor used to impute in S_{v_1} is strictly lower than the number of trees B_{v_2} used in imputation of S_{v_2}, i.e. $v_1 < v_2 \implies B_{v_1} < B_{v_2}$. The exact motivation of this requirement will be made clear in the proofs below.

Result 5.4.1. *Assume (H19) and (H20). Consider a sequence of uniform forest imputed estimators $\{\widehat{t}_{urf}^{(B)}\}$ described in Example 5.4.2. Assume also that:*

1. *The number of steps L_v increases as v increases such that $\lim_{v\to\infty} L_v = +\infty$ and $\lim_{v\to\infty} \frac{2^{L_v}}{n_v} = 0$.*
2. *The number of trees in the forest increases, without rate requirement, i.e. $\lim_{v\to\infty} B_v = +\infty$.*

Then, the forest estimator $\{\widehat{t}_{urf}^{(B)}\}$ is mean-square consistent for t_y, i.e.

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{urf}^{(B)} - t_y\right)\right)^2\right] = 0.$$

The conditions given in Result 5.4.1 follow from the conditions of results from Scornet (2016a).

Result 5.4.2. *Assume (H19) and (H20). Consider a sequence of Breiman's random forest imputed estimators $\{\widehat{t}_{brf}^{(B)}\}$ described in Example 5.4.3. Assume also that:*

1. *No additional split is performed in a node if it contains one element or if the maximal depth K_v is reached.*

2. *The regression function m is additive with each component bounded, i.e.*

$$m_v \in \mathcal{G}_v := \{g(\mathbf{x}) = \sum_{j=1}^{p_v} g_j(x_j),\ g_j \text{ is bounded and Borel measurable},\ j = 1, 2, ...p_v\},$$

and $||m_v||_{l_0} = o(\sqrt{K_v})$.

Then, if $\lim_{v\to\infty} K_v \left(p_v/p_{0v}\right) = +\infty$ *and* $\lim_{v\to\infty} 2^{K_v} \log\left(n_r p_v\right)/n_r = 0$, *the forest estimator* $\{\widehat{t}_{brf}^{(B)}\}$ *is mean-square consistent for* t_y, *i.e.*

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{brf}^{(B)} - t_y\right)\right)^2\right] = 0.$$

5.5 Variance estimation

It is well known that treating imputed values as observed values and using the naive variance estimator

$$\widehat{V}_{naive} := \sum_{k\in S}\sum_{\ell\in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{r_k y_k + (1-r_k)\widehat{m}_{rf}^{(B)}(\mathbf{x}_k)}{\pi_k} \frac{r_\ell y_\ell + (1-r_k)\widehat{m}_{rf}^{(B)}(\mathbf{x}_\ell)}{\pi_\ell} \tag{5.18}$$

may lead to a severe underestimation of the overall variance $\mathbb{V}(\widehat{t}_{rf}^{(B)})$. In this section, we derive two variance estimators that take into account all variation sources through the methods described in Särndal (1992) and Shao and Steel (1999); we shall call these approaches the two-phase and the reverse frameworks, respectively. For more details about variance estimation of imputed estimators in surveys, the reader is referred to Haziza and Vallée (2020). Proposition 5.5.1 illustrates how variance estimation of large forest estimators is similar to variance estimation of tree estimators. For simplicity of notation, we let $\mathbb{E}_\Theta$ and $\mathbb{V}_\Theta$ be the expectation and variance operators with respect to the random variables $\{\Theta_b\}_{b=1}^B$, conditionally on the every other random quantities.

Proposition 5.5.1. *Consider sequences of finite* $\{\widehat{t}_{rf}^{(B)}\}$ *and infinite* $\{\widehat{t}_{rf}^{(\infty)}\}$ *forest estimators. We have*

$$\mathbb{V}\left(\frac{\widehat{t}_{rf}^{(B)} - t_y}{N}\right) = \mathbb{V}\left(\frac{\widehat{t}_{rf}^{(\infty)} - t_y}{N}\right) + \mathbb{E}\left[\mathbb{V}_\Theta\left(\frac{\widehat{t}_{rf}^{(B)}}{N}\right)\right]. \tag{5.19}$$

Furthermore, there exists $C > 0$ *such that*

$$\mathbb{E}\left[\mathbb{V}_\Theta\left(N_v^{-1}\widehat{t}_{rf}^{(B)}\right)\right] \leqslant C \times \frac{n_v^2}{N_v^2 B_v}.$$

Corollary 5.5.1. *The contribution of* $\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{t}_{rf}^{(\infty)}\right)\right]$ *to the overall variance* $\mathbb{V}\left(\widehat{t}_{rf}^{(B)} - t_y\right)$ *given in (5.19) is at most of order* $O\left(\frac{n_v^2}{N_v^2} \times \frac{n_v}{B_v}\right)$.

Proposition 5.5.1 and Corollary 5.5.1 highlight that the contribution of the randomization variance can be made arbitrarily small by choosing a large value of B. More precisely, the contribution of the randomization variance is at most of order $O\left(f_v^2 \cdot n_v/B_v\right)$, which is small if either (or both): 1) the

sampling fraction f_v is small enough; 2) the number of trees is large enough. The other variance component is the variance of the infinite forest, which is, in some regards, a particular kind of regression tree. Hence, for conciseness, we describe the variance estimation procedures only for $\widehat{t}_{tree}$.

5.5.1 Variance estimation through the two-phase framework

Following Särndal (1992), we consider the decomposition

$$\widehat{t}_{tree} - t_y = \left(\widehat{t}_{tree} - \widehat{t}_\pi\right) + \left(\widehat{t}_\pi - t_y\right).$$

It follows that the overall mean-squared error of $\widehat{t}_{tree}$ can be written

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{t}_{tree} - t_y\right)^2\right] &= \mathbb{E}\left[\left(\widehat{t}_\pi - t_y\right)^2\right] + \mathbb{E}\left[\left(\widehat{t}_{tree} - \widehat{t}_\pi\right)^2\right] + 2\mathbb{E}\left[\left(\widehat{t}_{tree} - t_y\right)\left(\widehat{t}_\pi - t_y\right)\right], \\ &= V_{sam} + V_{nr} + 2V_{mix}.\end{aligned}$$

We propose, as described in Särndal (1992), to estimate these three terms separately. The first term is the sampling variance, the second corresponds the nonresponse variance and the third is a mixed component. Following the method of Beaumont and Bocci (2009), an estimator of the sampling variance is given by

$$\widehat{V}_{sam} := \widehat{V}_{naive} + \sum_{k \in S_m} d_k^2 \left(1 - \pi_k\right) \widehat{\sigma}^2,$$

where $\widehat{\sigma}^2$ is an estimator of σ^2. Usual sample variance estimators might be used on the data $\{\widehat{e}_k\}_{k \in S_r} := \{y_k - \widehat{m}_{tree}(\mathbf{x}_k)\}_{k \in S_r}$. If heteroscedasticity is suspected in these residuals, one might regress $\{\widehat{e}_k\}_{k \in S_r}$ on the covariates $\{\mathbf{x}_k\}_{k \in S_r}$ with a regression tree (or forest) to predict $\{\widehat{\sigma_k}\}_{k \in S_r}$; see e.g. Haziza and Vallée (2020) for more details about the procedure. An estimator of the nonresponse variance is given by

$$\widehat{V}_{nr} := \widehat{\sigma}^2 \sum_{k \in S} \gamma_k^2,$$

where $\gamma_k := r_k w_k - d_k$ for $k \in S$. An estimator of the mixed component is given by

$$\widehat{V}_{mix} := \sum_{k \in S} \gamma_k \left(d_k - 1\right) \widehat{\sigma}^2.$$

An estimator of the total variance is therefore given by

$$\widehat{V}_{sar} := \widehat{V}_{sam} + \widehat{V}_{nr} + 2\widehat{V}_{mix}. \tag{5.20}$$

Remark 5.5.1. *As noted by Beaumont and Bissonnette (2011) and in Haziza and Vallée (2020), the estimation of the nonresponse and mixed components is simplified when the imputation model is linear in the survey variable. However, as mentioned in Remark 5.3.1, a tree predictor (and estimator) is linear in the survey variable if and only if the partitioning algorithms has the X-property. For trees with the X-property, the overall variance of $\widehat{t}_{tree}$ is therefore taken into account by $\widehat{V}_{sar}$. However, when the partitioning algorithm does not have the X-property, a rigorous justification of the variance estimator given in (5.20) is beyond the scope of this article. The rationale behind it is based on the assumptions that $N_v^{-2}\mathbb{V}\left(\widehat{t}_{tree} - t_y\right) \approxeq N_v^{-2}\mathbb{V}\left(\widehat{t}_{tree} - t_y | \mathcal{P}\right)$ for large samples and that $\mathbb{E}\left[y_k^2 | \mathcal{P}\right] = \sigma^2 + o_{\mathbb{P}}(1)$. In words, the overall variance of $\widehat{t}_{tree}$ is taken into account by (5.20) if the influence of the variations produced by the partitions to the overall variations is asymptotically negligible. Since most splitting criteria can be shown to converge to their theoretical counterparts (see e.g., Scornet et al. (2015)), we expect these assumptions to hold. Simulations provided in Section 5.7 also seem to corroborate these assumptions.*

5.5.2 Variance estimation through the reverse framework

In the reverse framework (Fay, 1991, Shao and Steel, 1999), conditionally on the nonresponse mechanism, the variance of a regression tree imputed estimator can be decomposed as

$$\mathbb{V}\left(\widehat{t}_{tree} - t_y|\boldsymbol{r}\right) = \mathbb{E}\left[\mathbb{V}\left(\widehat{t}_{tree}|\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{X}\right)|\boldsymbol{r}\right] + \mathbb{V}\left[\mathbb{E}\left(\widehat{t}_{tree} - t_y|\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{X}\right)|\boldsymbol{r}\right] := V_1 + V_2. \tag{5.21}$$

Note that, if $\widehat{t}_{tree}$ is asymptotically unbiased, it follows that $\mathbb{V}\left(\widehat{t}_{tree} - t_y|\boldsymbol{r}\right) \approx \mathbb{V}\left(\widehat{t}_{tree} - t_y\right)$ for large samples. It is known that, for single-stage sampling designs, the contribution of V_2 to the overall variance is at most of order $O(n_v/N_v)$, see e.g., Shao and Steel (1999), Haziza and Vallée (2020). Hence, if the sampling fraction is negligible, its computation may be omitted. In what follows, we make this assumption.

Using a linearization of $\widehat{t}_{tree}$, it follows that

$$V_1 \approx \mathbb{E}\left[\mathbb{V}\left(\sum_{k\in S} d_k \xi_k |\boldsymbol{r}, \boldsymbol{X}, \boldsymbol{y}\right)\middle|\boldsymbol{r}\right],$$

where

$$\xi_k := \widetilde{m}_{tree}(\mathbf{x}_k) + r_k \cdot \frac{N\left(\mathbf{x}_k, U\right)}{N\left(\mathbf{x}_k, U_r\right)} \cdot \left(y_k - \widetilde{m}_{tree}(\mathbf{x}_k)\right), \qquad k \in S,$$

and

$$N(\mathbf{x}_k, U) := \sum_{k\in U} \mathbb{1}_{\mathbf{x}_k \in \widetilde{A}(\mathbf{x})}, \quad N(\mathbf{x}_k, U_r) := \sum_{k\in U_r} \mathbb{1}_{\mathbf{x}_k \in \widetilde{A}(\mathbf{x})}, \quad \widetilde{m}_{tree}(\mathbf{x}_k) := \sum_{k\in U} \frac{\mathbb{1}_{\mathbf{x}_k \in \widetilde{A}(\mathbf{x})}}{\sum_{\ell\in U} \mathbb{1}_{\mathbf{x}_\ell \in \widetilde{A}(\mathbf{x})}} \cdot y_k,$$

with $\widetilde{A}(\mathbf{x})$ denoting the node of the population partition containing $\mathbf{x}$ and U_r is the population of respondents. The quantities $\{\xi_k\}_{k\in S}$ are therefore unknown; they can be estimated by

$$\widehat{\xi}_k := \widehat{m}_{tree}(\mathbf{x}_k) + r_k \cdot \frac{\widehat{N}\left(\mathbf{x}_k, S\right)}{\widehat{N}\left(\mathbf{x}_k, S_r\right)} \cdot \left(y_k - \widehat{m}_{tree}(\mathbf{x}_k)\right), \qquad k \in S.$$

An estimator of V_1 is given by

$$\widehat{V}_{rev} := \sum_{k\in S}\sum_{\ell\in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\xi}_k}{\pi_k} \frac{\widehat{\xi}_\ell}{\pi_\ell}. \tag{5.22}$$

If n/N is negligible, $\widehat{V}_{rev}$ is an estimator of the total variance.

5.6 Mass imputation for data integration

In recent years, there has been a shift of paradigm in NSOs that can be explained by three main factors: (i) a dramatic decrease of response rates; (ii) increasing data collection costs; and (iii) the proliferation of nonprobabilistic data sources such as web survey panels, social media and satellite information. To meet these new challenges, NSOs face increasing pressure to utilize these convenient but often uncontrolled data sources. While such data sources provide timely data for a large number of variables and population units, they often fail to represent the target population of interest because of inherent selection biases. The integration of data from a nonprobability source to data from a probability survey is a topic that is currently being scrutinized by NSOs. The reader is referred to Beaumont and Rao (2021) and Yang and Kim (2020) for recent overviews on data integration methods in a survey sampling setting.

In this section, we show how our methodology and results can be applied to the problem of data integration. Consider a finite population U of size N. Two independent samples $S_A \subset U$ and $S_B \subset U$ are observed. On the one hand, the sample S_A of size n_A is selected from the sampling frame according to a probability sampling design $\mathcal{P}_A(\cdot)$ with first-order inclusion probabilities $\{\pi_k^{(A)}\}_{k\in U}$ known for all population units. On the other hand, the sample S_B of size n_B is a sample where the inclusion probabilities $\{\pi_k^{(B)}\}_{k\in U}$ are unknown. The survey variable is assumed to be observed only for the elements of S_B, whereas the vectors of covariates $\{\mathbf{x}_k\}_{k\in S_A}$ and $\{\mathbf{x}_k\}_{k\in S_B}$ are observed both samples. The framework is summarized in Table 15 below.

