Generalized Pareto Regression Trees for extreme events analysis

Sébastien Farkas, Antoine Heranval, Olivier Lopez, Maud Thomas

Generalized Pareto Regression Trees for extreme events analysis

Sébastien Farkas[1], Antoine Heranval[2,3], Olivier Lopez[3], and Maud Thomas[1]

[1] Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France,
[2] Mission Risques Naturels, 1 rue Jules Lefebvre 75009 Paris, France
[3] CREST Laboratory, CNRS, Groupe des Écoles Nationales d'Économie et Statistique, Ecole Polytechnique, Institut Polytechnique de Paris, 5 avenue Henry Le Chatelier 91120 Palaiseau, France E-mails : sebastien.farkas@sorbonne-universite.fr, antoine.heranval@ensae.fr, maud.thomas@sorbonne-universite.fr, olivier.lopez@ensae.fr

Abstract

This paper derives finite sample results to assess the consistency of Generalized Pareto regression trees introduced by Farkas et al. [2021] as tools to perform extreme value regression for heavy-tailed distributions. This procedure allows the constitution of classes of observations with similar tail behaviors depending on the value of the covariates, based on a recursive partition of the sample and simple model selection rules. The results we provide are obtained from concentration inequalities, and are valid for a finite sample size. A misspecification bias that arises from the use of a "Peaks over Threshold" approach is also taken into account. Moreover, the derived properties legitimate the pruning strategies, that is the model selection rules, used to select a proper tree that achieves a compromise between simplicity and goodness-of-fit. The methodology is illustrated through a simulation study, and a real data application in insurance for natural disasters.

Key words: Extreme value theory; Regression trees; Concentration Inequalities; Generalized Pareto Distribution.

1 Introduction

Extreme value theory (EVT) is the branch of statistics which has been developed and broadly used to handle extreme events, such as extreme floods, heat wave episodes or extreme financial losses [Katz et al., 2002, Embrechts et al., 2013]. One of the key results behind the success of this approach was proved by Balkema and de Haan [1974], who established the ability of the Generalized Pareto (GP) family to approximate the tail of a distribution. This property allows the statistician to find information from the largest observations of a random sample to extrapolate the tail. This yields the so-called Peaks over Threshold (PoT) method introduced by Smith [1984] which consists in fitting a GP distribution to the excesses above some (high) suitably chosen threshold. In a regression framework, the parameters of this GP distribution depend on covariates reflecting the fact that different values of these covariates may result in a different tail behavior of the response variable [see e.g. Davison and Smith, 1990, Smith, 1989]. In this paper, we study the use of regression trees to perform GP regression on the excesses for heavy-tailed distributions. This ensemble method, introduced by Breiman et al. [1984], determines clusters of similar tail behaviors depending on the value of the covariates, based on a recursive partition of the sample and simple model selection rules. In the present work, we provide theoretical results and empirical evidence on the consistency of such a procedure and of these selection rules. The result we provide are based on concentration inequalities, in order to hold for finite sample sizes. The main difficulty stands in the misspecification of the model and on handling the fact that the distributions are heavy tailed.

Tail regression is a challenging task. Several papers have been interested in extreme quantile regression, Chernozhukov [2005] and, Wang et al. [2012] derive extreme quantile estimators assuming a linear

form for the conditional quantile. Gardes and Stupfler [2019] and Velthoen et al. [2019] use conditional intermediate-level quantiles to extrapolate above the threshold and deduce estimators for extreme conditional quantiles. Another approach is to model the parameters of the GP distribution as functions of the covariates e.g. as local polynomials [Beirlant and Goegebeur, 2004] or as generalized additive models [Chavez-Demoulin et al., 2015, Youngman, 2019]. More and more approaches in extreme value regression use machine learning methods. Carreau and Vrac [2011] present a new class of stochastic downscaling models, the conditional mixture models (CMM) which builds on a neural network. CMM are mixture models whose parameters are functions of predictor variables. Rietsch et al. [2013] address the issue of the optimization of the spatial design of a network of existing weather stations by combining EVT with neural networks. Very recently, Velthoen et al. [2021] proposed a gradient boosting procedure to estimate conditional GP distribution. Several works [Richards and Huser, 2022, Pasche and Engelke, 2022, Allouche et al., 2022] have proposed methodologies based on neural networks for extreme quantile regression. Finally, Gnecco et al. [2022] have developed a method for extreme quantile regression using random forests. Their extremal random forest estimates the parameter of a GP distribution conditionally on the predictor vector using local likelihood maximization. Finally, two works consider piece-wise stationary marginal and dependence model to estimate the meteorological and oceanographic variables [Ross et al., 2018, Barlow et al., 2023].

Regression trees, introduced by Breiman et al. [1984] along with the CART algorithm (for Classification And Regression Trees), are flexible tools to perform a regression and clustering task simultaneously, with the ability to deal with discrete and smooth covariates simultaneously. They have been used in various fields, including industry [González et al., 2015], geology [see e.g. Rodriguez-Galiano et al., 2015], ecology [see e.g. De'ath and Fabricius, 2000], claim reserving in insurance [Lopez et al., 2016]. Through the iterative splitting algorithm used in CART, nonlinearities are introduced in the way the distribution is modeled, while furnishing an intelligible interpretation of the final classification of response variables. The splitting criterion—used to iteratively separate observations into clusters with similar behaviors—depends on the type of problems one is considering. While the standard CART algorithm relies on mean-squared criterion to perform mean-regression, alternative loss functions have been considered as in [Chaudhuri and Loh, 2002] for quantile regression, or in [Su et al., 2004] who used a log-likelihood based loss. Loh [2011, 2014] provide detailed descriptions of regression trees procedures and a review of their variants. In this paper, building on the result of Balkema and de Haan [1974] , we use a GP log-likelihood loss, as in [Farkas et al., 2021], to perform extreme value regression.

The rest of the paper is organized as follows. In Section 2, we introduce notations and describe the GP regression tree algorithm. Section 3 lists the main results of this paper, that is deviation bounds for the regression tree estimator for finite sample size, and consistency of the "pruning" (that is model selection) strategy. Empirical results are gathered in Section 4, which provides a simulation study, and a real data analysis in natural disaster insurance. Detailed proofs of the technical results are shown in the Appendix.

2 Regression trees for extreme value analysis

This section describes the estimation method (GP regression trees) that we consider in this paper, and which has already been introduced by Farkas et al. [2021]. Some classical results in EVT are given in Section 2.1 to motivate the GP approximation. Regression trees adapted to this context are described in Section 2.3.

2.1 Extreme value theory and regression

Let us consider independent and identically distributed observations $Y_1, Y_2, \ldots$ with an unknown survival function $\overline{F}$ (that is $\overline{F}(y) = P(Y_1 > y)$). A natural way to define extreme events is to consider the values of Y_i which have exceeded some high threshold u. The excesses above u are then defined as the variables $Y_i - u$ given that $Y_i > u$. The asymptotic behavior of extreme events is characterized by the distribution of the excesses which is given by

$$\overline{F}_u(z) = P[Y_1 - u > z \mid Y_1 > u] = \frac{\overline{F}(u+z)}{\overline{F}(u)},\ z > 0\,.$$

Pickands [1975] showed that, if $\overline{F}$ satisfies the following property

$$\lim_{t\to\infty} \frac{\overline{F}(ty)}{\overline{F}(y)} = y^{-1/\gamma_0}, \forall y > 0, \tag{1}$$

with $\gamma_0 > 0$, then

$$\lim_{u\to\infty} \sup_{z>0} |\overline{F}_u(z) - \overline{H}(z; \sigma_0, \gamma_0)| = 0 \tag{2}$$

for some $\sigma_0 > 0$ and $\overline{H}(\cdot; \sigma_0, \gamma_0)$ necessarily belongs to the Generalized Pareto (GP) distributions family which distribution function is of the form

$$\overline{H}(z; \sigma_0, \gamma_0) = \left(1 + \gamma_0 \frac{z}{\sigma_0}\right)^{-1/\gamma_0}, \ z > 0,$$

where $\sigma_0 > 0$ is a scale parameter and $\gamma_0 > 0$ is a shape parameter, which reflects the heaviness of the tail distribution. Especially, if $\gamma_0 \in (0, 1)$, the expectation of Y_1 is finite whereas if $\gamma_0 \geq 1$ the expectation of Y_1 is infinite. More details on these results can be found in e.g. [Coles, 2001, Beirlant et al., 2004].

Note that in full generality, the shape parameter $\gamma_0 \in \mathbb{R}$. However, the applications we have in mind, such as in Section 4, concern natural catastrophes which fall into the domain of heavy-tailed distributions, that is distributions for which $\gamma_0 > 0$. We therefore choose here to focus on the case $\gamma_0 > 0$. Besides, in this paper, we derive non-asymptotic results on the consistency of a procedure on the GP log-likelihood (see Section 3). The derivation of such results requires some smoothness on the GP log-likelihood, which is satisfied for $\gamma_0 > 0$, but not for all $\gamma_0 \in \mathbb{R}$.

