Submitted to the Annals of Statistics

ON THE CONSISTENCY OF A RANDOM FOREST ALGORITHM IN THE PRESENCE OF MISSING ENTRIES

BY IRVING GÓMEZ-MÉNDEZ[1] AND EMILIEN JOLY[2]

[1] Centro de Investigación en Matemáticas, AC (CIMAT), irving.gomez@cimat.mx

[2] Centro de Investigación en Matemáticas, AC (CIMAT), emilien.joly@cimat.mx

This paper tackles the problem of constructing a non-parametric predictor when the latent variables are given with incomplete information. The convenient predictor for this task is the random forest algorithm in conjunction to the so called CART criterion. The proposed technique enables a partial imputation of the missing values in the data set in a way that suits both a consistent estimation of the regression function as well as a probabilistic recovery of the missing values. A proof of the consistency of the random forest estimator that also simplifies the previous proofs of the classical consistency is given in the case where each latent variable is missing completely at random (MCAR).

1. Introduction. Random forests algorithms are widely used in applied computer science tasks as it fulfills two main purposes in extracting information of a possibly large scaled data set: regression and prediction. First introduced by Breiman [5], this non-parametric technique has various computational benefits over many regression/prediction techniques as it is fast and simple, therefore it is really common to see the complete random forest algorithm (or parts of it) as a building block of more intricate machine learning algorithms. The small number of hyper-parameters to be tuned in the random forests algorithm, along with the bootstrap-aggregation technique [4] -commonly called bagging- allow a straightforward parallelization and has made random forests one of the most popular tools for handling data sets of large size. It has been successfully involved in various practical problems, including chemioinformatics [35], ecology [24, 9], 3D object recognition [33], bioinformatics [10, 34] and econometrics [37].

On the theoretical side little is known about the mathematical properties of the method which has lead the community to underline a “gap” between theory and practice. Part of this gap can be explained by the bagging mechanism and the splitting criterion. Each of these processes introduces a source of randomness into the construction of the trees (the building blocks of the random forests algorithms) which makes the random forest algorithms very challenging to study in its full generality. One way to tackle the theoretical justifications of random forests is through simplified versions of the original procedure. This is often done by simply ignoring the bagging step and/or by replacing the splitting criterion with a more elementary splitting protocol (e.g. Breiman [6]). However, in recent years, important theoretical studies have been performed to analyze more elaborated models (e.g. Biau, Devroye and Lugosi [2], Ishwaran and Kogalur [18], Biau [1], Genuer [12], Zhu, Zeng and Kosorok [40]). Consistency and asymptotic normality for Breiman's infinite forests were proven by Wager and Athey [39] simplifying the splitting step, Mentch and Hooker [23] proved a similar result for finite forests and Scornet, Biau and Vert [31] proved a consistency result in the context of additive regression models. Biau and Scornet [3] review gives a very good introduction to those concepts and treat its inherent challenges which include the majority of those presented here.

MSC2020 subject classifications: Primary 62G05; secondary 62G20.

Keywords and phrases: Random Forest, Missing Data, Consistency, Partial Imputation.

The relative simplicity of random forests advocates for its use to handle missing data. As an initial solution, Breiman and Cutler [7] proposed an algorithm which takes advantage of the so-called proximity-matrix to impute the missing values and Ishioka [17] modified it in order to obtain a new version more robust to outliers. An interesting alternative may be the so-called surrogate splits (see Breiman et al. [8], Venables and Ripley [38], Ripley [27], Friedman, Hastie and Tibshirani [11], Hapfelmeier, Hothorn and Ulm [14], for which no imputation steps are required. Missing Incorporated in Attributes (MIA) [36] and conditional trees [15] are two other approaches that has gain attention in recent years. Josse et al. [20] perform a simulation study to compare different methods to handle missing data using trees and study the consistency of two approaches to estimate the prediction function with missing values in a framework of multiple imputation [30, 29], the use of a universally consistent algorithm and for a Missing At Random (MAR) mechanism. Many of these algorithms has been implemented in `R` in the package `partykit` [16]

The present paper gives a consistency result in the context of additive regression in presence of missing data accordingly to a Missing Completely at Random (MCAR) paradigm. For this purpose, we propose a new splitting mechanism suited to the mathematical proof given in Section 4. Unfortunately, this algorithm does not seems to scale with the data. A computation study would a be a subject of a future work.

The paper is organized as follows. In Section 2 we introduce the random forests, the concept of data-missing mechanisms and name some previous approaches to deal with missing data using random forests, an overview on different methodologies to handle missing data can be found in Josse et al. [19]. The main consistency result is presented in Section 3 and its proof is given in Section 4.

2. Random forests: aggregating decision trees. The random forests algorithms are named after the general technique that consists in building and exploiting the objects called decision trees. These objects are rooted trees in the mathematical sense and are, most of the time, binary trees. Each path starting from the root corresponds to a decision (sometime called prediction) and the descent mechanism is usually a key part in the decision. Our random forest algorithm is made of a set of decision trees that are later aggregated all together (with a simple mean idea). Each branch of the tree will be random (in its construction process) and represents a partition of the input space in two smaller regions. Moving along the path of the decision tree corresponds to a choice of one of the two possible regions. The main strength of random forests is that it aggregates the information of many different decision trees in a global predictor that ends to be a lot more stable (and then a lot more informative) than each specific tree.

Decision trees are conceptually simple yet powerful and attractive in practice for several reasons:

- They can model arbitrarily complex relations between the input and the output space.
- They handle categorical or numerical variables, or a mix of both.
- They can be used in regression or supervised classification problems.
- They are easy interpretable, even for non-statisticians.

To construct the partition of the input space, decision trees work in a recursive way. The root of the tree is the whole input space, $\mathcal{X}$, which is splitted into disjoint regions. Then each region is splitted into more regions, and this process is continued until some stopping rule is applied (see Figure 1). At each step of the tree construction, the partition performed over a cell (or equivalently its corresponding node) is determined by maximizing some splitting criterion.

In our framework, we have access to a training set $\mathcal{D}_n = (\mathbf{X}_i, Y_i)_{i=1,\ldots,n}$ where the response variables Y_i are real valued and the input variables $\mathbf{X}_i$ belong to some space $\mathcal{X}$.

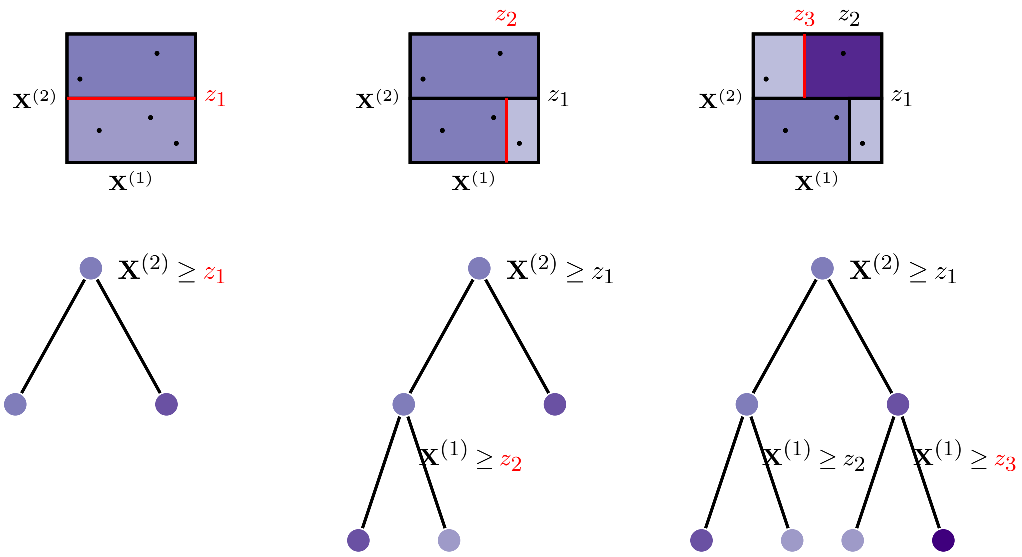
z_2
z_3
z_2
$\mathbf{X}^{(2)}$
z_1
$\mathbf{X}^{(2)}$
z_1
$\mathbf{X}^{(2)}$
z_1
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(1)}$
$\mathbf{X}^{(2)} \geq z_1$
$\mathbf{X}^{(2)} \geq z_1$
$\mathbf{X}^{(2)} \geq z_1$
$\mathbf{X}^{(1)} \geq z_2$
$\mathbf{X}^{(1)} \geq z_2$
$\mathbf{X}^{(1)} \geq z_3$

Fig 1: At each step of the tree construction a partition is performed over a cell (or equivalently its corresponding node) maximizing some split criterion.

For simplicity, we assume throughout the discussion that the vector $\mathbf{X}$ is a uniform random variable on the square $\mathcal{X}$. In most applications the space $\mathcal{X}$ is a compact portion of a p dimensional space. Hence in the sequel we assume, without loss of generality, that $\mathcal{X} = [0,1]^p$. The task is to predict the random variable Y, with respect to the input vector $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(p)})$ where $\mathbf{X}^{(j)} \in [0,1]$ (for $j = 1, \ldots, p$). For example, one could be interested in finding a function $f : [0,1]^p \to \mathbb{R}$ that minimizes the loss functional $\mathbb{E}_{\mathbf{X},Y}\left[\mathcal{L}(f(\mathbf{X}), Y)\right]$ where $\mathcal{L} : \mathbb{R}^2 \to \mathbb{R}$ is the squared error $\mathcal{L}(f(\mathbf{X}), Y) = (f(\mathbf{X}) - Y)^2$. The solution of this optimization problem

$$m = \underset{f:[0,1]^p \to \mathbb{R}}{\arg\min}\ \mathbb{E}_{\mathbf{X},Y}\left[(f(\mathbf{X}) - Y)^2\right]$$

is given by the regression function $m(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$. For practical problems, the distribution of $(\mathbf{X}, Y)$ and hence, the regression function, are unknown. The task is then to use the data $\mathcal{D}_n$ to construct a learning model, also called learner, predictor or estimator, $m_n : [0,1]^p \to \mathbb{R}$ that estimates the regression function m, and enables us to predict the outcome for new unseen objects. The focus of this present article is the estimation step and the authors leave the prediction study for upcoming publications. Two distinct strategies can be used to tackle the estimation problem, the parametric and the non-parametric estimation. In a nutshell, the non-parametric estimation uses a model that sums up to a set of functions $\mathcal{F}_n$, that serves the minimization (often referred to Empirical Risk Minimization)

$$m_n = \underset{f \in \mathcal{F}_n}{\arg\min}\, \mathbb{E}_{\mathcal{D}_n}\left[(f(\mathbf{X}) - Y)^2\right].$$

The focus of this work is on the non-parametric estimation of m through random forests algorithms. As mentioned above, a random forest is a predictor consisting of $M(> 1)$ randomized trees. The randomization is introduced in two different parts of the tree construction. Prior to the contruction of each tree, a_n observations are extracted at random with (or without) replacement from the learning data set $\mathcal{D}_n$. Only these a_n observations are taken into account in the tree construction. Then, at each cell, a split is performed by maximizing the split criterion over a number `mtry` of input variables chosen uniformly at random among the original ones. The tree construction is stopped when each final node contains less or equal than `nodesize` points or when the tree has t_n final nodes. Hence, the parameters of this algorithm are:

- $M > 1$, which is the number of trees in the forest.
- $a_n \in \{1, \dots, n\}$, which is the number of observations in each tree.
- `mtry` $\in \{1, \dots, p\}$, which is the number of directions (features) chosen, candidates to be splitted. We denote by $\mathcal{M}_{try}$ the features selected in each step.
- `nodesize` $\in \{1, \dots, a_n\}$, which is the maximum number of observations for a node to be a final cell.
- Instead of `nodesize` we can use the parameter $t_n \in \{1, \dots, a_n\}$, which is the number of leaves (final nodes) in each tree.