	π_k	$\mathbf{X}$	Y
S_A	Known	Observed	Unobserved
S_B	Unknown	Observed	Observed

Table 15: Summary of the data structure

Because the inclusion probabilities of the sample S_B are unknown, S_B cannot be used directly to produce reliable estimates of t_y. Moreover, a similar conclusion holds for S_A since the measurements of the survey variable are unobserved for those elements. In that framework, it is common to consider the methodology of mass imputation; that is, a model $\widehat{m}_{rf}^{(S_B)}$ is fitted on $\{(\mathbf{x}_k, y_k)\}_{k\in S_B}$ to define the following estimator of t_y:

$$\widehat{t}_{mi} := \sum_{k\in S_A} \frac{\widehat{m}_{rf}^{(S_B)}(\mathbf{x}_k)}{\pi_k^{(A)}}. \tag{5.23}$$

A similar mass imputation estimator can also be constructed with any other imputation model, including regression trees. In fact, the mass imputation estimator $\widehat{t}_{mi}$ can be viewed as an imputed estimator with $n_m = n$ and an imputation model coming from an auxiliary source. Thus, the mass imputation estimator inherits most of the properties proved for trees and forests imputed estimators. More precisely, our previous regularity conditions turn into the following in this framework.

The regularity conditions on the sampling design $\mathcal{P}_A(\cdot)$ are similar to those made for the sampling design $\mathcal{P}(\cdot)$ in the previous sections. Regarding the (unknown) sampling design $\mathcal{P}_B(\cdot)$, it is enough to assume that it is non-informative and that each element has a strictly positive probability of being in the sample S_B.

Under the assumptions mentionned in the above paragraph, the previous results of consistency hold for both tree and forest mass imputed estimators. Assuming that the sample S_B is of size much greater than S_A, i.e. $n_A/n_B \approx 0$, we suggest, using the reverse framework, the following variance estimator:

$$\widehat{\mathbb{V}}_{mi} = \sum_{k\in S_A}\sum_{\ell\in S_A} \frac{\pi_{k\ell}^{(A)} - \pi_k^{(A)}\pi_\ell^{(A)}}{\pi_{k\ell}^{(A)}} \frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k)}{\pi_k^{(A)}} \frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_\ell)}{\pi_\ell^{(A)}}. \tag{5.24}$$

5.7 Simulations

In this section, we present the results of several empirical studies to assess the behaviors of the methodologies introduced in this article. First, in Section 5.7.1, we study the empirical performances of trees and forest imputed estimators; we compare them with other state-of-the-art imputed estimators. In Section 5.7.2, we investigated the performances of the variance estimators suggested in Section 5.5. Section 5.7.3 provides the results of simulations investigating the performances of mass imputed

estimators. Section 5.7.4 focuses on the performances of the variance estimator suggested for mass imputed estimators.

5.7.1 Performances of point estimators

We generated a population U of size $N = 10\,000$ consisting of a set of covariates $X_1, X_2, ..., X_5$ and 5 survey variables. We begin by defining a matrix $\boldsymbol{Z} \in \mathbb{R}^{N \times p}$ with entries giving by $Z_{ij} \sim \mathcal{N}(5, 1)$. Next, We generated a design matrix $\boldsymbol{X} = \boldsymbol{A} + \boldsymbol{E} \in \mathbb{R}^{N \times p}$, with $\boldsymbol{A} = \mathrm{SVD}_2(\boldsymbol{Z})$ where SVD_2 denotes the rank-2 singular value decomposition operator and $\boldsymbol{E}$ with components such that:

- $E_{k\ell} \sim \mathcal{N}(0, 0.01)$ if $A_{k\ell} \leqslant Q_{0.25}$,
- $E_{k\ell} \sim \mathcal{N}(0, 0.8)$ if $Q_{0.25} < A_{k\ell} \leqslant Q_{0.5}$,
- $E_{k\ell} \sim \mathcal{N}(0, 1.6)$ if $Q_{0.5} < A_{k\ell} \leqslant Q_{0.75}$,
- $E_{k\ell} \sim \mathcal{N}(0, 2.4)$ if $Q_{0.75} < A_{k\ell}$,

where Q_α denotes the empirical quantile of order α of A_2, the second column of $\boldsymbol{A}$. The rationale behind this construction was to use a design matrix $\boldsymbol{X}$ whose colums represent the covariates, which was full rank, yet with correlations between the covariates and with an underlying structure of strata (given by the quantiles of the column A_2) with different variances in each stratum.

Using $X_1 - X_5$, we generated 5 survey variables according to:

- $Y_1 = 2 + X_1 + X_2 + X_3 + X_4 + \mathcal{N}(0, 1)$;
- $Y_2 = 2 + X_1^2 + X_2^2 + X_3^3 + X_4 + \mathcal{N}(0, 1)$;
- $Y_3 = 2 + \cos(X_1 + X_2) + \mathcal{N}(0, 1)$;
- $Y_4 = 2 + 2X_1 + 5X_2 + X_1^2 X_3^2 X_4^2 + \mathcal{N}(0, 1)$;
- $Y_5 = 2 + X_1 + 10 \exp\left(2\mathbb{1}_{X_5 > 5} - 3\mathbb{1}_{X_5 < 6}\right) + \mathcal{N}(0, 1)$.

The goal was to estimate the totals $t_{yj} := \sum_{k \in U} y_{kj}$ for $j = 1, ..., 5$, where y_{kj} denotes the measure of the survey variable Y_j for element $k \in U$. To this aim, we considered two scenarios: 1) we assumed that the underlying strata structure of the design matrix was unknown to the survey statistician; 2) the strata structure was known to the survey statistician, and thus this information could be incorporated into the sampling design. In the first scenario, we used simple random sampling without replacement of size $n = 1\,000$. In the second, we used the known strata structure to define a stratified sampling with X_2-optimal allocation. We note that the stratified sampling design was informative as correlations between the survey variables and the inclusion probabilities were between 0.3 and 0.5.

Nonresponse to $Y_1, Y_2, ..., Y_5$ was generated according to a MAR nonresponse mechanism, attributing response probabilities defined as follows

$$p_k = 0.1 + 0.8 \times \mathrm{logit}(5 - 0.25(x_{k1} + x_{k2} + x_{k3} + x_{k4})), \qquad k \in U,$$

and $r_k \sim \mathcal{B}(p_k)$. Missing values were imputed by using the following 5 different imputation procedures:

1) The imputed estimator (LR) based on linear regression.
2) The imputed estimator (CART) based on a regression tree.

3) The imputed estimator (RF) based on random forest with:

   - $B = 1000$ trees in the forests,
   - Bootstrap as resampling mechanism,
   - At least $n_0 = \lfloor n^{11/20} \rfloor$ elements in each terminal node,
   - The CART splitting criterion was optimized by selecting $m_{try} = \sqrt{p}$ + design variables with probability one at each split.

4) The imputed estimator (NN1) based on nearest neighbor imputation.

5) The imputed estimator (NN5) based on 5-nearest neighbors imputation.

A Monte-Carlo procedure of $R = 5\ 000$ iterations was used to evaluate the performances of these estimators. As a measure of bias, we used the Monte-Carlo relative bias (RB) defined as

$$RB(\widehat{t_{imp}}) = 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{(\widehat{t}^{(r)} - t_y)}{t_y}, \tag{5.25}$$

for an estimator $\hat{t}_{imp}$. The Monte-Carlo relative efficiency (RE) with respect to the Horvitz-Thompson estimator was also computed:

$$RE(\hat{t}_{imp}) = 100 \times \frac{\sum_{r=1}^{R} (\hat{t}^{(r)} - t_y)^2}{\sum_{r=1}^{R} (\hat{t}_{y\pi}^{(r)} - t_y)^2}. \tag{5.26}$$

The results in terms of relative efficiency and relative bias are reported in Table 16 and Table 17, respectively.

Survey variable	Design	LR	CART	NN1	NN5	RF
Y1	SRSWOR	108	143	131	159	118
	STRAT	113	159	130	126	124
Y2	SRSWOR	122	111	196	299	113
	STRAT	120	129	158	215	112
Y3	SRSWOR	206	168	256	339	160
	STRAT	209	153	227	244	145
Y4	SRSWOR	146	163	175	244	134
	STRAT	159	187	161	209	147
Y5	SRSWOR	139	101	188	257	103
	STRAT	144	103	172	210	104

Table 16: Relative efficiencies (%) of the imputed estimators.

Survey variable	Design	LR	CART	NN1	NN5	RF
Y1	SRSWOR	-0.0	-0.0	-0.1	-0.3	-0.00
	STRAT	0.0	0.09	-0.0	-0.1	0.1
Y2	SRSWOR	-0.1	-0.1	-2.4	-3.6	0.1
	STRAT	-0.0	1.2	-2.0	-2.9	0.6
Y3	SRSWOR	-0.0	-0.0	-1.0	-1.6	0.0
	STRAT	0.1	-0.1	-1.0	-1.4	-0.0
Y4	SRSWOR	-0.1	-0.8	-4.6	-6.7	-0.4
	STRAT	-0.3	1.0	-3.0	-5.3	1.1
Y5	SRSWOR	-0.0	-0.0	-2.7	-4.3	0.4
	STRAT	0.1	0.6	-2.4	-3.6	0.6

Table 17: Relative biases (%) of the imputed estimators.

Overall, the random forest imputed estimator stood out from other models with its good behavior, both in terms of bias and efficiency. It was followed by both the tree and linear regression estimators. Nearest neighbors estimators were less efficient and exhibited (small) negative biases in most scenarios. The conclusions are similar for both stratified and simple random sampling, as all methods behave in a similar manner in both designs. When the true relationship between the covariates and the survey variable was linear (e.g. Y_1), linear regression was the most efficient estimator. For instance, for the estimation of t_{y1}, the linear regression imputed estimator exhibited a relative efficiency of 108% for SRSWOR and 113% for STRAT, whereas the second best estimator was RF, with 118% and 124%. Thus, even in this scenario, the random forestimputed estimator was efficient and close to the performances of the linear regression imputed estimator. As expected, for non-linear relationships, linear regression was less efficient than random forest. For instance, for the estimation of Y_5, RF had relative efficiencies of 103% and 104%, where LR had relative efficiencies of 139% and 144%. We emphasize that, as noted in Dagdoug et al. (2022a), in order to make sure that unequal probability sampling designs remain uninformative with a random forest model, the design variables should be considered with high-probability at each split; this is especially true in high-dimensional scenarios.

5.7.2 Performance of variance estimators

We also investigated the performances of variance estimators. To that aim, we generated a larger design matrix $\boldsymbol{X} \in \mathbb{R}^{N \times p}$, with $N = 100\ 000$, $p = 5$ and the components X_{ij} were drawn i.i.d. from $\mathcal{N}(5, 1)$. We generated the survey variables $Y_1, ..., Y_5$ using the same relationships as in the previous section. The nonresponse mechanism was also the same. We used a regression tree imputed estimator with varying node sizes to better understand the impact of this hyper-parameter on variance estimators. We have included the following four different variance estimators in our simulations:

- The naive variance estimator (NAIVE) defined in (5.18);

- The two-phase variance estimator (SAR) defined in (5.20);
- The two-phase variance estimator (SAR-CV) with σ^2 estimated by means of a cross-validation procedure with $K = 5$-folds;
- The reverse variance estimator (REV) defined in (5.22).

For more details on the cross-validation procedure used, the reader is referred to Dagdoug et al. (2021b). A Monte-Carlo procedure of $R = 10\,000$ iterations was used to compute the Monte-Carlo relative biases of the variance estimators as well as the 95% coverage that they produce. The results are given in Table 18. For the estimation of Y_5, all variance estimators exhibited negligible biases and met the required coverage; these results are therefore omitted in Table 18.

		$n_0 = 30$		$n_0 = 90$		$n_0 = 150$	
Survey variable	Estimator	RB	Coverage	RB	Coverage	RB	Coverage
Y1	SAR	-19.05	0.921	-12.57	0.931	-10.18	0.936
	SAR_CV	9.82	0.958	6.03	0.953	4.82	0.953
	REV	-19.34	0.920	-12.90	0.930	-10.57	0.935
	NAIVE	-45.53	0.850	-55.59	0.811	-60.99	0.778
Y2	SAR	-2.21	0.946	-1.74	0.945	-0.746	0.950
	SAR_CV	2.17	0.951	3.18	0.950	10.76	0.960
	REV	-2.14	0.946	-1.74	0.945	-0.52	0.950
	NAIVE	-7.11	0.940	-12.34	0.932	-22.26	0.917
Y3	SAR	-13.86	0.928	-6.73	0.939	-6.17	0.941
	SAR_CV	13.87	0.962	10.27	0.958	6.43	0.954
	REV	-13.89	0.928	-6.87	0.938	-6.51	0.940
	NAIVE	-44.58	0.851	-54.95	0.813	-59.74	0.783
Y4	SAR	-13.25	0.925	-10.10	0.936	-7.26	0.941
	SAR_CV	7.98	0.953	7.55	0.958	7.81	0.957
	REV	-14.08	0.923	-10.79	0.935	-7.62	0.94
	NAIVE	-32.31	0.885	-48.32	0.842	-55.77	0.808

Table 18: Relative biases and coverage rates for the estimation of Y_j, for $j = 1, 2, ..., 4$.

