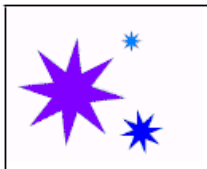
The Survey Statistician, 2024, Vol. 90, 18–29.



New and Emerging Methods

From traditional to modern machine learning estimation methods for survey sampling

Camelia Goga

Université de Franche-Comté, LMB, Besançon, France
camelia.goga@univ-fcomte.fr

Abstract

Modern parametric and nonparametric estimation methods based on machine learning are becoming increasingly popular in surveys. This paper intends presenting a synthetic review of different uses of recent modern parametric and nonparametric methods for estimating finite population totals by means of probabilistic surveys, with full data as well as with missing data.

Keywords: bagging, boosting, overfitting, penalized regression, random forests

1 Introduction

Since the first proposals for the use of probability surveys and estimation in the presence of auxiliary information in the 19th century, sample surveys have undergone significant developments. The objectives pursued, the data collected as well as their acquisition methods, and the statistical techniques used for their processing have all been deeply transformed. Over the past decade, due to the emergence of big data driven by advancements in technology and computational capabilities, we have witnessed a major transformation in this field, both in statistics in general and in sample surveys in particular, national statistical institutes being not spared from this constantly evolving reality. This paper intends giving an overview of recent machine learning methods used in survey sampling focusing on estimation and prediction issues with probability surveys.

We present in Section 2 the historical development of estimation methods in survey sampling, highlighting the major steps starting from the ratio estimator proposed during the 19th century to contemporary estimation methods, particularly non-parametric estimation methods. Section 3 focuses on modern machine learning estimation methods that have been proposed in survey sapling over the last decade. Finally, Section 4 concludes the paper and discusses new challenges associated with the use of machine learning methods in survey sampling as well as some caveats regarding the automatic use of them, including model interpretability and overfitting.

Copyright © 2024 Camelia Goga. Published by International Association of Survey Statisticians. This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

2 From parametric to traditional non-parametric estimation methods

We will consider as usual a target population U of size N. A probability sample $s \subseteq U$ is selected from U according to a sampling design $p(\cdot)$. Given $p(\cdot)$, each unit k from the population has a known inclusion probability $\pi_k = \mathbb{P}(k \in s)$ supposed to be strictly positive and a corresponding sampling design weight $d_k = 1/\pi_k$. In a survey, we are usually interested in estimating several study parameters. The simplest study parameter is the finite population total of the study variable y on U, $t_{yU} = \sum_{k \in U} y_k$. More complex study parameters such as means, ratios or quantiles as well as concentration measures (*i.e.* Gini index) may be also of interest but we devote our analysis to the finite population totals. Some discussions on more complex parameters are given in the conclusion.

With full data, the unknown total t_{yU} may be estimated by the Horvitz-Thomspon estimator (Horvitz and Thompson, 1952):

$$\hat{t}_{yd} = \sum_{k \in s} d_k y_k, \tag{1}$$

which is design unbiased, namely $\mathbb{E}_p(\hat{t}_{yd}) = t_{yU}$, where $\mathbb{E}_p(\cdot)$ is the expectation computed with respect to the sampling design $p(\cdot)$. The design-variance of $\hat{t}_{yd}$ is equal to $\mathbb{V}_p(\hat{t}_{yd}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l)(y_k/\pi_k)(y_l/\pi_l)$, where $\pi_{kl} = \mathbb{P}(k, l \in s)$ is the second-order inclusion probability of units k and l in the sample. If $\pi_{kl} > 0$ for all $k, l \in U$, then the variance $\mathbb{V}_p(\hat{t}_{yd})$ may be estimated unbiasedly by $\hat{\mathbb{V}}_p(\hat{t}_{yd}) = \sum_{k \in s} \sum_{l \in s} ((\pi_{kl} - \pi_k \pi_l)/\pi_{kl})(y_k/\pi_k)(y_l/\pi_l)$. Variance and variance estimation are important issues in the analysis of survey data, as national statistical or private institutes may desire computing confidence intervals.

2.1 First use of auxiliary information

If auxiliary information is present in the sampling frame, then it can be used to construct effective sampling strategies for the estimation of t_{yU}. When the auxiliary information is available prior to sampling, we may use it to build sampling designs, such as the stratified or balanced sampling, under which the Horvitz-Thompson estimator may be highly efficient (*i.e.* with small design variance). An alternative way is to build new estimators based on such auxiliary information and exhibiting low design variance. We focus on this paper on the second approach.
One of the first estimators to use auxiliary information was the ratio estimator (Laplace, 1814) used to estimate the total number of habitants from France in 1802:

$$\hat{t}_{yrat} = t_{xU} \frac{\hat{t}_{yd}}{\hat{t}_{xd}} = \sum_{k \in s} w_{ks} y_k, \tag{2}$$

