Chapter 55
Prediction Regions for Functional-valued Random Forests

Diego Serrano and Eduardo García-Portugués

Abstract We propose prediction regions for Random Forests (RFs) with functional output. Our approach is based on a metric specification and builds on the notion of Fréchet regression. It leverages the Out-Of-Bag (OOB) observations naturally generated during the training of RFs to estimate the uncertainty in the prediction, using the complete dataset. We outline the assumptions underpinning the construction of the prediction regions through OOB errors. A numerical experiment with quantile curves on the response and scalar predictor illustrates the prediction regions and shows that four types of nominal coverages are honored.

55.1 Introduction

Random Forests (RFs) [1] are a popular nonparametric regression method renowned for its adaptability and strong predictive power. Although the original formulation of RFs was designed exclusively for Euclidean data, several authors have generalized this algorithm for more complex data types. With this goal, Fréchet Random Forests (FRF) were developed in [5, 4] to extend RFs for metric data by incorporating the Fréchet mean in the splitting criterion and the aggregation of trees. This method can be applied to obtain predictions when the response and the predictors are functional variables, whenever a suitable distance is defined in the corresponding space of functions $\mathcal{F}$. A common example are Lebesgue spaces over an interval $T \subset \mathbb{R}$, defined as $L^q(T) := \{f : T \to \mathbb{R} : \int_T |f(t)|^q \, \mathrm{d}t < \infty\}$, with the L^q distance $d_q(f,g)^q := \int_T |f(t) - g(t)|^q \, \mathrm{d}t$. Building on the work in [5], a nonparametric

Diego Serrano (✉)
Department of Statistics, Universidad Carlos III de Madrid, Spain,
e-mail: dieserra@est-econ.uc3m.es

Eduardo García-Portugués
Department of Statistics, Universidad Carlos III de Madrid, Spain,
e-mail: edgarcia@est-econ.uc3m.es

locally adaptive kernel generated by random forests was developed in [8], along with theoretical guarantees for the consistency of the estimator.

Once a RF prediction is obtained, it is important to quantify its uncertainty to weight prediction statements. However, despite their predictive power, RFs have historically lacked strong statistical inference tools. This issue was addressed in [9] by leveraging the Out-Of-Bag (OOB) observations to develop confidence intervals for Euclidean data. The OOB observations arise naturally in the training of a RF and are a trustworthy estimate of its predictive performance, bypassing the need of sample splitting.

The goal of this work is to generalize the ideas from [9] to FDA in order to develop prediction regions for a RF with functional response. Our prediction regions are balls centered on the RF predicted function, with a radius given by the $1-\alpha$ quantile of the empirical distribution of the OOB errors. Both the prediction of the functional variable and the estimation of the uncertainty are obtained using the complete dataset, at the cost of fitting a single forest. We show through a numerical experiment in the Wasserstein space that these prediction regions respect the nominal coverage rates for four different types of coverage.

The rest of the work is organized as follows: Section 55.2 introduces basic concepts regarding RFs and their extension for FDA; Section 55.3 defines OOB prediction regions; and Section 55.4 gives a numerical experiment in the Wasserstein space.

55.2 Random forests and FDA

RFs are an ensemble method based on combining the predictions from multiple decision trees via bootstrap aggregation. We first introduce decision trees.

55.2.1 Fréchet decision trees with functional response

Let Y denote a response functional random variable defined in a functional space $(\mathcal{F}, d_{\mathcal{F}})$ where $d_{\mathcal{F}}$ denotes a distance function in $\mathcal{F}$. Let $\boldsymbol{X}$ be a p-dimensional random vector in $[0,1]^p$ acting as the predictor (see the end of this subsection for a note on the extension to metric-valued predictor). Consider a sample $\mathcal{L}_n := \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n \subset [0,1]^p \times \mathcal{F}$ with multivariate predictors and functional response. The application of RFs to functional data is achieved by adapting the prediction of results after the tree is grown. This requires the generalization of the mean to spaces where no inner product structure is defined (for example, L^q spaces with $q \neq 2$). Fréchet [6] defined the Fréchet mean $y_\oplus$ as the following extension of the mean to metric spaces:

$$y_\oplus := \arg\min_{y \in \mathcal{F}} \mathsf{E}\big(d_{\mathcal{F}}(Y, y)^2\big).$$

The Fréchet mean is based on the mean-squared error minimization property of the mean in Euclidean spaces. The Fréchet mean may not necessarily exist nor be

unique. Similarly, the classical regression model for Euclidean data is generalized in [7] to Fréchet regression, through the conditional Fréchet mean

$$m(\boldsymbol{x}) := \arg\min_{y \in \mathcal{F}} \mathsf{E}\big(d_{\mathcal{F}}(Y, y)^2 \mid \boldsymbol{X} = \boldsymbol{x}\big). \tag{55.1}$$

For Euclidean responses, $m(\boldsymbol{x}) = \mathsf{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$.