The so-called Peaks over Threshold (PoT) method is widely used [see Davison and Smith, 1990, Coles, 2001]. It consists in choosing a high threshold u and fitting a GP distribution on the excesses above that threshold u. The estimation of the parameters σ_0 and γ_0 may be done by maximizing the GP likelihood. The choice of the threshold u can be understood as a compromise between bias and variance: the smaller the threshold, the less valid the asymptotic approximation, leading to bias; on the other hand, a too high threshold will generate few excesses to fit the model, leading to high variance. In practice, threshold selection is a challenging task. The existing methods for the choice of the threshold u relies on graphical diagnostics or on computational approaches based on supplementary conditions (that depend on unknown parameters) on the underlying distribution function F [see Scarrott and MacDonald, 2012] . However, it should be mention that some recent works model GP distribution upper tail (with $\gamma_0 > 0$) and the remaining of the full distribution in one step, which allows one to overcome the challenging issue of threshold selection [Tencaliec et al., 2020, Huang et al., 2019].

In the present paper, we consider a regression framework, that is, our goal is to estimate the impact of some random covariates $\mathbf{X}$ on the tail of the distribution of a response variable Y. The previous convergence result (2) holds, but for quantities σ_0, γ_0 and u that may depend on $\mathbf{X}$. More precisely, this means that, if we assume that $\gamma_0(\mathbf{x}) > 0$ for all $\mathbf{x}$ (which is the assumption that we will make throughout this paper), then (1) becomes

$$\lim_{s\to\infty} \frac{\overline{F}(sy \mid \mathbf{x})}{\overline{F}(y \mid \mathbf{x})} = y^{-1/\gamma_0(\mathbf{x})}, \forall y > 0, \tag{3}$$

where $\overline{F}(y \mid \mathbf{x}) = \mathbb{P}(Y \geq y \mid \mathbf{X} = \mathbf{x})$ [see Beirlant et al., 2004, and references therein], and (2) becomes

$$\lim_{u(\mathbf{x})\to\infty} \sup_{z>0} |\overline{F}_{u(\mathbf{x})}(z \mid \mathbf{x}) - \overline{H}(z; \sigma_{0u(\mathbf{x})}(\mathbf{x}), \gamma_0(\mathbf{x}))| = 0. \tag{4}$$

where $\overline{F}_{u(\mathbf{x})}(z \mid \mathbf{x}) = P[Y - u(\mathbf{x}) > z \mid Y > u(\mathbf{x}), \mathbf{X} = \mathbf{x}]$.

Therefore, in this regression framework, the PoT approach consists now in the estimation of the function $\boldsymbol{\theta}_0(\mathbf{x}) = (\sigma_0(\mathbf{x}), \gamma_0(\mathbf{x}))^t$ (where a^t denotes the transpose of a vector a).

2.2 Framework

We now suppose that we have observed $(Y_i, \mathbf{X}_i)_{1\leq i\leq n}$ a n-sample of $(Y, \mathbf{X})$, where $\mathbf{X} = \left(X^{(1)}, \ldots, X^{(d)}\right)$ belongs to a compact set $\mathcal{X} \subset \mathbb{R}^d$ and $Y \in \mathbb{R}$. In the approach described thereafter, each covariate can be either discrete or smooth, and it is not necessary that they are all of the same nature. Recall that the PoT approach consists in considering observations such that $Y_i \geq u(\mathbf{X}_i)$.

In this paper, we will restrain ourselves to the case where the function $u(\mathbf{x}) = u$. To allow an adaptive choice of this parameter, our results hold uniformly for $u \in [u_{\min}(n), u_{\max}(n)]$ (see Section 3), with $u_{\min}(n)$ and $u_{\max}(n)$ such that

1. $u_{\min}(n)$ is defined as the $1 - k_n/n$ quantile of F, that is

$$\mathbb{P}(Y \geq u_{\min}(n)) = \frac{k_n}{n},$$

   where k_n be an intermediate sequence, that is $k_n \to \infty$ and $k_n/n \to 0$, as $n \to \infty$,

2. $u_{\max}(n)$ is defined such that

$$\mathbb{P}(Y \geq u_{\max}(n)) = \frac{u_0 k_n}{n},$$

   for some constant $u_0 \leq 1$.

   Note that $u_{\min}(n)$ and $u_{\max}(n)$ are functions of n.

Here, k_n denote the average number (up to some constant) of observations on which the model is fitted. It is hence related to the rate of convergence of the procedure.

Remark 1. *Our results easily extend to the case where* $u(\mathbf{x}) = \sum_{j=1}^{m} u_j \mathbf{1}_{\mathbf{x} \in \mathcal{X}_j}$, *where* $(\mathcal{X}_j)_{1 \leq j \leq m}$ *are subsets of the space of covariates. Another possible extension would be to assume that* $u(\mathbf{x}) = f(\beta, \mathbf{x})$ *for some parameter* β *and* f *a known function. Nevertheless, a choice of such a particular threshold function seems hard to justify. Hence, we restrain ourselves to the simplest case.*

In the next section, we introduce a regression tree approach adapted to both smooth and discrete covariates, and relying on few assumptions (since the estimated regression function $\boldsymbol{\theta}_0$ does not need to be smooth).

2.3 GPD regression trees

Regression trees are a convenient tool to capture heterogeneous behaviors in the data [see Breiman et al., 1984]. These models aim at constituting classes of observations which have a relatively similar behavior in terms of the response variable Y. These classes are defined by "rules", which affect an observation to one of these classes according to the values of its covariates $\mathbf{X}$. These rules are obtained from the data through the CART (Classification And Regression Tree) algorithm, and the non-linearity of the procedure allows for an adaptation to the estimation of large classes of regression functions.

Fitting regression trees relies on a so-called "growing phase", described in our context in Section 2.3.1, which corresponds to the determination of these splitting rules, and explains how an estimator of the regression function $\boldsymbol{\theta}_0$ can be deduced from such a tree. The "pruning step", which can be understood as a model selection procedure, is described in Section 2.3.2.

2.3.1 Growing step: construction of the maximal tree

The ultimate goal of the CART algorithm is to optimize some objective function $\boldsymbol{\theta}^*(\mathbf{x})$ (also referred to as splitting criterion). This function $\boldsymbol{\theta}^*(\mathbf{x})$ can be seen as the minimizer of a certain risk function over a class of target functions, that is

$$\boldsymbol{\theta}^*(\mathbf{x}) = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}[\phi(Y, \boldsymbol{\theta}) \mid \mathbf{X} = \mathbf{x}],$$

where $\boldsymbol{\Theta} \subset \mathbb{R}^d$ represents the parameter space and ϕ a loss function whose choice depends on the quantity to be estimated. For instance, if ϕ is the quadratic (absolute) loss, then $\boldsymbol{\theta}^*$ corresponds to the conditional mean (median) of Y given $\mathbf{X}$.

The procedure of the CART algorithm consists in determining iteratively a set of "rules" $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)}) \to R_\ell(\mathbf{x})$ to split the data into two more homogeneous classes by finding at each step an appropriate simple rule (that is a condition on the value of some covariate). A set of rules $(R_\ell)_\ell$ is a set of maps such that, for all $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$, $R_\ell(\mathbf{x}) = 1$ or 0 depending on whether some conditions are satisfied by $\mathbf{x}$, with $R_{\ell 1}(\mathbf{x}) R_{\ell 2}(\mathbf{x}) = 0$ for $\ell 1 \neq \ell 2$ and $\sum_\ell R_\ell(\mathbf{x}) = 1$. In case of regression trees, these partitioning rules have a particular structure, since they can be written, for quantitative covariates (the case of $\mathbf{x}$ containing qualitative variables is described in Remark 2 below), as $R_\ell(\mathbf{x}) = \mathbf{1}_{\mathbf{x}_1 \leq \mathbf{x} < \mathbf{x}_2}$ for

some $\mathbf{x}_1 \in \mathcal{X}$ and $\mathbf{x}_2 \in \mathcal{X}$, with comparison symbols to be understood as component-wise comparisons. In other terms, if $d = 1$, rules can be identified as partitioning segments, if $d = 2$ they are rectangles (hyper-rectangles in the general case).