This randomization (independent from the original source of randomness in the sample $\mathcal{D}_n$) is represented in a symbolic random variable Θ. To each tree – randomized with the random variable Θ_k – is associated a predicted value at a query point $\mathbf{x}$, denoted $m_n(\mathbf{x}; \Theta_k)$. The different trees are constructed by the same procedure but with independent randomization so that the random variables $\Theta_1, \dots, \Theta_M$ are i.i.d. with common law Θ. The nature and dimension of Θ depends on its use in the tree construction. In our choice of construction rules, Θ consists of the observations selected for the tree and the candidate variables to split at each step. At the end of the tree construction, a partition of the space $\mathcal{X}$ is returned in a form of a collection of cells $(A_{n,i})_{i \geq 1}$. Each of these cells corresponds to a leaf of the tree and they are called final cells to emphasize their difference with the cells that we consider during the construction of the tree. Finally, the kth tree estimate is defined as

$$m_n(\mathbf{x}; \Theta_k) = \sum_{i \in \mathcal{I}_{n,\Theta_k}} \frac{Y_i \mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_k)}}{N(A_n(\mathbf{x}; \Theta_k))}$$

where $\mathcal{I}_{n,\Theta_k}$ is the set of the a_n observations selected prior to the construction of the kth tree, $A_n(\mathbf{x}; \Theta_k)$ is the unique final cell that contains $\mathbf{x}$, and $N(A_n(\mathbf{x}; \Theta_k))$ is the number of observations which belong to the cell $A_n(\mathbf{x}; \Theta_k)$. The aggregation of the trees forms the finite random forest estimate given by

$$m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M) = \frac{1}{M} \sum_{k=1}^{M} m_n(\mathbf{x}; \Theta_k).$$

It is known from the work of Breiman [5] that the random forest estimate does not overfit when M tends to infinity. This makes the parameter M only restricted by the computational power. Besides being important in practice, we do not make big case of the choice of M since the following results show the consistency of each of the tree estimates, hence the consistency of the random forest for any number M.

Different split criteria have been proposed depending on the statistical problem and the nature of the input space. For regression purposes, the most commonly used is the CART criterion [8]. For supervised classification, the split criterion takes the form of an impurity function, like the misclassification error, the Gini index [13] or the criteria known as ID3 and C4.5 [25, 26] which minimize the Shannon entropy [32] and replace binary splits on categorical variables with multiway splits.

The present work focuses on the CART split criterion. We first introduce some important notations.

- A denotes a general node (or cell).
- $N(A)$ holds for the number of points in A.

- The notation $d = (h, z)$ denotes a cut in A, where

  h is a direction, $h \in \{1, \ldots, p\}$, and
  z is the position of the cut in the hth direction, between the limits of A.
- $\mathcal{C}_A$ is the set of all possible cuts in the node A. It means that h do belong to the set $\mathcal{M}_{try}$ and that for any chosen h, z, it lies between the bounds of the cell A in that specific direction.
- A cell A is split into two cells denoted $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(h)} < z\}$, $A_R = \{\mathbf{x} \in A : \mathbf{x}^{(h)} \geq z\}$.
- $\bar{Y}_A$ (resp. $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$) is the empirical mean of the response variable Y_i for the indexes such that $\mathbf{X}_i$ belongs to the cell A (resp. A_L, A_R).

Then, the empirical version of the CART split criterion for a generic cell A and over the full sample $\mathcal{D}_n$ is defined as

$$\begin{aligned} L_n(A,d) =& \frac{1}{N(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_A\right)^2 \mathbb{1}_{\mathbf{X}_i \in A} \\ & - \frac{1}{N(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(h)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(h)} \geq z}\right)^2 \mathbb{1}_{\mathbf{X}_i \in A} \end{aligned}$$

with the convention $0/0 = 0$.

Intuitively, the CART criterion compares the empirical variance inside the cell A with the sum of the two empirical variances of the sub-cells A_R and A_L. At each step of the creation of the random tree, a current cell A is selected and cut by choosing the best empirical cut $\widehat{d} = (\widehat{h}, \widehat{z})$ so that $L_n(A,d)$ is maximal over the set $\mathcal{C}_A$, that is

$$\widehat{d} = (\widehat{h}, \widehat{z}) \in \underset{d \in \mathcal{C}_A}{\arg\max}\, L_n(A,d).$$

The theoretical counterpart of the empirical CART criterion is given by

$$\begin{aligned} L^\star(A,d) =& \mathbb{V}[Y | \mathbf{X} \in A] - \mathbb{V}[Y | \mathbf{X}^{(h)} < z, \mathbf{X} \in A] \mathbb{P}[\mathbf{X}^{(h)} < z | \mathbf{X} \in A] \\ & - \mathbb{V}[Y | \mathbf{X}^{(h)} \geq z, \mathbf{X} \in A] \mathbb{P}[\mathbf{X}^{(h)} \geq z | \mathbf{X} \in A]. \end{aligned}$$

Analogously to the empirical case, we define the best theoretical cut $d^\star = (h^\star, z^\star)$ in A as

$$d^\star = (h^\star, z^\star) \in \underset{d \in \mathcal{C}_A}{\arg\max}\, L^\star(A,d)$$

By the classical strong law of large numbers, $L_n(A,d)$ converges almost surely to $L^\star(A,d)$ as n tends to infinity, for all cuts $d \in \mathcal{C}_A$. This fact leads to the interpretation that the chosen cuts at each step of the tree construction tend to decrease the variability of the sets of data points corresponding to the child nodes. This implies that the empirical mean in each cell tends to stabilize around a value that correspond to the conditional mean for the cell. Applying basic algebra, it is easy to get an equivalent expression for the empirical CART criterion, given by

$$L_n(A,d) = \frac{N(A_L) N(A_R)}{N(A) N(A)} \left(\bar{Y}_{A_L} - \bar{Y}_{A_R}\right)^2 \tag{1}$$

From Equation (1), we can get an alternative expression to the theoretical CART criterion, given by

$$\begin{aligned} L^\star(A,d) =& \mathbb{P}[\mathbf{X}^{(h)} < z | \mathbf{X} \in A] \mathbb{P}[\mathbf{X}^{(h)} \geq z | \mathbf{X} \in A] \\ & \times \Big(\mathbb{E}[Y | \mathbf{X} < z, \mathbf{X} \in A] - \mathbb{E}[Y | \mathbf{X} \geq z, \mathbf{X} \in A]\Big)^2 \end{aligned}$$

2.1. *Definition of Missigness.* The concept of data-missing mechanisms (introduced by Rubin [28]) establishes the relationship between missingness and the data. It is common to define the data-missing mechanisms through the data matrix. However, to have a useful definition from a theoretical point of view, we formally define them using the random variables $\mathbf{X}$ and Y. First, let us define a new variable, called the indicator of missing value,

$$\mathbf{M}^{(h)} = \begin{cases} 1 \text{ if } \mathbf{X}^{(h)} \text{ is missing} \\ 0 \text{ otherwise} \end{cases}, \quad 1 \leq h \leq p$$

We are assuming throughout this work that the response Y has no missing values (there is not point to include uninformative data in the sample), so it is not necessary to define an indicator of missing variable for Y. Then, the mechanisms are fully characterized by the information of the conditional distribution of $\mathbf{M}^{(h)}$ given $(\mathbf{X}, Y)$. One of the data-missing mechanisms identified by Rubin [28] is the so-called Missing Completely at Random (MCAR). We say that the variable $\mathbf{X}^{(h)}$ is MCAR if $\mathbf{M}^{(h)} \perp (\mathbf{X}, Y)$. In other words, under the MCAR assumption, a coordinate $\mathbf{X}^{(h)}$ has some probability to be missing in the sample and this probability does not depend on the value of $\mathbf{X}$ nor the response variable Y. Note that it is allowed to get missing rates that differ from one coordinate to another. In this work, we allow the probability of missingness to depend on the sample size n. Let us define $p_n^{(h)} = \mathbb{P}[\mathbf{M}^{(h)} = 1 | \mathcal{D}_n]$. See Theorem 1 for the condition on the values of $p_n^{(h)}$.

2.1.1. *Previous Approaches Implementing Imputation of Missing Values.* Many methods proposed in the literature to handle missing data using random forests operate through imputation in a recursive way. First, they use the original training data set, $\mathcal{D}_n$, to fill the blank spaces in a roughly way. For example, with the median of the observed values in the variable. We denote this new data set as $\mathcal{D}_{n,t_1}$.

The imputed data set $\mathcal{D}_{n,t_1}$ is used to build a random forest. Then, some structures of the forest are exploited, like the so-called proximity matrix, improving the imputation and resulting in a new data set $\mathcal{D}_{n,t_2}$. The procedure follows iteratively until some stopping rule is applied, for example when there is little change between the imputed values or when a fix number of iterations is achieved. More formally, let us define

$$\mathbf{X}_{i,t_\ell}^{(h)} = \begin{cases} \mathbf{X}_i^{(h)} \text{ if } \mathbf{M}_i^{(h)} = 0 \\ \widehat{\mathbf{X}}_{i,t_\ell}^{(h)} \text{ if } \mathbf{M}_i^{(h)} = 1 \end{cases}$$

where $\widehat{\mathbf{X}}_{i,t_\ell}^{(h)}$ is the imputation of $\mathbf{X}_i^{(h)}$ at time t_ℓ, $\ell \geq 1$, and let $\mathbf{X}_{i,t_\ell} = \left(\mathbf{X}_{i,t_\ell}^{(1)}, \ldots, \mathbf{X}_{i,t_\ell}^{(p)}\right)$.

In order to properly introduce previous methods that handle missing data through imputation, we need to define the connectivity between two points in a tree and the proximity matrix of the forest. Let $K_{\Theta,n}(\mathbf{X}, \mathbf{X}') = \mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}'}$ be the indicator that $\mathbf{X}$ is in the same final cell that $\mathbf{X}'$ in the tree designed with $\mathcal{D}_n$ and the parameter Θ. If $K_{\Theta,n}(\mathbf{X}, \mathbf{X}') = 1$ we say that $\mathbf{X}$ and $\mathbf{X}'$ are connected in the tree $m_n(\cdot; \Theta)$. The proximity of two points is the average of times in which they were connected in the forest, it measures the similarity of the observations in the eyes of the random forest. Formally, let us define the proximity between $\mathbf{X}$ and $\mathbf{X}'$ in the finite forest, $m_{M,n}(\cdot; \Theta_1, \ldots, \Theta_M)$, as

$$K_{M,n}(\mathbf{X}, \mathbf{X}') = \frac{1}{M} \sum_{k=1}^{M} K_{\Theta_k,n}(\mathbf{X}, \mathbf{X}')$$

To simplify the notation, let $K_{M,t_\ell}(i,j)$ be the proximity between $\mathbf{X}_i$ and $\mathbf{X}_j$ at time t_ℓ. That is, $K_{M,t_\ell}(i,j)$ is the proximity between $\mathbf{X}_i$ and $\mathbf{X}_j$ in the random forest constructed with the data set $\mathcal{D}_{n,t_\ell}$. We also define $\mathbf{i}_{miss}^{(h)} \subseteq \{1, \ldots, n\}$ as the indexes where $\mathbf{X}^{(h)}$ is missing, and $\mathbf{i}_{obs}^{(h)} = \{1, \ldots, n\} \setminus \mathbf{i}_{miss}^{(h)}$ as the indexes were $\mathbf{X}^{(h)}$ is observed.

Breiman's Approach. . This method was proposed by Breiman and Cutler [7]. If $\mathbf{X}^{(h)}$ is a continuous variable, $\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}}$ is the weighted mean of the observed values in $\mathbf{X}^{(h)}$, where the weights are defined by the proximity matrix of the previous random forest, that is

$$\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}} = \frac{\sum_{i \in \mathbf{i}^{(h)}_{obs}} K_{M,t_\ell}(i,j)\mathbf{X}^{(h)}_i}{\sum_{i \in \mathbf{i}^{(h)}_{obs}} K_{M,t_\ell}(i,j)}, \qquad \begin{array}{l} \ell \geq 1 \\ j \in \mathbf{i}^{(h)}_{miss} \end{array}$$

On the other hand, if $\mathbf{X}^{(h)}$ is a categorical variable, $\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}}$ is given by

$$\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}} = \underset{\mathbf{x} \in \mathcal{X}^{(h)}}{\arg\max} \sum_{i \in \mathbf{i}^{(h)}_{obs}} K_{M,t_\ell}(i,j) \mathbb{1}_{\mathbf{X}^{(h)}_i = \mathbf{x}}, \qquad \begin{array}{l} \ell \geq 1 \\ j \in \mathbf{i}^{(h)}_{miss} \end{array}$$

That is, $\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}}$ is the class that maximizes the sum of the proximity considering the observed values in the class.