The relative biases of the naive variance estimators were, for some scenarios, very large. For example, for the estimation of the totals of Y_1 and Y_3, the naive estimator displayed a relative bias of about 45%. Therefore, as expected, these variance estimators lead to important undercoverages. In most scenarios, the variance estimators SAR and REV behave similarly. For the estimation of Y_1, the exhibited negative biases ranging between approximately -20% and -10% which lead to coverages between 0.921-0.936. For the estimation of the totals of $Y_2, ..., Y_5$, SAR and REV were close to the 95% required coverage with more than 94% of coverage when enough elements were in the terminal nodes of the tree. We note that, for these two variance estimators, the best results were always obtained for larger values of n_0, that is, when there were many elements in the terminal nodes. However, independently of the number of elements in the terminal nodes and across all scenarios, the estimator SAR_CV was very efficient, with only negligible positive biases and coverages ranging between 95.1% and 96.1%.

5.7.3 Empirical performances of tree and forest mass imputed estimators

We generated a finite population of size $N = 10\,000$ with two sets of auxiliary variables and four survey variables. The first set of auxiliary variables consisted of $X_1 \sim \mathcal{N}(2,1)$ and $X_2 \sim \mathcal{N}(2,1)$. The second set of auxiliary variables consisted of $Z_1 \sim \mathcal{N}(0,1)$, $Z_2 \sim \text{Beta}(3,1)$, $Z_3 \sim 2 \times \text{Gamma}(3,2)$, $Z_4 \sim \text{Bernoulli}(0.7)$, and $Z_5 \sim \text{Multinomial}(0.4, 0.3, 0.3)$. Given these variables, we generated the survey variables Y_1-Y_4 according to the following models:

- $Y_1 = 2 + X_1 + X_2 + \mathcal{N}(0,1)$;
- $Y_2 = 1 + 2X_1^3 + \mathcal{N}(0, 0.5)$;
- $Y_3 = 2 + Z_1^2 + Z_2 + Z_3^2 + 1,5\mathbb{1}_{\{Z_5=1\}} + \mathcal{N}(0,1)$;
- $Y_4 = 2 + (Z_1 + Z_2 + Z_3)^2 + \mathcal{N}(0,1) + \text{Beta}(3,1)$.

To estimate the population totals for the survey variables Y_1 and Y_2, we performed 5,000 iterations of the following process: we selected a probability sample, S_A, of size $n_A = 500$, according to simple random sampling without replacement. Independently, we selected a nonprobability sample S_B, of size $n_B = 500$, as follows: we partitioned the population into two strata : Stratum 1 consisted of the units k with $x_{k1} < 2$ and Stratum 2 contained the remaining units. In Stratum 1, we selected $n_1 = 0,7 \times n_B$ units according to simple random sampling without replacement. In Stratum 2, we selected $n_2 = 0,3 \times n_B$ units according, again, to simple random sampling without replacement. For the survey variables Y_3 and Y_4, we used a similar procedure with a slight difference: the stratification was performed using the variable Z_1 instead of X_1.

We mass imputed the missing elements $\{y_k\}_{k \in S_A}$ according to the same imputation models as those described in Section 5.7.1, with the exception of the random forest algorithm for which we used $B = 500$ and $n_0 = 10$. For each imputation procedures, the imputations were obtained using the set of predictors described in Table 19.

Survey variable	Vector of explanatory variable $\mathbf{X}$ used in the working model
Y_1	$X_1 - X_2$
Y_2	$X_1 - X_2$
Y_3	$Z_1 - Z_5$
Y_4	$Z_1 - Z_5$

Table 19: Working models used.

We were interested in estimating the population total $t_{y_j} = \sum_{k \in U} y_{kj}$, $j = 1, \cdots, 4$. For each imputation procedure (i)-(v), we computed the corresponding mass imputed estimator given by (5.23). In addition, we computed the naive estimator, $\widehat{t}_{naive} = n_B^{-1} \sum_{k \in S_B} y_k$ (NAIVE).

As a measure of efficiency, we computed the relative bias given by (5.25) and the relative efficiency given by (5.26), using the unfeasible Horvitz-Thompson estimator, $\widehat{t}_\pi = \sum_{k \in S_A} y_k / \pi_{k^{(A)}}$, as the reference.

The results are displayed in Table 20. As expected, the naive estimator is considerably biased in all the scenarios with value of absolute RB ranging from 15.3% to 60.9%. This can be explained that the

Population	Criterion	NAIVE	LR	CART	NN	5NN	RF
Y_1	RB	-16.7	0.0	-0.8	-0.2	0.0	-0.7
	RE	17971	131	269	199	162	177
Y_2	RB	-60.9	7.0	-4.9	-3.4	-6.9	-4.8
	RE	12948	376	210	128	242	163
Y_3	RB	29.3	10.1	0.0	-3.1	-4.1	0.8
	RE	11251	1528	130	236	301	92.6
Y_4	RB	15.3	5.0	4.2	-6.9	-10.2	2.6
	RE	1085	299	273	273	471	111

Table 20: Monte Carlo percent relative bias (RB) and Monte Carlo efficiency (RE) of several mass imputation estimators.

participation mechanism depended on the variable X_1 (for Y_1 and Y_2) and the variable Z_1 (for Y_3 and Y_4) but neither of X_1 nor Z_1 was used in the estimation procedure. In the case of Y_1, all the procedures led to negligible biases and LR was the most efficient. We note that CART was significantly less efficient than the other procedures with a value of RE equal to 269%. For the estimation of the total of Y_2, all the procedures exhibited some moderate bias with values of absolute RB ranging from 3.4% (for NN) to 7.0%. In terms of efficiency the best procedure as NN followed by RF. The other procedures were significantly less efficient. In the case of the survey variables Y_3 and Y_4, the best procedure in terms of both RB and RE was RF. The imputation procedure was considerably inefficient. In the case of Y_4, both NN and 5NN showed a significant bias.

5.7.4 Empirical performances of variance estimators for mass imputed estimators

In this section, we investigate the performance of $\widehat{V}_{\text{rf}}$ given by (5.24) in terms of bias and coverage of normal confidence intervals. We generated a population of size $N = 50\,000$ consisting of a survey variable $Y = 0,3 + 2X + \mathcal{N}(0, 0.4)$, where $X \sim \mathcal{N}(0, 1)$. From the population, $5,000$ probability samples and nonprobability samples were selected using the setup in Section 7.2 for the survey variables Y_1 and Y_2. We used $n_A = 500; 2000$ and $n_B = 500; 2000; 10000$.

In each sample, we computed (i) the imputed estimator (5.23) based on the random forest algorithm of Breiman; (ii) the variance estimator given by (5.24); and (iii) a 95% confidence interval of the form $\widehat{t}_{mi} \pm 1.96\sqrt{\widehat{\mathbb{V}}_{rf}}$.

Table 21 reports the Monte Carlo percent relative bias of $\widehat{\mathbb{V}}_{rf}$ and the Monte Carlo coverage probability of the confidence intervals. For $n_A = 500$, the variance estimator showed a small bias. The bias decreased as the ratio n_A/n_B decreased. For $n_A/n_b = 500/10000 = 0.05$, the variance estimator was virtually unbiased. The coverage rates ranged from 94.1% to 94.7%. For $n_A = 2000$, the variance estimator

Sample size	Criterion	$n_B = 500$	$n_B = 2000$	$n_B = 10000$
$n_A = 500$	RB	-5.5	-3.6	0.0
	Coverage	94.1	94.1	94.7
$n_A = 2000$	RB	-25.1	-7.7	-1.5
	Coverage	91.5	94.5	94.5

Table 21: Coverage rates and relative biases of $\widehat{V}\text{rf}$ in percentage

was biased for $n_B = 500$ with a value of absolute RB of about 25.1%. Again, the bias decreased as n_B increased. For $n_B = 10000$, the absolute RB was approximately equal to 1.5% and the coverage rate was about 94.5%. The results suggest that the coverage rate is close to the nominal rate when $n_B \geq n_A$.

5.8 Appendix

Result 5.2.1. *Assume (H19) and (H20). Consider a sequence of predictors $\{\widehat{m}\}$ fitted on D_{n_r} and its population counterparts $\{\widetilde{m}\}$ fitted on $D_N := \{(\mathbf{x}_k, y_k)\,; k \in U\}$. If*

i) The sequence of population predictors $\{\widetilde{m}\}$ satisfies

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\widetilde{m}(\mathbf{x}) - m(\mathbf{x})\right)^2\right] = 0,$$

with a convergence rate denoted γ_v.

ii) There exists a positive constant C, independent of v, such that

$$\mathbb{E}\left\{\left(\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right)^2 \,\middle|\, \boldsymbol{r}, \boldsymbol{X}, \boldsymbol{I}\right\} \leqslant C. \qquad a.s.$$

Then, the sequence of imputed estimators $\{\widehat{t}_{\widehat{m}}\}$ is L^2-consistent with rate

$$\mathbb{E}\left[\left(\frac{1}{N_v}\left\{\widehat{t}_{\widehat{m}} - t_y\right\}\right)^2\right] = O\left(\gamma_v\right).$$

Proof. Write

$$\mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{\widehat{m}} - t_y\right)\right)^2\right] \leqslant 2\mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{\widehat{m}} - \widehat{t}_{\pi}\right)\right)^2\right] + 2\mathbb{E}\left[\left|\frac{1}{N_v^2}\left(\widehat{t}_{\pi} - t_y\right)^2\right|\right], \tag{5.27}$$

where $\widehat{t}_{\pi}$ denotes the HT estimator on complete data defined in (5.1). We turn into the second term of the right-hand side of (5.27). Write

$$\mathbb{E}\left[\left|\frac{1}{N_v^2}\left(\widehat{t}_{\pi} - t_y\right)^2\right|\right] = \mathbb{E}\left[\frac{1}{N_v^2}\sum_{k\in U}\alpha_k^2 y_k^2\right] + \mathbb{E}\left[\frac{1}{N_v^2}\sum_{\substack{k,\ell\in U\\ k\neq\ell}}\alpha_k\alpha_\ell y_k y_\ell\right]$$

where $\alpha_k := I_k \pi_k^{-1} - 1$. Under (H20) each of these term converge to zero with the rate $O(n_v^{-1})$. It remains to show that the first term of (5.27) is converging to 0 with the rate $O(\gamma_v)$.

Recall that

$$\frac{1}{N_v}\left(\widehat{t}_{\widehat{m}} - \widehat{t}_{\pi}\right) = \frac{1}{N_v}\sum_{k\in S}\left\{\frac{(1-r_k)}{\pi_k}\left(\widehat{m}(\mathbf{x}_k) - y_k\right)\right\},$$

hence

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{\widehat{m}} - \widehat{t}_{\pi}\right)\right)^2\right] \leqslant\ & 2\mathbb{E}\left[\left(\frac{1}{N_v}\sum_{k\in S}\frac{(1-r_k)}{\pi_k}\left(\widehat{m}(\mathbf{x}_k) - m(\mathbf{x}_k)\right)\right)^2\right] \\ & + 2\mathbb{E}\left[\left(\frac{1}{N_v}\sum_{k\in S}\frac{(1-r_k)}{\pi_k}\left(m(\mathbf{x}_k) - y_k\right)\right)^2\right]\end{aligned} \tag{5.28}$$

We now establish the consistency of the second term of (5.28) with the rate $O(n_v^{-1})$. Write

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{N_v}\sum_{k\in S}(1-r_k)\left(m(\mathbf{x}_k) - y_k\right)\right)^2\right] =\ & \mathbb{E}\left[\frac{1}{N_v^2}\sum_{\substack{k,\ell\in S\\ \ell\neq k}}\frac{(1-r_k)}{\pi_k}\frac{(1-r_\ell)}{\pi_l}\times\epsilon_k\epsilon_\ell\right] \\ & + \mathbb{E}\left[\frac{1}{N_v^2}\sum_{k\in S}\left(\frac{(1-r_k)}{\pi_k}\right)^2\epsilon_k^2\right]\end{aligned} \tag{5.29}$$

For the first term of (5.29), we use the law of total expectation as follows:

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{N_v^2}\sum_{\substack{k,\ell\in S\\ l\neq k}}\frac{(1-r_k)}{\pi_k}\frac{(1-r_l)}{\pi_l}\times\epsilon_k\epsilon_\ell\right] \\ &= \mathbb{E}\left[\frac{1}{N_v^2}\sum_{\substack{k,\ell\in S\\ \ell\neq k}}\frac{(1-r_k)}{\pi_k}\frac{(1-r_\ell)}{\pi_\ell}\mathbb{E}\left[\epsilon_k\epsilon_\ell\middle|\boldsymbol{X},\boldsymbol{I},\boldsymbol{r}\right]\right]\end{aligned}$$

Notice now that the random variables ϵ_k and ϵ_ℓ are independent for all $k \neq \ell$. Furthermore, recall that, for all $k \in U$, $\mathbb{E}[\epsilon_k|\boldsymbol{X}] = 0$. Thus, it follows that

$$\mathbb{E}\left[\epsilon_k\epsilon_\ell\middle|\boldsymbol{X},\boldsymbol{I},\boldsymbol{r}\right] = \mathbb{E}\left[\epsilon_k\middle|\boldsymbol{X},\boldsymbol{I},\boldsymbol{r}\right]\mathbb{E}\left[\epsilon_\ell\middle|\boldsymbol{X},\boldsymbol{I},\boldsymbol{r}\right] = 0. \tag{5.30}$$

Therefore, the first term in (5.29) is 0. For the second term, a similar derivation does not work. However, we have that

$$\mathbb{E}\left[\frac{1}{N_v^2}\sum_{k\in S}\left(\frac{(1-r_k)}{\pi_k}\right)^2\epsilon_k^2\right] \leqslant \frac{N_v}{\lambda^2 N_v^2}\max_{k\in U}\mathbb{E}\left[(m(\mathbf{x}_k) - y_k)^2\right] = O(N_v^{-1})$$

since, for all $k \in U$, $\mathbb{E}\left[\epsilon_k^2 \middle| \boldsymbol{X}, \boldsymbol{I}, \boldsymbol{r}\right] = \sigma^2 < \infty$.

