# Random forests in surveys: from model-assisted estimation to imputation

*Mehdi Dagdoug*[a]

Joint work with *Camelia Goga*[a] and *David Haziza*[b]

(a) Université de Bourgogne Franche-Comté, LmB, Besançon, France

(b) University of Ottawa, Department of mathematics and statistics, Ottawa, Canada

JMS2022

29-31 March 2022

# Outline

**1)** Basic set-up and prediction models in surveys.

**2)** An introduction to regression trees and random forests.

**3)** Model-assisted estimation with random forests *(JASA, 2021)*.

**4)** Imputation with random forests (to be submitted).

**5)** Conclusion and future works.

*This research was supported by grants of the region of Franche-Comté and Médiamétrie.*

# Set-up

- $U = \{u_1, u_2, ..., u_N\}$ : finite population of size $N$.
- $Y$: survey variable.
- **Goal:** Estimate
$$t_y := \sum_{k \in U} y_k,$$
with $y_k$ the measurement of $Y$ for element $k$ of $U$.
- $S$: probability sample with, for $k, l \in U$,
$$\pi_k := \mathbb{P}(k \in S) > 0, \qquad \text{and} \qquad \pi_{kl} := \mathbb{P}(k, l \in S) > 0.$$
- If $Y$ is fully observed (no nonresponse), we have access to
$$D_y := \{y_k; k \in S\}.$$
- Horvitz-Thompson estimator $\widehat{t}_{ht}$ of $t_y$:
$$\widehat{t}_{ht} := \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} d_k y_k.$$

# Model-assisted estimation

- $X_1, X_2, ..., X_p$: auxiliary information.
- If, for all $k \in U$, the vectors $\mathbf{x}_k := [x_{k1}, ..., x_{kp}]^\top$ are observed, we have access to

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

- Model-assisted estimator $\widehat{t}_{ma}$ of $t_y$:

$$\widehat{t}_{ma} := \sum_{k \in U} \widehat{m}_1(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \widehat{m}_1(\mathbf{x}_k)}{\pi_k}, \quad (1)$$

  with $\widehat{m}_1 : \mathbb{R}^p \to \mathbb{R}$, a prediction method which may depend on $D_{ma}$.

- The estimator $\widehat{t}_{ma}$ might improve on $\widehat{t}_{ht}$.

# Nonresponse

- In most surveys, the variable $Y$ is prone to nonresponse.
- Let $r_k$ be the response indicator for $Y$, i.e.

$$r_k = \begin{cases} 1, & \text{if } y_k \text{ is observed,} \\ 0, & \text{if } y_k \text{ is missing.} \end{cases}$$

  and define $S_r = \{k \in S; r_k = 1\}$, $S_m = \{k \in S; r_k = 0\}$.

- We thus have access to

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

- Nonresponse mechanism is assumed to be **missing at random** (Rubin, 1976):

$$\mathbb{P}\{r_k = 1 | y_k, \mathbf{x}_k\} = \mathbb{P}\{r_k = 1 | \mathbf{x}_k\}.$$

# Imputation

- Imputed estimator of $t_y$:

$$\widehat{t}_{imp} = \sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_m} \frac{\widehat{m}_2(\mathbf{x}_k)}{\pi_k},$$

with $\widehat{m}_2 : \mathbb{R}^p \to \mathbb{R}$, a prediction method which may depend on $D_{imp}$.

- The estimator $\widehat{t}_{imp}$ might reduce the undesirable effects of nonresponse.

- It is possible to write $\widehat{t}_{imp}$ as

$$\widehat{t}_{imp} = \sum_{k \in S} \frac{\widehat{m}_2(\mathbf{x}_k)}{\pi_k} + \sum_{k \in S_r} \frac{y_k - \widehat{m}_2(\mathbf{x}_k)}{\pi_k}.$$

- Many properties of $\widehat{t}_{ma}$ will also be shared by $\widehat{t}_{imp}$.