Ishioka's Approach. This method, proposed by Ishioka [17], is an improvement to outliers of the previous Breiman's approach. If $\mathbf{X}^{(h)}$ is a continuous variable, $\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}}$ is the weighted mean of the k nearest neighbors, according to the proximity matrix, over all the values, both imputed and observed. The k closest values are chosen to make more robust the method and avoid values which are outliers.

$$\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}} = \frac{\sum_{\substack{i \in \text{neigh}_k \\ i \neq j}} K_{M,t_\ell}(i,j)\widehat{\mathbf{X}}^{(h)}_{i,t_\ell}}{\sum_{\substack{i \in \text{neigh}_k \\ i \neq j}} K_{M,t_\ell}(i,j)}, \qquad \begin{array}{l} \ell \geq 1 \\ j \in \mathbf{i}^{(h)}_{miss} \end{array}$$

For categorical variables, it is not necessary to see only the k closest values because the outliers of $\mathbf{X}$ will have few attention. Meanwhile the proximity with missing values should have more attention, specially when the missing rate is high. Hence, if $\mathbf{X}^{(h)}$ is a categorical variable, $\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}}$ is given by

$$\widehat{\mathbf{X}}^{(h)}_{j,t_{\ell+1}} = \underset{\mathbf{x} \in \mathcal{X}^{(h)}}{\arg\max} \sum_{i \neq j} K_{M,t_\ell}(i,j) \mathbb{1}_{\mathbf{X}^{(h)}_{i,t_\ell} = \mathbf{x}}, \qquad \begin{array}{l} \ell \geq 1 \\ j \in \mathbf{i}^{(h)}_{miss} \end{array}$$

MissForest. This algorithm, proposed by Stekhoven and Bühlmann [34], handle the missing data as a supervised learning problem itself, where the target variable is the input variable with missing values. The MissForest consists in iteratively building a random forest from the observed data and the previous imputations to predict the missing values of the input variable.

2.1.2. *Previous Approaches Without Implementing Imputation of Missing Values.*

"Missing" Category. For categorical variables a simple way to deal with the missing data is to create the new category "missing" [25, 11]. However, this approach can lead to anomalous situations as exemplified by Quinlan [25].

Ternary decision trees. We can change the structure of the trees and consider ternary splits instead of binary, where the third child contains all observations where the feature is missing, as Louppe [22] comments.

Propagate in both child nodes. We can weight the samples, giving less weight to the missing data and propagate it in both child nodes dividing the observation into fractional objects. However, there is not an obvious methodology to calculate the weights, but as long as the method to calculate them does not involve the target variable, this method can be used even for prediction with missing predictors [25, 22].

Split the observations. The observations with missing values could be assigned to some child node according to some weight that estimates the probability of belonging to each one [27, 38].

As far as it will go. The idea of this method is to use all the observations and dropping out those that have the splitted variable missing. An obvious problem of this alternative is that we can finish with a relatively small amount of data very quickly. For prediction, it consists in dropping the case down the tree as far as it will go, until we cannot longer assign the observation to one of the child nodes. We predict with the node reached by the observation [27, 38].

Surrogate splits. This popular alternative consists in ordering the directions that best split the node, if the first direction is missing we take the second surrogate split, if the second is missing we take the third and so on. The surrogate split is such that maximizes the probability of making the same decision as the primary split [8, 27, 38, 11].

Missingn Incorporated in Attributes (MIA). The Missing Incorporated in Attributes (MIA) approach was introduced by Twala, Jones and Hand [36]. It consists in keeping all the missing values together when a split is performed. Thus, the splits with this approach assign the values according to one of the following rules:

- $\{\mathbf{X}^{(h)} < z$ and $\mathbf{M}^{(h)} = 1\}$ versus $\{\mathbf{X}^{(h)} \geq z\}$
- $\{\mathbf{X}^{(h)} < z\}$ versus $\{\mathbf{X}^{(h)} \geq z$ and $\mathbf{M}^{(h)} = 1\}$
- $\{\mathbf{M}^{(h)} = 0\}$ versus $\{\mathbf{M}^{(h)} = 1\}$

3. Main Result.

3.1. *Introduction of our algorithm.* Let us emphasize, through an example, the issues that one faces with the original CART criterion in a context of missing values. Consider the space $\mathcal{X} = [0,1]^2$ and two observations, $\mathbf{X}_1$ and $\mathbf{X}_2$, that belong to a cell $A \subset \mathcal{X}$. We assume that a direction and location for a cut of A have been chosen and we denote by A_L and A_R the two resulting cells (see Table 1, Figure 2 for an illustration). Since its value is missing, $\mathbf{X}_1^{(1)}$ is represented as a dashed line [1] over the interval $[0,1]$.

It is clear that $\mathbf{X}_2 \in A_R$, however it is not possible to decide, without any further operation, if $\mathbf{X}_1 \in A_R$ or $\mathbf{X}_1 \in A_L$ or even if $\mathbf{X}_1 \in A$. The CART criterion is then untraceable since the quantities, $N(A)$, $N(A_L)$, $N(A_R)$, $\bar{Y}_A$, $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$, $\mathbb{1}_{\mathbf{X}_i \in A}$, $\mathbb{1}_{\mathbf{X}_i^{(h)} < z}$ and $\mathbb{1}_{\mathbf{X}_i^{(h)} \geq z}$ can not be computed.

The approach proposed in this paper keeps the form of the CART criterion and uses adapted imputations of the untraceable parts that allow us to compute the CART criterion. Unlike most of the so-called imputation techniques, the imputation step is not performed independently of the evaluation of the CART criterion but is integrated to its later optimization. As a cut was chosen as a maximiser of the CART criterion in the normal set up, a couple (cut,

[1] In our illustrations we represent observations with missing values with dashed lines.

	$x^{(1)}$	$x^{(2)}$
$\mathbf{X}_1$		0.5
$\mathbf{X}_2$	0.75	0.25
A	[0.3,0.9]	[0.2,0.7]
A_L	[0.3,0.6]	[0.2,0.7]
A_R	[0.6,0.9]	[0.2,0.7]

TABLE 1
Example data

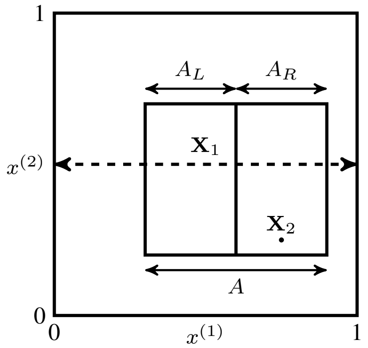

FIG 2
Illustration of the example data.

imputation) is chosen at each split in the creation of the random tree. The idea is that, for a cut, the observations with missing values are assigned to the child node that maximizes our CART criterion. Then, we move on to the next cell and we proceed in the same way until a stopping rule is achieved.

This algorithm looks similar to other approaches that perform imputation. As an initialization, we "impute" the missing values with all the possible values of the variable, then at each split step, the missing values locations are updated to belong to one of the node children. Despite this similarity, our algorithm does not impute the missing places with a punctual value, instead each tree "imputes" the missing places with an interval. Note also, that the algorithm proposed in this work has the advantage of not having to calculate extra structures like the proximity matrix, for example. To avoid confusion in the sequel, we make the subtle difference between the imputation of the missing values at the start (referred to as *in*) of an iteration and the imputation at the end (referred to *out*) of the iteration. The mathematical formalism is as follows.

- $\widehat{\mathbf{X}}_{i,in} = \left(\widehat{\mathbf{X}}_{i,in}^{(1)}, \ldots, \widehat{\mathbf{X}}_{i,in}^{(p)}\right)$ be the current imputation of $\mathbf{X}_i$.
- $\widehat{N}(A)$ is the number of points assigned to the cell A.
- $\widehat{Y}_A$ is the empirical mean of the response variable Y_i such that $\widehat{\mathbf{X}}_{i,in}$ belongs to the cell A.
- $\widehat{N}_{miss}^{(h)}(A) = \sum_{i=1}^{n} \mathbb{1}_{\widehat{\mathbf{X}}_{i,in} \in A, \mathbf{M}_i^{(h)}=1}$ is the number of observations assigned to the cell A whose variable $\mathbf{X}^{(h)}$ is missing. We denote by $\widehat{N}_{miss}(A)$ the number of observations assigned to A such that at least one coordinate is missing and by $\widehat{N}_{obs}(A)$ the number of observations with no missing values inside the cell A.
- $\mathbf{i}_{A,miss}^{(h)} = \{j_1, \ldots, j_{\widehat{N}_{miss}^{(h)}(A)}\}$ is the set of indexes of the observations assigned to the cell A whose variable $\mathbf{X}^{(h)}$ is missing.
- For each direction h, let $\mathcal{W}_A^{(h)} = \{0,1\}^{\widehat{N}_{miss}^{(h)}(A)}$ be the collection of binary vectors w with the convention that $w_k = 1$ means that the observation $(\mathbf{X}_{j_k}, Y_{j_k})$ is assigned to the left child node and $w_k = 0$ means that the observation $(\mathbf{X}_{j_k}, Y_{j_k})$ is assigned to the right child node, for all $j_k \in \mathbf{i}_{A,miss}^{(h)}$.

Finally, the variables $\mathbf{X}_i^{(h)}$ with $i \in \mathbf{i}_{A,miss}^{(h)}$ are updated by

$$\widehat{\mathbf{X}}_{i,out}^{(h)} = \begin{cases} \left[a^{(h)}, z\right] \text{ if } w_k = 1 \\ \left[z, b^{(h)}\right] \text{ if } w_k = 0 \end{cases}$$

while $\widehat{\mathbf{X}}^{(h)}_{i,out} = \widehat{\mathbf{X}}^{(h)}_{i,in}$ for all $i \in \{1,\ldots,n\}\backslash \mathbf{i}^{(h)}_{A,miss}$. Every other coordinate is kept unchanged in the process, so that, for all $h' \neq h$, $\forall i$, $\widehat{\mathbf{X}}^{(h')}_{i,out} = \widehat{\mathbf{X}}^{(h')}_{i,in}$. To keep the notation fairly simple, we omit the dependence on w in the notation of $\widehat{\mathbf{X}}_{out}$ even though the two notions are deeply linked by definition. See Figure 3 for an illustration with $p = 2$.

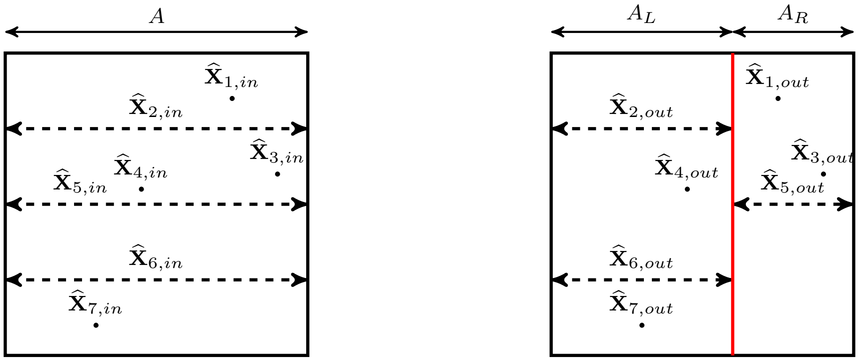
A
$\widehat{\mathbf{X}}_{1,in}$
$\widehat{\mathbf{X}}_{2,in}$
$\widehat{\mathbf{X}}_{3,in}$
$\widehat{\mathbf{X}}_{4,in}$
$\widehat{\mathbf{X}}_{5,in}$
$\widehat{\mathbf{X}}_{6,in}$
$\widehat{\mathbf{X}}_{7,in}$
A_L
A_R
$\widehat{\mathbf{X}}_{1,out}$
$\widehat{\mathbf{X}}_{2,out}$
$\widehat{\mathbf{X}}_{3,out}$
$\widehat{\mathbf{X}}_{4,out}$
$\widehat{\mathbf{X}}_{5,out}$
$\widehat{\mathbf{X}}_{6,out}$
$\widehat{\mathbf{X}}_{7,out}$

Fig 3: We perform a cut and assignation of points were the variable is missing, maximizing the CART criterion $L_n(A,d,w)$.