It remains to show that the first term of (5.28) is $O(\gamma_v)$. Bounding arguments ensures that

$$\mathbb{E}\left[\left(\frac{1}{N_v}\sum_{k\in S}\frac{(1-r_k)}{\pi_k}\left(\widehat{m}(\mathbf{x}_k)-m(\mathbf{x}_k)\right)\right)^2\right] \leqslant \frac{n_v}{\lambda^2 N_v}\frac{1}{N_v}\sum_{k\in U}\mathbb{E}\left[\left(\widehat{m}(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2\right].$$

Now, Condition ii) implies that there exists a positive constant $C > 0$, independent of v, such that

$$\mathbb{E}\left[\left(\widehat{m}(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2 \middle| \boldsymbol{r}, \boldsymbol{X}, \boldsymbol{I}\right] \leqslant C, \qquad \text{a.s.}$$

and it follows from Condition i) and Lemma 7 that, for all $k \in U$,

$$\mathbb{E}\left[\left(\widehat{m}(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2 \middle| \boldsymbol{r}, \boldsymbol{X}, \boldsymbol{I}\right] \xrightarrow{\mathbb{P}} 0.$$

Hence, by the Lebesgues dominated convergence theorem,

$$\lim_{v\to\infty}\mathbb{E}\left[\mathbb{E}\left[\left(\widehat{m}(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2 \middle| \boldsymbol{r}, \boldsymbol{X}, \boldsymbol{I}\right]\right] = 0,$$

with the rate $O(\gamma_v)$. Moreover, $\max(\gamma_v, 1/n_v) = \gamma_v$. The result follows. ∎

Proposition 5.3.1. *The tree estimator $\widehat{t}_{tree}$ defined in (5.8) can be written as*

$$\widehat{t}_{tree} = \sum_{k\in S_r} w_k y_k,$$

where the estimation weights $\{w_k\}_{k\in S_r}$ are given by

$$w_k = \frac{1}{\pi_k} + \sum_{\ell\in S_m}\frac{\widehat{W}_k(\mathbf{x}_\ell)}{\pi_\ell} = \frac{1}{\pi_k} + \frac{\widehat{N}(\mathbf{x}_k, S_m)}{N(\mathbf{x}_k, S_r)}, \qquad k \in S_r, \tag{5.31}$$

with $\widehat{N}(\mathbf{x}_k, S_m) := \sum_{\ell\in S_m}\pi_\ell^{-1}\mathbb{1}_{\mathbf{x}_\ell\in A(\mathbf{x}_k)}$ denoting the Horvitz-Thompson estimator of the number of elements of $A(\mathbf{x}_k)$ with elements of S_m; accordingly, $N(\mathbf{x}_k, S_r) := \sum_{\ell\in S_r}\mathbb{1}_{\mathbf{x}_\ell\in A(\mathbf{x}_k)}$ is used to denote the cardinal of elements in S_r that fall in $A(\mathbf{x}_k)$.

Proof. Write

$$\begin{aligned}
\widehat{t}_{tree} &= \sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\left(\frac{1}{\pi_k}\sum_{\ell\in S_r}\widehat{W}_\ell(\mathbf{x}_k)y_\ell\right)\\
&= \sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{\ell\in S_r}\left(\sum_{k\in S_m}\frac{1}{\pi_k}\widehat{W}_\ell(\mathbf{x}_k)\right)y_\ell\\
&= \sum_{k\in S_r}\left\{\frac{1}{\pi_k} + \sum_{\ell\in S_m}\frac{\widehat{W}_k(\mathbf{x}_\ell)}{\pi_\ell}\right\}y_k.
\end{aligned}$$

To prove the last equality, see that

$$\sum_{\ell \in S_m} \frac{\widehat{W}_k(\mathbf{x}_\ell)}{\pi_\ell} \overset{(1)}{=} \frac{\sum_{\ell \in S_m} \pi_\ell^{-1} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x}_k)}}{\sum_{i \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x}_k)}} := \frac{\widehat{N}(\mathbf{x}_k, S_m)}{N(\mathbf{x}_k, S_r)},$$

where equality (1) follows from the symmetry property given in Technical lemma 1 item iv). ■

Proposition 5.3.2. *The estimation weights $\{w_k\}_{k \in S_r}$ in (5.31) have the following properties.*

i) The weights are calibrated to the population size N whenever the original weighting system $\{d_k\}_{k \in U}$ is:

$$\sum_{k \in S_r} w_k = \sum_{k \in S} d_k := \widehat{N}.$$

ii) If there are at least n_0 elements in each node, the weights $\{w_k\}_{k \in S_r}$ are bounded,

$$d_k \leqslant w_k \leqslant d_k \left(1 + \frac{n_m}{n_0}\right), \quad a.s. \qquad k \in S_r.$$

The bounds are sharp, i.e. each of the bounds can be attained.

Proof. For i), recall that, for all $\mathbf{x} \in \mathbb{R}^p$, we have $\sum_{k \in S_r} W_k(\mathbf{x}) = 1$, as stated in Technical lemma 1 item i). Thus,

$$\sum_{k \in S_r} w_k = \sum_{k \in S_r} \frac{1}{\pi_k} + \sum_{\ell \in S_m} \frac{\sum_{k \in S_r} \widehat{W}_k(\mathbf{x}_\ell)}{\pi_\ell} = \sum_{k \in S_r} \frac{1}{\pi_k} + \sum_{\ell \in S_m} \frac{1}{\pi_\ell} = \sum_{k \in S} \frac{1}{\pi_k}.$$

The bounds follow as a consequence of Technical lemma 1 item iii). ■

Proposition 5.3.3. *If the sampling design is such that $\pi_k = \pi$ for all $k \in U$, then $\widehat{t}_{tree}$ can be written in projection form, that is,*

$$\widehat{t}_{tree} = \sum_{k \in S} \frac{\widehat{m}_{tree}(\mathbf{x}_k)}{\pi_k}.$$

Proof. Note that

$$\widehat{t}_{tree} = \sum_{k \in S} \frac{\widehat{m}_{tree}(\mathbf{x}_k)}{\pi_k} + \sum_{k \in S_r} \frac{y_k - \widehat{m}_{tree}(\mathbf{x}_k)}{\pi_k}.$$

It is therefore enough to show that

$$\sum_{k \in S_r} \frac{y_k - \widehat{m}_{tree}(\mathbf{x}_k)}{\pi_k} = 0.$$

Hence, write

$$\begin{aligned}
\sum_{k \in S_r} \frac{y_k - \widehat{m}_{tree}(\mathbf{x}_k)}{\pi_k} &= \sum_{k \in S_r} \pi_k^{-1} \left(y_k - \sum_{\ell \in S_r} \widehat{W}_\ell(\mathbf{x}_k) y_\ell\right) \\
&= \sum_{k \in S_r} \pi_k^{-1} y_k - \sum_{k \in S_r} \pi_k^{-1} \sum_{\ell \in S_r} \widehat{W}_\ell(\mathbf{x}_k) y_\ell \\
&\overset{(2)}{=} \sum_{k \in S_r} \pi_k^{-1} y_k - \sum_{\ell \in S_r} \pi_\ell^{-1} \left(\sum_{k \in S_r} \widehat{W}_k(\mathbf{x}_\ell)\right) y_\ell
\end{aligned}$$

$$= \sum_{k \in S_r} \pi_k^{-1} y_k - \sum_{\ell \in S_r} \pi_\ell^{-1} y_\ell$$
$$= 0.$$

Equality (2) follows from the fact that $\pi_k = \pi_\ell$, for all $k, \ell \in S_r$ and the symmetry property given in Technical lemma 1 item iv). ■

Result 5.3.1. *Assume (H20) and (H19). Consider a sequence of tree imputed estimators $\{\widehat{t}_{tree}\}$ based on the CART criterion described in Example 5.3.1. Assume also that:*

1. *There is no more split in a node if there is either only one element in it or if the maximal depth K_v is reached.*

2. *The regression function m is additive and bounded, i.e.*

$$m_v \in \mathcal{G}_v := \{g(\mathbf{x}) = \sum_{j=1}^{p_v} g_j(x_j),\ g_j \text{ is bounded and Borel measurable},\ j = 1, 2, ...p_v\},$$

and $||m_v||_{l_0} = \#\{j = 1, 2, ..., p_v;\ m_j \text{ non-constant}\} = o(\sqrt{K_v})$.

Then, if $\lim_{v\to\infty} K_v = +\infty$ and $\lim_{v\to\infty} 2^{K_v} \log(n_r p_v)/n_r = 0$, the tree estimator $\{\widehat{t}_{tree}\}$ is mean-square consistent for t_y, i.e.

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{tree} - t_y\right)\right)^2\right] = 0.$$

Proof. We begin by noting that, from Corollary 4.3 of Klusowski (2021), it follows that the sequence $\{\widetilde{m}_{tree,v}\}$ of tree predictors fitted on D_{N_v} is universally consistent in L^2 for m, meaning

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\widetilde{m}_{tree}(\mathbf{x}) - m(\mathbf{x})\right)^2\right] = 0,$$

which is Condition i) of Result 5.2.1. Since we are in a framework in which Y is almost surely bounded, it follows that

$$\mathbb{E}\left\{\left(\widehat{m}_{tree}(\mathbf{x}) - m(\mathbf{x})\right)^2 \ \middle|\ \boldsymbol{r}, \boldsymbol{X}, \boldsymbol{I}\right\} \leqslant C. \qquad \text{a.s.}$$

Therefore, Condition ii) of Result 5.2.1 holds as well. Hence, Result 5.2.1 guarantees the mean-square consistency of $\{\widehat{t}_{tree}\}$ is proved. ■

Lemma 6. *Consider sequences of finite $\{\widehat{t}_{rf}^{(B)}\}$ and infinite $\{\widehat{t}_{rf}^{(\infty)}\}$ forest estimators.*

There exists C such that

$$0 \leqslant \mathbb{E}\left[\left(\frac{\widehat{t}_{rf}^{(B)} - t_y}{N_v}\right)^2\right] - \mathbb{E}\left[\left(\frac{\widehat{t}_{rf}^{(\infty)} - t_y}{N_v}\right)^2\right] \leqslant \frac{C}{B}.$$

We also obtain that

$$\frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{rf}^{(B)} - t_y\right) = \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{rf}^{(\infty)} - t_y\right) + O_{\mathbb{P}}\left(\sqrt{\frac{n_v}{B}}\right).$$

Proof. The proof essentially follows ideas described in Scornet (2016a). Write

$$\begin{aligned}\left(\widehat{t}_{rf}^{(B)} - t_y\right)^2 &= \left(\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)} + \widehat{t}_{rf}^{(\infty)} - t_y\right)^2 \\ &= \left(\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)}\right)^2 + \left(\widehat{t}_{rf}^{(\infty)} - t_y\right)^2 + 2\left(\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)}\right)\left(\widehat{t}_{rf}^{(\infty)} - t_y\right). \end{aligned} \tag{5.32}$$

Next, notice that

$$\mathbb{E}\left[\left(\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)}\right)\left(\widehat{t}_{rf}^{(\infty)} - t_y\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)}\right)\middle| \boldsymbol{r}, \boldsymbol{X}, \boldsymbol{I}, \boldsymbol{y}\right]\left(\widehat{t}_{rf}^{(\infty)} - t_y\right)\right] = 0.$$

Therefore, taking expectations on both sides of (5.32) gives

$$\mathbb{E}\left[\left(\widehat{t}_{rf}^{(B)} - t_y\right)^2\right] = \mathbb{E}\left[\left(\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)}\right)^2\right] + \mathbb{E}\left[\left(\widehat{t}_{rf}^{(\infty)} - t_y\right)^2\right], \tag{5.33}$$

so that

$$\mathbb{E}\left[\left(\widehat{t}_{rf}^{(B)} - t_y\right)^2\right] - \mathbb{E}\left[\left(\widehat{t}_{rf}^{(\infty)} - t_y\right)^2\right] = \mathbb{E}\left[\left(\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)}\right)^2\right] \geqslant 0. \tag{5.34}$$

Next, write

$$\frac{\widehat{t}_{rf}^{(B)}}{N_v} - \frac{\widehat{t}_{rf}^{(\infty)}}{N_v} = \frac{1}{N_v}\sum_{k \in S_m} \frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k},$$

so that

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\widehat{t}_{rf}^{(B)}}{N_v} - \frac{\widehat{t}_{rf}^{(\infty)}}{N_v}\right)^2\right] &= \frac{1}{N_v^2} \cdot \mathbb{E}\left[\left(\sum_{k \in S_m} \frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k}\right)^2\right] \\ &\leqslant \frac{n_v}{N_v^2} \cdot \mathbb{E}\left[\sum_{k \in S_m} \frac{\left(\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)\right)^2}{\pi_k^2}\right] \\ &\leqslant \frac{n_v N_v}{N_v^2 \lambda^2} \cdot \max_{k \in U} \mathbb{E}\left[\left(\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)\right)^2\right] \end{aligned}$$

Now, using Theorem 3.3 of Scornet (2016a), there exists a positive constant C such that, for all $k \in U$,

$$\mathbb{E}\left[\left(\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)\right)^2\right] \leqslant \frac{C}{B},$$

leading to

$$\mathbb{E}\left[\left(\frac{\widehat{t}_{rf}^{(B)}}{N_v} - \frac{\widehat{t}_{rf}^{(\infty)}}{N_v}\right)^2\right] \leqslant \frac{C n_v N_v}{N_v^2 \lambda^2 B_v} = O\left(\frac{1}{B_v}\right).$$

■

Lemma 7. *Assume (H19). Let $\{\widetilde{m}\}$ be a sequence of L^2 consistent regression function estimates and let $\{\widehat{m}\}$ be the corresponding estimates fitted on $D_{n_r} = \{(\mathbf{x}_k, y_k); k \in S_r\}$. Then, $\{\widehat{m}\}$ is such that,*

$$\mathbb{E}\left[\left(\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right)^2 \middle| \boldsymbol{X}, \boldsymbol{I}, \boldsymbol{r}\right] \xrightarrow{\mathbb{P}} 0.$$

Proof. By assumption,

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\widetilde{m}(\mathbf{x}) - m(\mathbf{x})\right)^2\right] = 0.$$

From which it follows that

$$\lim_{v\to\infty} \mathbb{E}\left[\left|\mathbb{E}\left[\left(\widetilde{m}(\mathbf{x}) - m(\mathbf{x})\right)^2 \middle| \boldsymbol{X}\right] - 0\right|\right] = 0.$$

In other words, the random variable $\mathbb{E}\left[\left(\widetilde{m}(\mathbf{x}) - m(\mathbf{x})\right)^2 \middle| \boldsymbol{X}\right] := g(\boldsymbol{X})$ converges in L^1 towards 0, which implies that $g(\boldsymbol{X}) \xrightarrow{\mathbb{P}} 0$. Furthermore, note that, under (H20) and (H19), we have almost sure equality of the two random measures $\mathbb{P}_{Y|X}$ and $\mathbb{P}_{Y|X,I,r}$. That is, the nonresponse mechanism and the sampling design are ignorable. Therefore, fixing the sample of respondents and using the equality of conditional distribution gives

$$\mathbb{E}\left[\left(\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right)^2 \middle| \boldsymbol{X}, \boldsymbol{I}, \boldsymbol{r}\right] \xrightarrow{\mathbb{P}} 0.$$

We can note that this lemma is somewhat similar to that proven in Doob (1953) for the case of almost sure convergence.