where $\hat{t}_{xd} = \sum_{k \in s} d_k x_k$ is the Horvitz-Thomspon estimator of the x-total, $t_{xU} = \sum_{k \in U} x_k$. The ratio estimator only needs the total, t_{xU}, of the univariate x-variable on U, without needing values of x for the non-sampled individuals, which is particularly interesting when the auxiliary information is accessible only in aggregate form. In addition, y_k and x_k must be available for all $k \in s$. Laplace considered as auxiliary information the number of births, with known total thanks to the national birth registers. From (2), the ratio estimator is a weighted sum of the y-values recorded for the sampled individuals with weights $w_{ks} = d_k t_{xU}/\hat{t}_{xd}$ depending only on the x-variable and independent of the study y-variable. The ratio estimator is no longer unbiased for t_{yU}, as it is a non-linear function of finite population totals, but it can be proven that it is asymptotically unbiased under mild asymptotic assumptions (Särndal et al., 1992). It is highly efficient, namely its asymptotic variance is low, if the relationship between y and x may be modeled by a straight line through the origin with the variance around the line increasing proportionally to x. Laplace had visionary ideas since the ratio estimator is

one of the most widely used estimators for a population total and many more complicated estimators are in fact based on the ratio estimator.

2.2 Traditional linear regression based estimators

Since the ratio estimator, several estimators have been suggested in order to improve the estimation of t_{yU} under a given sampling design $p(\cdot)$ by using several auxiliary variables $x_1, \ldots, x_p$. Usually, we know only the finite population total of $\mathbf{x}_k = (x_{kj})_{j=1}^p$ denoted by $t_{\mathbf{x}U}$. With multipurpose surveys, the main goal is to derive a unique system of weights w_{ks} for each unit $k \in s$ with w_{ks} independent of the study variables, making so possible the simultaneous estimation of any linear combination of totals and other finite population parameters.There are mainly two ways to incorporate auxiliary information at the estimation stage: the *model-assisted* (Särndal et al., 1992) or the *model-based* (Valliant et al., 2000) approaches if a model is considered, and the *calibration* approach (Deville and Särndal, 1992) otherwise.

We assume that the y_k values are realizations from an infinite super-population model ξ relating y_k to the vector $\mathbf{x}_k$, as follows:

$$\xi: \quad y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \quad k \in U, \tag{3}$$

where the error terms ε_k are supposed to be independent, with zero mean, $\mathbb{E}_\xi(\varepsilon_k) = 0$ and variance $\mathbb{V}_\xi(\varepsilon_k) = v_k$. The model-assisted approach is based on the generalized difference estimator (Cassel et al., 1976):

$$t_{y,\mathbf{x}}^{\text{diff}} \;=\; \hat{t}_{yd} - (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}U})^\top \boldsymbol{\beta} = \sum_{k\in s} d_k (y_k - \mathbf{x}_k^\top \boldsymbol{\beta}) + \sum_{k\in U} \mathbf{x}_k^\top \boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\beta}$ is the true regression coefficient. It is in fact the difference between the Horvitz-Thompson estimator $\hat{t}_{yd}$ and the bias of $\hat{t}_{yd} - t_{yU}$ under the model ξ. It can be also seen as the prediction of t_{yU} under the model ξ plus a design-bias adjustment. The unknown true $\boldsymbol{\beta}$ is estimated by design-based weighted least square criterion as follows:

$$\hat{\boldsymbol{\beta}} \;=\; \arg\min_{\boldsymbol{\beta}\in\mathbf{R}^p} \sum_{k\in s} d_k v_k^{-1} (y_k - \mathbf{x}_k^\top \boldsymbol{\beta})^2. \tag{5}$$

The solution is given by $\hat{\boldsymbol{\beta}} = \left(\sum_{k\in s} d_k v_k^{-1} \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\sum_{k\in s} d_k v_k^{-1} \mathbf{x}_k y_k\right)$, assuming that the matrix $\mathbf{X}_s = (\mathbf{x}_k^\top)_{k\in s}$ is of full rank. The model-assisted estimator of t_{yU} is obtained by plugging $\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}$ in (4), we get $\hat{t}_{y,\mathbf{x}}^{\text{ma}} = \hat{t}_{yd} - (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}U})^\top \hat{\boldsymbol{\beta}} = \sum_{k\in s} d_k (y_k - \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}) + \sum_{k\in U} \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}$. For univariate x variable and variance function given by $v_k = \sigma^2 x_k, k \in U$, we get the ratio estimator (2). The widely used *poststratified* estimator of t_{yU} is obtained for $\mathbf{x}_k = (\mathbf{1}_{\{k\in U_g\}})_{g=1}^G$, where $\mathbf{1}_{\{k\in U_g\}} = 1$ if the unit k belongs to U_g and zero otherwise, $U_g, g = 1, \ldots, G$ being a partition of the population U according to some classification criterion. The variance function is supposed to be constant over the whole population, namely $v_k = v$ for all $k \in U$. The regression coefficient estimator is in this case given by $\hat{\boldsymbol{\beta}} = (\hat{\bar{y}}_g)_{g=1}^G$, where $\hat{\bar{y}}_g = \sum_{k\in s_g} d_k y_k / \hat{N}_g$, with $\hat{N}_g = \sum_{k\in s_g} d_k$ and $s_g = s \cap U_g$ for all $g = 1, \ldots, G$. The poststratified estimator reduces to the sum of the estimated predictions of y_k under the super-population model and given by $\hat{t}_{y,\mathbf{x}}^{\text{post}} = \sum_{g=1}^G N_g \hat{\bar{y}}_g$, where N_g is the size of U_g. The poststratified estimator is the sum of G ratio estimators of totals of y over the poststrata U_g, for $g = 1, \ldots, G$.
The model-based estimator of t_{yU} is built on a prediction approach,