For multivariate predictors, given a node $A \subset \mathcal{L}_n$, a split variable index $j \in \{1, \ldots, p\}$, and a threshold value c_j, the left and right child nodes are defined, respectively, as

$$A_{j,\ell} := \big\{(\boldsymbol{X}_i, Y_i) \in A : X_{ij} \leq c_j\big\} \quad \text{and} \quad A_{j,r} := \big\{(\boldsymbol{X}_i, Y_i) \in A : \boldsymbol{X}_{ij} > c_j\big\}.$$

The standard way of measuring the quality of a split is the Classification And Regression Trees (CART) criterion [2]. For each split variable X_j, the best threshold c_j is selected over the observed values of X_{ij}, $i = 1, \ldots, n$, as the maximizer of the variance decrease in the response due to the split. The optimal split for node A is the value that maximizes the variance decrease among the p selected thresholds. Once the tree is built, the prediction of a new observation $\boldsymbol{x}$ is calculated as the Fréchet mean of the responses from the training set contained in the terminal node of $\boldsymbol{x}$.

The regression setting can be extended to cases in which the predictors lie on a metric space (in particular, functional spaces). For that, the RF splitting criterion can be generalized using the 2-means algorithm as a split function, and assigning each datum to the closest of the two obtained centroids. The CART criterion is adapted for metric predictors by replacing the sample variance with an empirical version of the Fréchet variance, which is defined as the expected squared distance from the Fréchet mean (see [4]). For easiness in the exposition, we keep restricting to the case where the predictor is multivariate and bounded (as in [8]).

55.2.2 Random forests with functional response

The aggregation of predictions of the training responses through the Fréchet mean used in Fréchet trees can be used to build FRFs through the aggregation of multiple tree predictions [5]. Every forest is comprised of a fixed number $B \in \mathbb{N}$ of trees, each trained with a random resample $\mathcal{L}_n^{*b}$ of $\mathcal{L}_n$, $b = 1, \ldots, B$. The absence of a theoretical analysis of the consistency of FRFs in [5, 4] motivated the development of a locally adaptive kernel generated by random forests with theoretical guarantees in [8]. This estimator of (55.1) is

$$m_n(\boldsymbol{x}) := \arg\min_{y \in \mathcal{F}} \sum_{i=1}^{n} \omega_i(\boldsymbol{x}) d_{\mathcal{F}}(Y_i, y)^2, \tag{55.2}$$

where the weights $\omega_i(\boldsymbol{x})$ are generated during the Fréchet tree-building procedure. Specifically, for a given value of the predictors $\boldsymbol{x} \in [0, 1]^p$, the prediction of the bth tree provides a sequence of weights $\omega_i^b(\boldsymbol{x}, \boldsymbol{\theta}_b) := |\tau_{\boldsymbol{x}}^b|^{-1} 1_{\{(\boldsymbol{x}_i, y_i) \in \tau_{\boldsymbol{x}}^b\}}$, where $|\tau_{\boldsymbol{x}}^b|$

denotes the number of elements in the corresponding terminal node, $b = 1, \ldots, B$. Here, $\boldsymbol{\theta}_b$ represents the randomization parameter vector that governs the growth of the bth tree, determining which variables are considered for each split. The weights computed with the individual trees are then averaged to obtain a single weight $\omega_i(\boldsymbol{x})$ for the ith sample observation, $\omega_i(\boldsymbol{x}) := B^{-1} \sum_{b=1}^{B} \omega_i^b(\boldsymbol{x}, \boldsymbol{\theta}_b)$.