Finally, the CART criterion in the context of missing values is defined on a cell A by the formula

$$\begin{aligned} L_n(A,d,w) = & \frac{1}{\widehat{N}(A)} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_A\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,in}\in A} \\ & - \frac{1}{\widehat{N}(A)} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_{A_L}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,in}\in A,\, a^{(h)} \le \widehat{\mathbf{X}}^{(h)}_{i,out} < z} \\ & - \frac{1}{\widehat{N}(A)} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_{A_R}\right)^2 \mathbb{1}_{\widehat{\mathbf{X}}_{i,in}\in A,\, z \le \widehat{\mathbf{X}}^{(h)}_{i,out} \le b^{(h)}} \end{aligned}$$

where $\widehat{Y}_{A_L}$ (resp. $\widehat{Y}_{A_R}$) is the empirical mean of the Y_i such that $\widehat{\mathbf{X}}_{i,out}$ belongs to the cell A_L (resp. A_R). The assignation vector w plays a role in the definition of the output imputation vector $\widehat{\mathbf{X}}_{out}$. For a cell A and an input imputation vector $\widehat{\mathbf{X}}_{in}$, the algorithm chooses a cut and assignation $(\widehat{d},\widehat{w})$ by maximizing $L_n(A,d,w)$ over $\mathcal{C}_A \times \mathcal{W}^{(h)}_A$,

$$(\widehat{d},\widehat{w}) \in \underset{\substack{d\in\mathcal{C}_A \\ w\in\mathcal{W}^{(h)}_A}}{\arg\max}\, L_n(A,d,w).$$

Finally, the “imputed” intervals are updated, which we symbolize by $\widehat{\mathbf{X}}^{(\widehat{h})}_{i,in} \leftarrow \widehat{\mathbf{X}}^{(\widehat{h})}_{i,out}$.

Note that, if we know the mechanism of missingness, the possible values for w might change. For example, assume that there are missing values just in $\mathbf{X}^{(1)}$, and its value is missing when it is bigger than τ, that is, we have right censoring. Also consider that we know the value of τ, so the missing mechanism is known. if we perform a cut at τ in $\mathbf{X}^{(1)}$, then the only admissible assignation for $\widehat{\mathbf{X}}_i$, where $i \in \mathbf{i}^{(h)}_{miss}$, is to the right child node.

Another interesting observation corresponds to the relation between our procedure and Missing Incorporated in Attributes (MIA), which assigns all the observations where the splitted variable is missing to the same child node. Our method goes further than MIA, since not only considers these assignations, but also all possible combinations for the assignation of the observations with missing values. Unfortunately, the propose algorithm does not scale well when the percentage proportion of missing data increases.

3.2. *Hypothesis and Main Theorem.* We consider an additive regression model satisfying the following properties

HYPOTHESIS 1 (H1). *The response variable Y is of the form*

$$Y = \sum_{j=1}^{p} m_j(\mathbf{X}^{(j)}) + \varepsilon$$

where $\mathbf{X}$ is uniformly distributed over $[0,1]^p$, ε is an independent Gaussian centered noise with finite variance $\sigma^2 > 0$ and each component m_j is continuous.

HYPOTHESIS 2 (H2). *The random variables $\mathbf{X}_i^{(h)}$ are not observed (missing) by following an MCAR mechanism. The probability of missingness $p_n^{(h)} = \mathbb{P}\left[\mathbf{M}^{(h)} = 1\right]$ only depends on the size n of the sample $\mathcal{D}_n$ and $\lim_{n\to\infty} p_n^{(h)} = c^{(h)}$ where $0 < c^{(h)} < \infty$ is constant for all $h \in \{1, \dots, p\}$.*

THEOREM 1. *Assume that H1 and H2 hold. Denote by q_n the minimum number of points in each final cell of the random trees. Then, under the condition $q_n \to \infty$, the random forest estimator with missing values is consistent in probability, i.e., for all $\xi, \rho > 0$ there exists $N \in \mathbf{N}^\star$, such that for all $n > N$*

$$\mathbb{P}\Big[|m_n(\mathbf{X}) - m(\mathbf{X})| \leq \xi\Big] \geq 1 - \rho$$

The fact that $q_n \to \infty$ implies that the number of points selected in each tree, a_n tends to infinity. Hence, in the sequel, we assume that $a_n \to \infty$ and omit the dependence of the trees over a_n. The condition that $q_n \to \infty$ is sub-optimal, and we think that with some extra effort the results presented in this work can be adapted to a more classical (and more powerful) condition like $a_n/t_n \to \infty$. In this case it is possible that not every tree is consistent but the random forest converges to the regression function.

4. Proof of Theorem 1. Remember that, in our procedure, the CART criterion takes “imputed” intervals which are updated after a cut is selected. From a theoretical point of view we do not longer have observations. Hence, we introduce the notions of the input assigned random variable $\widehat{\mathbf{X}}_{in}$ and output assigned random variable $\widehat{\mathbf{X}}_{out}$. The random variable $\widehat{\mathbf{X}}_{in}$ correspond to a prior assignation whereas $\widehat{\mathbf{X}}_{out}$ corresponds to the distribution of the assignations after the theoretical cut is performed. Furthermore, the binary assignations w are translated into probabilities.

Formally, let $\mathcal{W}$ be the collection of functions from $\mathbb{R}$ to $[0,1]$, and $\widehat{\mathbf{X}}_{in} = (\widehat{\mathbf{X}}_{in}^{(1)}, \dots, \widehat{\mathbf{X}}_{in}^{(p)})$ be the input distribution of imputation, then $\widehat{\mathbf{X}}_{out}^{(h)} | \widehat{\mathbf{X}}_{in}$ is defined as

$$\widehat{\mathbf{X}}_{out}^{(h)} | \widehat{\mathbf{X}}_{in} \in A \overset{d}{=} \begin{cases} \mathbf{X}^{(h)} | \mathbf{X} \in A \text{ if } \mathbf{M}^{(h)} = 0 \\ \mathbf{B}^{(h)} \qquad\quad\;\; \text{if } \mathbf{M}^{(h)} = 1 \end{cases}$$

where

$$\mathbf{B}^{(h)} = \begin{cases} \left(a^{(h)}, z\right) \text{ if } \mathrm{Ber}(w(Y)) = 1 \\ \left(z, b^{(h)}\right) \text{ if } \mathrm{Ber}(w(Y)) = 0 \end{cases}, \quad w \in \mathcal{W} \tag{2}$$

The imputation variable $w(Y)$ is the (random) probability that $\widehat{\mathbf{X}}_{out}^{(h)} < z$ conditionally to $\mathbf{M}^{(h)} = 1$ and Y. Note that $\widehat{\mathbf{X}}_{in}$ always belongs to a cell A, so the above definition of $\widehat{\mathbf{X}}_{out}^{(h)} | \widehat{\mathbf{X}}_{in} \in A$ is well defined. We define the theoretical CART over a cut $d = (h, z)$ and a function $w \in \mathcal{W}$ as

$$\begin{aligned} L^{\star}(A, d, w) =& \mathbb{V}[Y | \widehat{\mathbf{X}}_{in} \in A] - \mathbb{V}[Y | \widehat{\mathbf{X}}_{out}^{(h)} < z, \widehat{\mathbf{X}}_{in} \in A] \mathbb{P}[\widehat{\mathbf{X}}_{out}^{(h)} < z | \widehat{\mathbf{X}}_{in} \in A] \\ & - \mathbb{V}[Y | \widehat{\mathbf{X}}_{out}^{(h)} \geq z, \widehat{\mathbf{X}}_{in} \in A] \mathbb{P}[\widehat{\mathbf{X}}_{out}^{(h)} \geq z | \widehat{\mathbf{X}}_{in} \in A] \end{aligned}$$

and the best cut and assignation $(d^{\star}, w^{\star})$ is selected by maximizing $L^{\star}(A, d, w)$ over $\mathcal{C}_A \times \mathcal{W}$, that is

$$(d^{\star}, w^{\star}) \in \underset{\substack{d \in \mathcal{C}_A \\ w \in \mathcal{W}}}{\arg\max} \, L^{\star}(A, d, w)$$

We define, for any subset $A \subset \mathcal{X}$, the variation of m within A as

$$\Delta(m, A) = \sup_{\mathbf{x}, \mathbf{x}' \in A} |m(\mathbf{x}) - m(\mathbf{x}')|$$

Furthermore, we denote by $A_{s(n)}(\mathbf{X}, \Theta)$ the final cell of the tree built with the random variable Θ that contains $\mathbf{X}$, where $s(n)$ is the number of cuts necessary to construct the cell in the tree. After the last step of the tree construction, we end with a collection of imputed values that corresponds to the last imputed sample $(\widehat{\mathbf{X}}_{1,out}, \ldots, \widehat{\mathbf{X}}_{n,out})$. Each of these vectors are non ambiguously assigned to a specific final cell of the tree. In all that follows, when we write an imputation $\widehat{\mathbf{X}}$ without any specification of *in* or *out*, we refer to the final imputation. The proof of Theorem 1 relies on Proposition 1 below, which states that the variation of the regression function within a cell of a random forest is small for n large enough and helps us to control the error of our predictor. In the sequel, we will abuse of the notation and will not write the dependence of the cells over $\mathbf{X}$ and Θ.

PROPOSITION 1. *Assume that H1 and H2 hold. Then,*

$$\Delta(m, A_{s(n)}) \to 0 \quad \textit{almost surely.}$$

For all $\mathbf{x} \in [0,1]^p$, we denote by $L^{\star}\left(A_{s(n)}, d, w\right)$ the theoretical CART criterion over the cell $A_{s(n)}$ evaluated at a cut $d \in \mathcal{C}_{A_{s(n)}}$ and assignation $w \in \mathcal{W}$. Let $(d^{\star}_{s(n)}, w^{\star}_{s(n)})$ be the optimal couple (cut, assignation) of the cell $A_{s(n)}$ for the theoretical criterion and let $(\widehat{d}_{s(n)}, \widehat{w}_{s(n)})$ be the optimal couple (cut, assignation) of the cell $A_{s(n)}$ for the empirical criterion.

LEMMA 1. *Assume that H1 and H2 are satisfied and fix $\mathbf{x} \in [0,1]^p$. Then for all $\rho, \xi > 0$, there exists $N \in \mathbb{N}^{\star}$ such that, for all $n \geq N$ if $\mathbb{P}\left[L^{\star}\left(A_{s(n)}, d^{\star}_{s(n)}, w^{\star}_{s(n)}\right) \leq \xi\right] \geq 1 - \rho$, then*

$$\Delta(m, A_{s(n)}(\mathbf{x})) \to 0 \quad \textit{almost surely.}$$

LEMMA 2. *Assume that H1 and H2 are satisfied and fix $\mathbf{x} \in [0,1]^p$. Then for all $\rho, \xi > 0$, there exists $N \in \mathbb{N}^{\star}$ such that, for all $n \geq N$*

$$\mathbb{P}\left[L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) \leq \xi\right] \geq 1 - \rho$$

4.1. *Proof of Lemma 1.*

TECHNICAL LEMMA 1. *Assume that H1 and H2 are satisfied and for a cell A, $L^{\star}(A,d,\tilde{w}_{A,d}) \equiv 0$ for all cuts $d=(h,z)\in\mathcal{C}_A$, where*

$$\tilde{w}_{A,d} = \mathbb{P}\left[a^{(h)} \leq \mathbf{X}^{(h)} < z | \mathbf{X}\in A, \mathbf{M}^{(h)}=1\right].$$

Then, m is constant on the cell A.