# Regression trees

## Definition. (Regression trees)

A regression tree algorithm fitted on $D_U = \{(\mathbf{x}_k, y_k)\}_{k \in U}$ can be defined as follows:

Step 1: Choose a splitting criterion and a stopping criterion (e.g. a minimum of $n_0$ elements per node).

Step 2: Split recursively $[0;1]^p$ to obtain a partition $\widetilde{\mathcal{P}} = \left\{\widetilde{\mathcal{A}}_1, ..., \widetilde{\mathcal{A}}_T\right\}$ of $[0;1]^p$.

Step 3: For a prediction at the point $\mathbf{x}$, compute

$$\widetilde{m}_{tree}\left(\mathbf{x}, D_U\right) := \sum_{k \in U} \frac{\mathbb{1}_{\mathbf{x}_k \in \widetilde{\mathcal{A}}(\mathbf{x})}}{\sum_{l \in U} \mathbb{1}_{\mathbf{x}_l \in \widetilde{\mathcal{A}}(\mathbf{x})}} y_k,$$

with $\widetilde{\mathcal{A}}(\mathbf{x})$ the node containing $\mathbf{x}$.

# Example 1: Regression trees

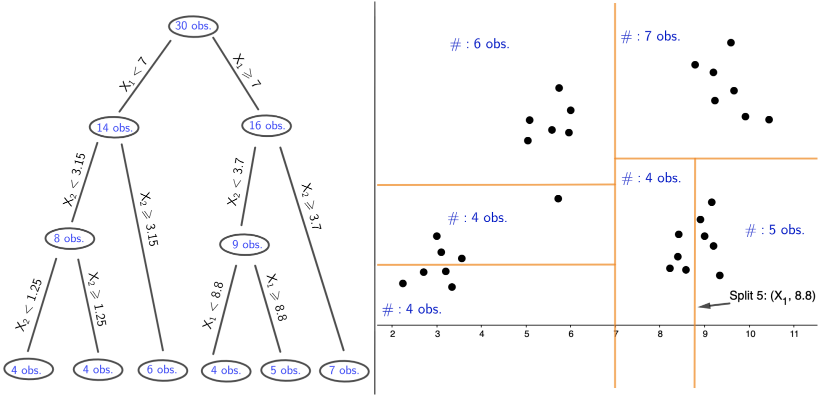


Figure: A regression tree (left) and its corresponding partition (right).

$\hookrightarrow$ The prediction at a point $\mathbf{x} \in \widetilde{A}_j$ is given by the **average** of the $\{y_k\}_{k:\mathbf{x}_k \in \widetilde{A}_j}$.

# Breiman's random forests (Breiman, 2001)

**Random forests are ensemble methods based on a large collection of regression trees.** These can be defined by the following steps.

Step 1: Select $B$ bootstrap samples (samples of $N$ elements from $D_U$, with replacement) $D_U(\Theta_1), ..., D_U(\Theta_B)$ from $D_U$.

Step 2: On $D_U(\Theta_b)$, fit $\widetilde{m}_{tree}^{(b)}$ using the randomized CART criterion optimized on $p_0$ covariates chosen **uniformly at random**, without replacement, at each split.