■

Proposition 5.4.3. *Fix $B \in \mathbb{N}$ and $\epsilon > 0$. The probability (with respect to $\mathbb{P}_\Theta$) that the finite forest estimator is not in an ϵ-neighborhood of the infinite forest estimator is bounded by*

$$\mathbb{P}_\Theta\left(|\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)}| > \epsilon\right) \leqslant 2\exp\left(\frac{-B\epsilon^2}{2n_m^2\left(\frac{\sup_{\omega\in\Omega_Y} Y(\omega) - \inf_{\omega\in\Omega_Y} Y(\omega)}{\min_{k\in U}\pi_k}\right)^2}\right),$$

where Ω_Y denotes the sample space of the random variable Y.

Proof. Observe that

$$\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)} = \sum_{k\in S_m} \frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k},$$

so that

$$\mathbb{P}_\Theta\left(|\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)}| > \epsilon\right) = \mathbb{P}_\Theta\left(\left|\frac{1}{B}\sum_{b=1}^{B}\left\{\sum_{k\in S_m} \frac{\widehat{m}_{tree}^{(b)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)}{\pi_k}\right\}\right| > \epsilon\right).$$

Define $\widehat{d}^{(b)} := \sum_{k\in S_m} \pi_k^{-1}\left(\widehat{m}_{tree}^{(b)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k)\right)$. Note that, given the covariates, the sample membership indicators, the survey variable and the nonresponse indicators, the sequence $\{\widehat{m}_{tree}^{(b)}\}_{b=1}^B$ is a sequence of independently and identically distributed (according to $\mathbb{P}_\Theta$) random variables. The same holds therefore for the sequence $\{\widehat{d}^{(b)}\}_{b=1}^B$. Moreover, in our framework, these are zero mean bounded

random variables. To see that, first note that $\inf_{\omega\in\Omega_Y} Y(\omega)$ and $\sup_{\omega\in\Omega_Y} Y(\omega)$ are finite constants. Hence, for all $b \in \{1, 2, ..., B\}$ and $k \in S_m$,

$$\inf_{\omega\in\Omega_Y} Y(\omega) - \sup_{\omega\in\Omega_Y} Y(\omega) \leqslant \widehat{m}_{tree}^{(b)}(\mathbf{x}_k) - \widehat{m}_{rf}^{(\infty)}(\mathbf{x}_k) \leqslant \sup_{\omega\in\Omega_Y} Y(\omega) - \inf_{\omega\in\Omega_Y} Y(\omega). \qquad a.s.$$

Therefore, noting that $\inf_{\omega\in\Omega_Y} Y(\omega) - \sup_{\omega\in\Omega_Y} Y(\omega) < 0$, it follows that

$$n_m \cdot \frac{\inf_{\omega\in\Omega_Y} Y(\omega) - \sup_{\omega\in\Omega_Y} Y(\omega)}{\min_{k\in U} \pi_k} \leqslant \widehat{d}^{(b)} \leqslant n_m \cdot \frac{\sup_{\omega\in\Omega_Y} Y(\omega) - \inf_{\omega\in\Omega_Y} Y(\omega)}{\min_{k\in U} \pi_k}, \qquad a.s.$$

Thus, for $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}_\Theta\left(|\widehat{t}_{rf}^{(B)} - \widehat{t}_{rf}^{(\infty)}| > \epsilon\right) &= \mathbb{P}_\Theta\left(\left|\sum_{b=1}^{B} \widehat{d}^{(b)}\right| > B\epsilon\right) \\ &\overset{(3)}{\leqslant} 2\exp\left(\frac{-2B\epsilon^2}{4n_m^2\left(\frac{\sup_{\omega\in\Omega_Y} Y(\omega) - \inf_{\omega\in\Omega_Y} Y(\omega)}{\min_{k\in U} \pi_k}\right)^2}\right),\end{aligned}$$

where (3) follows from Hoeffding inequality for bounded random variables. ■

Result 5.4.1. *Assume (H20) and (H19). Consider a sequence of uniform forest imputed estimators $\{\widehat{t}_{urf}^{(B)}\}$ described in Example 5.4.2. Assume also that:*

1. *The number of steps L_v increases as v increases such that $\lim_{v\to\infty} L_v = +\infty$ and $\lim_{v\to\infty} \frac{2^{L_v}}{n_v} = 0$.*
2. *The number of trees in the forest increases, without rate requirement, i.e. $\lim_{v\to\infty} B_v = +\infty$.*

Then, the forest estimator $\{\widehat{t}_{urf}^{(B)}\}$ is mean-square consistent for t_y, i.e.

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\frac{1}{N_v}\left(\widehat{t}_{urf}^{(B)} - t_y\right)\right)^2\right] = 0.$$

Proof. Similar arguments than those used in the proof of Result 5.3.1 in coordination with Corollary 1 of Scornet (2016a) leads to the consistency of the infinite forest estimator $\widehat{t}_{urf}^{(\infty)}$, that is,

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\frac{\widehat{t}_{urf}^{(\infty)} - t_y}{N_v}\right)^2\right] = 0.$$

Moreover, from Lemma 6, we have

$$0 \leqslant \mathbb{E}\left[\left(\frac{\widehat{t}_{urf}^{(B)} - t_y}{N_v}\right)^2\right] - \mathbb{E}\left[\left(\frac{\widehat{t}_{urf}^{(\infty)} - t_y}{N_v}\right)^2\right] \leqslant \frac{C}{B}.$$

Thus, if we consider large forests (i.e. with an increasing number of trees), the sequences $\mathbb{E}\left[N_v^{-2}\left(\widehat{t}_{urf}^{(B)} - t_y\right)^2\right]$ and $\mathbb{E}\left[N_v^{-2}\left(\widehat{t}_{urf}^{(\infty)} - t_y\right)^2\right]$ must have the same limit. Hence,

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\frac{\widehat{t}_{urf}^{(B)} - t_y}{N_v}\right)^2\right] = 0,$$

which concludes the proof. ■

Proposition 5.5.1. *Consider sequences of finite* $\{\widehat{t}_{rf}^{(B)}\}$ *and infinite* $\{\widehat{t}_{rf}^{(\infty)}\}$ *forest estimators. We have*

$$\mathbb{V}\left(\frac{\widehat{t}_{rf}^{(B)} - t_y}{N}\right) = \mathbb{V}\left(\frac{\widehat{t}_{rf}^{(\infty)} - t_y}{N}\right) + \mathbb{E}\left[\mathbb{V}_\Theta\left(\frac{\widehat{t}_{rf}^{(B)}}{N}\right)\right].$$

Furthermore, there exists $C > 0$ *such that*

$$\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{t}_{rf}^{(B)}\right)\right] \leqslant C \times \frac{n_v^2}{N_v^2 B_v}.$$

Proof. By the law of iterated variance,

$$\mathbb{V}\left(\widehat{t}_{rf}^{(B)} - t_y\right) = \mathbb{V}\left(\mathbb{E}_\Theta\left[\widehat{t}_{rf}^{(B)} - t_y\right]\right) + \mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{t}_{rf}^{(B)} - t_y\right)\right].$$

From (5.16), it follows that

$$\mathbb{V}\left(\widehat{t}_{rf}^{(B)} - t_y\right) = \mathbb{V}\left(\widehat{t}_{rf}^{(\infty)} - t_y\right) + \mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{t}_{rf}^{(B)} - t_y\right)\right].$$

Relation (5.19) is proved. Next, using Proposition 5.4.1, we have

$$\mathbb{V}_\Theta\left(\widehat{t}_{rf}^{(B)} - t_y\right) = \mathbb{V}_\Theta\left(\widehat{t}_{rf}^{(B)}\right) = \mathbb{V}_\Theta\left(\frac{1}{B}\sum_{b=1}^{B}\widehat{t}_{tree}^{(b)}\right) \overset{(4)}{=} \frac{1}{B}\cdot\mathbb{V}_\Theta\left(\widehat{t}_{tree}^{(1)}\right),$$

where equality (4) follows from the fact that, as detailed in the proof of Proposition 5.4.3, conditionally on everything but $\{\Theta_b\}_{b=1}^B$, $\{\widehat{t}_{tree}^{(b)}\}_{b=1}^B$ is a sequence of i.i.d. random variables. Now, for any $b \in \{1, 2, ..., B\}$,

$$\mathbb{V}_\Theta\left(\widehat{t}_{tree}^{(b)}\right) = \mathbb{V}_\Theta\left(\sum_{k\in S_m}\frac{\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)}{\pi_k}\right) \leqslant \mathbb{E}_\Theta\left[\left(\sum_{k\in S_m}\frac{\widehat{m}_{tree}^{(b)}(\mathbf{x}_k)}{\pi_k}\right)^2\right] \leqslant n^2\left(\sup_{\omega\in\Omega_Y}|Y(\omega)|\max_{k\in U}d_k\right)^2.$$

This concludes the proof. ■

Technical lemma 1. *Consider the weights of a regression tree as defined in (5.7). The following hold:*

i) If there is at least one element per terminal node, then, for all $\mathbf{x} \in \mathbb{R}^p$,

$$\sum_{k\in S_r}\widehat{W}_k(\mathbf{x}) = 1.$$

ii) The weights of the tree can be seen as a the images of a weight function from $\mathbb{R}^p \times \mathbb{R}^p$ *to* $[0; 1]$, *that is,*

$$\widehat{W}_k(\mathbf{x}_\ell) := \widehat{W}(\mathbf{x}_k, \mathbf{x}_\ell).$$

iii) If there is at least n_0 *elements per terminal node, then the range of* $\widehat{W}$ *reduces to* $[0; n_0^{-1}]$.

iv) The weight function is symmetrical in its arguments, that is, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$$\widehat{W}(\mathbf{x}, \mathbf{y}) = \widehat{W}(\mathbf{y}, \mathbf{x}).$$

Proof. For i), fix $\mathbf{x} \in \mathbb{R}^p$. Using the definition of (5.7), we have

$$\sum_{k \in S_r} \widehat{W}_k(\mathbf{x}) = \sum_{k \in S_r} \frac{\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} = \frac{\sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} = 1.$$

Point ii) follows directly from the definition. To prove, iii), write

$$\frac{\mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} \leqslant \frac{1}{\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})}} \leqslant \frac{1}{n_{0v}}$$

by noting that $\sum_{\ell \in S_r} \mathbb{1}_{\mathbf{x}_\ell \in A(\mathbf{x})} \geqslant n_0$. To see iv), let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Observe that

$$\widehat{W}(\mathbf{x}, \mathbf{y}) = \begin{cases} \dfrac{1}{\sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A(\mathbf{y})}} & \text{if } \mathbf{x} \in A(\mathbf{y}), \\ 0 & \text{otherwise.} \end{cases}$$

Noting that the conditions $\mathbf{x} \in A(\mathbf{y})$ and $\mathbf{y} \in A(\mathbf{x})$ are the same, it is enough to split cases to prove the equality. Assuming that $\mathbf{x} \in A(\mathbf{y})$, it follows that $A(\mathbf{y}) = A(\mathbf{x})$, so that

$$\widehat{W}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A(\mathbf{y})}} = \frac{1}{\sum_{k \in S_r} \mathbb{1}_{\mathbf{x}_k \in A(\mathbf{x})}} = \widehat{W}(\mathbf{y}, \mathbf{x}).$$

In cases where, $\mathbf{x} \notin A(\mathbf{y})$, then $\widehat{W}(\mathbf{x}, \mathbf{y}) = 0$ and $\widehat{W}(\mathbf{x}, \mathbf{y}) = 0$ so that the equality also holds. ∎

6 CONCLUSION AND FUTURE WORKS

In this PhD thesis, I investigated the use of statistical learning procedures for the estimation of population totals in presence of a large number of covariates. Two frameworks were considered: full response and item nonresponse.

In the first, we assumed full response of the sampled elements and we studied the behavior of model-assisted estimators in presence of a large number of covariates. Convergence rates involving conditions of the ratio of the number of covariates over the sample size were established for linear and penalized linear model-assisted estimators. We also suggested the use of random forests to face high-dimensional data and analyzed their design-based properties.

In the second, we considered a more general framework in which the sampled elements might refuse to respond, thus leading to missing data. We performed a large empirical study aiming at examining which statistical learning predictors are particularly promising for imputation, in a wide variety of scenarios. Finally, we performed an in-depth analysis of regression tree and random forest imputed estimators.

In what follows, I discuss some thoughts on the use of machine learning predictors for survey statistics, I highlight some of the limitations of the work presented in this thesis and introduce a few ideas which might serve as future works.