55.2.3 Out-of-bag errors

For a given tree built with a resample $\mathcal{L}_n^{*b}$, we say that $(\boldsymbol{X}_i, Y_i)$ is OOB if $(\boldsymbol{X}_i, Y_i) \in \mathcal{L}_n \setminus \mathcal{L}_n^{*b}$. We denote by $\widehat{Y}_{(i)}$ the OOB prediction of Y_i, which is based only on trees for which $(\boldsymbol{X}_i, Y_i)$ is OOB. Our goal is to estimate the uncertainty in the prediction $\widehat{Y}$ of a functional response Y using the OOB errors

$$\widehat{R}_i^{\text{oob}} := d_{\mathcal{F}}(Y_i, \widehat{Y}_{(i)}), \quad i = 1, \ldots, n, \tag{55.3}$$

as an estimate of the actual prediction errors $d_{\mathcal{F}}(Y_i, m(\boldsymbol{X}_i))$, $i = 1, \ldots, n$. Notice that $\widehat{Y}_{(i)}$ is quite different from the leave-one-out cross-validation prediction, denoted by $\widehat{Y}_{-i}$, which is formed by the prediction on $\boldsymbol{X}_i$ of the forest trained with $\mathcal{L}_n \setminus \{(\boldsymbol{X}_i, Y_i)\}$.

Once a RF is trained using [8], we outline the procedure to obtain a specific OOB prediction $\widehat{Y}_{(i_0)}$. By construction, certain observations are left out of each random resample $\mathcal{L}_n^{*b}$, and hence are OOB for the bth tree. Let $\mathcal{B}_{i_0}$ denote the indices of the trees for which Y_{i_0} is OOB. If $b \in \mathcal{B}_{i_0}$, calculate $\omega_i^b(\boldsymbol{x}_{i_0}, \boldsymbol{\theta}_b)$ as explained in Section 55.2.2. For each $b \notin \mathcal{B}_{i_0}$, set $\omega_i^b(\boldsymbol{x}_{i_0}, \boldsymbol{\theta}_b) = 0$ for every $i = 1, \ldots, n$. Finally, aggregate the weights as $\omega_i(\boldsymbol{x}_{i_0}) = |\mathcal{B}_{i_0}|^{-1} \sum_{b=1}^{B} \omega_i^b(\boldsymbol{x}_{i_0}, \boldsymbol{\theta}_b)$, and plug them in (55.2) to obtain $\widehat{Y}_{(i_0)}$. The procedure described shows that OOB predictions estimate the predictive power of the RF at no additional training cost, since it is only required to select among the already grown trees.

55.3 Out-of-bag prediction regions for FDA

We develop confidence regions $\mathcal{P}_{1-\alpha}$ that contain the true functional variable Y within a given probability $1 - \alpha$, as a generalization for functional response and multivariate predictors of the confidence intervals in [9]. Since the OOB errors (55.3) provide a reliable estimate of the RF errors, it is natural to measure the uncertainty in a RF prediction through the quantile of the empirical distribution function of the OOB errors. This idea can be applied to FDA in the following way.

Definition 55.1 (Prediction region). The prediction region for functional predictors $\boldsymbol{x} \in [0, 1]^p$, with significance level $\alpha \in (0, 1)$ is defined as

$$\mathcal{P}_{1-\alpha}(\boldsymbol{x}, \mathcal{L}_n) = \left\{ y \in \mathcal{F} : d_{\mathcal{F}}(\widehat{m}(\boldsymbol{x}), y) < \widehat{R}_{[1-\alpha, n]} \right\}, \tag{55.4}$$

where $\widehat{m}(\boldsymbol{x})$ is a RF estimation of the conditional Fréchet mean (55.1), and $\widehat{R}_{[1-\alpha,n]}$ denotes the $(1-\alpha)$-quantile of the empirical distribution based on $\widehat{R}_1^{\text{oob}}, \ldots, \widehat{R}_n^{\text{oob}}$.

We consider four probability coverage types, each suited to different statistical contexts. For a significance level $\alpha \in (0,1)$:

- Type I: $\mathsf{P}\{Y \in \mathcal{P}_{1-\alpha}(\boldsymbol{X}, \mathcal{L}_n)\}$.
- Type II: $\mathsf{P}\{Y \in \mathcal{P}_{1-\alpha}(\boldsymbol{X}, \mathcal{L}_n) \mid \mathcal{L}_n\}$.
- Type III: $\mathsf{P}\{Y \in \mathcal{P}_{1-\alpha}(\boldsymbol{X}, \mathcal{L}_n) \mid \boldsymbol{X} = \boldsymbol{x}\}$.
- Type IV: $\mathsf{P}\{Y \in \mathcal{P}_{1-\alpha}(\boldsymbol{X}, \mathcal{L}_n) \mid \mathcal{L}_n,\ \boldsymbol{X} = \boldsymbol{x}\}$.