The determination of these rules from one step to another can be represented as a binary tree, since each rule R_ℓ at step k generates two rules $R_{\ell 1}$ and $R_{\ell 2}$ (with $R_{\ell 1}(\mathbf{x}) + R_{\ell 2}(\mathbf{x}) = 0$ if $R_\ell(\mathbf{x}) = 0$) at step $k+1$. The list of rules (R_ℓ) are identified with the leaves of the tree at step k, and the number of leaves of the tree is increasing from step k to step $k+1$. The algorithm stops when each leaf contains only one observation or when the observations in the same leaf have the same characteristics. The stopping rule can also be slightly modified to ensure that there is a minimal number of points of the original data in each leaf of the tree at each step.

From a given set of K rules $\mathcal{R} = (R_\ell)_{\ell=1,\dots,K}$, let $\mathcal{T}_\ell = \{\mathbf{x} : R_\ell(\mathbf{x}) = 1\}$, the ℓ–th leaf of the corresponding tree. The estimator $\widehat{\boldsymbol{\theta}}^K(\mathbf{x})$ associated with the set of leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ is obtained as

$$\widehat{\boldsymbol{\theta}}^K(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\boldsymbol{\theta}}^K(R_\ell) R_\ell(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\boldsymbol{\theta}}_\ell^K \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}.$$

The tree is obtained when the previous algorithm stops is referred to as the maximal tree and denoted $\widehat{T}_{\max}$ with the set of leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K_{\max}}$, where $K_{\max}$ denotes its number of leaves. It corresponds to a trivial estimator of the objective function $\boldsymbol{\theta}^*(\mathbf{x})$ since for each leaf, either the number of observations is equal to one, or all observations in this leaf share the same characteristics $\mathbf{x}$. The tree $\widehat{T}_K$ is thus identified by its leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ and the list of parameter values $\widehat{\boldsymbol{\theta}}_\ell^K$ associated with each leaf $\mathcal{T}_\ell$.

In our case, ϕ will be chosen as the negative GP log-likelihood, that is

$$\phi(z, \boldsymbol{\theta}) = \log(\sigma) + \left(\frac{1}{\gamma} + 1\right) \log\left(1 + \frac{\gamma z}{\sigma}\right), \quad z > 0$$

where $\boldsymbol{\theta} = (\sigma, \gamma)^t \in \boldsymbol{\Theta}$. Thus, this objective function $\boldsymbol{\theta}^*(\mathbf{x})$ is given by

$$\boldsymbol{\theta}^*(\mathbf{x}) \quad = \quad \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}[\phi(Y - u, \boldsymbol{\theta}) \mathbf{1}_{Y>u} \mid \mathbf{X} = \mathbf{x}],$$

and, the estimator $\widehat{\boldsymbol{\theta}}^K(\mathbf{x})$ associated with the set of leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ corresponds to

$$\widehat{\boldsymbol{\theta}}^K(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\boldsymbol{\theta}}_\ell^K \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} = \sum_{\ell=1}^{K} \begin{pmatrix} \widehat{\sigma}_\ell^K \\ \widehat{\gamma}_\ell^K \end{pmatrix} \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}.$$

Note that, in this case, the CART algorithm is applying only to the observations Y_i such that $Y_i > u$, and that all the quantities defined may depend on u. The algorithm can be described as follows:

Step 1: $R_1(\mathbf{X}_i) = 1$ for all $i = 1, \dots, n$ (corresponds to the root of the tree), and let $n_1 = 1$ the number of rules at Step 1.

Step k+1: Let n_k be the number of rules at Step k and let $(R_1, ...R_{n_k})$ denote the rules obtained at step k. For $\ell = 1, \dots, n_k$,

- if all observations i such that $R_\ell(\mathbf{X}_i) = 1$ have the same characteristics, then keep rule ℓ as it is no longer possible to split the data;
- else, rule R_ℓ is replaced by two new rules $R_{\ell 1}$ and $R_{\ell 2}$ determined in the following way: for each component $X^{(j)}$ of $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$, define the best threshold $x_{\ell\star}^{(j)}$ to split the data, such that

$$\begin{aligned} x_{\ell\star}^{(j)} \quad = \quad & \arg\min_{x^{(j)}} \left\{ \sum_{i=1}^{n} \phi(Y_i, \widehat{\boldsymbol{\theta}}_{j-}(x^{(j)}, R_\ell)) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} \le x^{(j)}} R_\ell(\mathbf{X}_i) \right. \\ & \left. + \sum_{i=1}^{n} \phi(Y_i, \widehat{\boldsymbol{\theta}}_{j+}(x^{(j)}, R_\ell)) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_\ell(\mathbf{X}_i) \right\}, \end{aligned}$$

where

$$\begin{cases} \widehat{\boldsymbol{\theta}}_{j-}(x^{(j)}, R_\ell) & = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{n} \phi(Y_i, \boldsymbol{\theta}) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} \le x^{(j)}} R_\ell(\mathbf{X}_i), \\ \widehat{\boldsymbol{\theta}}_{j+}(x^{(j)}, R_\ell) & = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{n} \phi(Y_i, \boldsymbol{\theta}) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_\ell(\mathbf{X}_i). \end{cases}$$

Then, select the best splitting component index :

$$j_\star = \arg\min_j \left\{ \sum_{i=1}^n \phi(Y_i, \widehat{\boldsymbol{\theta}}_{j-}(x_{\ell\star}^{(j)}, R_\ell)) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} \le x_{\ell\star}^{(j)}} R_\ell(\mathbf{X}_i) + \sum_{i=1}^n \phi(Y_i, \widehat{\boldsymbol{\theta}}_{j+}(x_{\ell\star}^{(j)}, R_\ell)) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} > x_{\ell\star}^{(j)}} R_\ell(\mathbf{X}_i) \right\}$$

Define the two new rules: $R_{\ell 1}(\mathbf{x}) = R_\ell(\mathbf{x}) \mathbf{1}_{x^{(j_\star)} \le x_{\ell\star}^{(j_\star)}}$, and $R_{\ell 2}(\mathbf{x}) = R_\ell(\mathbf{x}) \mathbf{1}_{x^{(j_\star)} > x_{\ell\star}^{(j_\star)}}$.

- Let $n_{k+1} = n_k + 2$ denote the new number of rules.

Stopping rule: Stop if $n_{k+1} = n_k$.

Remark 2. *In this version of the CART algorithm, all covariates are smooth or $\{0,1\}-$valued. For qualitative variables with more than two modalities, they must be transformed into binary variables, or the algorithm must be slightly modified so that the splitting step of each R_ℓ should be done by finding the best partition into two groups on the values of the modalities that minimizes the loss function. This can be done by ordering the modalities with respect to the average value—or the median value—of the response for observations associated with this modality.*

The procedure of the growing phase is summarized in Algorithm 1.

Algorithm 1 Growing phase

Input: Observations $(Y_i, \mathbf{X}_i)_{i=1,\dots,n}$ such that $Y_i > u$

$n_1 \leftarrow 1$, $R_1(\mathbf{X}_i) \leftarrow 1 \ \forall i = 1, \dots, n$ ▷ Root of the tree

for $\ell = 1, \dots, n_k$ **do**

  if All observations i such that $R_\ell(\mathbf{X}_i) = 1$ have the same characteristics **then**

    $R_\ell \leftarrow R_\ell$ ▷ Do not change R_ℓ

  else

    for $j = 1, \dots, d$ **do**

      for $x^{(j)} \in \mathbb{R}$ **do** ▷ via a grid search

        $\boldsymbol{\theta}_{j-}(x^{(j)}, R_\ell) \leftarrow \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} \le x^{(j)}} R_\ell(\mathbf{X}_i)$

        $\boldsymbol{\theta}_{j+}(x^{(j)}, R_\ell) \leftarrow \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_\ell(\mathbf{X}_i)$

        $x_{\ell\star}^{(j)} \leftarrow \arg\min_{x^{(j)}} \{ \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}_{j-}(x^{(j)}, R_\ell)) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} \le x^{(j)}} R_\ell(\mathbf{X}_i)$
        $+ \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}_{j+}(x^{(j)}, R_\ell)) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} > x^{(j)}} R_\ell(\mathbf{X}_i) \}$

      end for

      $j_\star \leftarrow \arg\min_j \{ \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}_{j-}(x_{\ell\star}^{(j)}, R_\ell)) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} \le x_{\ell\star}^{(j)}} R_\ell(\mathbf{X}_i)$
      $+ \sum_{i=1}^n \phi(Y_i, \boldsymbol{\theta}_{j+}(x_{\ell\star}^{(j)}, R_\ell)) \mathbf{1}_{Y_i>u} \mathbf{1}_{X_i^{(j)} > x_{\ell\star}^{(j)}} R_\ell(\mathbf{X}_i) \}$

    end for

    $R_{\ell 1}(\mathbf{x}) \leftarrow R_\ell(\mathbf{x}) \mathbf{1}_{x^{(j_\star)} \le x_{\ell\star}^{(j_\star)}}$

    $R_{\ell 2}(\mathbf{x}) \leftarrow R_\ell(\mathbf{x}) \mathbf{1}_{x^{(j_\star)} > x_{\ell\star}^{(j_\star)}}$

    $n_{k+1} \leftarrow n_k + 2$

  end if

end for

Output: $K_{\max}$, $(R_\ell)_{\ell=1,\dots,K_{\max}}$, $(\widehat{\boldsymbol{\theta}}_\ell)_{\ell=1,\dots,K_{\max}}$

2.3.2 Selection of a subtree: pruning step

The pruning step, presented in the next section, consists in extracting from the maximal tree $\widehat{T}_{\max}$ a subtree, that is a tree with the same root as $\widehat{T}_{\max}$ and all of its nodes in $\widehat{T}_{\max}$, that achieves a compromise between simplicity and goodness-of-fit.