Proof. Without loss of generality, we will assume that the cut $d=(1,z)$ is performed in the first direction so that $h=1$ and that the bounds of the cell A are a on the left and b on the right. We omit the direction h and simply note X instead of $\mathbf{X}^{(1)}$. Note that if $\widehat{\mathbf{X}}_{out}$ is assigned to the left child node using $\tilde{w}_{A,d}$, which does not depend on Y, then $\widehat{\mathbf{X}}_{out}$ follows the same distribution than $\mathbf{X}$ and then our theoretical CART criterion is similar to the usual one for every cut $d=(h,z)\in\mathcal{C}_A$, that is,

$$L^{\star}\left(d,\tilde{w}_{A,d}\right) = \mathbb{P}\left[a \leq X < z|\mathbf{X}\in A\right]\mathbb{P}\left[z \leq X \leq b|\mathbf{X}\in A\right]\left(\mu_{A_L}^{Y_{in}} - \mu_{A_R}^{Y_{in}}\right)^2$$

where the notation $\mu_{A_L}^{Y_{in}}$ (resp. $\mu_{A_R}^{Y_{in}}$) holds for the conditional expected value of Y given $\widehat{\mathbf{X}}_{in}\in A$ and $a\leq X<z$ (resp. $z\leq X\leq b$). Because $\mathbf{X}$ is uniformly distributed over $[0,1]^p$,

$$\mathbb{P}\left[a \leq X < z|\mathbf{X}\in A\right] = \frac{z-a}{b-a}$$

and

$$\mathbb{P}\left[z \leq X \leq b|\mathbf{X}\in A\right] = \frac{b-z}{b-a}$$

Next, we need to understand the distribution of Y conditionally to $\widehat{\mathbf{X}}_{in}\in A$. This random variable is a mixture of the values of Y such that $\mathbf{X}\in A$ and the ones that where assigned to the cell A through the notion of the variable $\widehat{\mathbf{X}}_{in}$. Since the vector $(\widehat{\mathbf{X}}_{in}, Y)$ has a density, we can give a precise meaning to the quantity

$$\tilde{m}(x) = \mathbb{E}\left[Y|\widehat{\mathbf{X}}_{in}=\mathbf{x}\right], \tag{3}$$

for all $\mathbf{x}\in A$. By assumption H1, we introduce the notation $Y^{(j)} = m_j(X^{(j)}) + \epsilon^{(j)}$, where $\epsilon^{(j)} \sim \mathcal{N}(0,\sigma^2/p)$ so that we have $Y \sim \sum_{j=1}^{p} Y^{(j)}$. Since under the condition $\mathbf{X}\in A$, the random variable $\mathbf{X}$ is also uniformly distributed on the cell A, we have that

$$\begin{aligned}
\mu_{A_L}^{Y_{in}} &= \mathbb{E}\left[\sum_{j=1}^{p} Y^{(j)}|a\leq X<z, \widehat{\mathbf{X}}_{in}\in A\right] \\
&= \sum_{j\geq 2}\mathbb{E}\left[Y^{(j)}|\widehat{\mathbf{X}}_{in}\in A\right] + \mathbb{E}\left[Y^{(1)}|a\leq X<z, \widehat{\mathbf{X}}_{in}\in A\right] \\
&= \tilde{K} + \frac{1}{z-a}C_a^z
\end{aligned}$$

where $\tilde{K}$ does not depend on z, $C_x^y = \int\limits_x^y \tilde{m}_1(t)dt$ and $\tilde{m}_1(t) = \mathbb{E}\left[Y^{(1)}|\widehat{\mathbf{X}}_{in}\in A,\ \mathbf{X}^{(1)}=t\right]$.
Analogously, we have that

$$\mu_{A_R}^{Y_{in}} = \tilde{K} - \frac{1}{b-z}C_a^z + \frac{1}{b-z}C_a^b$$

Therefore,

$$L^\star\left(A,d,\tilde{w}_{A,d}\right)=\left(\frac{z-a}{b-a}\right)\left(\frac{b-z}{b-a}\right)\left(\frac{1}{z-a}C_a^z+\frac{1}{b-z}C_a^z-\frac{1}{b-z}C_a^b\right)^2$$

$$=\frac{1}{(z-a)(b-z)}\left(C_a^z-\frac{z-a}{b-a}C_a^b\right)^2$$

Since $L^\star\left(h,z,\tilde{w}_{A,d}\right)=0$ by assumption, we obtain that for any $a\leq z\leq b$,

$$C_a^z=\frac{z-a}{b-a}C_a^b.$$

This proves that $z\mapsto C_a^z$ is linear in z and thus, $\tilde{m}_1$ is constant on $[a,b]$. Since the law of Y_1 is a mixture of distribution such that $M^{(1)}=0$ and $M^{(1)}=1$, we have that

$$\tilde{m}_1(t)=(1-p^{(1)})m_1(t)+p^{(1)}\mathbb{E}\left[Y^{(1)}|\widehat{\mathbf{X}}_{in}\in A,\ \mathbf{X}^{(1)}=t,\ M^{(1)}=1\right]$$

where the second term does not depend on t (since the value is lost). This forces the function m_1 to be constant on the interval $[a,b]$. And, by additivity, the function m is constant on A.

We need a second technical lemma that states that any cell within a cell for which each split gives a small value of $L^\star$ has also a uniformly small value of the CART criterion $L^\star$.

TECHNICAL LEMMA 2. *Assume constructed the imputation $\widehat{\mathbf{X}}_{in}$. Let $\epsilon>0$, then there exists $\delta>0$ such that for all cell A and cut $d\in\mathcal{C}_A$ satisfying $L^\star(A,d,\tilde{w}_{A,d})\leq\delta$, we have*

$$L^\star(B,d,\tilde{w}_{A,d})\leq\epsilon$$

for all cell $B\subset A$, where both notions of CART criteria are constructed with the same vector of imputation $\widehat{\mathbf{X}}_{in}$.

Proof. We first prove the unidimensional case. Without loss of generality, we can assume that the cell A is the interval $[0,1]$ and we first consider the cell B of the form $[0,a]$ with $0\leq a\leq 1$. Following the lines of Technical Lemma 1, we have that

$$L^\star(A,z,\tilde{w}_{A,d})=\frac{1}{z(1-z)}\left(C_0^z-C_0^1z\right)^2$$

where for all x and y, $C_x^y=\int_x^y\tilde{m}(t)dt$ and $\tilde{m}$ is defined as in Equation (3). Evaluating the CART criterion at the value $z=a$, we have

$$\frac{1}{a(1-a)}\left(C_0^a-C_0^1a\right)^2\leq\delta.$$

For the cell B, at a cut level $0<z<a$, we have that

$$\begin{aligned}L^\star(B,z,\tilde{w}_{A,d})&=\frac{1}{z(a-z)}\left(C_0^z-C_0^a\frac{z}{a}\right)^2\\&=\frac{1}{z(a-z)}\left(C_0^z-C_0^1z+C_0^1z-C_0^a\frac{z}{a}\right)^2\\&\leq\frac{2}{z(a-z)}\left(C_0^z-C_0^1z\right)^2+\frac{2z}{a-z}\left(C_0^1-\frac{C_0^a}{a}\right)^2\\&\leq 2\frac{1-z}{a-z}L^\star(A,z,\tilde{w})+\frac{2z(1-a)}{a(a-z)}\delta\\&\leq 2\frac{a(1-z)+z(1-a)}{a(a-z)}\delta\leq\frac{2}{a(a-z)}\delta.\end{aligned}$$

But since the function $z \mapsto C_0^z$ is differentiable, we have that $C_0^z = C_0^a - (a-z)\tilde{m}(a) + o(a-z)$ when $z \to a$ by a Taylor expansion. This shows that $L^\star(B, z, \tilde{w}_{A,d}) \to 0$ when $z \to a$ and then there exists a $\delta_0 > 0$ such that $L^\star(B, z, \tilde{w}_{A,d}) \leq \epsilon$ when $(a-z) \leq \delta_0$. Then, using the previous inequalities for $0 < z < a - \delta_0$, we have that for $\delta = \epsilon\delta_0 a/2$, $L^\star(B, z, \tilde{w}_{A,d}) \leq \epsilon$. In the same way, we generalize the previous arguments for $A = [0,1]^p$ and for a specific type of cell

$$B = \{\mathbf{x} \in A : \mathbf{x}^{(1)} \leq a\}$$

with $0 < a < 1$. The values C_y^z are then replaced by $\int_y^z \int_0^1 \cdots \int_0^1 \tilde{m}(t) dt_1 dt_2 \ldots dt_p$. The result can be repeated for the case $B = \{\mathbf{x} \in A : \mathbf{x}^{(1)} \geq a\}$.

For the general case of $B \subset A$ we see that any B can be obtained by a finite sequence of $B = B_k \subset B_{k-1} \subset \ldots B_1 \subset A$ of subset constructed by the scheme described above. This finishes the proof.

Proof of Lemma 1. We will show that $\Delta(m, A_{s(n)}(\mathbf{x})) \to 0$ a.s. by contradiction. We assume that with positive probability, there exists a positive constant $c > 0$ and a subsequence $\phi(n)$ of cells $A_{s(\phi(n))}$ such that $\Delta(m, A_{s(\phi(n))}(\mathbf{x})) > c$. This means that in each set $A_{s(\phi(n))}(\mathbf{x})$ one can find a pair of elements (x_n, y_n) such that

$$|m(x_n) - m(y_n)| > c.$$

The sequences $(x_n)_n$ and $(y_n)_n$ belong to the compact set $[0,1]^p$ so one can extract a subsequence $\psi(n)$ such that $(x_n)_n$ and $(y_n)_n$ converge respectively to two points x and y. By continuity of m, we have that

$$|m(x) - m(y)| > c.$$

At the cost of taking an x' and y' close to x and y, satisfying

$$|m(x') - m(y')| > \frac{c}{2},$$

and such that, for n large enough, all the cells $A_{s(\phi(n))}$ contain the pair of points x' and y'. By hypothesis, we know that

$$\sup_{d \in \mathcal{C}_{A_{s(\phi\circ\psi(n))}}} L^\star(A_{s(\phi\circ\psi(n))}, d, \tilde{w}) \to 0 \quad \text{in probability}$$

where we just wrote $\tilde{w}$ for the choice of the assignation given in Technical Lemma 1. So one can extract a subsequence $\chi(n)$ such that the $\sup_d L^\star(A_{s(\phi\circ\psi\circ\chi(n))}, d, \tilde{w})$ converges to 0 almost surely. For simplicity, we keep denoting n for the sub-sequence $\phi \circ \psi \circ \chi(n)$ in the following part of the proof. Lastly, we define the sequence of cells $(C_i)_{i \geq 1}$ such that

$$C_i = \bigcap_{k=1}^{i} A_{s(k)}.$$

These cells form a non increasing sequence for the inclusion order and for each i, $C_i \subset A_{s(i)}$. The cells C_i inherit the same vector of imputation as for $A_{s(i)}$. We can use Technical Lemma 2 to get that $\sup_{d \in \mathcal{C}_{A_{s(n)}}} L^\star(C_n, d, \tilde{w}) \to 0$ a.s. Since the sequence of cells $(C_i)_i$ is a non increasing sequence, there exists a cell, denoted C_∞ that is the limit of the cells C_i when $i \to \infty$ in the sense

$$C_\infty = \bigcap_{i \geq 1} C_i.$$

To see that, one can write $C_i = \prod_h [a_i^{(h)}, b_i^{(h)}]$ and take the limits of the sequences $(a_i^{(h)})_i$ and $(b_i^{(h)})_i$. The objective is to show that there is no other possibilities than having $L^\star(C_\infty, d, \tilde{w}) = 0$ for every cut d. For a cell A, the CART criterion $L^\star(A, d, \tilde{w}_{A,d})$ has a continuous behavior with respect to A. Indeed, we see that

$$L^\star\left(A, d, \tilde{w}_{A,d}\right) = \mathbb{P}\left[a^{(h)} \le \mathbf{X}^{(h)} < z | \mathbf{X} \in A\right] \mathbb{P}\left[z \le \mathbf{X}^{(h)} \le b^{(h)} | \mathbf{X} \in A\right] (\widehat{\mu}_{A_L} - \widehat{\mu}_{A_R})^2$$

which is a product of three terms that are uniformly continuous in A (since m is a continuous function) with respect to the distance between cells given by $\Delta(A, B) = \max_h |x_A^{(h)} - x_B^{(h)}| + \max_h |y_A^{(h)} - y_B^{(h)}|$ where $x_A^{(h)}$ and $y_A^{(h)}$ are defined as $A = \bigcap_h [x_A^{(h)}, y_A^{(h)}]$. Since $C_i \to C_\infty$ for the distance Δ, we have that for all $\epsilon > 0$, for all i large enough, for all cut d of the final cell C_∞,

$$|L^\star\left(C_i, d, \tilde{w}\right) - L^\star\left(C_\infty, d, \tilde{w}\right)| \le \epsilon.$$