In the Euclidean case, guarantees of asymptotic convergence of these probabilities to $1-\alpha$ as n diverges to infinity are provided in [9].

The following conditions are assumed for the application of the prediction regions:

(*c.1*) $(\boldsymbol{X}, Y), (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \overset{\text{iid}}{\sim} \mathbb{G}$.
(*c.2*) The data generation process $\mathbb{G}$ is such that:

(*c.2.1*) There exists a unique conditional Fréchet mean (55.1) for every $\boldsymbol{x} \in [0,1]^p$ $\mathbb{G}$-a.s.
(*c.2.2*) The radial error $d_{\mathcal{F}}(Y, m(\boldsymbol{X}))$ is independent from the predictor $\boldsymbol{X}$.

These assumptions generalize some of the fundamental conditions defined in [9]. Assumption (*c.1*) ensures the identical distribution of the OOB radial errors (55.3). Existence and uniqueness of the conditional Fréchet mean from assumption (*c.2.1*) is required for the regression model to be well defined, and is a common assumption that ensures the consistency of the RF estimators, see [8] or [3]. We conjecture that, similarly to [9], the prediction regions (55.4) have correct asymptotic coverage under conditions (*c.1*)–(*c.2*), also under metric-valued predictor. We leave this proof for a future work, and assess the empirical coverage of the prediction regions through simulations in the following section.

55.4 A numerical experiment

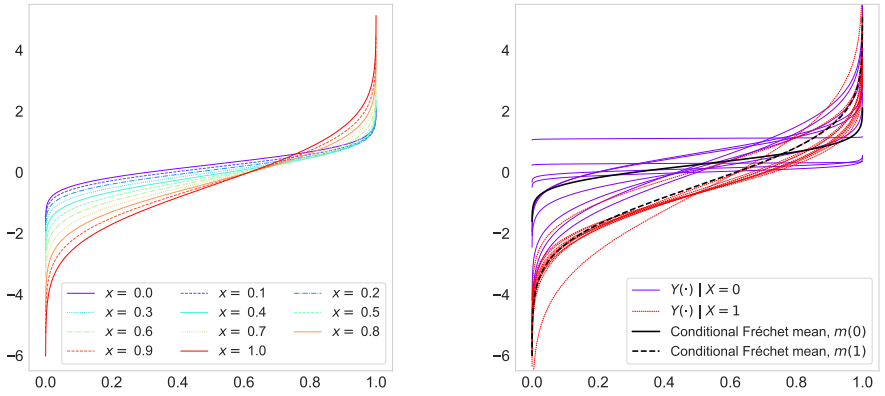
We consider a simple simulation scenario to illustrate the prediction regions and test their empirical coverage for Types I–IV. The functional space $\mathcal{F}$ considered is the 2-Wasserstein space $\mathcal{W}_2(\mathbb{R})$ of probability distributions on the real line with finite second moment and endowed with the 2-Wasserstein metric $d_{\mathcal{W}_2}$. This space is isomorphic to the subset of $L^2[0,1]$ of quantile functions, so the distance $d_{\mathcal{W}_2}$ between two probability distributions P and Q of $\mathcal{W}_2(\mathbb{R})$ can be expressed through the respective quantile functions F_{P}^{-1} and F_{Q}^{-1}: $d_{\mathcal{W}_2}(\mathrm{P},\mathrm{Q})^2 = \int_0^1 \left|F_{\mathrm{P}}^{-1}(u) - F_{\mathrm{Q}}^{-1}(u)\right|^2 \mathrm{d}u$. This isomorphism motivates the definition of regression models in $\mathcal{W}_2$ in terms of the corresponding quantile functions. Following [7, Section 6.1], we consider the regression function

$$m(x)(\cdot) = \mathsf{E}(Y(\cdot) \mid X = x) = \frac{1}{4} - \log(1+x) + \left(\frac{1}{2} + x^2\right)\Phi^{-1}(\cdot),$$

for $x \in [0,1]$, where Φ^{-1} denotes the quantile function of a $\mathcal{N}(0,1)$ distribution. To generate the response, consider

$$Y(\cdot) = C - \log(1+X) + (S + X^2)\Phi^{-1}(\cdot).$$

with $C \sim \Gamma(1/2, 1/2)$ and $S \sim \text{Exp}(2)$ independent of X, so that the model satisfies (*c.2.2*) and the conditional Fréchet mean of $Y(\cdot) \mid X = x$ is $m(x)(\cdot)$ (*c.2.1*). We consider that the predictor X follows a $U(0,1)$ distribution. The data generation process of $(X, Y(\cdot))$, which we denote by $\mathbb{G}$, is illustrated in Figure 55.1.