$$t_{y,\mathbf{x}}^{\text{pred}} = \sum_{k\in s} y_k + \sum_{k\in U-s} \mathbf{x}_k^\top \boldsymbol{\beta} = \sum_{k\in s} (y_k - \mathbf{x}_k^\top \boldsymbol{\beta}) + \sum_{k\in U} \mathbf{x}_k^\top \boldsymbol{\beta}. \tag{6}$$

The unknown $\boldsymbol{\beta}$ is estimated as in (5) but without considering the sampling weights in the optimisation criterion leading to the *model-based* estimator $\hat{t}_{y,\mathbf{x}}^{\mathrm{mb}} = \sum_{k\in s}(y_k - \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}) + \sum_{k\in U} \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}$. Finally, the calibration approach consists in finding weights $(w_{ks}^{\mathrm{cal}})_{k\in s}$, such that they are as close as possible (from a pseudo distance point of view) to the sampling weights $(d_k)_{k\in s}$ while satisfying the *calibration constraints*: $\sum_{k\in s} w_{ks}^{\mathrm{cal}}\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k$. The calibration estimator is equal to the model-assisted estimator under some conditions (Deville and Särndal, 1992).

Several important properties are shared by the above three-type estimators. All estimators need only the population total of the auxiliary variables contained in the vector $t_{\mathbf{x}U}$. They may be written as weighted sums $\sum_{k\in s} w_{ks}y_k$ of sampled y-values with weights $w_{ks}, k \in s$, depending only on $\mathbf{x}_k$ recorded for sampled units and importantly, they do not depend on the y-variable. As for the ratio estimator, such weights are useful in multipurpose surveys.

2.3 Modern linear regression estimators

In the context of big data, the number p of auxiliary variables may be very large with respect to the sample size and the efficiency of estimators based on the whole set of auxiliary information may be highly deteriorated. The first issues appeared in a model-based approach since the estimators are model-dependent and many auxiliary variables may be considered to protect from model misspecification, while the model-assisted or the calibration estimators are more robust to model misspecification since they are asymptotically design unbiased and consistent for t_{yU}, whether the model is correct or not.

The weights of estimators built on the traditional linear model (3) with a large number of auxiliary variables become very instable (very large or very small) and they did not meet the predefined upper and lower range limits; they hardly satisfy a large number of calibration constraints. Finally, the design-based precision of estimators may be deteriorated when p is large with respect to the sample size, as it was noticed by Silva and Skinner (1997) by means of simulation studies and shown recently theoretically by Goga and Chauvet (2022). To correct these drawbacks, ridge-type penalized optimization criteria were suggested to relax the weight constraints (Bardsley and Chambers, 1984, Rao and Singh, 1997) leading to penalized estimators of t_{yU}. Beaumont and Bocci (2008) studied the properties of penalized calibration estimators and Guggemos and Tillé (2010) suggested a new optimisation criterion to ensure partial penalized calibration, namely a small number of important calibration equations are exactly satisfied while the other ones are approximately satisfied. As shown in Goga (2024), these penalized estimators for t_{yU} may be also obtained by considering ridge-type penalized optimization criterion to compute the regression coefficient as in classical statistics:

$$\hat{\boldsymbol{\beta}}^{\mathrm{pen}} \quad = \quad \arg\min_{\boldsymbol{\beta}\in\mathbf{R}^p} \sum_{k\in s} c_k(y_k - \mathbf{x}_k^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{7}$$