For the pruning step, a standard way to proceed is to use a penalized criterion to select the appropriate subtree of $\widehat{T}_{\max}$ [see Breiman et al., 1984, Gey and Nedelec, 2005]. To determine this subtree, it is not necessary to compute all the subtrees of $\widehat{T}_{\max}$. It is sufficient to determine, among all the subtrees with K leaves for $K \leq K_{\max}$, the subtree $\widehat{T}_K$ that minimizes the following criterion

$$\frac{1}{k_n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \phi(Y_i - u, \widehat{\boldsymbol{\theta}}^K(\mathbf{X}_i)) \mathbf{1}_{Y_i > u} \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell} + \lambda K, \tag{5}$$

where $\lambda > 0$ denotes a penalisation constant, that can be chosen using cross-validation [see e.g. Allen, 1974, Stone, 1974]. Recall that k_n is the average number of observations such that $Y_i > u$, that is the number of observations on which the CART procedure is performed. Then, it only remains to determine the final tree among the obtained list of $K_{\max}$ admissible subtrees. The trees $\widehat{T}_K$, $K = 1, \ldots, K_{\max}$, are easy to determine, since $\widehat{T}_K$ is obtained by removing one leaf from the tree $\widehat{T}_{K+1}$ [see Breiman et al., 1984, p.284–290].

The number of leaves of the selected tree is thus obtained as the minimizer of the penalised criterion (5), that is

$$\widehat{K} = \min \left\{ \arg \min_{K=1,\ldots,K_{\max}} \left\{ \frac{1}{k_n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \phi(Y_i - u, \widehat{\boldsymbol{\theta}}^K(\mathbf{X}_i)) \mathbf{1}_{Y_i > u} \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell} + \lambda K \right\} \right\},$$

and the selected tree is denoted by $\widehat{T}_K = \widehat{T}_{\widehat{K}}$.

3 Main results

In this section, we show that the GP regression tree procedure defined in Section 2.3 is consistent. Notations and assumptions used throughout this section are listed in Section 3.1. We then state our first main results on the consistency of a fixed tree with K leaves, by separating the stochastic part of the error (Section 3.2) from the misspecification part (Section 3.3) caused by the GP approximation. The consistency of the pruning methodology is studied in Section 3.4.

3.1 Notations and assumptions

In order to derive our consistency results, we need the following assumptions.

Assumption 1. *1. $k_n = O(n^{a_1})$, with $a_1 > 0$*

2. The number of leaves $K_{\max}$ of the maximal tree $\widehat{T}_{\max}$ is such that $K_{\max} \leq \kappa k_n$ with $0 < \kappa \leq 1$

3. The parameter space $\boldsymbol{\Theta}$ is compact, that is

$$\boldsymbol{\Theta} = [\sigma_{\min}, \sigma_n] \times [\gamma_{\min}, \gamma_{\max}],$$

where $\gamma_{\min}, \gamma_{\max}, \sigma_{\min} > 0$ and $\sigma_n = O(n^{a_2})$ with $a_2 > 0$.

Consider a threshold $u \in [u_{\min}, u_{\max}]$ (defined in Section 2.2) and a tree $\widehat{T}_K$, recall that the trees and the estimators all depend on u. We denote $\widehat{\boldsymbol{\theta}}_\ell^K = \left(\widehat{\sigma}_\ell^K, \widehat{\gamma}_\ell^K\right)^t$ the estimated parameter in each leaf $\mathcal{T}_\ell$, that is, for $\ell = 1, \ldots, K$

$$\widehat{\boldsymbol{\theta}}_\ell^K = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \frac{1}{k_n} \sum_{i=1}^{n} \phi(Y_i - u, \boldsymbol{\theta}) \mathbf{1}_{Y_i > u} \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell} \right\}.$$

For each $\ell = 1, \ldots, K$, this estimator is expected to be close to $\boldsymbol{\theta}_\ell^{*K} = \left(\sigma_\ell^{*K}, \gamma_\ell^{*K}\right)^t$ defined by

$$\boldsymbol{\theta}_\ell^{*K} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}\left[\phi(Y - u, \boldsymbol{\theta}) \mathbf{1}_{Y > u} \mathbf{1}_{\mathbf{X} \in \mathcal{T}_\ell}\right]. \tag{6}$$

However, this quantity is not exactly our target: ideally, we wish to estimate, for $\ell = 1, \ldots, K$,

$$\boldsymbol{\theta}_{0,\ell}^K = (\sigma_{0,\ell}^K, \gamma_{0,\ell}^K),$$

such that

$$\lim_{t\to\infty}\sup_{z>0}|\overline{F}_t(z \mid \mathcal{T}_\ell) - \overline{H}(z;\sigma^K_{0,\ell}(t),\gamma^K_{0,\ell})| = 0,$$

where $\overline{F}_t(z \mid \mathcal{T}_\ell) = \mathbb{P}(Y - t \geq z \mid \mathbf{X} \in \mathcal{T}_\ell, Y \geq t)$.

Hence, $\widehat{T}_K$ denotes the tree with leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ and with parameters $\widehat{\boldsymbol{\theta}}^K(u) = (\widehat{\boldsymbol{\theta}}^K_\ell)_{\ell=1,\dots,K}$. Similarly, we denote by T^*_K (resp. $T_{0,K}$) the tree with the same leaves as $\widehat{T}_K$ but with parameters $\boldsymbol{\theta}^{*K} = (\boldsymbol{\theta}^{*K}_\ell)_{\ell=1,\dots,K}$ (resp. $\boldsymbol{\theta}^K_0 = (\boldsymbol{\theta}^K_{0,\ell})_{\ell=1,\dots,K}$).

For any sequence of parameters $\boldsymbol{\theta}^K = (\boldsymbol{\theta}^K_\ell)_{\ell=1,\dots,K}$ of a tree with K leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$, we denote $\boldsymbol{\theta}^K(\mathbf{x})$ the regression function defined as the following step-wise function

$$\boldsymbol{\theta}^K(\mathbf{x}) = \sum_{\ell=1}^K \boldsymbol{\theta}^K_\ell \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}.$$

In the next section, we will also need some regularity assumptions on the negative log-likelihood $y \to \phi(y-u,\boldsymbol{\theta})\mathbf{1}_{y>u}$.

Assumption 2. *For $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\theta}_4 \in \boldsymbol{\Theta}$, let*

$$H^\ell_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\theta}_4}(y-u) = \begin{pmatrix} \partial^2_\sigma\phi(y-u,\boldsymbol{\theta}_1) & \partial_\sigma\partial_\gamma\phi(y-u,\boldsymbol{\theta}_2) \\ \partial_\sigma\partial_\gamma\phi(y-u,\boldsymbol{\theta}_3) & \partial^2_\gamma\phi(y-u,\boldsymbol{\theta}_4) \end{pmatrix} \mathbf{1}_{y\geq u}.$$

Assume that there exists a constant $\mathfrak{C}_1 > 0$ such that

$$\inf_{\substack{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\theta}_4\in\boldsymbol{\Theta}\\ a,b\in\mathbb{R}}} \inf_{\substack{\ell=1,\dots,K\\ u_{\min}\leq u\leq u_{\max}}} \left|\mathbb{E}\left[H^\ell_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3,\boldsymbol{\theta}_4}(Y-u)\begin{pmatrix} a \\ b\end{pmatrix} \mid \mathbf{X}\in\mathcal{T}_\ell\right]\right| \geq \mathfrak{C}_1\|(a,b)\|_\infty,$$

where $\|(a,b)^t\|_\infty = \max(|a|,|b|)$.