Step 3: The prediction at $\mathbf{x} \in [0;1]^p$ is given by

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

# Exemple 2: Estimation of a regression function

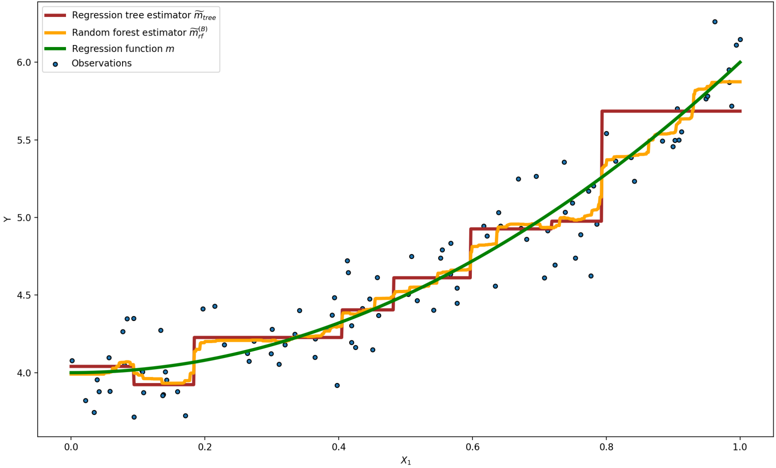


Figure: Regression function estimation with a tree and a forest, with $Y = m(X_1) + \mathcal{N}(0; 0.2)$, such that $m : x \mapsto 4 + 2x^2$, and $X_1 \sim \mathcal{U}[0; 1]$.

# Random forest model-assisted estimator

- At the sample level, we define

$$\widehat{m}_{rf1}(\mathbf{x}) := \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S(\Theta_b)}\frac{\pi_k^{-1}\mathbb{1}_{\mathbf{x}_k\in\widehat{A}_b(\mathbf{x})}}{\sum_{l\in S(\Theta_b)}\pi_l^{-1}\mathbb{1}_{\mathbf{x}_l\in\widehat{A}_b(\mathbf{x})}}y_k.$$

- Proposed random forest model-assisted estimator of $t_y$:

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

- Taking the particular case of $B = 1$, and no random mechanism, we obtain a regression tree model-assisted estimator, as in Toth and McConville (2019).

# The random forest weighting system

- We can write $\widehat{t}_{rf1}$ as

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

with

$$w_{k1} = \frac{1}{\pi_k} \left\{ 1 + \frac{1}{B} \sum_{b=1}^{B} \psi_k^{(b)} \frac{N_b(\mathbf{x}_k, U) - \widehat{N}_b(\mathbf{x}_k, S)}{\widehat{N}_b(\mathbf{x}_k, S(\Theta_b))} \right\}, \qquad k \in S,$$

where:

- $\psi_k^{(b)} = 1$ if $k \in S(\Theta_b)$, 0 otherwise,
- $N_b(\mathbf{x}_k, U)$ denoting the number of elements of $U$ belonging to the node $\widehat{\mathcal{A}}_b(\mathbf{x}_k)$,
- $\widehat{N}_b(\mathbf{x}_k, S)$ denoting the Horvitz-Thompson estimator of the number of elements of $U$ with elements of $S$ belonging to the node $\widehat{\mathcal{A}}_b(\mathbf{x}_k)$.

# Behavior of the weighting system

- Considering the case of a regression tree, we have

$$w_{k1} = d_k \times \frac{N(\mathbf{x}_k, U)}{\widehat{N}(\mathbf{x}_k, S)}, \qquad k \in S.$$

- It follows that:
  - If the original weighting system **estimates correctly** the number of elements similar to $u_k$, then $w_{k1} \approx d_k$.
  - If the original weighting system **underestimates** the number of elements similar to $u_k$, then $w_{k1} >> d_k$.
  - If the original weighting system **overestimates** the number of elements similar to $u_k$, then $w_{k1} << d_k$.
- The weights satisfy $\sum_{k \in S} w_{k1} = N$, for all $S \in \mathcal{S}$.

# Asymptotic properties and variance estimation

In the framework of Isaki and Fuller (1982), under mild conditions, the following asymptotic properties hold.

- There exists constants $C_1, C_2$ such that

$$\mathbb{E}_p \left[ \left| \frac{1}{N} \left( \widehat{t}_{rf1} - t_y \right) \right| \right] \leqslant \frac{C_1}{\sqrt{N}} + \frac{C_2}{n_0}. \quad \text{a.s.}$$

- The asymptotic variance of $\widehat{t}_{rf1}$ is given by

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

- It is possible to estimate this asymptotic variance consistently.

- The estimator $\widehat{t}_{rf1}$ is asymptotically gaussian for common sampling designs.