where c_k are positive constants; with $c_k = d_k$, we get the penalized calibration estimator. The solution of (7) is a ridge-type regression coefficient estimator, $\hat{\boldsymbol{\beta}}^{\mathrm{pen}} = \left(\sum_{k\in s} c_k\mathbf{x}_k\mathbf{x}_k^\top + \lambda\mathbf{I}_p\right)^{-1}\left(\sum_{k\in s} c_k\mathbf{x}_k y_k\right)$, where $\mathbf{I}_p$ is the identity matrix of size p. The ridge-type penalized estimators of t_{yU} are next obtained by plugging $\hat{\boldsymbol{\beta}}^{\mathrm{pen}}$ in (4) or (6). The resulting ridge-type estimator of t_{yU} holds the same properties as the non-penalized estimator, namely it needs only t_{xU} and it is a weighted sum of y-values, with weights not depending on the study variable. Different penalty functions in (7) lead to different penalized estimators of $\boldsymbol{\beta}$ and so, to different penalized estimators of t_{yU}. McConville et al. (2017) used the penalty $\lambda\sum_{j=1}^{p}|\beta_j|$ in (7), leading to the lasso estimator of $\boldsymbol{\beta}$ (Tibshirani, 1996) and studied the lasso-penalized estimator of t_{yU}. This penalty has the effect of shrinking the β-coefficients to zero and, unlike the ridge, it can set some of them to zero, acting as a variable selection method. Dagdoug et al. (2023b) used $\lambda[\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2]$ in (7) leading to the elastic-net estimator of $\boldsymbol{\beta}$

(Zou and Hastie, 2005), which can be viewed as a trade-off between the ridge estimator and the lasso estimator, realizing variable selection and regularization simultaneously. Alternatively, dimension reduction methods based on principal component analysis may be used to estimate $\boldsymbol{\beta}$ and to build new class of improved model-assisted or calibration estimators in presence of high-dimensional auxiliary information (Cardot et al., 2017).

The penalized estimators of t_{yU} may exhibit better efficiency than the non-penalized estimators; however, their efficiency highly depends on the tuning parameter values, which are sample data dependent. Several algorithms have been suggested to compute λ in the case of ridge regression, such as the bisection method (Beaumont and Bocci, 2008), the Fisher algorithm (Guggemos and Tillé, 2010) or simply, choosing the value for which all the weights are positive (Bardsley and Chambers, 1984, Cardot et al., 2017). More research is needed in this field.

With the emergence of smart connected objets, variables may be recorded at a very fine scale leading to another kind of high-dimensional data. The study objects are functions or curves now and called *functional data*. Cardot et al. (2013) considered the functional linear model, $y_k(t) = \mathbf{x}_k^\top \boldsymbol{\beta}(t) + \varepsilon_k(t)$, for $t \in [0, \mathcal{T}]$ and extended the model-assisted estimator to estimate the total curve of some function y over the target population. New goals and challenges appear in this new setting, such as computing global confidence bands, and Lardin-Puech et al. (2014) give a review of works related to these issues.

2.4 Traditional non-parametric model-based estimators

Estimation methods presented in sections 2.1-2.3 are all related to a linear relationship between the study variable and the auxiliary ones. Datasets are nowadays more and more complex and nonparametric models are more flexible to model the relationship between y and the x-variables:

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

where the regression function $m(\cdot)$ is unknown, but supposed to be a smooth function. Again, it was in a model-based approach that nonparametric models have been employed for the first time as a protection against model misspecification (Kuo, 1988). In model-assisted or calibration approaches, nonparametric methods have emerged later, at the beginning of the 2000's, with the seminal work of Breidt and Opsomer (2000).

With nonparametric models, we need to estimate the unknown regression function $m(\cdot)$. Traditional nonparametric methods consist in estimating $m(\cdot)$ by using kernel functions or by projecting onto a known basis function such as the truncated polynomials or the B-spline functions. Both approaches need to specify some tuning parameters, the bandwidth for kernel-based methods or the number of knots and the polynomial degree for the latter one. Once m is estimated by $\hat{m}$, the nonparametric model-assisted estimator is built from the difference estimator: $\hat{t}_{y,x}^{\text{np}} = \sum_{k\in s} d_k(y_k - \widehat{m}(\mathbf{x}_k)) + \sum_{k\in U} \widehat{m}(\mathbf{x}_k)$ and the nonparametric model-based from the prediction estimator: $\hat{t}_{y,x}^{\text{np}} = \sum_{k\in s}(y_k - \widehat{m}(\mathbf{x}_k)) + \sum_{k\in U} \widehat{m}(\mathbf{x}_k)$. As usual, the design weights are included in $\widehat{m}$ for the model-assisted case, while they are neglected for the model-based one. The nonparametric model-assisted estimators based on spline functions (Breidt et al., 2005, Goga, 2005, McConville and Breidt, 2013) inherit many desirable properties from the linear case. They may be written as traditional model-assisted estimators with explicative variables given by the basis functions, they are weighted sums of y-values with weights depending only on the x-values. However, the nonparametric estimators need x_k to be known for all the population units as we need to compute $\sum_{k\in U} \widehat{m}(\mathbf{x}_k)$. For several auxiliary variables, the additive models are the simplest way to incorporate them; for exemple, with two variables, the model is $y_k = m_1(x_{k1}) + m_1(x_{k2}) + \epsilon_k$ and $m_1(\cdot)$ and $m_2(\cdot)$ may be estimated by using one of the above method. Breidt and Opsomer (2017) give a recent review of nonparametric model-assisted

estimation techniques and Goga (2021) of B-spline nonparametric estimation methods in surveys.