Remark 3. *The condition on the infimum can be relaxed: Assumption 2 comes naturally in using a Taylor expansion. Hence, the infimum with respect of $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_4$ can be restricted to $\boldsymbol{\theta}_2$ to $\boldsymbol{\theta}_3$ belonging to a small neighborhood of $\boldsymbol{\theta}_1$ (and not to the whole set $\boldsymbol{\theta}$).*

We will first focus on the difference $\widehat{T}_K$ and T^*_K in Section 3.2, which is the stochastic part of the error. Section 3.3 concerns the difference between T^*_K and $T_{0,K}$ (and ultimately the difference between the regression functions $\widehat{\boldsymbol{\theta}}^*(\mathbf{x})$ and $\boldsymbol{\theta}_0(\mathbf{x})$) that can be understood as a misspecification term, caused by the fact that the excesses above the threshold are not exactly GP distributed. Finally, the consistency of the pruning step is shown in Section 3.4.

3.2 Deviation bounds for our estimator

In this section, we study the consistency of a fitted tree $\widehat{T}_K$ with K leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$, a subtree of the maximal tree $\widehat{T}_{\max}$. For this first result, K is fixed. Selection results for K are provided in Theorem 3 in Section 3.4. The leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ of $\widehat{T}_K$ are supposed to be fixed sets, as it is classically assumed to derive consistency of regression trees, [see e.g. Chaudhuri, 2000, Chaudhuri and Loh, 2002]. Recall that the tree $\widehat{T}_K$ is identified by its leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ and the list of parameter values $\widehat{\boldsymbol{\theta}}^K_\ell$ associated with each leaf $\mathcal{T}_\ell$. Considering a leaf $\mathcal{T}_\ell$, $\widehat{\boldsymbol{\theta}}^K_\ell$ should ideally be close to its limit value $\boldsymbol{\theta}^{*K}_\ell$, as n tends to ∞. Hence, we introduce the "oracle" tree $\widehat{T}^*_K$ which is defined by the same subdivision $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ as $\widehat{T}_K$ but differs via the value of the parameters in each leaf (which is taken as $\boldsymbol{\theta}^{*K}_\ell$ for leaf ℓ). We denote $\boldsymbol{\theta}^{*K}(\mathbf{x})$ the regression function associated with $\widehat{T}^*_K$.

To compare $\widehat{T}_K$ and T^*_K, the first step is to define a distance between trees. Let us define for two trees T and T' associated with the regression functions $\boldsymbol{\theta}(\mathbf{x}) = (\sigma(\mathbf{x}),\gamma(\mathbf{x}))^t$ and $\boldsymbol{\theta}'(\mathbf{x}) = (\sigma'(\mathbf{x}),\gamma'(\mathbf{x}))^t$ respectively,

$$\|T - T'\|_2 = \left(\int \|\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}'(\mathbf{x})\|^2_\infty \mathrm{d}P_{\mathbf{X}}(\mathbf{x})\right)^{1/2},$$

where $P_{\mathbf{X}}$ denotes the distribution of the covariates $\mathbf{X}$ and $\|\boldsymbol{\theta}(\mathbf{x}) - \boldsymbol{\theta}'(\mathbf{x})\|_\infty = \max\left(|\sigma(\mathbf{x}) - \sigma'(\mathbf{x})|, |\gamma(\mathbf{x}) - \gamma'(\mathbf{x})|\right)$.

The main result of this section is a deviation bound for $\|\widehat{T}_K - T^*_K\|_2$, which is Theorem 1 below.

Theorem 1. *Under Assumptions 1 and 2, there exists $\rho_0 > 0$ such that for $\beta \geq 10/(\rho_0 a_1)$ and $t \geq c_1 K(\log k_n)k_n^{-1}$, with $c_1 > 0$,*

$$\begin{aligned}\mathbb{P}&\left(\sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K - T_K^*\|_2^2 \geq t\right)\\ &\leq \quad 2\left(\exp\left(-\frac{\mathcal{C}_1 k_n t}{K\beta^2(\log k_n)^2}\right) + \exp\left(-\frac{\mathcal{C}_2 k_n t^{1/2}}{K^{1/2}\beta \log k_n}\right)\right) + \frac{\mathcal{C}_3 K}{k_n^{5/2} t^{3/2}},\end{aligned} \tag{7}$$

where $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ are positive constants.

Moreover,

$$\mathbb{E}\left[\sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K - T_K^*\|_2^2\right] \leq \mathcal{C}_4 \frac{K\beta^2(\log k_n)^2}{k_n}. \tag{8}$$

The proof of Theorem 1 is postponed to the appendix section (Section A.3). The exponential terms on the right-hand side of (7) come from concentration inequalities proved by Einmahl et al. [2005], while the polynomially decreasing term is related to the fact that the log-likelihood is an unbounded quantity, but that can still controlled when considering its expectation.

As a by-product, we obtain (8) (by integration of the bound of (7)). From (8), one can see that the L^2-norm of the stochastic part of the error, $\mathbb{E}\left[\sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K - T_K^*\|_2^2\right]^{1/2}$, is proportional to $K^{1/2}$, and, as expected, increases with the complexity of the tree. On the other hand, the error decreases almost at rate $k_n^{1/2}$ (up to some logarithmic factor), which is the convergence rate of standard estimators used to estimate the parameters of a GP distribution in absence of covariates.

Let us note that we do not explicitly take into account the dimension d of the covariate $\mathbf{X}$ in the result of Theorem 1, in order to simplify the notations. However, it is possible to retrieve the contribution of the dimension through the results contained in the Appendix: it appears inside the covering numbers obtained in Lemma 10 and then can be tracked through all the proofs below. All the constants provided in the results are increasing functions of d. From an asymptotic point of view, they could modify the rate of consistency if d were allowed to go to infinity with n. This is not a situation when regression trees are traditionally used, since a too high dimension for d would lead to a too important computation time.

3.3 Misspecification bias

For $\mathbf{X} = \mathbf{x}$, the ultimate goal is to estimate the parameter set $\boldsymbol{\theta}_0(\mathbf{x}) = (\sigma_0(\mathbf{x}), \gamma_0(\mathbf{x}))^t$, introduced in (2), by maximization of the GP likelihood, and from the fact that the true function $\boldsymbol{\theta}_0(\mathbf{x})$ is not necessarily piecewise constant as $\boldsymbol{\theta}^*(\mathbf{x})$. The difference between $\boldsymbol{\theta}_0(\mathbf{x})$ and $\boldsymbol{\theta}^*(\mathbf{x})$ can be understood as a misspecification term due to the fact that the observations above the threshold are not exactly distributed according to a GP distribution. This bias term can be controlled under second order conditions which are standard in Extreme Value Analysis [see e.g. Beirlant et al., 2004].

Indeed, recall that assuming that the underlying distribution $\overline{F}(\cdot \mid \mathbf{x})$ satisfies Condition (3) guarantees that asymptotically the associate excesses above the threshold u are GP distributed. For finite samples, the excesses are thus not exactly GP distributed which introduces some bias term. In order to control this bias term, a second-order condition is needed, that is a condition to control the rate of convergence in Condition (3). There exist numerous ways to express this second-order condition. Here, we consider the same condition as Condition C.6 in [Beirlant and Goegebeur, 2004]. First, Condition (3) can be translated into

$$\overline{F}(y \mid \mathbf{x}) = y^{-1/\gamma_0(\mathbf{x})}\eta(y \mid \mathbf{x}), \forall y > 0, \tag{9}$$

where η is a slow-varying function, that is $\eta(ty \mid \mathbf{x})/\eta(t \mid \mathbf{x}) \to 1$ as $t \to \infty$, for all $y > 0$.

Assumption 3. *Assume that for all $\mathbf{x}$, there exist a constant c and a function ψ such that*

$$\eta(ty \mid \mathbf{x})/\eta(t \mid \mathbf{x}) = 1 + c\psi(t)\int_1^t v^{\rho-1}\mathrm{d}v + o(\psi(t))$$

as $t \to \infty$ for each $y > 0$ with $\psi(t) > 0$ and $\psi(t) \to 0$ as $t \to \infty$ and $\rho \leq 0$.

Let us note that we could also consider the case of c, ψ and ρ depending on $\mathbf{x}$, and then assume some uniform bound over x of these quantities. We chose this more restrictive formulation to simplify the notations.

The next result guarantees that the bias term tends to 0 as $u \to \infty$.

Proposition 2. *Under Assumptions 2 and 3, there exist a constant c and a function ψ such that $\psi(u) > 0$, $\psi(u) \to 0$ as $u \to \infty$, and such that, for $\mathbf{X} = \mathbf{x}$,*

$$\|\boldsymbol{\theta}_0(\mathbf{x}) - \boldsymbol{\theta}^*(\mathbf{x})\|_\infty \quad \leq \quad \mathfrak{C}_2 \frac{k_n}{n} \left(1 + c\gamma_{\max}\psi(u) + o(\psi(u))\right) ,$$

where $\mathfrak{C}_2$ is a constant depending on u, $\gamma_{\min}$ and $\gamma_{\max}$.