(a) Conditional Fréchet mean $m(x)$

(b) Realizations of $Y(\cdot) \mid X$

Fig. 55.1: Description of the data generation process $\mathbb{G}$. In Panel (a), the regression function (conditional Fréchet mean) is illustrated for different fixed values of the predictor. Panel (b) contains ten realizations of the functional response $Y(\cdot) \mid X = 0$ and ten of $Y(\cdot) \mid X = 1$. In each case, the realizations of $Y(\cdot) \mid X = x$ are centered around the regression function $m(x)$.

We generated $N = 500$ samples of size $n = 400$ according to the aforementioned data generation process, and used the estimator (55.2) in [8] for constructing the RF. Every forest is formed by $B = 200$ trees, and each tree is trained with a subsample of size n generated with nonparametric bootstrap with replacement, using the implementation from [3] in the `pyfréchet` package for Python. Three different values are considered for the significance level α: 0.01, 0.05, and 0.10.

For Types I–IV, each probability is estimated using $M = 500$ Monte Carlo samples. In the case of Type I, the probability is estimated as follows:

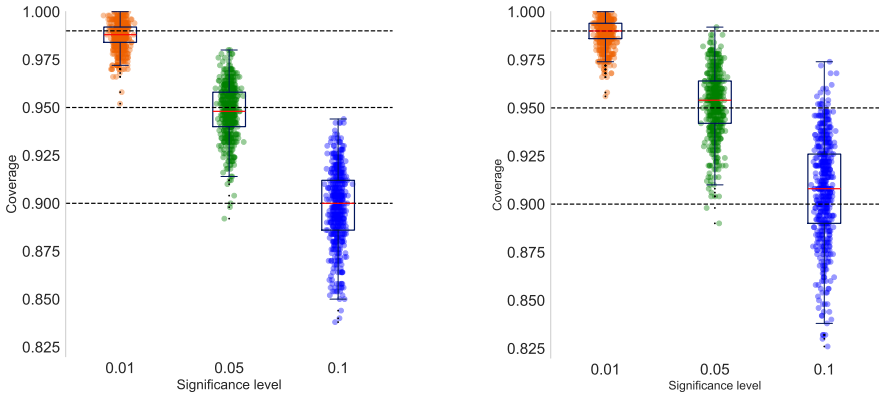
$$\mathsf{P}\left\{Y \in \mathcal{P}_{1-\alpha}(X, \mathcal{L}_n)\right\} \approx \frac{1}{M}\sum_{j=1}^{M} 1_{\left\{Y_j \in \mathcal{P}_{1-\alpha}\left(X_j, \mathcal{L}_n^{(j)}\right)\right\}}$$

for $\left((X_j, Y_j), \mathcal{L}_n^{(j)}\right)$ such that, for every $j = 1, \ldots, M$, (X_j, Y_j) is independent of $\mathcal{L}_n^{(j)}$ and $(X_j, Y_j) \sim \mathbb{G}$, $\mathcal{L}_n^{(j)} \sim \mathbb{G}^n$. For Type IV, the probabilities are estimated as

$$\mathsf{P}\left\{Y \in \mathcal{P}_{1-\alpha}(X, \mathcal{L}_n) \mid X = x, \mathcal{L}_n\right\} \approx \frac{1}{M} \sum_{j=1}^{M} 1_{\left\{Y_j \in \mathcal{P}_{1-\alpha}(x, \mathcal{L}_n)\right\}}, \tag{55.5}$$

for $Y_j \mid X = x$ induced by $\mathbb{G}$, and $x \in \{0, 0.5, 1\}$. The estimation of the probabilities in Types II and III is analogous.

For Type I, coverages of 0.992, 0.946, and 0.91 were obtained, respectively, for $\alpha = 0.01, 0.05, 0.10$. For Type III with $X = 0$ the simulations showed coverages of 0.998, 0.956, and 0.928. With $X = 0.5$, the coverages were 0.992, 0.958, and 0.916. Finally, with $X = 1$, the coverages were 0.988, 0.916, and 0.864. The reported coverages of Types II and IV are shown in Figure 55.2. The results show that our prediction regions achieve overall correct coverage rates across the three significance levels considered.