In case of the traditional calibration estimator, the underlying relationship between y and $\mathbf{x}_k$ is implicitly a linear one, so it is not adapted to account for nonlinear relationships. Montanari and Ranalli (2005) suggest the *nonparametric model-calibration* estimator, which consists in finding weights satisfying $\sum_{k\in s} w_{ks}\hat{m}(x_k) = \sum_{k\in U} \hat{m}(x_k)$, while Goga (2021) suggests calibrating directly on the basis functions (B-spline functions) instead of the regression function estimator, allowing in this way to obtain weights not depending on the y-variable, property not owned by the nonparametric model-calibration.

3 Modern non-parametric estimation methods

The traditional nonparametric models based on kernel or spline smoothing are relatively easy to use and efficient if the number of auxiliary variables is low, at most three or four. As the number p of auxiliary variables is becoming large, these models tend to breakdown as they need extremely large sample sizes, phenomenon known as the curse of dimensionality. Moreover, additive models do not account for interactions between the x-variables. Semiparametric models (Breidt et al., 2007), containing linear terms as well as nonlinear terms, may be used as a tradeoff between completely parametric and nonparametric models. Alternately, the K-nearest neighbors may be used as it is a simple nonparametric method which can be used with multivariate auxiliary information (Baffeta et al., 2010).

Modern machine learning algorithms, as suggested lately in the statistical literature, are nonparametric methods which can handle easily a large number of auxiliary variables. Broadly speaking, these methods may be classified into two classes as they are based on *bagging* or on *boosting* (Hastie et al., 2011). Bagging produces a large number B of predictions and combines them to produce more accurate predictions than a single model would do:

$$\widehat{m}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \widehat{m}^{(b)}(\mathbf{x}), \tag{8}$$

where $\widehat{m}^{(b)}$ is the prediction of m obtained by some nonparametric method. Any nonparametric method may be used to obtain $\widehat{m}^{(b)}$, however bagging is particularly interesting for regression trees. To obtain B different models, initial dataset is bootstrapped (with replacement). Boosting works differently, it 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:

$$\widehat{m}^{(b)}(\mathbf{x}) = \widehat{m}^{(b-1)}(\mathbf{x}) + \widehat{m}(\mathbf{x}, \varepsilon^{(b-1)}), \quad b = 1, \ldots, B,$$

where $\widehat{m}(\mathbf{x}, \varepsilon^{(b-1)})$ is the prediction based on data $\mathbf{x}_k$ and the residuals $\varepsilon_k^{(b-1)} = y_k - \widehat{m}^{(b-1)}(\mathbf{x}_k)$ computed from the previous model. While B should be very large to improve the efficiency of bagging methods, it should be small for boosting to avoid overfitting. To cope with overfitting issues, a large value of B is considered and a penalty term is added in the boosting algorithm (Hastie et al., 2011).

3.1 Tree-based estimation methods

Regression trees based on CART algorithm as suggested by Breiman et al. (1984) are simple to use in practice and useful for interpretation. Toth and Eltinge (2011) studied the asymptotic behavior of regression trees for survey data and McConville and Toth (2019) used them in a model-assisted context. Regression tree prediction of x in some point $\mathbf{x}$ is obtained in two steps. The predictor space spanned by the x-variables measured on data is partitioned, according to some criterion, into

A_j, $j = 1, \ldots, J$ disjointed zones called *terminal nodes* and the unknown regression function in a point $\mathbf{x}$ is approximated as $m(\mathbf{x}) \simeq \beta_1 \mathbf{1}_{\{\mathbf{x} \in A_1\}} + \ldots + \beta_J \mathbf{1}_{\{\mathbf{x} \in A_J\}}$. The β-coefficients are estimated with survey data by weighted least-square criterion (McConville and Toth, 2019) leading to $\widehat{\beta}_j = \sum_{k \in A_j} d_k y_k / \hat{N}_j = \hat{\bar{y}}_j$, where $\hat{N}_j = \sum_{k \in A_j} d_k$. For every unit from A_j, tree-based predictions are the same and equal to the weighted mean of y-values of units belonging to A_j, namely $\widehat{m}(\mathbf{x}) = \widehat{\beta}_j$ for all $\mathbf{x} \in A_j$. The tree model-assisted estimator $\hat{t}_{y,\mathbf{x}}^{\text{tree}}$ of t_{yU} is obtained by plugging $\widehat{m}(\mathbf{x}_k)$ in (4). As $(A_j)_{j=1}^J$ is a partition of the predictor space, then $\sum_{k \in s} d_k (y_k - \widehat{m}(\mathbf{x}_k)) = 0$ leading to $\hat{t}_{y,\mathbf{x}}^{\text{tree}} = \sum_{k \in U} \widehat{m}(\mathbf{x}_k) = \sum_{j=1}^J N_j \hat{\bar{y}}_j$, which is a poststratified-type estimator with random poststrata A_j of size N_j built on the sample data $(\mathbf{x}_k, y_k)_{k \in s}$ (see section 2.2 for the traditional poststratified estimator).