# Random forest imputed estimators

- Let $\widehat{m}_{rf2}$ denote a random forest estimator (unweighted) fitted on $\{(\mathbf{x}_k, y_k)\,; k \in S_r\}$, that is,

$$\widehat{m}_{rf2}(\mathbf{x}) := \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S_r(\Theta_b)} \frac{\mathbb{1}_{\mathbf{x}_k \in \widehat{A}_b(\mathbf{x})}}{\sum_{l\in S_r(\Theta_b)} \mathbb{1}_{\mathbf{x}_l \in \widehat{A}_b(\mathbf{x})}} y_k.$$

- The forest imputed estimator $\widehat{t}_{rf2}$ is defined by

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

- The forest $\widehat{t}_{rf2}$ estimator can be written as

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

where the estimation weights $\{w_{k2}\}_{k\in S_r}$ are given by

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

# Understanding the behavior of the weighting system

Consider the case of a regression tree. Then,

- Assuming equality of first order inclusion probabilities, we have

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

- It follows that:
  - If most people similar to $u_k$ **did not answer**, then $R_{mr}(\mathbf{x}_k)$ is large and $w_{k2}$ is large.
  - If most people similar to $u_k$ **did answer**, then $R_{mr}(\mathbf{x}_k)$ is close to 0 and $w_{k2}$ is close to $d_k$, the original weight.

# Instability of small forest estimators

- The weights of unselected elements are such that

$$w_{k2} = d_k, \qquad k \in \bigcap_{b=1}^{B} S_r(\Theta_b).$$

- The weights are calibrated to the population size $N$ whenever the original weighting system is:

$$\sum_{k \in S_r} w_{k2} = \sum_{k \in S} d_k := \widehat{N}.$$

- **Unselected elements have low weights, forcing selected elements to have large weights.**

- For all $k \in S_r$ and $n_r \geqslant 1$

$$\mathbb{P}\left\{ k \in \bigcap_{b=1}^{B} S_r(\Theta_b) \,\middle|\, n_r \right\} = \left( \frac{n_r - 1}{n_r} \right)^B \xrightarrow{B \to \infty} 0.$$

Hence, **stability is recovered for large forests.**

# Asymptotic properties and variance estimation

- **Forests with a large number of trees are more efficient** than forests with a small number of trees.

- For large forests with Breiman's algorithm, we have

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

- The randomization variance is controlled by

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

  $\hookrightarrow$ For large forests, the randomization variance can be neglected.

- Variance estimators are suggested using both the two-phase and reverse approaches.

# Some empirical considerations

- Simulations show the good behavior of model-assisted and imputed random forests estimators, particularly in high-dimensional frameworks.
- Most packages do not provide the option of weighting the predictions.

  $\hookrightarrow$ We recommend adding design variables to the set of covariates, while forcing these additional covariates to always be considered.
- Variance estimators are approximately unbiased for large choices of $n_0$; for small values of $n_0$, however, the variance might be under-estimated.

  $\hookrightarrow$ We recommend using a cross-validated variance estimator for small choices of $n_0$.

# Final remarks

- Statistical learning prediction procedures provide highly flexible tools for survey practitioners and can be used in many areas:
  - Model-assisted estimation,
  - Imputation,
  - Propensity score adjustment,
  - Model-based estimation,
  - Definition of the sampling design (e.g. adaptive sampling).
- Most machine learning procedures are not yet fully understood. Problems in surveys may arise:
  - Model-assisted variance underestimated by the usual variance estimator for complex models.
  - Important bias in forest estimators when design design variables are not considered for splitting.
- There is an important need for additional research in this area.

# Short list of references


Breidt, F.-J. and Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1023–1053.

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

Dagdoug, M., Goga, C., and Haziza, D. (2020). Model-assisted estimation through random forests in finite population sampling. *arXiv preprint arXiv:2002.09736*.

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.

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, 77:49–61.

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.

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

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.