6.1 Some thoughts on the use of machine learning algorithms in surveys

The increasing attention for the field of statistical learning has permitted the emergence of numerous highly complex predictive models and algorithms, often called machine learning methods. As illustrated with various applications throughout this thesis, the use of such tools may be particularly useful in surveys for the estimation of finite population totals. Yet, an important question arises: *is it always wise, profitable even, to incorporate highly complex predictive models, whose efficiency sometimes relies entirely on empirical clues, into survey strategies?* It is often objected to statistical learning models that they behave like "black boxes", thus producing results that we do not really know how to explain or interpret. The articles Dagdoug et al. (2021b) and Dagdoug et al. (2022b) presented in this thesis are an attempt at shedding some light towards the intepretability of random forests in surveys. However, the intepretability of random in surveys remains partial only. More generally, additional research is required for a better intepretability of these machine learning methods in a survey framework.

Moreover, recent empirical investigations (e.g., Dagdoug et al. (2021a), Larbi et al. (2022)) showed that highly complex predictive algorithms such as Cubist (Quinlan et al., 1992), XGBoost (Chen and Guestrin, 2016), BART (Chipman et al., 2010) may be superior in many scenarios to traditional methodologies in surveys. However, these methods should be used with caution as there are still questions regarding the mechanisms that enable them to be that effective; in particular, they may not always be highly effective when applied in a survey sampling framework. For example, in the article Dagdoug et al. (2022a), simulation studies put into evidence that the model-assisted estimator based on the original random forest algorithm was biased and thus particularly inefficient in case of informative sampling designs and high-dimensional scenarios. This inefficiency was due to the fact that important design variables might not be selected by the random mechanism used in random forest algorithm used at each

split. Forcing the algorithm to consider the design variables with probability one at each split solves the problem and the random forest model-assisted estimator implemented in this way recovers its usual efficiency.

Most often, these unsatisfactory behaviors arise from unexpected interactions between statistical learning methods and survey sampling. Detecting these undesirable phenomena is therefore of great interest and there is a need for both empirical and theoretical in-depth investigations of these methodologies in a survey statistics framework. This thesis has provided answers to several issues related to the use of machine learning techniques with survey data, but there are still many questions which require further investigations. This work has also revealed new phenomena, sometimes unexpected. I discuss and illustrate below several such phenomena as well as some personal thoughts about machine learning methodologies combined with survey statistics.

Estimation in surveys is more than a problem of prediction

Studying the asymptotic properties of finite population total estimators with respect to the sampling design, as we did in Chapter 2 and Chapter 3, can be particularly insightful. Indeed, the fact that some model-assisted estimators converge in L^1 in the design space without restrictions on the number of covariates reveals that, while predictive models can improve substantially the efficiency of point estimators in surveys, *estimation of finite population parameters is not a problem of prediction.*

A particularly striking example of that statement is given by the Horvitz-Thompson estimator. Indeed, observe that the Horvitz-Thompson estimator is a model-assisted estimator based on the constant function $f = 0$, that is,

$$\widehat{t}_{ma}(f) := \sum_{k \in U} f(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - f(\mathbf{x}_k)}{\pi_k} = \sum_{k \in U} 0 + \sum_{k \in S} \frac{y_k - 0}{\pi_k} = \sum_{k \in S} \frac{y_k}{\pi_k} = \widehat{t}_{\pi}.$$

Yet, under mild conditions, the Horvitz-Thompson is L^2 consistent for the joint distribution, independently of the true regression function. Of course, it is possible to find examples of regression function for which the function $f = 0$ is a particularly poor estimate. This example illustrates that, even in a scenario in which the predictor f used in the model-assisted estimator is a mediocre estimate of the regression function, the estimator might still satisfy the usual square-root consistency.

The objectives of survey statistics and predictions are sometimes not aligned

In some cases, efficient predictions might lead to estimators less efficient than if they were based on "worse" predictions. An example of this situation is provided by the propensity score adjusted (PSA) estimator in presence of nonresponse. Assume the framework of Chapter 4 and Chapter 5, in which the observed data are $\{(\mathbf{x}_k; y_k); k \in S_r\} \cup \{\mathbf{x}_k; k \in S_m\}$ and the aim is to estimate t_y is presence of nonresponse. One possibility, studied in Chapter 3 and Chapter 4, is to use an imputed estimator to counterbalance the negative effects of nonresponse. Another possibility is to model the unknown response probabilities (Haziza, 2009) with propensity score adjusted estimator (PSA) defined as

$$\widehat{t}_{psa}(\widehat{r}) := \sum_{k \in S_r} \frac{y_k}{\pi_k \widehat{p}_k},$$

where $\widehat{p}_k := \widehat{r}(\mathbf{x}_k)$ is an estimator of the unknown response probability p_k of element k and $\widehat{r} : \mathbb{R}^p \mapsto [0;1]$ is a predictor for the response probabilities. The rationale behind the PSA estimator follows from the fact that, if the true response probabilities were known, one could use the unbiased estimator of t_y

$$\widehat{t}^{*}_{psa} := \sum_{k \in S_r} \frac{y_k}{\pi_k p_k}.$$

Since the response probabilities $\{p_k\}_{k \in S_r}$ are unknown, the predictor $\widehat{r}$ is used to estimate them. At first, it would seem that, if $\widehat{r}_1$ is more efficient for modelling the response probabilities $\{p_k\}_{k \in S_r}$ than $\widehat{r}_2$ (with respect to some criterion, e.g. mean squared error), then the estimator $\widehat{t}_{psa}(\widehat{r}_1)$ should be more effective than $\widehat{t}_{psa}(\widehat{r}_2)$ to estimate t_y, in terms of mean squared error. However, simulations performed in Larbi et al. (2022) proved that this statement might not always be true. Indeed, the "best" predictor is the predictor which uses best the information explaining how the response indicators behave, given the covariates. The "best" estimator of t_y, however, is the estimator based on the predictor of the response probabilities which best uses the information both related to the response indicators and the survey variable.

Imputation is not more difficult than regression

In the previous paragraphs, we illustrated through practical examples that the goal of survey sampling is the estimation of finite population parameters and not prediction. In some cases, these two goals are even pointing in different directions. In some other cases, however, efficient predictions lead to efficient estimation.

In Chapter 5, we established a result formalizing conditions about the predictor $\widehat{m}$ used for imputation such that, whenever satisfied, the resulting imputed estimator $\widehat{t}_{\widehat{m}}$ would be L^2-consistent for t_y with respect to the joint distribution. More precisely, if the predictor $\widehat{m}$ is L^2-consistent for the regression function m when fitted on i.i.d. data and if its L^2 risk is uniformly integrable, then the imputed estimator $\widehat{t}_{\widehat{m}}$ is L^2-consistent for t_y.

This result reveals that, in order to build a consistent imputed estimator, it is enough to use a predictor consistent for the regression function as imputation procedure. In that respect, imputation is not more difficult than regression. A parallel can be made with the problem of binary classification, in which one aims at predicting a label, say 0 or 1. In that case, it can be shown that if a predictor $\widehat{m}$ is L^2 consistent for the regression function, then the decision function $\mathbb{1}_{\widehat{m}>1/2}$ is consistent in the sense that it minimizes the Bayes risk, see Devroye et al. (2013) for definitions and details.

We emphasize, however, that the conditions that we found on $\widehat{m}$ for the consistency of $\widehat{t}_{\widehat{m}}$ towards t_y are sufficient, but probably not required for the consistency of the imputed estimators.

6.2 Open questions, extensions and future works

High-dimensional asymptotics for survey data

Contrary to the case of linear models studied in high dimensions (see Ta et al. (2020), Chauvet and Goga (2022), Dagdoug et al. (2022a)), no condition on the rate of divergence of the number of covariates is required for the design asymptotic properties of model-assisted estimators built upon tree-based methods. This phenomenon seems to be explained by two reasons: 1) as explained above, a model-assisted estimator does not need to be based on a consistent predictor (i.e. consistent for the regression function) to be consistent for t_y; 2) the "asymptotic structure" of a linear estimator (e.g. GREG) versus a tree-based estimator is different in essence. Indeed, a linear estimator can be seen as a calibrated estimator (see Deville and Särndal (1992) for details) on the p covariates. Therefore, when

$p = p_v$ is allowed to increase to infinity, it is actually a matter of imposing an ever-increasing number of calibration constraints. This is not the case for estimators built on regression trees: these can also be seen as calibrated estimators, but on the T covariates formed by the indicators of the tree nodes, and not on the p covariates. Typically, T is a function of n_0 (and of n) rather than p, and therefore the number of calibration constraints imposed on an estimator built on a tree remains fixed as p tends to infinity.

Several open questions remain in this research area. First, the conditions that we obtained in Chapter 2 for the high-dimensional consistency of model-assisted estimators may not be optimal, in the sense it might be possible to obtain weaker conditions. Whether or not these model-assisted estimators are consistent or unconsistent for t_y when the conditions that we found are not satisfied remains an open question. A more refined asymptotic analysis would be required to bring additional insight towards this question. Moreover, the asymptotic results of Chapter 2 were established for the consistency of point estimators. However, to our knowledge, the asymptotic properties of variance estimators in high-dimensional settings are yet to be determined. Furthermore, while the equivalence of asymptotic distribution holds for the random forest generalized estimator and the random forest model-assisted estimator in high-dimensional settings, a central limit theorem for the difference estimator in a high-dimensional scenario is yet to be established and might require a substantial amount of additional research. Similar research questions for the case of imputation in presence of a large number of covariates remain open as well.

Variance estimation for model-assisted estimators

In the simulation studies presented in Chapter 3, we discovered that the choice of the minimal number of elements per terminal node of the random forest model-assisted estimator is crucial for the variance estimator

$$\widehat{V}_{rf1}^{(B)} = \frac{1}{N_v^2} \sum_{k \in S_v} \sum_{\ell \in S_v} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}} \frac{y_k - \widehat{m}_{rf1}^{(B)}(\mathbf{x}_k)}{\pi_k} \frac{y_\ell - \widehat{m}_{rf1}^{(B)}(\mathbf{x}_\ell)}{\pi_\ell}$$

to be efficient. The simulations performed have shown that, if n_0 is too small, then $\widehat{V}_{rf1}^{(B)}$ might suffer from an important negative bias. We discovered that the problem encountered is more general and may happen to any model-assisted estimator based on a flexible predictor. To illustrate the issue, consider the naive predictor $\widehat{m}_{naive}$ defined as:

$$\widehat{m}_{naive} : \begin{cases} \mathbb{R}^p \longrightarrow \mathbb{R}, \\ \mathbf{x} \longmapsto y_k \mathbb{1}_{\{\mathbf{x}_k;\, k \in S\}}(\mathbf{x}). \end{cases}$$

It is easily seen that the estimated variance $\widehat{V}_{ma}(\widehat{t}_{naive})$ of $\widehat{t}_{naive}$ is zero:

$$\widehat{V}_{ma}(\widehat{t}_{naive}) = \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{y_k - \widehat{m}_{naive}(\mathbf{x}_k)}{\pi_k} \frac{y_\ell - \widehat{m}_{naive}(\mathbf{x}_\ell)}{\pi_\ell} = \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{y_k - y_k}{\pi_k} \frac{y_\ell - y_\ell}{\pi_\ell} = 0.$$

Yet, there is obviously no reason for the true variance of $\widehat{t}_{naive}$ to be zero. The problem follows from the fact that an overfitted predictor $\widehat{m}$ will produce a set of underestimated residuals $\{y_k - \widehat{m}(\mathbf{x}_k)\}_{k \in S}$, which are then used in the traditional variance estimator.

To cope with this issue, in Dagdoug et al. (2021b), we proposed to use a general *K-fold cross-validated variance estimator*, defined by the following procedure, to estimate the variance of a model-assisted estimator based on a predictor $\widehat{m}$. We randomly split the sample S into K groups $S_\kappa, \kappa = 1, \ldots, K$, of

approximately equal size. For $k \in S_\kappa$, let $\widehat{m}^{(-\kappa)}(\mathbf{x}_k)$ denote the prediction at the point $\mathbf{x}_k$ built on $S - S_\kappa$ and $\widehat{\epsilon}_k^{(-\kappa)} = y_k - \widehat{m}^{(-\kappa)}(\mathbf{x}_k)$ the associated residual. The proposed K-fold variance estimator is given by

$$\widehat{V}_{ma}^{(cv,K)} := \sum_{\kappa_1=1}^{K} \sum_{\kappa_2=1}^{K} \sum_{k \in S_{\kappa_1}} \sum_{\ell \in S_{\kappa_2}} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\epsilon}_k^{(-\kappa_1)}}{\pi_k} \frac{\widehat{\epsilon}_\ell^{(-\kappa_2)}}{\pi_\ell} = \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\epsilon}_k^{(cv)}}{\pi_k} \frac{\widehat{\epsilon}_\ell^{(cv)}}{\pi_\ell}, \tag{6.1}$$

where $\widehat{\epsilon}_k^{(cv)}$ is the uniquely defined residual for element $k \in S$. The estimator $\widehat{V}_{ma}^{(cv,K)}$ can be seen as a generalization of the estimator suggested in Opsomer and Miller (2005) for local polynomial regression. The estimator that we suggest also has connections with the jackknife variance estimators discussed in Duchesne (2000) and Valliant (2002).

Simulations suggest that, in case of random forests, the estimator $\widehat{V}_{ma}^{(cv,K)}$ is almost unbiased, independently of the minimal number of elements per terminal node. However, the theoretical properties of the estimator (6.1) are yet to be determined, further research in this area would therefore be required. Moreover, since this variance estimator is very general and can be used for the variance estimation of any model-assisted estimator, it is of interest to study the theoretical properties of (6.1) not only for the particular case of random forest, but in a general setting; i.e. to find conditions on the predictor $\widehat{m}$ which are required for the good behavior of $\widehat{V}_{ma}^{(cv,K)}$.