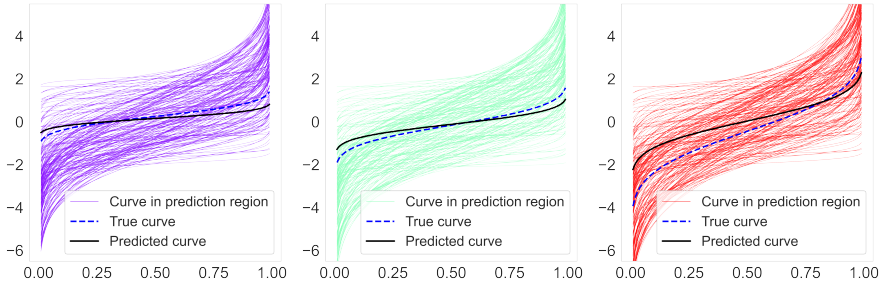
(a) Type II

(b) Type IV ($X = 0.5$)

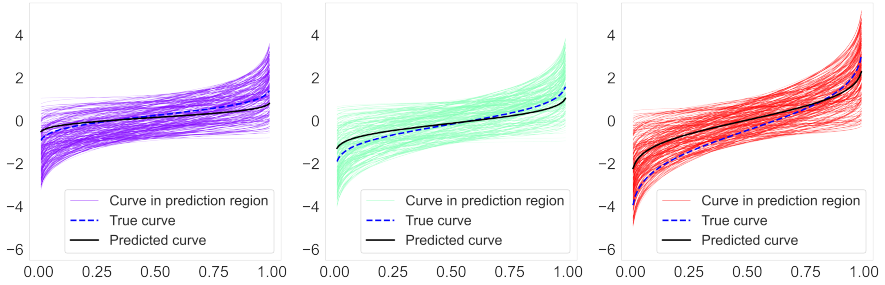
Fig. 55.2: Reported coverage (Types II and IV) across the $N = 500$ simulated datasets. For Type IV, the value of the predictor $X = 0.5$ was considered. Each dot in the boxplots is associated to a dataset $\mathcal{L}_n^{(k)}$, $k = 1, \ldots, N$, and its vertical position measures the estimated coverage probability (as in (55.5)) conditioned to that dataset.

Figure 55.3 shows the shape of the prediction regions for different values of the predictor X and the significance level α. The prediction regions adapt to the shape of the predicted curve as X varies between 0 and 1.

Acknowledgements The first author acknowledges support from the INFLUENTIA-CM-UC3M project. The second author acknowledges support from "Convocatoria de la Universidad Carlos III de Madrid de Ayudas para la recualificación del sistema universitario español para 2021–2023", funded by Spain's Ministerio de Ciencia, Innovación y Universidades.



(a) $X = 0$, $\alpha = 0.01$ (b) $X = 0.5$, $\alpha = 0.01$ (c) $X = 1$, $\alpha = 0.01$



(d) $X = 0$, $\alpha = 0.10$ (e) $X = 0.5$, $\alpha = 0.10$ (f) $X = 1$, $\alpha = 0.10$

Fig. 55.3: Examples of prediction regions for different values of X and significance levels α. To represent graphically the shape of each prediction region, 100 Monte Carlo samples of $Y \mid X = x$ inside the prediction region are shown.

References

1. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
2. Breiman, L., Friedman, J., Stone, C., Olshen, R.: Classification and Regression Trees. Taylor & Francis (1984)
3. Bulté, M., Sørensen, H.: Medoid splits for efficient random forests in metric spaces. Computational Statistics & Data Analysis (2024)
4. Capitaine, L., Bigot, J., Thiébaut, R., Genuer, R.: Fréchet random forests for metric space valued regression with non Euclidean predictors. Journal of Machine Learning Research **25**(355), 1–41 (2024)
5. Capitaine, L., Genuer, R., Thiébaut, R.: Fréchet random forests (2019)
6. Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. Annales de l'Institut Henri Poincaré **10**(4), 215–310 (1948)
7. Petersen, A., Müller, H.G.: Fréchet regression for random objects with Euclidean predictors. The Annals of Statistics **47**(2), 691–719 (2019)
8. Qiu, R., Yu, Z., Zhu, R.: Random forest weighted local Fréchet regression with random objects. Journal of Machine Learning Research **25**(107), 1–69 (2024)
9. Zhang, H., Zimmerman, J., Nettleton, D., Nordman, D.J.: Random forest prediction intervals. The American Statistician **74**(4), 392–406 (2020)