But since the $L^\star\left(C_i, d, \tilde{w}\right)$ converges almost surely uniformly in d to 0, and ϵ is arbitrary, for every cut d inside the valid cuts of C_∞, we have that $L^\star(C_\infty, d, \tilde{w}) = 0$. But then by Technical Lemma 1, the function m has to be constant in the cell C_∞. But the points x' and y' do belong to the cell C_∞ since they belong to each of the cells in the intersection. So $m(x') = m(y')$ which contradicts the fact that $|m(x') - m(y')| > c/2$. This proves that

$$\Delta(m, A_{s(n)}(\mathbf{x})) \to 0 \quad \text{almost surely.}$$

4.2. *Proof of Lemma 2.* Remember that $A_{s(n)}$ denotes the final cell of the tree built with the random variable Θ that contains $\mathbf{X}$, where $s(n)$ is the number of cuts necessary to construct the cell. Similarly, A_k is the same cell but where only the first k cuts have been performed.

TECHNICAL LEMMA 3. *Assume that H1 and H2 hold and fix $\mathbf{x} \in [0,1]^p$. For all $\rho, \xi > 0$ there exists $N \in \mathbb{N}^\star$ such that for all $n \ge N$ there exists $k_0(n) \in \mathbb{N}^\star$ such that*

$$\mathbb{P}\left[\left|L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) - L_n\left(A_k, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right)\right| \le \xi\right] \ge 1 - \rho, \quad \textit{for all } k \ge k_0(n)$$

Proof. Fix $\alpha, \rho > 0$ and consider the following standard inequality on a Gaussian tail

$$\mathbb{P}[\varepsilon_1 \ge \alpha] \le \frac{\sigma^2}{t\sqrt{2\pi}} \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

then, simple calculations show that, for all $n \in \mathbf{N}^\star$

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} \varepsilon_i\right| \ge n\alpha\right\} \le \frac{\sigma}{\alpha\sqrt{n}} \exp\left\{-\frac{\alpha^2 n}{2\sigma^2}\right\} \tag{4}$$

Note that there are at most $n(n+1)/2$ sets of the form $\{i : \mathbf{X}_i^{(h)} \in [a_n, b_n], \mathbf{M}_i^{(h)} = 0\}$ for $0 \le a_n < b_n \le 1$. On the other hand, let $(Y_{(1)}, \ldots, Y_{(n)})$ be the order vector of Y, since missing observations are assigned to the cell using Y and maximizing the CART criterion, this implies that close values of Y must be assigned to the same cell, thus once again note that there are at most $n(n+1)/2$ sets of the form $\{i : Y_{(i)} \in [a_n, b_n], \mathbf{M}_i^{(h)} = 1\}$ for $0 \le a_n < b_n \le 1$.

We deduce from Equation (4) and the union bound, that there exists $N_1 \in \mathbf{N}^\star$ such that, with probability at least $1-\rho$, for all $n \geq N_1$ and all $0 \leq a_n < b_n \leq 1$ satisfying $\widehat{N}\left(\prod_{h=1}^{p}[a_n^{(h)}, b_n^{(h)}]\right) \geq q_n$,

$$\left|\frac{1}{\widehat{N}\left(\prod_{h=1}^{p}[a_n^{(h)}, b_n^{(h)}]\right)}\sum_{i=1}^{n}\varepsilon_i \mathbb{1}_{\widehat{\mathbf{X}}_i \in A}\right| \leq \frac{\sigma n^{4p}}{\alpha\sqrt{n}}\exp\left\{-\frac{\alpha^2 n}{2\sigma^2}\right\} \leq \alpha \tag{5}$$

Furthermore, making use of the same ideas and applying the inequality $\mathbb{P}[\chi^2(n) \geq 5n] \leq e^{-n}$ (for interested readers, see Laurent and Massart [21]), there exists $N_2 \in \mathbb{N}^\star$ such that, with probability at least $1-\rho$ for all $n \geq N_2$ and all $0 \leq a_n < b_n \leq 1$ satisfying $\widehat{N}\left(\prod_{h=1}^{p}[a_n^{(h)}, b_n^{(h)}]\right) \geq q_n$

$$\frac{1}{\widehat{N}\left(\prod_{h=1}^{p}[a_n^{(h)}, b_n^{(h)}]\right)}\sum_{i=1}^{n}\varepsilon_i^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A} \leq \tilde{\sigma}^2 \tag{6}$$

where $\tilde{\sigma}^2$ is a positive constant, depending only on ρ. Since $(A_k)_k$ is a decreasing sequence of compact sets, for every $\xi > 0$ there exists k_0 such that, for all $k \geq k_0$

$$\max\left(\|\mathbf{a}_k - \mathbf{a}_{s(n)}\|_\infty, \|\mathbf{b}_k - \mathbf{b}_{s(n)}\|_\infty\right) \leq \xi \tag{7}$$

where $\mathbf{a}_k = \left(a_k^{(1)}, \ldots, a_k^{(p)}\right) \in [0,1]^p$ and $\mathbf{b}_k = \left(b_k^{(1)}, \ldots, b_k^{(p)}\right) \in [0,1]^p$ such that

$$A_k = \prod_{j=1}^{p}\left[a_k^{(j)}, b_k^{(j)}\right]$$

We define $\mathbf{a}_{s(n)}$ and $\mathbf{b}_{s(n)}$ analogously for the cell $A_{s(n)}$. Since the variable $\mathbf{X}$ is uniformly distributed in the hypercube $[0,1]^p$ and the missing entries follow the MCAR mechanism we have that, for any cell A, $\widehat{N}_{obs}(A)/\widehat{N}_{obs}([0,1]^p) \to \mathrm{Vol}(A)$ almost surely as $n \to \infty$. Furthermore, for any $k \geq k_0(n)$, Equation (7) implies that $\mathrm{Vol}(A_k) \leq (1+2\xi)^p \mathrm{Vol}(A_{s(n)})$. Then there exists $N_3 \in \mathbf{N}^\star$ such that for all $n \geq N_3$, $\widehat{N}_{obs}(A_k) \leq (1+3\xi)^p \widehat{N}_{obs}(A_{s(n)})$, and

$$\widehat{N}_{obs}(A_k \setminus A_{s(n)}) \leq \xi' \widehat{N}_{obs}(A_{s(n)}) \leq \xi' \widehat{N}_{obs}(A_k) \tag{8}$$

where $\xi' = (1+3\xi)^p - 1$. On the other hand, for any fixed n, the quantity $u_k = \widehat{N}_{miss}(A_k) - \widehat{N}_{miss}(A_{s(n)})$ converges to 0 almost surely as $k \to \infty$. Now taking $n \geq N_3$ fixed we have that there exists a $k_1(n)$ such that for all $k \geq k_1(n)$, with probability at least $1-\rho$, we have

$$\begin{aligned}
\widehat{N}(A_k \backslash A_{s(n)}) &= (\widehat{N}_{obs}(A_k) + \widehat{N}_{miss}(A_k)) - (\widehat{N}_{obs}(A_{s(n)}) + \widehat{N}_{miss}(A_{s(n)})) \\
&= \widehat{N}_{obs}(A_k \setminus A_{s(n)}) + u_k \\
&\leq 2\xi' \widehat{N}_{obs}(A_{s(n)}) \qquad (9)
\end{aligned}$$

where we used Equation (8) and the fact that convergence almost sure implies convergence in probability. For the rest of the proof, we take $k \geq \max\{k_0(n), k_1(n)\}$ and assume that Equations (5),(6) and (9) are satisfied, which occurs with probability at least $1-3\rho$ for every $n > N$ with $N = \max\{N_1, N_2, N_3\}$. Note that $A_k \setminus A_{s(n)}$ is either a final node, contains at least one final node or is empty (in which case the result holds trivially). Since each final node contains at least q_n points then for q_n sufficiently large, using equation (5) we conclude that $|\widehat{Y}_{A_k}| \leq \|m\|_\infty + \alpha$, $|\widehat{Y}_{A_{s(n)}}| \leq \|m\|_\infty + \alpha$ and $|\widehat{Y}_{A_k \setminus A_{s(n)}}| \leq \|m\|_\infty + \alpha$. We use the following decomposition

$$\left| L_n\left(A_k, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) - L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) \right| \leq K_0 + K_L + K_R$$

where the three terms K_0, K_L and K_R are given by

$$K_0 = \Big| \frac{1}{\widehat{N}(A_k)} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_k})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_k} - \frac{1}{\widehat{N}(A_{s(n)})} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_{s(n)}})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} \Big|,$$

$$\begin{aligned} K_L = \Big| & \frac{1}{\widehat{N}(A_k)} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_{L,k}})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_k, \widehat{\mathbf{X}}_i^{(\widehat{h})} < \widehat{z}} \\ & - \frac{1}{\widehat{N}(A_{s(n)})} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_{L,s(n)}})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}, \widehat{\mathbf{X}}_i^{(\widehat{h})} < \widehat{z}} \Big|, \end{aligned}$$

$$\begin{aligned} K_R = \Big| & \frac{1}{\widehat{N}(A_k)} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_{R,k}})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_k, \widehat{\mathbf{X}}_i^{(\widehat{h})} \geq \widehat{z}} \\ & - \frac{1}{\widehat{N}(A_{s(n)})} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_{R,s(n)}})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}, \widehat{\mathbf{X}}_i^{(\widehat{h})} \geq \widehat{z}} \Big|. \end{aligned}$$

We first tackle the term K_0 that can be upper bounded once again by using a similar split in $K_0 \leq K_{0,1} + K_{0,2} + K_{0,3}$ where

$$K_{0,1} = \Big| \frac{1}{\widehat{N}(A_k)} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_k})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} - \frac{1}{\widehat{N}(A_k)} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_{s(n)}})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} \Big|$$

$$K_{0,2} = \Big| \frac{1}{\widehat{N}(A_k)} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_{s(n)}})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} - \frac{1}{\widehat{N}(A_{s(n)})} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_{s(n)}})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} \Big|$$

$$K_{0,3} = \Big| \frac{1}{\widehat{N}(A_k)} \sum_{i=1}^n (Y_i - \widehat{Y}_{A_k})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_k \setminus A_{s(n)}} \Big|$$

For $K_{0,1}$, observe that

$$\begin{aligned} |\widehat{Y}_{A_k} - \widehat{Y}_{A_{s(n)}}| &= \Big| \frac{1}{\widehat{N}(A_k)} \sum_{i=1}^n Y_i \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_k \setminus A_{s(n)}} + \frac{1}{\widehat{N}(A_k)} \sum_{i=1}^n Y_i \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} - \widehat{Y}_{A_{s(n)}} \Big| \\ &\leq \frac{\widehat{N}(A_k \setminus A_{s(n)})}{\widehat{N}(A_k)} |\widehat{Y}_{A_k \setminus A_{s(n)}} - \widehat{Y}_{A_{s(n)}}| \\ &\leq 4\xi'(\|m\|_\infty + \alpha) \end{aligned}$$