To determine the terminal nodes, we may use the greedy CART algorithm (Breiman et al., 1984) which 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 sum of squares. More exactly, let $\mathcal{C}_A$ be the set of all possible pairs (j, z) in A and $A_L(j, z) = \{\mathbf{x}_k \in A; x_{kj} < z\}$, $A_R(j, z) = \{\mathbf{x}_k \in A; x_{kj} \geq z\}$. The best split (j^*, z^*) in a region A is $(j^*, z^*) = \arg\min_{(j,z) \in \mathcal{C}_A} \{\sum_{k \in s: \mathbf{x}_k \in A_L(j,z)} (y_k - \bar{y}_{A_L})^2 + \sum_{k \in s: \mathbf{x}_k \in A_R(j,z)} (y_k - \bar{y}_{A_R})^2\}$, where $\bar{y}_{A_L}$ (respectively $\bar{y}_{A_R}$) is the average of the y-values of units belonging to the node $A_L(j, z)$ (respectively $A_R(j, z)$). The procedure continues until a stopping criterion is reached. The random non-overlapping regions obtained by the CART algorithm depend on the sample data $(\mathbf{x}_k, y_k), k \in s$. Nalenz et al. (2024) suggest a CART criterion based on design-based estimation of the residual sum of square and Beaumont et al. (2024) adapt the CART criterion to a classification problem and data integration issues by considering well-chosen stopping criteria.

Regression trees are simply to use and interpret, however they are not appropriate with high-dimensional data and deep trees are known to have large variance and to lead to overfitting. The random forest algorithm (Breiman, 2001) is an ensemble method that corrects the tree defaults by a large number of randomized deep decorrelated trees. More exactly, the random forest prediction of $m(\cdot)$ is a bagging estimator as in (8), where each $\widehat{m}^{(b)}(\cdot)$ is a regression tree prediction of $m(\cdot)$ built on a bootstrap sample data and selecting randomly at each split in the tree a new set of p_0 auxiliary variables from the p initial variables. In this way, a fresh set of variables is considered at each tree split. Random forests are very popular methods due to their predictive performances and ability to handle large data sets, however their theoretical properties have been proved only recently for particular algorithms (Scornet et al., 2015, Klusowski and Tian, 2024). Random forest algorithms have been only recently used with survey data starting with Tipton et al. (2013), Buskirk and Kolenikov (2015) for missing data, De Moliner and Goga (2018) for small area estimation. Very recently, Dagdoug et al. (2023c) suggested the random forest model-assisted estimator and studied its asymptotical properties. The random forest model-assisted estimator may be also written as a weighted sum of y-values, however the weights depend now on the study variable as the partitions are built on sample data $(\mathbf{x}_k, y_k), k \in s$. With multipurpose surveys, the user has the choice between two options: use random forest algorithms not depending on the study variable (Devroye et al., 2013) or use a model-calibration procedure as suggested in Dagdoug et al. (2023c) to determine weights for estimating simultaneously several totals. Dagdoug et al. (2023c) use a without replacement bootstrap resampling procedure and they show that the random forest model-assisted estimator can be written as the total of the estimated prediction of $m(\cdot)$ plus a correction term equal to the weighted sum of residuals computed for the non-resampled units, also called the *out-of-bag* individuals, from each of the B trees. This correction term brings additional information from the units not used in computing the prediction and preventing in this way from overfitting. Nalenz et al. (2024) suggest bootstrapping individuals with unequal probabilities.

The random forest algorithms depend on several hyper parameters: the number B of trees, the number p_0 of the selected variables and the number n_0 of individuals from the terminal nodes, the

hyper parameter values affecting the precision of the model-assisted estimators and finding the best values of such hyper parameters may be difficult with complex sampling designs (Dagdoug et al., 2023c,b). Another issue with random forests, and even with nonparametric methods in general, is the result interpretability. These methods are known to have high predictive performances, however the predictions are difficult to interpret and this may be a problem with surveys conducted by national institutes.

3.2 Missing data

The estimators presented above supposed that all the sampled individuals respond, so we have complete sample data $y_k, k \in s$. In practice however, due to various reasons, some individuals respond only partially (*item nonresponse*) or do not respond to the survey questionnaire (*unit nonresponse*). Item nonresponse is treated by imputation while unit nonresponse is treated by weighting methods.