3.4 Consistency of the pruning step

The previous results cover the case of a tree with a fixed number of leaves K. In practice, the question is to select the proper subtree of $\widehat{T}_{\max}$, the maximal tree obtained once the previous step of the CART procedure has stopped, with some "optimal" number of leaves, which is the objective of the pruning step described in Section 2.3.2.

As seen in Theorem 1 Equation (8), the stochastic part of the error put to the square increases proportionally to K. This is closely related to the natural inflation of the log-likelihood (which is locally quadratic) when the number of leaves increases, justifying a penalty proportional to K, as in [Breiman et al., 1984, Gey and Nedelec, 2005]. The aim of Theorem 3 is to corroborate this choice.

Let K^* denote the optimal number of leaves, that is

$$K^* = \min \left\{ \arg \min_{K=1,\dots,K_{\max}} \mathbb{E}\left[\phi(Y - u, \boldsymbol{\theta}^{*K}(\mathbf{X}))\mathbf{1}_{Y>u}\right] \right\}.$$

In words, $T^* = T^*_{K^*}$ is the subtree of $T^*_{\max}$ that achieves the closest proximity to the objective function $\mathbf{x} \to \boldsymbol{\theta}^*(\mathbf{x})$ in the sense that it maximizes the expectation of the (pseudo)-log-likelihood.

Second of all, as explained in Section 2.3.2, the selected number of leaves is defined by

$$\widehat{K} = \arg \min_{K=1,\dots,K_{\max}} \left\{ \frac{1}{k_n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \phi(Y_i - u, \widehat{\boldsymbol{\theta}}^K(\mathbf{X}_i))\mathbf{1}_{Y_i>u}\mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell} + \lambda K \right\} ,$$

and $\widehat{T} = T_{\widehat{K}}$ the corresponding selected tree.

The following Theorem 3 shows that the pruning methodology selects a tree $\widehat{T}$ which approximately achieves the same rate of convergence as $\widehat{T}_{K^*}$, even if K^* is unknown, provided that the penalty constant λ belongs to some reasonable interval.

In Theorem 3, $\Delta L(T^*, T^*_K)$ denotes the expectation of the difference of the likelihoods associated with the trees T^* and T^*_K (for a formal definition see Section A.5).

Theorem 3. *Let $\mathfrak{D} = \inf_u \inf_{K<K^*} \Delta L(T^*, T^*_K)$ and suppose that there exists a constant $c_2 > 0$ such that the penalization constant λ satisfies*

$$c_2\{\log k_n\}^{1/2} k_n^{-1/2} \leq \lambda \leq \mathfrak{D} - 2c_2\{\log(k_n)\}^{1/2} k_n^{-1/2},$$

assuming that the right-hand side is positive. Then, under Assumptions 1 and 2, for all $u \in [u_{\min}, u_{\max}]$,

$$\mathbb{E}\left[\|\widehat{T} - T^*\|_2^2\right] \leq \frac{\mathcal{C}_5 K^* (\log k_n)^2}{k_n},$$

where $\mathcal{C}_5$ is a constant depending on T^*.

The proof is given in Section A.5.

4 Simulation study and real data analysis

This section is devoted to the illustration of the GP regression procedure on simulated data (Section 4.1) and on a real dataset (Section 4.2).

For both the simulations and the real data application, we used the `R` package `rpart` package for the GP CART procedure. The function `rpart` allows to fix the tuning parameter `minbucket`, which represents the minimal number of observations allowed in each leaf, that is the stopping rule. This tuning parameter was set to 50 for the simulations and 20 for the real data applications.

4.1 Simulations

In this section, we assess the performance of the GP regression procedure on simulated data and compare it with the competing approach proposed by Chavez-Demoulin et al. [2015]. They propose a semi-parametric framework to separate the smooth covariates from the discrete ones. Smoothing splines are used to estimate non-parametrically the smooth part, while the influence of discrete covariates is captured by a parametric function. This framework relies on a stronger assumption on the shape of the function $\boldsymbol{\theta}_0$.

We now describe the two cases considered in the simulation framework and then discuss the experiments results. In both cases, conditionally on the covariates $\mathbf{X} = \mathbf{x}$, the response variable Y is assumed to be distributed according to a Burr distribution of parameters $(\sigma_0(\mathbf{x}), \gamma_0(\mathbf{x}))$ whose survival function is given by

$$\overline{F}(y \mid \mathbf{x}) = \frac{1}{1 + (y/\sigma_0(\mathbf{x}))^{1/\gamma_0(\mathbf{x})}},$$

with $\sigma_0(x) > 0$ and $\gamma_0(x) > 0$ for all $\mathbf{x}$. Note that $\overline{F}(\cdot \mid \mathbf{x})$ satisfies Property (3).

Step-wise case In this first case, X is assumed to be an one-dimensional variable uniformly distributed on $[0, 1]$, the function γ_0 is taken as

$$\gamma_0(x) = \begin{cases} 0.8 & \text{if } 0 \leq x < 0.3 \\ 0.4 & \text{if } 0.3 \leq x < 0.7 \\ 0.2 & \text{if } 0.7 \leq x \leq 1, \end{cases}$$

and then we consider two settings for the function $\sigma_0(x)$:

1. $\sigma_0(x) = 1 - \gamma_0(x)$. This guarantees that the mean of the GP distribution is constant.
2. $\sigma_0(x) = (2^{\gamma_0(x)} - 1)/\gamma_0(x)$, here the median of the GP distribution is constant.

For some x, the function $\gamma(x)$ exceeds 0.5, which corresponds to the case where the conditional variance is not defined. This case is important for risk management: if the variable Y corresponds to the loss associated to a given risk, the mean-variance paradigm traditionally used by risk managers does not hold.

Smooth case In this second case, $\mathbf{X}$ is no longer assumed to be an one-dimensional variable uniformly distributed on $[0, 1]$, we consider a two-dimensional variable $\mathbf{X} = (X^{(1)}, X^{(2)})$. The functions γ_0 and σ_0 are then taken as

$$\begin{aligned} \gamma_0(x) &= 1 + \frac{\tanh(10(x - 1/4))}{4} + \frac{\tanh(10(x - 3/4))}{4} \\ \sigma_0(x) &= \begin{cases} 1 & \text{if } x \leq 0.5 \\ 0.5 & \text{if } x > 0.5 \end{cases} \end{aligned}$$

where $x = tx^{(1)} + (1 - t)x^{(2)}$ for $t \in [0, 1]$.

We simulate 1,000 replicates of samples of size n, with n =1,000; 2,500; 5,000; 10,000 and 25,000 according to the described framework for all the cases. For each sample, we consider the excesses above the threshold u =0.90-empirical quantile, which corresponds to k_n =100; 250; 500; 1,000 and 2,500 excesses above u. For each simulated sample, we compute the regression tree procedure (GP CART), and the method based on generalized additive model (GAM) proposed by Chavez-Demoulin et al. [2015]. Next, we compare the models by computing the three different empirical root-mean-square errors (RMSE) obtained by averaging the following quantities over the 1,000 replicates.

1. for the estimation of $\boldsymbol{\theta}_0(x) = (\sigma_0(x), \gamma_0(x))^t$, that is

$$\left(\int \left\{ (\widehat{\sigma}(x) - \sigma_0(x))^2 + (\widehat{\gamma}(x) - \gamma_0(x))^2 \right\} \mathrm{d}x \right)^{1/2}$$

Results are shown in Table 1 and the corresponding boxplots in Figures 1, 2, 3.

Table 1: Empirical RMSE for the estimation of $\boldsymbol{\theta}(\mathbf{x})$ for the GP regression tree procedure (GP CART), and the GAM model for different values of k_n for a) the step-wise case with the constant mean (setting 1), b) the step-wise case with the constant median (setting 2), and c) the smooth case.

k_n	100	250	500	1,000	2,500
GP CART	0.8101	0.8058	0.8032	0.8026	0.8021
GAM	0.8054	0.7777	0.7618	0.7541	0.7484

a)

k_n	100	250	500	1,000	2,500
GP CART	0.8099	0.8058	0.8032	0.8026	0.8021
GAM	0.8051	0.7777	0.7618	0.7541	0.7484

b)

k_n	100	250	500	1,000	2,500
GP CART	1.2029	1.2615	1.2345	1.2081	1.1971
GAM	1.2546	1.2768	1.2736	1.2220	1.2417

c)

2. for the conditional survival function $\overline{F}(Y \mid x)$, that is

$$\left(\frac{1}{k_n} \sum_{i=1}^{k_n} \left(\overline{F}_u(Z_i \mid x_i) - \overline{H}(Z_i; \widehat{\sigma}(x_i), \widehat{\gamma}(x_i)) \right)^2 \right)^{1/2}$$

   Results are shown in Table 2 and the corresponding boxplots are presented in Section A of the supplementary material.