Hence,

$$\begin{aligned} K_{0,1} &\leq \frac{2}{\widehat{N}(A_k)} |\widehat{Y}_{A_{s(n)}} - \widehat{Y}_{A_k}| \left| \sum_{i=1}^n \left(Y_i + \frac{\widehat{Y}_{A_{s(n)}} + \widehat{Y}_{A_k}}{2} \right) \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} \right| \\ &\leq \frac{8\xi'(\|m\|_\infty + \alpha)}{\widehat{N}(A_k)} \left[\left| \widehat{N}(A_{s(n)}) \widehat{Y}_{A_{s(n)}} \right| + \left| \frac{\widehat{Y}_{A_{s(n)}} + \widehat{Y}_{A_k}}{2} \widehat{N}(A_{s(n)}) \right| \right] \\ &\leq 16\xi'(\|m\|_\infty + \alpha)^2 \end{aligned}$$

For the term $K_{0,2}$, with the help of Equation (6) observe that

$$\begin{aligned}
K_{0,2} &\le \frac{\widehat{N}(A_k \setminus A_{s(n)})}{\widehat{N}(A_k)} \Big| \frac{1}{\widehat{N}(A_{s(n)})} \sum_{i=1}^{n} (Y_i - \widehat{Y}_{A_{s(n)}})^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} \Big| \\
&\le 2\xi' \Big| \frac{1}{\widehat{N}(A_{s(n)})} \sum_{i=1}^{n} Y_i^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} + \widehat{Y}^2_{A_{s(n)}} \Big| \\
&\le 2\xi' \Bigg[(\|m\|_\infty + \alpha)^2 + \frac{1}{\widehat{N}(A_{s(n)})} \sum_{i=1}^{n} m^2(\mathbf{X}_i) \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} \\
&\qquad + \Big| \frac{2}{\widehat{N}(A_{s(n)})} \sum_{i=1}^{n} m(\mathbf{X}_i)\varepsilon_i \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} \Big| + \frac{1}{\widehat{N}(A_{s(n)})} \sum_{i=1}^{n} \varepsilon_i^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}} \Bigg] \\
&\le 2\xi' \left[(\|m\|_\infty + \alpha)^2 + \|m\|_\infty^2 + 2\|m\|_\infty \alpha + \tilde{\sigma}^2 \right]
\end{aligned}$$

Regarding $K_{0,3}$, observe that

$$\begin{aligned}
K_{0,3} &\le 2\xi' \Big| \frac{1}{\widehat{N}(A_k \setminus A_{s(n)})} \sum_{i=1}^{n} (Y_i^2 + 2Y_i \widehat{Y}_{A_k} + \widehat{Y}^2_{A_k}) \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_k \setminus A_{s(n)}} \Big| \\
&\le 2\xi' \Bigg[\Big| \frac{1}{\widehat{N}(A_k \setminus A_{s(n)})} \sum_{i=1}^{n} (m(\mathbf{X}_i) + \varepsilon_i)^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_k \setminus A_{s(n)}} \Big| \\
&\qquad + \frac{2}{\widehat{N}(A_k \setminus A_{s(n)})} (\|m\|_\infty + \alpha) \Big| \sum_{i=1}^{n} (m(\mathbf{X}_i) + \varepsilon_i) \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_k \setminus A_{s(n)}} \Big| \\
&\qquad + (\|m\|_\infty + \alpha)^2 \Bigg] \\
&\le 2\xi' \Bigg[\|m\|_\infty^2 + 2\|m\|_\infty \Big| \frac{1}{\widehat{N}(A_k \setminus A_{s(n)})} \sum_{i=1}^{n} \varepsilon_i \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_k \setminus A_{s(n)}} \Big| \\
&\qquad + \frac{1}{\widehat{N}(A_k \setminus A_{s(n)})} \sum_{i=1}^{n} \varepsilon_i^2 \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_k \setminus A_{s(n)}} + 2(\|m\|_\infty + \alpha)^2 + (\|m\|_\infty + \alpha)^2 \Bigg] \\
&\le 2\xi' [3(\|m\|_\infty + \alpha)^2 + \|m\|_\infty^2 + 2\|m\|_\infty \alpha + \tilde{\sigma}^2]
\end{aligned}$$

Therefore, there exists a universal constant $C > 0$ such that $K_0 \le C\xi$ and with similar arguments we can show that $K_L \le C\xi$ and $K_R \le C\xi$. Which concludes that with probability at least $1 - 3\rho$,

$$\left| L_n\left(A_k, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) - L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) \right| \le 3C\xi$$

Proof of Lemma 2. Assume that H1 and H2 are satisfied, fix $\mathbf{x} \in [0,1]^p$ and Θ. Then, let us show by contradiction that for all $\xi > 0$, there exists $N \in \mathbb{N}^\star$ such that, with probability at least $1 - \rho$ for all $n \ge N$

$$L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) \le \xi.$$

So, assume that there exists $c > 0$, $0 < p_0 < 1$ and a sub-sequence $\phi(n)$ such that

$$L_{\phi(n)}\left(A_{s(\phi(n))}, \widehat{d}_{s(\phi(n))}, \widehat{w}_{s(\phi(n))}\right) > c$$

with probability at least p_0. To keep the notation simple, we omit to write $\phi(n)$ and still write n for the indexes of the sub-sequence. Additionally, assume that k is sufficiently large so that the conclusion of Technical Lemma 3 is satisfied, hence

$$\left|L_n\left(A_k, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) - L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right)\right| \le \xi \tag{10}$$

and equation (7) is satisfied. Note that all the feasible cuts d in A_k must be performed in $A_k \setminus A_{s(n)}$, otherwise d would split $A_{s(n)}$ and $(A_k)_k$ would not converge to $A_{s(n)}$ (see figure 4 for an illustration in $p = 2$). From here we conclude that

$$L_n\left(A_k, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) \le \sup_{\substack{d \in \mathcal{C}_{A_k} \cap \mathcal{C}_{A_{s(n)}} \\ w \in \mathcal{W}_{A_k}}} L_n\left(A_k, d, w\right) \le \sup_{\substack{d \in \mathcal{C}_{A_k} \\ w \in \mathcal{W}_{A_k}}} L_n\left(A_k, d, w\right) \le \xi$$

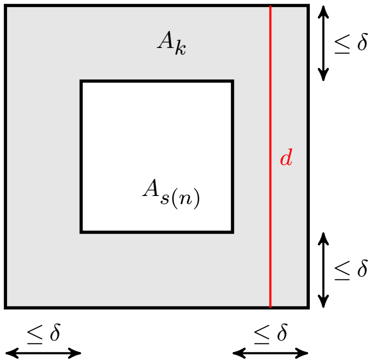

Fig 4: All feasible cuts in A_k must be performed in the $A_k \setminus A_{s(n)}$, like the cut d in the figure, otherwise the cut would split $A_{s(n)}$.

On the other hand, from equation (10), with probability at least p_0, we have

$$c - \xi \le L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) - \xi^2 \le L_n\left(A_k, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right)$$

Hence, we have, with probability at least p_0,

$$c - C\xi \le L_n\left(A_k, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) \le \xi$$

which is absurd, since $c > 0$ is fixed and ξ is arbitrarily small. Thus the result follows.

4.3. *Proof of Proposition 1.* For a cell A, fix a cut $d \in \mathcal{C}_A$ and consider a function $w \in \mathcal{W}$, we need to define $L_n(A, d, w)$. This is done according to the following procedure, first create a random vector W of dimension $\widehat{N}^{(h)}_{miss}(A) = \text{Card}(\mathbf{i}^{(h)}_{A,miss})$, where $W_k = \text{Ber}(w(Y_{j_k}))$ for $j_k \in \mathbf{i}^{(h)}_{A,miss}$, then assign the observations $\widehat{\mathbf{X}}_{j_k}$ to the child nodes according to the random vector W. Once we have assigned the observations to the child nodes, we evaluate the empirical CART criterion L_n considering these assignations. Note that in this case L_n is a random variable and the assignations are independent to each other so $L_n(A,d,w)$ is a sum of independent random variables with the same distribution. Hence, by the strong law of large numbers $L_n(A,d,w) \to L^\star(A,d,w)$ almost surely as $n \to \infty$ for all cuts $d \in \mathcal{C}_A$ and all functions $w \in \mathcal{W}$.

We prove the almost sure convergence of $\Delta(m, A_{s(n)})$ towards 0 by showing that the theoretical CART criterion of the sequence $(A_{s(n)})_n$ tends to 0 and use of Lemmas 1 and 2. Note that

$$\begin{aligned}
&L^\star\left(A_{s(n)}, d^\star_{s(n)}, w^\star_{s(n)}\right) - L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right)\\
&= L^\star\left(A_{s(n)}, d^\star_{s(n)}, w^\star_{s(n)}\right) - L_n\left(A_{s(n)}, d^\star_{s(n)}, w^\star_{s(n)}\right)\\
&\qquad\qquad + L_n\left(A_{s(n)}, d^\star_{s(n)}, w^\star_{s(n)}\right) - L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right)\\
&\le L^\star\left(A_{s(n)}, d^\star_{s(n)}, w^\star_{s(n)}\right) - L_n\left(A_{s(n)}, d^\star_{s(n)}, w^\star_{s(n)}\right)
\end{aligned}$$

Where the last inequality comes from noting that $L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) \ge L_n\left(A_{s(n)}, d, w\right)$ for all cut $d \in \mathcal{C}_{A_{s(n)}}$ and assignation $w \in \mathcal{W}^{(\widehat{h})}_{A_{s(n)}}$, where $\widehat{d}_{s(n)} = (\widehat{h}, \widehat{z})$. As discussed above, by strong law of large numbers $L^\star\left(A_{s(n)}, d^\star_{s(n)}, w^\star_{s(n)}\right) - L_n\left(A_{s(n)}, d^\star_{s(n)}, w^\star_{s(n)}\right) \to 0$, almost surely. Fix $\xi, \rho > 0$, for n sufficiently large, we have

$$L^\star\left(A_{s(n)}, d^\star_{s(n)}, w^\star_{s(n)}\right) - L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) \le \xi \quad \text{almost surely.}$$

On the other hand, by Lemma 2, there exists N_1 such that for all $n \ge N_1$, with probability at least $1-\rho$

$$L_n\left(A_{s(n)}, \widehat{d}_{s(n)}, \widehat{w}_{s(n)}\right) \le C\xi$$

Hence, with the same probability,

$$L^\star\left(A_{s(n)}, d^\star_{s(n)}, w^\star_{s(n)}\right) \le \xi$$

And by Lemma 1, we conclude that

$$\Delta(m, A_{s(n)}) \xrightarrow{a.s.} 0$$

4.4. *Proof of Theorem 1.* Let

$$m_n(\mathbf{X}) = \frac{1}{\widehat{N}(A_{s(n)}(\mathbf{X}))} \sum_{i=1}^{n} Y_i \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}(\mathbf{X})}$$

our tree estimator and define two other quantities. The first one takes our partition of the cells $A_{s(n)}$ built up using the imputed variables but considers the local means using the complete (unseen) observations $\mathbf{X}_i$,

$$m'_n(\mathbf{X}) = \frac{1}{N(A_{s(n)}(\mathbf{X}))} \sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_{s(n)}(\mathbf{X})},$$

while the second one takes $m(\mathbf{X}_i)$ for the prediction and the complete observations $\mathbf{X}_i$,

$$m''_n(\mathbf{X}) = \frac{1}{N(A_{s(n)}(\mathbf{X}))} \sum_{i=1}^{n} m(\mathbf{X}_i) \mathbb{1}_{\mathbf{X}_i \in A_{s(n)}(\mathbf{X})}.$$

By Equation (5), we know that for all $\alpha, \xi > 0$ there exists $N \in \mathbf{N}^\star$, such that for all $n \geq N$,

$$\begin{aligned}\mathbb{P}\left[|m'_n(\mathbf{X}) - m''_n(\mathbf{X})| \geq \alpha\right] &= \mathbb{P}\left[\left|\frac{1}{N(A_{s(n)}(\mathbf{X}))}\sum_{i=1}^{n}\Big(Y_i - m(\mathbf{X}_i)\Big)\mathbb{1}_{\mathbf{X}_i \in A_{s(n)}(\mathbf{X})}\right| \geq \alpha\right] \\ &= \mathbb{P}\left[\left|\frac{1}{N(A_{s(n)}(\mathbf{X}))}\sum_{i=1}^{n}\varepsilon_i \mathbb{1}_{\mathbf{X}_i \in A_{s(n)}(\mathbf{X})}\right|\right] \\ &\leq \xi\end{aligned}$$