With item nonresponse, the imputed estimator $\hat{t}_I$ of t_{yU} is obtained from the Horvitz-Thompson estimator given in (1) by replacing the missing values y_k by predicted values $\hat{y}_k$, $\hat{t}_I = \sum_{k\in s_r} d_k y_k + \sum_{k\in s_m} d_k \hat{y}_k$, where s_r is the respondent subset and s_m, the subset of s containing the nonrespondents. The imputed values are obtained by fitting an imputation model, assuming usually that the response mechanism is MAR (*missing at random*). Recently, Dagdoug et al. (2023a) conduct a large simulation study of parametric as well as nonparametric and machine-learning imputation procedures in terms of bias and efficiency in a wide variety of settings, including high-dimensional data sets. They considered methods such as B-spline additive model and K-nearest neighbor as well as regression trees, random forest and boosting with the XGBoost algorithm (Chen and Guestrin, 2016), Bayesian additive regression trees (Chipman et al., 2010), cubist algorithm (Quinlan, 1993). Their simulation results show that, in general, the non-parametric imputation models are superior to parametric models to capture the non-linear trend in the data. However, in high-dimension settings (*i.e.* a large number p of auxiliary variables), the K-nearest neighbor or the additive models are out-performed by machine learning methods which are more robust in such contexts. Dagdoug et al. (2024) study the asymptotic properties of the regression tree and the random forest imputed estimator.

With unit nonresponse, the weighted estimator of t_{yU} is $\hat{t}_w = \sum_{k\in s} d_k y_k/\hat{p}_k$, where $\hat{p}_k$ is the estimated response probability of unit k. The response probabilities p_k may be estimated by parametric logistic regression, or through nonparametric regression. Da Silva and Opsomer (2006) studied the kernel smoothing and Da Silva and Opsomer (2009) extended it to local polynomial regression. Very recently, Larbi et al. (2023) make a large simulation study of nonparametric and machine learning methods for estimating the response probabilities.

3.3 Variance estimation

Variance estimation with survey data is a very important but difficult issue. In a design-based approach, the variance of estimators derived under the sampling design is desired in order to deduce next estimated confidence intervals. All estimators presented in this paper are non linear estimators, so their variances are not computable and in the best case, only asymptotic variances may be deduced. Using linearization techniques and adapted asymptotic framework including assumptions on the sampling design, the study and the auxiliary variable, the asymptotic variance of model-assisted or calibration estimators are equal to Horvitz-Thompson variance applied to residuals $y_k - m(\mathbf{x}_k)$, $A\mathbb{V}(\hat{t}^{\mathrm{ma}}_{y\mathbf{x}}) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)((y_k - m(\mathbf{x}_k))/\pi_k)((y_l - m(\mathbf{x}_l))/\pi_l)$ and estimated by the Horvitz-Thompson variance estimator applied to estimated residuals $\widehat{\epsilon}_k = y_k - \widehat{m}(\mathbf{x}_k), k \in s$. With nonparametric methods, overfitting usually happens leading to underestimated residuals $\widehat{\epsilon}_k, k \in s$, so confidence intervals based on such variance estimator will not have the desired rate. This issue was already raised by Opsomer and Miller (2005) in the context of local polynomial regression. To

cope with this issue, Dagdoug et al. (2023c) suggested a variance estimator based on a K-fold cross-validation criterion, widely used in machine learning community for determining for example the tuning hyper parameters. More specifically, the sample s is split randomly 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 $\mathbf{x}_k$ fitted 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 $\widehat{\mathbb{V}}^{(K)} = \sum_{\kappa_1=1}^{K} \sum_{\kappa_2=1}^{K} \sum_{k \in s_{\kappa_1}} \sum_{l \in s_{\kappa_2}} ((\pi_{kl} - \pi_k \pi_l)/\pi_{k\ell})(\widehat{\epsilon}_k^{(-\kappa_1)}/\pi_k)(\widehat{\epsilon}_l^{(-\kappa_2)}/\pi_l)$. In practice, the number of groups (or folds) is often set to $K = 5$ or $K = 10$. This variance estimation procedure allowed to greatly improve the symmetric confidence interval rates for the random forest model-assisted estimator and Dagdoug et al. (2024) adapted the method to account for the item-nonresponse.