3. for the estimation of 0.95-quantile $q_{0.95}(x)$, that is

$$\left(\int (\hat{q}_{0.95}(x) - q_{0.95}(x))^2 \mathrm{d}x \right)^{1/2}$$

   Results are shown in Table 3 and the corresponding boxplots are presented in Section A of the supplementary material.

Tables 1, 2 and 3 show that the GAM and the GP CART procedures present similar results. Results on the RMSE for the estimation of $\boldsymbol{\theta}_0(x)$ and of 0.95-quantile $q_{0.95}(x)$, the GAM procedure seems to perform slightly better in the step-wise case and the GP CART in the smooth case while results for the conditional survival function, the GP CART seems to have a better performance in the step-wise case and the GAM procedure in the smooth case. The boxplots on the quadratic errors present the same conclusion. The simulation study shows that the GP CART procedure can be applied in various situations, is thus very flexible and an easy interpretation of the results.

4.2 Prediction of the cost of flooding events in France

In order to improve the knowledge and the management of natural catastrophes, France Assureurs (FA, French Federation of Insurance) is interested in the prediction of the cost of such events, especially of the most severe ones, shortly after their occurrence. These catastrophic events present some heterogeneity in their intensity depending on their characteristics, such as the affected meteorological region or the number of individual houses in flood risk area. The prediction of their cost thus becomes a challenging task. In this section, we illustrate how the GP regression tree procedure can be used to gain further

Table 2: Empirical RMSE for the conditional survival function for the GP regression tree procedure (GP CART), and the GAM model for different values of k_n for a) the step-wise case with the constant mean (setting 1), b) the step-wise case with the constant median (setting 2), and c) the smooth case.

k_n	100	250	500	1,000	2,500
GP CART	0.5186	0.5180	0.5182	0.5180	0.5182
GAM	0.5191	0.5196	0.5189	0.5192	0.5200

a)

k_n	100	250	500	1,000	2,500
GP CART	0.4891	0.4881	0.4883	0.4881	0.4883
GAM	0.5191	0.5196	0.5189	0.5192	0.5200

b)

k_n	100	250	500	1,000	2,500
GP CART	0.1179	0.1330	0.1308	0.1288	0.1273
GAM	0.0949	0.1091	0.1149	0.1162	0.1170

c)

Table 3: Empirical square root mean squared errors for the estimation of the 0.95-quantile $q_{0.95}(x)$ for the GP regression tree procedure (GP CART), and the GAM model for different sample sizes for a) the step-wise case with the constant mean (setting 1), b) the step-wise case with the constant median (setting 2), and c) the smooth case.

k_n	100	250	500	1,000	2,500
GP CART	0.382	0.382	0.382	0.382	0.382
GAM	0.378	0.374	0.373	0.371	0.371

a)

k_n	100	250	500	1,000	2,500
GP CART	0.382	0.382	0.382	0.382	0.382
GAM	0.378	0.374	0.373	0.369	0.369

b)

k_n	100	250	500	1,000	2,500
GP CART	1.382	1.414	1.360	1.311	1.285
GAM	1.622	1.526	1.469	1.338	1.261

c)

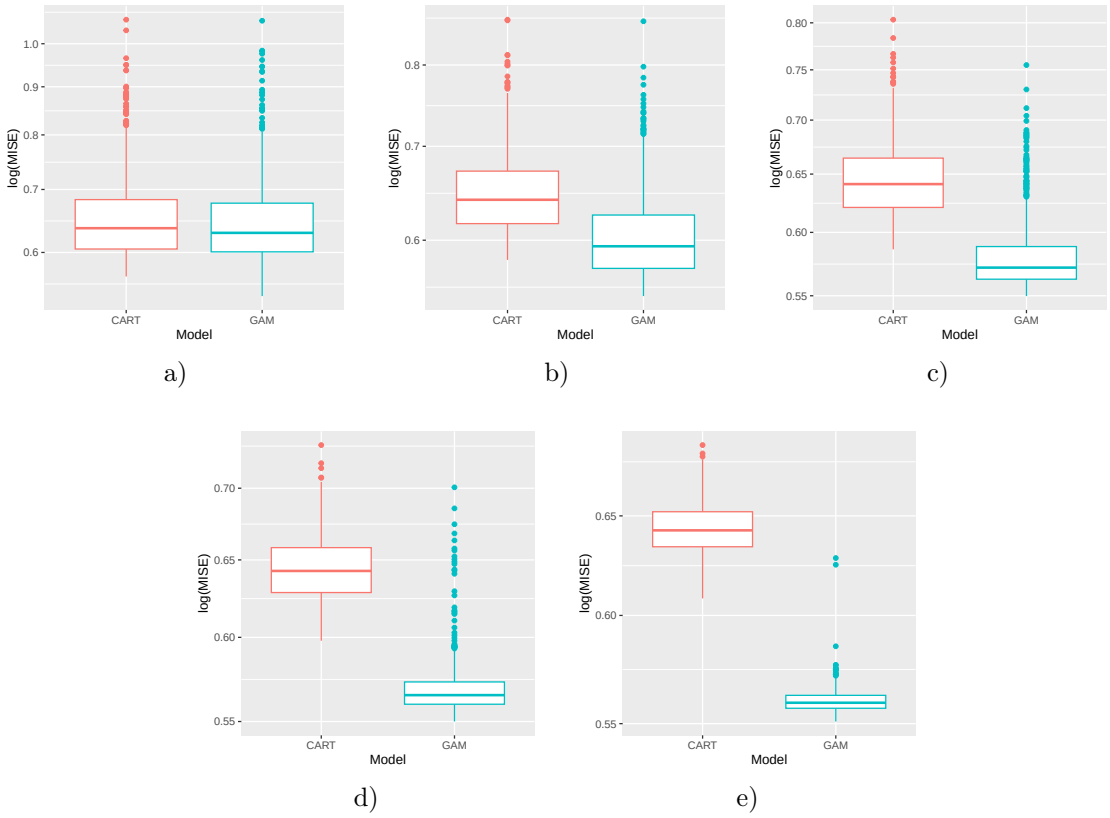

Figure 1: Boxplots (in logarithm scale) of the quadratic errors for the estimation of $\boldsymbol{\theta}$ for each model in the step-wise case (setting 1) for a) 100 b) 250 c) 500 d) 1,000 and e) 2,500 excesses.

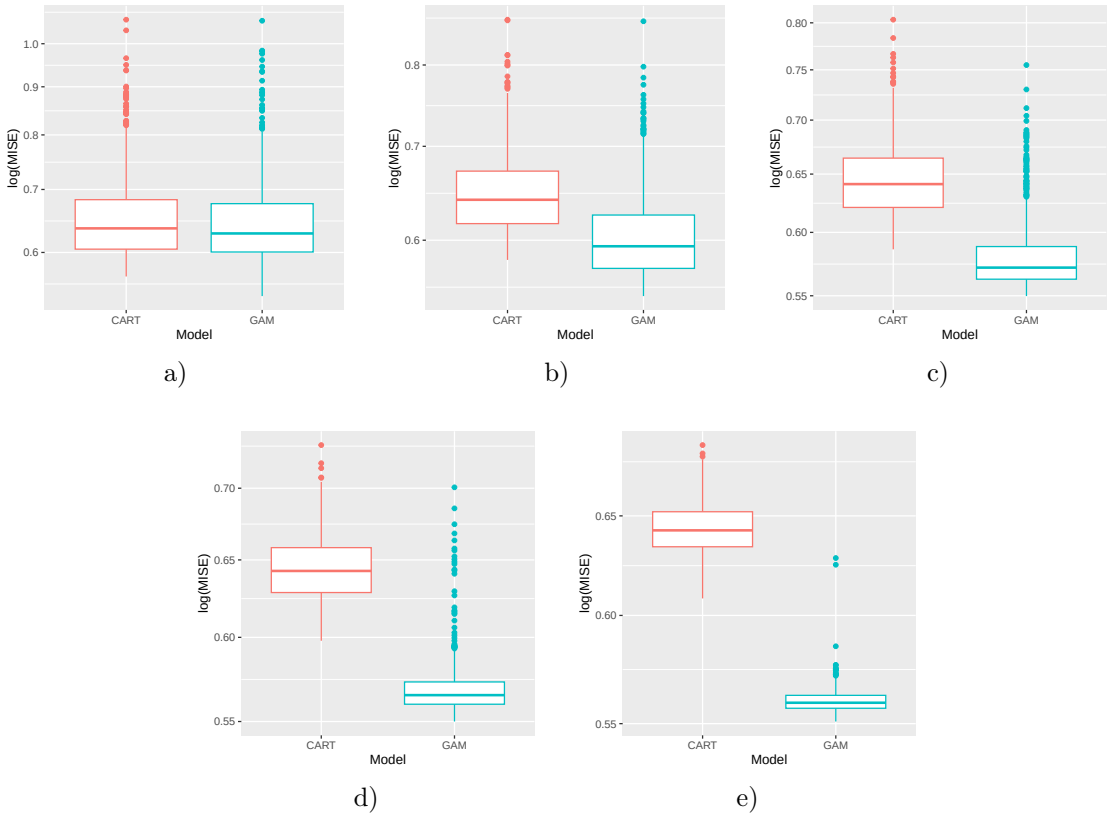

Figure 2: Boxplots (in logarithm scale) of the quadratic errors for the estimation of $\boldsymbol{\theta}$ for each model in the step-wise case (setting 2) for a) 100 b) 250 c) 500 d) 1,000 and e) 2,500 excesses.