On the other hnd, note that $m''_n(\mathbf{X}) = \sum_{i=1}^{n} W_{n,i}(\mathbf{X}) m(\mathbf{X}_i)$, where

$$W_{n,i}(\mathbf{X}) = \frac{1}{N(A_{s(n)}(\mathbf{X}))}\mathbb{1}_{\mathbf{X}_i \in A_{s(n)}(\mathbf{X})}$$

Then,

$$\begin{aligned}\mathbb{E}[m''_n(\mathbf{X}) - m(\mathbf{X})]^2 &= \mathbb{E}\left[\sum_{i=1}^{n} W_{n,i}(\mathbf{X}) m(\mathbf{X}_i) - m(\mathbf{X})\right]^2 \\ &= \mathbb{E}\left[\sum_{i=1}^{n}\sqrt{W_{n,i}(\mathbf{X})}\sqrt{W_{n,i}(\mathbf{X})}(m(\mathbf{X}_i) - m(\mathbf{X}))\right]^2 \\ &\quad \text{(Applying Cauchy-Schwartz's inequality)} \\ &\leq \mathbb{E}\left[\sum_{i=1}^{n} W_{n,i}(\mathbf{X})\sum_{i=1}^{n} W_{n,i}(\mathbf{X})(m(\mathbf{X}_i) - m(\mathbf{X}))^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} W_{n,i}(\mathbf{X})(m(\mathbf{X}_i) - m(\mathbf{X}))^2 \mathbb{1}_{\mathbf{X}_i \in A_{s(n)}(\mathbf{X})}\right]\end{aligned}$$

Note that $(m(\mathbf{X}_i) - m(\mathbf{X}))^2 \mathbb{1}_{\mathbf{X}_i \in A_{s(n)}(\mathbf{X})} \leq \Delta(m, A_{s(n)}(\mathbf{X}))^2$, hence

$$\mathbb{E}[m''_n(\mathbf{X}) - m(\mathbf{X})]^2 \leq \mathbb{E}\left[\Delta(m, A_{s(n)}(\mathbf{X}))^2\right]$$

Since $\Delta(m, A_{s(n)}(\mathbf{X})) \leq \Delta(m, [0,1]^p) < \infty$, we can use the dominated convergence theorem, and conclude that, using Proposition 1,

$$\lim_{n\to\infty} \mathbb{E}[m''_n(\mathbf{X}) - m(\mathbf{X})]^2 = 0.$$

Hence we have the consistency $m'_n(\mathbf{X}) \xrightarrow{\mathbb{P}} m(\mathbf{X})$. This means that a (fictive) estimator built upon the empirical partition but where all the values are observed to compute the empirical mean step is consistent. We will use this fact to show $m_n(\mathbf{X}) \xrightarrow{\mathbb{P}} m(\mathbf{X})$.

First, consider the case in dimension 1. We use the specific case where a cut d and an assignation w (of the cell $A_{s(n)}$) leaves all the observed points to the left and assigns all the missing observations to the right. By Lemma 2, we know that for n sufficiently large $L_n(A_{s(n)}, d, w) \leq \xi$. As already seen in Equation (1),

$$L_n(A_{s(n)}, d, w) = \frac{\widehat{N}_{obs}(A_{s(n)})\widehat{N}_{miss}(A_{s(n)})}{\widehat{N}(A_{s(n)})\widehat{N}(A_{s(n)})}\left(\widehat{Y}_{obs} - \widehat{Y}_{miss}\right)^2$$

By the convergence of $m'_n(\mathbf{X})$, we have, in particular that $\widehat{Y}_{obs} \xrightarrow{\mathbb{P}} m(\mathbf{X})$.

Using the same ideas as in Lemma 2, there exists $N \in \mathbf{N}^\star$ such that, with probability at least $1-\rho$ for all $n \geq N$, $\widehat{N}_{obs}(A_{s(n)})/\widehat{N}(A_{s(n)}) \geq c$, where $c > 0$ is a constant. On the other hand, if $\widehat{N}_{miss}(A_{s(n)})/\widehat{N}(A_{s(n)}) \xrightarrow{\mathbb{P}} 0$ then, trivially $m_n(\mathbf{X}) \xrightarrow{\mathbb{P}} m'_n(\mathbf{X})$ so let us consider the case $\widehat{N}_{miss}(A_{s(n)})/\widehat{N}(A_{s(n)}) \geq c'$, hence with probability at least $1-\rho$,

$$\left(\widehat{Y}_{obs} - \widehat{Y}_{miss}\right)^2 \leq \frac{\xi}{cc'}.$$

This shows that the random variable $\widehat{Y}_{obs} - \widehat{Y}_{miss}$ converges to 0 in probability. Since $\widehat{Y}_{obs} \xrightarrow{\mathbb{P}} m(\mathbf{X})$, we obtain that $\widehat{Y}_{miss} \xrightarrow{\mathbb{P}} m(\mathbf{X})$. Finally, using the following formula for $m_n(\mathbf{X})$

$$m_n(\mathbf{X}) = \frac{\widehat{N}_{obs}(A_{s(n)}(\mathbf{X}))}{\widehat{N}(A_{s(n)}(\mathbf{X}))}\widehat{Y}_{obs} + \frac{\widehat{N}_{miss}(A_{s(n)}(\mathbf{X}))}{\widehat{N}(A_{s(n)}(\mathbf{X}))}\widehat{Y}_{miss}$$

we conclude that $m_n(\mathbf{X}) \xrightarrow{\mathbb{P}} m(\mathbf{X})$. For dimension bigger than 1, denote again $Y^{(j)} = m_j(X^{(j)}) + \epsilon^{(j)}$, where $\epsilon^{(j)} \sim \mathcal{N}(0, \sigma^2/p)$ so that we have $Y \sim \sum_{j=1}^{p} Y^{(j)}$ and define

$$m_n^{(j)}(\mathbf{X}) = \frac{1}{\widehat{N}(A_{s(n)}(\mathbf{X}))}\sum_{i=1}^{n} Y_i^{(j)} \mathbb{1}_{\widehat{\mathbf{X}}_i \in A_{s(n)}(\mathbf{X})}.$$

Consider the cut in the direction 1 which leaves all the observations where $\mathbf{M}^{(1)} = 0$ to the left and assigns the observations where $\mathbf{M}^{(1)} = 1$ to the right. We denote $\widehat{Y}_{obs(1)}$ and $\widehat{Y}_{miss(1)}$ the respective (on the left and on the right) empirical means. By the arguments in dimension 1, we have that $\widehat{Y}_{obs(1)} - \widehat{Y}_{miss(1)} \to 0$ in probability. By definition,

$$\widehat{Y}_{obs(1)} - \widehat{Y}_{miss(1)} = \widehat{Y}_{obs}^{(1)} - \widehat{Y}_{miss}^{(1)} + \left(\sum_{j=2}^{p} \widehat{Y}_{obs(1)}^{(j)} - \sum_{j=2}^{p} \widehat{Y}_{miss(1)}^{(j)}\right).$$

Since the random variable $Y^{(j)}$ $(j \neq 1)$ is independent of the random variable $\widehat{\mathbf{X}}^{(1)}$ conditionally to $\widehat{\mathbf{X}} \in A_{s(n)}$, the distributions of the two sums on the right hand side are equal. Since each random variable $\widehat{Y}^{(j)}$ converges (see Technical Lemma 1) we conclude that the difference of the two sums converges in probability to 0. Hence, $\widehat{Y}_{obs}^{(1)} - \widehat{Y}_{miss}^{(1)} \to 0$ in probability. This finally shows that $m_n^{(1)}(\mathbf{X}) \xrightarrow{\mathbb{P}} m_1(\mathbf{X})$. Similarly, we show that for all $j \geq 1$, $m_n^{(j)}(\mathbf{X}) \xrightarrow{\mathbb{P}} m_j(\mathbf{X})$ and then $m_n(\mathbf{X}) \xrightarrow{\mathbb{P}} m(\mathbf{X})$ by summation of the p previous convergences which concludes the proof of the Theorem.

REFERENCES

[1] Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research* **13** 1063–1095.

[2] Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research* **9** 2015–2033.

[3] Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test* **25** 197–227.

[4] Breiman, L. (1996). Bagging predictors. *Machine learning* **24** 123–140.

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

[6] Breiman, L. (2004). Consistency for a simple model of random forests.

[7] Breiman, L. and Cutler, A. (2003). Manual–setting up, using, and understanding random forests V4.0.2003. *URL https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.*

[8] Breiman, L., Friedman, J. H., Stone, C. and Olshen, R. A. (1984). *Classification and regression trees*. Chapman and Hall/CRC.

[9] Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology* **88** 2783–2792.

[10] DÍAZ-URIARTE, R. and DE ANDRES, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics* **7** 3.

[11] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2009). *The elements of statistical learning*, 2 ed. Springer series in statistics New York.

[12] GENUER, R. (2012). Variance reduction in purely random forests. *Journal of Nonparametric Statistics* **24** 543–562.

[13] GINI, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*.

[14] HAPFELMEIER, A., HOTHORN, T. and ULM, K. (2012). Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. *Computational Statistics & Data Analysis* **56** 1552–1565.

[15] HOTHORN, T., HORNIK, K. and ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics* **15** 651–674.

[16] HOTHORN, T. and ZEILEIS, A. (2015). partykit: A Modular Toolkit for Recursive Partytioning in R. *Journal of Machine Learning Research* **16** 3905-3909.

[17] ISHIOKA, T. (2013). Imputation of missing values for unsupervised data using the proximity in random forests. In *International Conference on Mobile, Hybrid, and On-line Learning. Nice* 30–36.

[18] ISHWARAN, H. and KOGALUR, U. B. (2010). Consistency of random survival forests. *Statistics & probability letters* **80** 1056–1064.

[19] JOSSE, J., REITER, J. P. et al. (2018). Introduction to the special section on missing data. *Statistical Science* **33** 139–141.

[20] JOSSE, J., PROST, N., SCORNET, E. and VAROQUAUX, G. (2019). On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*.

[21] LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 1302–1338.

[22] LOUPPE, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

[23] MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research* **17** 841–881.

[24] PRASAD, A. M., IVERSON, L. R. and LIAW, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9** 181–199.

[25] QUINLAN, J. R. (1986). Induction of decision trees. *Machine learning* **1** 81–106.

[26] QUINLAN, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[27] RIPLEY, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.

[28] RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

[29] RUBIN, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association* **91** 473–489.

[30] RUBIN, D. B. (2004). *Multiple imputation for nonresponse in surveys* **81**. John Wiley & Sons.

[31] SCORNET, E., BIAU, G. and VERT, J.-P. (2015). Consistency of random forests. *The Annals of Statistics* **43** 1716–1741.

[32] SHANNON, C. E. (1948). A mathematical theory of communication. *Bell system technical journal* **27** 379–423.

[33] SHOTTON, J., FITZGIBBON, A., COOK, M., SHARP, T., FINOCCHIO, M., MOORE, R., KIPMAN, A. and BLAKE, A. (2011). Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* 1297–1304. Ieee.

[34] STEKHOVEN, D. J. and BÜHLMANN, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28** 112–118.

[35] SVETNIK, V., LIAW, A., TONG, C., CULBERSON, J. C., SHERIDAN, R. P. and FEUSTON, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences* **43** 1947–1958.

[36] TWALA, B., JONES, M. and HAND, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters* **29** 950–956.

[37] VARIAN, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives* **28** 3–28.

[38] VENABLES and RIPLEY (2002). Modern Applied Statistics with S. *Springer, New York* **1228** 1229.

[39] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113** 1228–1242.

[40] ZHU, R., ZENG, D. and KOSOROK, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association* **110** 1770–1784.