4 Conclusion

This paper provides a synthetic presentation of estimation methods for totals in surveys, as their objectives and type of data have evolved over time. Modern machine learning methods are becoming increasingly popular in surveys. However, their automatic use in survey sampling comes with its own set of challenges and caveats, including concerns about model interpretability, overfitting, and bias amplification. The implementation of such modern estimation methods is not straightforward with complex sampling designs. Most of machine learning algorithms have been implemented for non survey data and they do not allow considering the sampling weights in the predictions $\widehat{m}(\cdot)$, leading to potentially biased estimators for unequal and complex survey designs (Dagdoug et al., 2023b). As such, it is essential for researchers and users to exercise caution and rigor in applying these techniques and to complement them with traditional estimation methods to ensure the validity and reliability of survey estimates. Many research perspectives open up: the choice of hyper-parameters to use in machine learning algorithms, the bootstrapping of individuals or the estimation of variance are really important questions that need to be explored further more deeply. The estimation of non-linear study parameters with high-dimensional auxiliary information may be also of interest, however there is little research on this field. Goga and Ruiz-Gazen (2014) used B-spline nonparametric estimation for nonlinear functions such as median, Gini index but we are not aware of use of machine learning methods for the estimation of such parameters. This paper treated the estimation issues with probabilistic samples. Non-probabilistic surveys are used more and more often nowadays and recent works started treating estimation issues with such samples by using also machine learning methods. Another item not treated in this paper is the use of machine learning methods for small-area estimation, the reader is referred to the excellent paper of Krennmair et al. (2022) for a review on this area.

References

Baffeta, F., Corona, P. and Fattorini, L. . (2010). Design-based diagnostics for k-nn estimators of forest resources. *Canadian Journal of Forest Research*, 41:59–72.

Bardsley, P. and Chambers, R. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33:290–299.

Beaumont, J.-F. and Bocci, C. (2008). Another look at ridge calibration. *Metron-International Journal of Statistics*, LXVI:260–262.

Beaumont, J. F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *To appear in Survey Methodology*.

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*, pages 190–205.

Breidt, F. J., Opsomer, J. D., Johnson, A. A. and Ranalli, M. G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology*, 33:35–44.

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

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Advanced Books and Software, Belmont, CA., MR0726392.

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.

Cardot, H., Goga, C. and Lardin, P. (2013). 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.

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.

Da Silva, D. N. and Opsomer, J. D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *Canadian Journal of Statistics*, 4:563–579.

Da Silva, D. N. and Opsomer, J. D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology*, 35(2):165–176.

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

Dagdoug, M., Goga, C. and Haziza, D. (2023b). Model-assisted estimation in high-dimensional settings for survey data. *Journal of Applied Statistics*, 50(3):761–785.

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

Dagdoug, M., Goga, C. and Haziza, D. (2024). Statistical inference in the presence of imputed survey data through regression trees and random forests. *Submitted*.

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, 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.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire: une approche non paramétrique par splines de régression. *Canad. J. Statist.*, 33(2):163–180.

Goga, C. (2021). B-spline estimation in a survey sampling framework. In Daouia, A. and Ruiz-Gazen, A., editors, *Advances in Contemporary Statistics and Econometrics, Festschrift in Honor of Christine Thomas-Agnan*, pages 79–99. Springer.

Goga, C. (2024). High-dimensional estimation in a survey sampling framework, model-assisted and calibration points of view. *Accepted for publication in Metron*.

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

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.

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.

Hastie, T., Tibshirani, R. and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

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.

Klusowski, J. and Tian, P. (2024). Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119:525–537.

Krennmair, P., Wurz, N. and Schmid, T. (2022). Tree-based machine learning in small area estimation. *The Survey Statistician*, 86:22–31.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. In Association, A. S., editor, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 280–285.

Laplace, P. (1814). Essai philosophique sur les probabilités. *Paris: MME VE Courcier, Imprimeur-Libraire pour les Mathématiques, quai des Augustins, no. 57*.

Larbi, K., Tsang, J., Haziza, D. and M., D. (2023). Treatment of unit nonresponse in surveys through machine learning methods: an empirical comparison. *Submitted*.

Lardin-Puech, P., Cardot, H. and Goga, C. (2014). Analysing large datasets of functional data: a survey sampling point of view. *Journal de la Société Française de Statistique*, 155(4):70–94.

McConville, K. and Breidt, F. J. (2013). Survey design asymptotics for the model-assisted penalized spline regression estimator. *Journal of Nonparametrics Statistics*, 25:745–763.

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. E. and Ranalli, M. G. (2005). Nonparametric model calibration in survey sampling. *J. Amer. Statist. Assoc.*, 100:1429–1442.

Nalenz, M., Rodemann, J. and Augustin, T. (2024). Learning de-biased regression trees and forests from complex samples. *Machine learning*, https://doi.org/10.1007/s10994-023-06439-1.

Opsomer, J. and Miller, C. (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *Nonparametric Statistics*, 17(5):593–611.

Quinlan, J. (1993). Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning*, pages 236–243.

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*.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

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. 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., Dorfman, A. and Royall, R. M. (2000). *Finite Population Sampling and Inference*. Wiley Series in Probability and Statistics.

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.