As suggested by Yves Tillé, another possibility could be to take into account, in the variance estimator a correction based on the degrees of freedom of the predictor used in a given model-assisted estimator. This approach has the advantage of being less computationally intensive than the cross-validated variance estimator as defined in (6.1); however, this approach also presents the drawback of having to specify the degrees of freedom of a given predictor, which might a delicate task for complex machine learning predictors.

Estimator selection in surveys with full response

The use of complex algorithms such as random forest, boosting, Cubist, neural networks or BART forces statisticians to choose a certain number of hyper-parameters. The choice of some of these parameters may highly influence the performances of the resulting estimators. More generally, as many estimators have been suggested in the literature (e.g., model-assisted, calibration, ...), the following question arises: *among a list of candidates, which estimator should be chosen, given an observed sample?* In many areas of statistics, procedures and methodologies have been developed to address this issue; for example, information criteria such as the Akaike Information Criterion (AIC, Akaike (1998)), the Bayesian Information Criterion (BIC, Schwarz (1978)) among others, for parametric models, cross-validation for statistical learning, the global Box and Jenkins methodology in time series (Box et al., 2015), to name a few. In surveys, however, there is, to our knowledge, no general methodology to be applied by practitioners. I believe that an important research area for the future in survey statistics is to elaborate a methodology that practitioners can use to guide their choice. Two possibilities come to mind for that purpose: 1) choose the best estimator $\widehat{t}^*$ among a list of possible candidates $\{\widehat{t}_1, \widehat{t}_2, ..., \widehat{t}_J\}$; 2) build an aggregate estimator $\widehat{t}_{agg}$, function of the J candidates. Both options present advantages and drawbacks. Estimator selection is particularly efficient if the estimator selected is itself efficient, but a wrong choice might lead to inefficient strategies. Aggregation has the advantage of being robust as it might use each of J candidates, but as such might suffer from additional variations. In a context of imputation, Chen and Haziza (2017) suggested using multiply robust estimators which aggregates a list of predictors to produce a robust estimator, converging towards the parameter of interest if there is one predictor in the list which is correctly specified. Simulations of a similar procedure in a context of model-assisted estimation seem to produce efficient estimators as well, but additional research in this

area is required.

From an estimator selection point of view, the cross-validated variance estimator (6.1) can also be used in that setting. Initially, it was in that setting that Opsomer and Miller (2005) proposed a similar estimator. They suggested a weighted version of the usual variance estimator, which corresponds to the estimator in (6.1) with $K = n - 1$; however, the weighted version that they suggested can be used only with linear predictors[1], whereas the variance estimator in (6.1) can be used with any predictor. The authors proposed to minimize a particular case of the cross-validated variance estimator (6.1) in order to choose the bandwidth of the local polynomial predictor. This idea can be generalized to choose not only an optimal hyper-parameter for a particular model-assisted estimator, but the optimal estimator among a list of model-assisted estimators built on different predictors. Consider, as in the previous paragraph, a (finite) list of candidates $\{\widehat{t}_1^{\alpha_1}, \widehat{t}_2^{\alpha_2}, ..., \widehat{t}_J^{\alpha_J}\}$, where $\alpha_j \in \mathbb{R}^{d_j}$ is a vector of d_j parameters to be chosen for the candidate j. This representation is rather general and includes most model-assisted estimators. For each of the candidates, we suggest choosing the optimal set of parameters solving

$$\alpha_j^* := \underset{\alpha \in \mathbb{R}^{d_j}}{\arg\min}\, \widehat{V}_{ma}^{(cv,K)}\left(\widehat{t}_1^{\alpha}\right).$$

In practice, the problem should be discretized using a grid of possible values; in case of ties, random selection could be used. We follow this procedure by selecting the best estimator among the list of candidates defined as the estimator $\widehat{t}^*$ satisfying

$$\widehat{V}_{ma}^{(cv,K)}\left(\widehat{t}^*\right) = \min\left\{\widehat{V}_{ma}^{(cv,K)}\left(\widehat{t}_1^{\alpha_j^*}\right), j = 1, 2, ..., J\right\}.$$

Simulations seem to suggest that the proposed method select the best estimator with high probability. However, rigorous proofs of that statement in a general setting would require substantial additional research, which we believe would be interesting as a future work.

Calibrated matrix completion for covariates imputation in survey sampling

As explained in the introduction, originally, the aim of this PhD thesis was to develop and investigate matrix completion algorithms for imputation in survey statistics. Eventually, other research areas were explored, although, along the way, we did investigate some of the existing algorithms and investigated their use in surveys. To be more precise, some additional notations are required.

Let $\boldsymbol{X}_S$ denote the sample restriction of $\boldsymbol{X}_U$, and consider a set of survey variables $y_1, y_2, ..., y_q$, concatenated in a population matrix $\boldsymbol{Y}_U$ with its sample restriction $\boldsymbol{Y}_S$. We denote by $\boldsymbol{S} := [\boldsymbol{X}_S, \boldsymbol{Y}_S] \in \mathbb{R}^{n \times d}$ the sampled data, where $d := p + q$. In this framework, nonresponse is allowed for both the survey variables and the covariates, and we assume to have access to the population totals of $X_1, X_2, ..., X_p$ denoted by the vector $\mathbf{t}_{\mathbf{x}}$. We denote by $\Omega_S \subset \{1, 2, ..., n\} \times \{1, 2, ..., d\}$ the set of indexes containing the elements of $\boldsymbol{S}$ which are not subject to nonresponse. That is, $z_{ij} = (i, j) \in \Omega_S$ if and only if S_{ij}, the element in the i-th row and j-th column of $\boldsymbol{S}$, is observed. Following Candès and Tao (2010), denote by $P_{\Omega_S} : \mathbb{R}^{n \times d} \mapsto \mathbb{R}^{n \times d}$ the orthogonal projection onto the subspace of rectangular matrices which vanishes outside of Ω_S; that is, for $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, the matrix $P_\Omega(\boldsymbol{A})$ has for coefficients

$$P_\Omega(\boldsymbol{A})_{ij} = \begin{cases} A_{ij} & \text{if } (i, j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

[1] A predictor $\widehat{m}$ is said to be linear if there exists a set of weights $\{w_k\}_{k \in S}$, independent of the survey variable, such that, for all $\mathbf{x}$, $\widehat{m}(\mathbf{x}) = \sum_{k \in S} w_k y_k$.

$P_{\Omega|Y}$ and $P_{\Omega|X}$ are defined similarly for the restriction of Ω to the first q columns, and the next p columns, respectively. With the projection operator above, the information available at the sample level is contained in $P_\Omega(\boldsymbol{S})$. Lastly, denote by $||\boldsymbol{A}||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} A_{ij}^2}$ the Frobenius norm of $\boldsymbol{A}$ and by $||\boldsymbol{A}||_* = \sum_{j=1}^{\min(n,d)} \sigma_j(\boldsymbol{A})$ its nuclear norm, with $\sigma_j(\boldsymbol{A})$ denoting the j-th largest singular value of $\boldsymbol{A}$.

The original matrix completion problem can be stated as

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times d}}{\text{minimize}} \ \text{rank}(\boldsymbol{Z}) \quad \text{subject to } P_{\Omega_S}(\boldsymbol{Z}) = P_{\Omega_S}(\boldsymbol{S}). \tag{6.3}$$

However, the rank minimization problem in (6.3) is not convex and is known to be NP-hard (Candès and Tao, 2010); in particular no algorithm is known to solve such a problem in a reasonable time when, say, $n \geqslant 10$. In Fazel et al. (2001), the authors considered a convex relaxation of the rank minimization problem in (6.3) by using the nuclear norm,

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times d}}{\text{minimize}} \ ||\boldsymbol{Z}||_* \quad \text{subject to } P_{\Omega_S}(\boldsymbol{Z}) = P_{\Omega_S}(\boldsymbol{S}). \tag{6.4}$$

This convex relaxation has been widely studied in the literature, see e.g. Candès and Recht (2009), Candès and Tao (2010). Several authors (e.g. Mazumder et al. (2010)) considered a problem in which it was assumed that the observed values are noisy and thus suggested solving the following problem instead:

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times d}}{\text{minimize}} \ ||\boldsymbol{Z}||_* \quad \text{subject to } ||P_{\Omega_S}(\boldsymbol{Z}) - P_{\Omega_S}(\boldsymbol{S})||_F^2 \leqslant \delta, \tag{6.5}$$

with $\delta > 0$, a given tolerance. Equivalently, (6.5) can be rewritten

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times d}}{\text{minimize}} \ ||\boldsymbol{Z}||_* + \lambda_\delta ||P_{\Omega_S}(\boldsymbol{Z}) - P_{\Omega_S}(\boldsymbol{S})||_F^2, \tag{6.6}$$

where λ_δ is a given constant, depending on δ only. Mazumder et al. (2010) suggested the soft-impute algorithm, a recursion based on iterative SVDs, to solve this problem. They have shown, among other things, that Soft-Impute is convergent for the solution of (6.6).

In a survey framework, the simulations that we performed using the Soft-Impute algorithm to recover the matrix $\boldsymbol{S}$ and estimate the totals of the survey variables $Y_1, Y_2, ..., Y_q$ seem to show that this approach leads to severely biased and inefficient total estimators. However, we found that applying the Soft-Impute algorithm to the matrix of covariates $\boldsymbol{X}_S$ instead, and then using the imputed design matrix $\widehat{\boldsymbol{X}}_S$ with traditional imputation procedures as described in Chapter 4 and Chapter 5 resulted in rather efficient estimators of t_y. Yet, this approach does not make use of the totals $\mathbf{t_x}$ assumed to be known. We thus suggest solving a *calibrated matrix completion* problem:

$$\widehat{\boldsymbol{X}}_S = \underset{\boldsymbol{Z}\in\mathbb{R}^{n\times p}}{\text{argmin}} \ ||\boldsymbol{Z}||_* + \lambda_2 ||P_{\Omega|X}(\boldsymbol{Z}) - P_{\Omega|X}(\boldsymbol{X}_S)||_F^2 \quad \text{subject to } ||HT_{\boldsymbol{\pi}}(\boldsymbol{Z}) - \mathbf{t_x}||_2^2 \leqslant \gamma, \tag{6.7}$$

for a given tolerance γ, and where $HT_{\boldsymbol{\pi}}$ denotes the Horvitz-Thompson operator $HT_{\boldsymbol{\pi}} : \mathbb{R}^{n\times p} \to \mathbb{R}^p$ defined as

$$HT_{\boldsymbol{\pi}}(\boldsymbol{Z}) = \left[\sum_{k\in S}\frac{Z_{k1}}{\pi_k}, \quad \sum_{k\in S}\frac{Z_{k2}}{\pi_k}, \quad ..., \quad \sum_{k\in S}\frac{Z_{kp}}{\pi_k}\right]. \tag{6.8}$$

This optimization problem recovers a low rank matrix, approximating $\boldsymbol{X}_S$ at the observed entries and such that, the Horvitz-Thomspon estimators of $\mathbf{t_x}$ based on $\widehat{\boldsymbol{X}}_S$ are close to be calibrated. Using techniques of Cai et al. (2010), Mazumder et al. (2010) and an iterative algorithm, we could implement an algorithm which seems to converge to the solution of (6.7), but additional research is required to

prove the properties of this algorithm. Moreover, an additional extension would be to investigate the properties of survey estimators when the completed matrix $\widehat{\boldsymbol{X}}_S$ is used as auxiliary information.

It seems to me that this approach could be a valuable extension to the work presented in this thesis; indeed, in Chapter 4 and 5, we examined the properties of imputed estimators in a framework in which a set of p covariates were assumed to be fully observed at the sample level. However, in practice, nonresponse might also be present in the covariates. As such, the methodologies detailed in 4 and 5 cannot be applied directly: an adaptation must be made. Matrix completion procedures could be an interesting possibility for this adaptation.

Bibliography

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.

Arnould, L., Boyer, C., and Scornet, E. (2020). Analyzing the tree-layer structure of deep forests. *arXiv preprint arXiv:2010.15690*.

Bardsley, P. and Chambers, R. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33:290–299.

Beaumont, J.-F. and Bissonnette, J. (2011). Variance estimation under composite imputation: The methodology behind sevani. *Survey Methodology*, 37(2):171–179.

Beaumont, J.-F. and Bocci, C. (2008). Another look at ridge calibration. *Metron-International Journal of Statistics*, LXVI:260–262.

Beaumont, J.-F. and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canad. J. Statist.*, 37:400–416.

Beaumont, J.-F. and Rao, J. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? surv. *Stat*, 83:11–22.

Berger and Tillé (2009). Sampling with unequal probabilities. In *Handbook of statistics*, volume 29, pages 39–54. Elsevier.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24:2546–2554.

Bertail, P., Chautru, E., and Clémençon, S. (2013). Empirical processes in survey sampling. Preprint HAL.

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

Biau, G. and Devroye, L. (2014). Cellular tree classifiers. In *International Conference on Algorithmic Learning Theory*. Springer.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The annals of statistics*, pages 470–482.

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

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

Boistard, H., Lopuhaä, H. P., and Ruiz-Gazen, A. (2017). Functional central limit theorems for single-stage sampling designs. *The Annals of Statistics*, 45(4):1728–1758.

Bonnéry, D. (2011). *Propriétés asymptotiques de la distribution d'un échantillon dans le cas d'un plan de sondage informatif*. PhD thesis, Université Rennes 1.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92:831–846.

Breidt, F.-J. and Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1023–1053.

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

Breiman, L. (1984). *Classification and regression trees*. Routledge.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

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

Buskirk, T. D. and Kolenikov, S. (2015). Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Methods: Insights from the Field*, pages 1–17.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080.

Cardot, H., Cénac, P., and Zitt, P.-A. (2013a). Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43.

Cardot, H., Chaouch, M., Goga, C., and Labruère, C. (2010). Properties of design-based functional principal components analysis. *J. of Statistical Planning and Inference*, 140:75–91.

Cardot, H., Dessertaine, A., Goga, C., Josserand, E., and Lardin, P. (2013b). Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption. *Survey Methodology*, 39(2):283–301.

Cardot, H., Goga, C., and Lardin, P. (2013c). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic journal of statistics*, 7:562–596.

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

Cassel, C., Särndal, C., and Wretman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63:615–620.

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

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

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of official statistics*, 16(2):113.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, 96(453):260–269.

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

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

Chen, S., Haziza, D., Léger, C., and Mashreghi, Z. (2019). Pseudo-population bootstrap methods for imputed survey data. *Biometrika*, 106(2):369–384.

Chen, T. and Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. ACM Press.

Chipman, H., George, E., and McCulloch, R. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.

Choley, F. (2018). *Deep learning with Python*. Manning.

Conti, P. L. and Mecatti, F. (2022). Resampling under complex sampling designs: Roots, development and the way forward. *Stats*, 5(1):258–269.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Creel, D. and Krotki, K. (2006). Creating imputation classes using classification tree methodology. In *In Proc. Surv. Res. Methods Sect., Am. Stat. Assoc.*, pages pp. 2884–2887.

Dagdoug, M., Goga, C., and Haziza, D. (2021a). Imputation Procedures in Surveys Using Nonparametric and Machine Learning Methods: an Empirical Comparison. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smab004.

Dagdoug, M., Goga, C., and Haziza, D. (2021b). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, pages 1–18. https://doi.org/10.1080/01621459.2021.1987250.

Dagdoug, M., Goga, C., and Haziza, D. (2022a). Model-assisted estimation in high-dimensional settings for survey data. *Journal of Applied Statistics*. https://doi.org/10.1080/02664763.2022.2047905.

Dagdoug, M., Goga, C., and Haziza, D. (2022b). Random forest imputation in surveys and application to data integration. *In work*.

De Moliner, A. and Goga, C. (2018). Sample-based estimation of mean electricity consumption curves for small domains. *Survey Methodology*, 44(2):193–214.

Deville and Tillé (2004). Efficient balanced sampling: the cube method. *Biometrika*, 91(4):893–912.

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

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Díaz-Uriarte, R. and de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3.

Dierckx, P. (1993). *Curves and Surface Fitting with Splines*. Oxford: Clarendon.

Doob, J. L. (1953). *Stochastic processes*. John Wiley & Sons, Inc., New York; Chapman & Hall, Limited, London.

Duchesne, P. (2000). A note on jackknife variance estimation for the general regression estimator. *Journal of Official Statistics*, 16(2):133.

Erdos, P. and Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 4:49–61.

Fan, Muller, and Rezucha (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*, 57(298):387–402.

Fay, R. (1991). *A design-based perspective on missing data variance*. US Census Bureau.

Fazel, M., Hindi, H., and Boyd, S. P. (2001). A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, volume 6, pages 4734–4739. IEEE.

Firth, D. and Bennett, K. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):3–21.

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., and Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, 108(1):10–19.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Fuller, W.-A. (2009a). *Introduction to statistical time series*, volume 428. John Wiley & Sons.

Fuller, W. A. (2009b). *Sampling statistics*. John Wiley & Sons.

Gelein, B. (2017). *Handling missing data with superpopulation model, design-based approach and machine learning*. PhD thesis, Université Bretagne Loire.

Genuer, R. (2012). Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562.

Genuer, R. and Poggi, J.-M. (2019). *Les forêts aléatoires avec R*. Presses universitaires de Rennes.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Giraud, C. (2021). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC.

Godambe (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):269–278.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliarie: Une approache non paramétrique par splines de régression. *Canadian Journal of Statistics*, 33(2):163–180.

Goga, C., Haziza, D., and Dagdoug, M. (2019). B-spline based imputation procedures for the treatment of item nonresponse in surveys. In work.

Goga, C. and Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society, B*, 76:113–140.

Goga, C. and Shehzad, M. A. (2010). Overview of ridge regression estimators in survey sampling. *Université de Bourgogne: Dijon, France*.

Goga, C., Shehzad, M. A., and Vanheuverzwyn, A. (2011). Principal component regression with survey data. application on the french media audience. In *Proceedings of the 58th World Statistics Congress of the International Statistical Institute, Dublin, Ireland*, pages 3847–3852.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on barro colorado island — digital soil mapping using random forests analysis. *Geoderma*, 146(1-2):102–113.

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

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523.

Hamza, M. and Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, 75(8):629–643.

Han, Q. and Wellner, J. A. (2021). Complex sampling designs: Uniform limit theorems and applications. *The Annals of Statistics*, 49(1):459–485.

Han, T., Jiang, D., Zhao, Q., Wang, L., and Yin, K. (2018). Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Transactions of the Institute of Measurement and Control*, 40(8):2681–2693.

Han, X. and Clemmensen, L. (2014). On weighted support vector regression. *Quality and Reliability Engineering International*, pages 891–903.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, New York.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C., editors, *Handbook of statistics*, volume 29A, pages 215–246. Elsevier.

Haziza, D. and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75(1):25–43.

Haziza, D. and Vallée, A.-A. (2020). Variance estimation procedures in the presence of singly imputed survey data: a critical review. *Japanese Journal of Statistics and Data Science*, 3(2):583–623.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Hoerl, A. E. and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Journal of the American Statistical Association*, 42:80–86.

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

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, 77:49–61.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2015). *An Introduction to Statistical Learning with Applications in R.* Springer Texts in Statistics.

Kane, M., Price, N., Scotch, M., and Rabinowitz, P. (2014). Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC Bioinformatics*, 15(1).

Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based machine learning methods for survey research. In *Survey Research Methods*, volume 13, pages 73–93.

Klusowski, J. M. (2021). Universal consistency of decision trees in high dimensions.

Kott, P. S. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89(426):693–696.

Krewski, D. and Rao, J. N. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, pages 1010–1019.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modelling*. Springer.

Larbi, K., Haziza, D., and Dagdoug, M. (2022). Treatment of unit nonresponse in surveys through machine learning methods: an empirical comparison. unpublished.

Lee, D., Song, J.-H., Song, S.-O., and Yoon, E. S. (2005). Weighted support vector machine for quality estimation in the polymerization process. *Ind. Eng. Chem. res.*, pages 2101–2105.

Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24:51–56.

Little, R. J. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review/Revue Internationale de Statistique*, pages 139–157.

Lohr, S., Hsu, V., and Montaquila, J. (2015). Using classification and regression trees to model survey nonresponse. In *JSM Proceedings, Survey Research Methods Section, Alexandria, VA: American Statistical Association*, pages 2071–2085.

Lohr, S. L. (2021). *Sampling: design and analysis (3rd ed.)*. Chapman and Hall/CRC.

Madow, W. G. (1948). On the limiting distributions of estimates based on samples from finite universes. *The Annals of Mathematical Statistics*, pages 535–545.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322.

McConville, K. and Breidt, F. J. (2013). Survey design asymptotics for the model-assisted penalised spline regression estimator. *Journal of Nonparametric Regression*, 25(3):745–763.

McConville, K. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2):389–413.

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

Montanari, G. and Ranalli, M. G. (2009). Multiple and ridge model calibration for sample surveys. unpublished.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration in survey sampling. *J. Amer. Statist. Assoc.*, 100:1429–1442.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3(2):169–175.

Neyman (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Nobel, A. (1996). Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105.

Opsomer, J. and Miller, C. (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *Nonparametric Statistics*, 17(5):593–611.

Opsomer, J. D., Breidt, F. J., Moisen, 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.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, pages 317–337.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. In *Handbook of statistics*, volume 29, pages 455–487. Elsevier.

Qi, Y. (2012). *Random forests for bioinformatics*, pages 307–323. Springer.

Quinlan, J. (1993). Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning*, pages 236–243.

Quinlan, J. et al. (1992). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. World Scientific.

Rao, J. and Singh, A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Robinson, P. M. and Särndal, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā Ser. B*, 45(2):240–248.

Rogez, G., Rihan, J., Ramalingam, C., Orrite, C., and Torr, P. (2008). Randomized trees for human pose detection. In *Computer Vision and Pattern Recognition, CVPR, IEEE Conference on.*, pages 1–8.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin-Bleuer, S. and Kratina, I. S. (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33(6):2789–2810.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*, volume 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

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

Särndal, C. E. (1980). On pi-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67(3):639–650.

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. (2007). The calibration approach in survey theory and practice. *Survey methodology*, 33(2):99–119.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Särndal, C.-E., Thomsen, I., Hoem, J. M., Lindley, D., Barndorff-Nielsen, O., and Dalenius, T. (1978). Design-based and model-based inference in survey sampling [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 27–52.

Särndal, C.-E. and Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian J. of Statistics*, 11:146–156.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. New York: Wiley.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Scornet, E. (2016a). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83.

Scornet, E. (2016b). Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62:1485–1500.

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, 60:144–162.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91(435):1278–1288.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94(445):254–265.

Shao, J. and Tu, D. (2012). *The jackknife and bootstrap*. Springer Science & Business Media.

Smola, A. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222.

Stekhoven, D. J. and Buhlmann, P. (2011). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC BioInformatics*, 25(8).

Ta, T., Shao, J., Li, Q., and Wang, L. (2020). Generalized regression estimators with high-dimensional covariates. *Statistica Sinica*.

Thompson, M. (1997). *Theory of sample surveys*, volume 74. CRC Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tillé (2011). Ten years of balanced sampling with the cube method: an appraisal. *Survey methodology*, 37(2):215–226.

Tillé, Y. (2020). *Sampling and estimation from finite populations*. John Wiley & Sons.

Tipton, J., Opsomer, J., and Moisen, G. (2013). Properties of endogenous post-stratified estimation using remote sensing data. *Remote sensing of environment*, 139:130–137.

Toth, D. (2021). *rpms: Recursive Partitioning for Modeling Survey Data*. R package version 0.5.1.

Toth, D. and Eltinge, J. L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106(496):1626–1636.

Valliant, R. (2002). Variance estimation for the general regression estimator. *Survey methodology*, 28(1):103–108.

Vapnik, V. (1998). *Statistical Learning Theory*. WILEY.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Springer New York.

Wager, S. (2014). Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*.

Wang, J. C. and Opsomer, J. D. (2011). On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika*, 98(1):91–106.

Wang, L. and Wang, S. (2011). Nonparametric additive modelassisted estimation for survey data. *Journal of Multivariate Analysis*, 102:1126–1140.

Wright, M. and Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*.

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

Yang, S. and Kim, J. K. (2017). Predictive mean matching imputation in survey sampling. *arXiv preprint arXiv:1703.10256*.

Yang, S. and Kim, J. K. (2019). Nearest neighbor imputation for general parameter estimation in survey sampling. In *The Econometrics of Complex Survey Data: Theory and Applications*, pages 209–234. Emerald Publishing Limited.

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *arXiv preprint arXiv:2001.03259*.

Zhong, P.-S. and Chen, S. (2014). Jackknife empirical likelihood inference with regression imputation and survey data. *Journal of Multivariate Analysis*, 129:193–205.

Zhou, S., Shen, X., and Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, 26(5):1760–1782.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Résumé. Dans cette thèse, nous considérons le problème de l'estimation de totaux en population finie en présence d'un grand nombre de variables auxiliaires. Les scénarios de réponse totale et de non-réponse partielle sont étudiés. Nous examinons les propriétés théoriques et empiriques d'estimateurs assistés par modélisation et d'estimateurs imputés, construits à partir de modèles prédictifs. Les modèles considérés sont des modèles de type machine learning réputés pour être peu sensibles au fléau de la dimension, fréquemment étudiés dans la littérature de l'apprentissage statistique.

Dans un cadre de réponse totale, nous examinons les propriétés de différents estimateurs assistés par modélisation en considérant un cadre asymptotique dans lequel le nombre de covariables tend vers l'infini. Des conditions suffisantes sont obtenues pour la convergence d'estimateurs par modélisation assistée basés sur des modèles linéaires et linéaires pénalisés tels que Ridge, Lasso ou Elastic-net. De plus, une nouvelle classe d'estimateurs des totaux par modélisation assistée basée sur des algorithmes de forêts aléatoires est suggérée. Leurs propriétés en échantillons finis et asymptotiques sont étudiées. Des estimateurs de la variance, classique et basé sur la validation croisée, sont également proposés. L'efficacité des estimateurs est testée sur des données simulées et des données réelles d'audience fournies par Médiamétrie.

En présence de nonréponse partielle, nous avons réalisé une large étude par simulation pour comparer des estimateurs imputés basés sur différents modèles prédictifs provenant de l'apprentissage statistique. Nous avons de plus étudié théoriquement les propriétés des arbres de régression et des forêts aléatoires pour l'imputation. Les propriétés en échantillons finis et asymptotiques de ces modèles ont été examinées et leur efficacité a été testée sur des simulations.

Mots-clés: Théorie des sondages; données manquantes; apprentissage statistique; statistique en grande dimension; forêts aléatoires.

Abstract. In this thesis, we consider the problem of estimating finite population totals in presence of a large number of auxiliary variables. The scenarios of full response and missing data are both investigated. To that aim, we examine the theoretical and empirical properties of model-assisted and imputed estimators based on statistical learning predictors deemed efficient in high-dimensional scenarios.

In case of full response, we examined the properties of existing model-assisted estimators in a high-dimensional asymptotic framework in which the number of covariates increases to infinity. Conditions for the convergence of model-assisted estimators based on linear and penalized linear models such as ridge, Lasso or Elastic-net are obtained. A new class of model-assisted estimators of finite population totals based on random forest algorithms is also suggested. Their finite sample and asymptotic properties are examined. We also suggested a classic and a cross-validated variance estimators. The performances of the estimators suggested are tested via a large simulation study on simulated and Médiamétrie data.

In presence of item nonresponse, we conducted a large-scale simulation study to compare imputed estimators based on many statistical learning predictors. We also investigated theoretically the use of regression trees and random forests predictors for imputation in surveys. Both their finite sample and asymptotic properties are studied and their properties are investigated by means of simulation studies.

Keywords: Survey sampling; missing data; statistical learning; high-dimensional statistics; random forests.