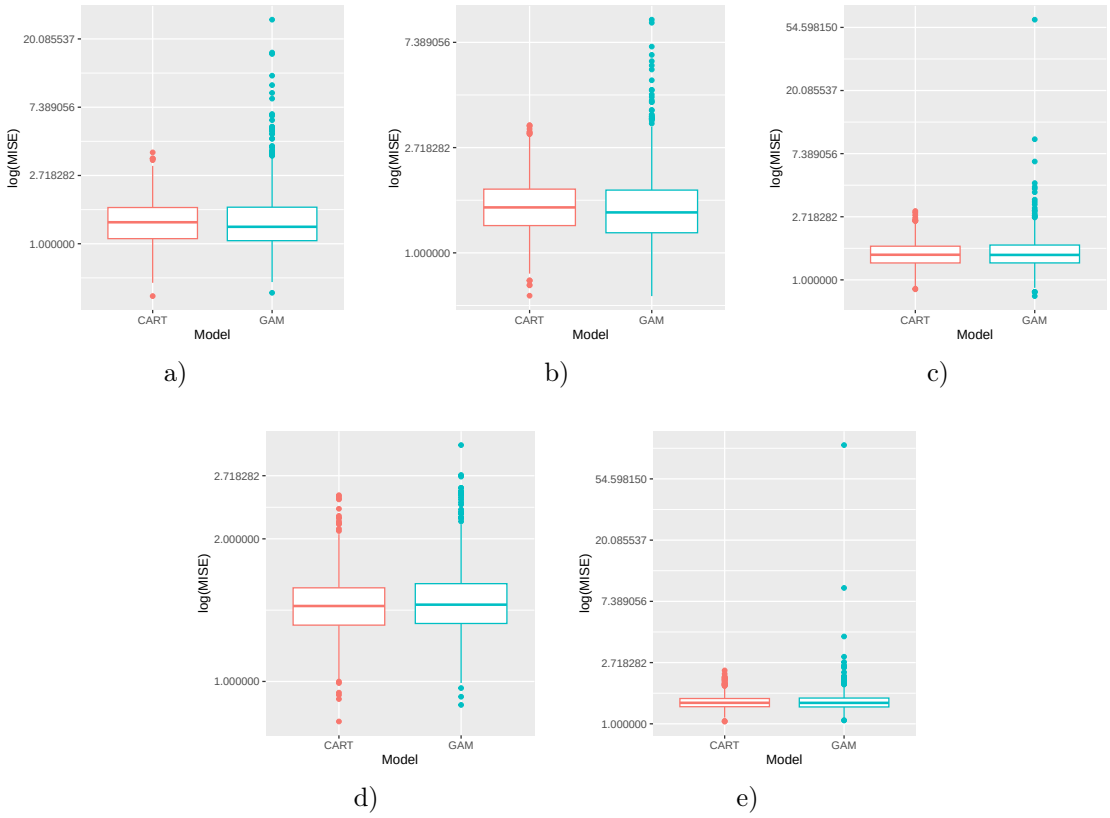

Figure 3: Boxplots (in logarithm scale) of the quadratic errors for the estimation of $\boldsymbol{\theta}$ for each model in the smooth case for a) 100 b) 250 c) 500 d) 1,000 and e) 2,500 excesses.

insight in this heterogeneity. The ability of the procedure to design classes of events that are more homogeneous (in view of analyzing the tail of their distribution) is an appealing property in view of operation applications in insurance.

The database we consider was obtained through a partnership with the FA, in particular with one of its dedicated technical body, the association of French insurance undertaking for natural risk knowledge and reduction (Mission Risques Naturels, MRN). It consists of all 4,300 flooding events that have been granted the status of natural catastrophe in France from 1999 to 2021 (let us note that the status "natural catastrophe" is a French specificity, with some legal consequences when an event receives this label [see Charpentier et al., 2021, MRN, 2016]). This database is fed by 12 contributors including the major French insurance companies, allowing this database to cover 70% of French non-life insurance market. The database gathers information regarding each flooding event (its cost, the meteorological region, the season, the number of affected hydrological regions, the number of individual houses and the number of professional business premises in flood-risk area). Note that, since the purpose of this database is the fast prediction of the cost of a flooding event (as soon as possible after its occurrence), the variables that are registered correspond to quantities that are available before the event, or soon after it.

The variable of interest, the total cost of a flooding event, is highly volatile. Indeed, it ranges between 0 and $394,376,000$ euros with an empirical variance equal to $1.77e+14$. Figure 4 shows the average of the costs of the 10% most onerous flooding events within each meteorological region. This highlights the heterogeneity of the severity of the most severe events. Furthermore, the top ten most onerous events represent 43% of the total cost of this database and the top hundred 80%.

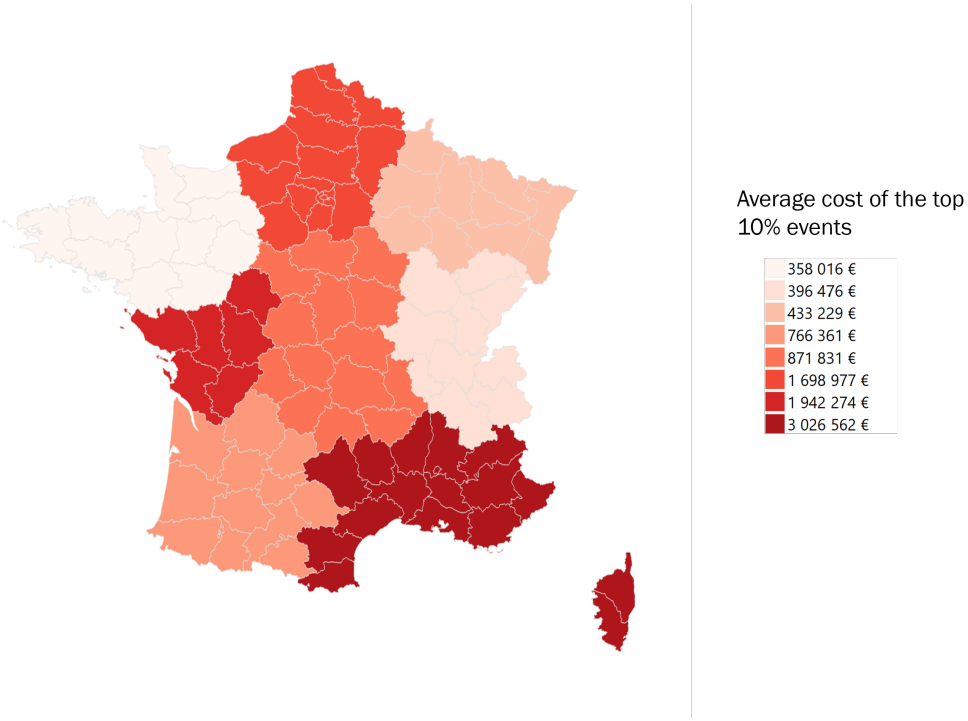

Figure 4: Cartography of the cost of flooding events in France from 1999 to 2019. For each meteorological region, the average of the costs of the 10% more onerous events is shown. The lighter red color suggesting a small cost while a darker color suggests a large cost.

Now, let us recall that our goal is to understand the heterogeneity of the total cost of the most severe flooding events, that is of extreme flooding events. As explained in Section 2.1, the definition of extreme events consists in choosing a threshold u, which should be chosen as a bias-variance trade-off. We chose a value of $u = 100,000$ based practical considerations and validated by sensitivity analyses (shown in the supplementary material, Section D). This yields 1,100 extreme events, that is for which the cost is larger

than u.

The GP regression tree was performed on the database corresponding to the flooding events extracted from the original database for which the total cost is larger than u (=100 000 euros). The variables of this database and their characteristics are summarized in Table 4. Again, it can be noticed that the cost, the variable of interest, is highly volatile.

Table 4: List of quantitative and categorical variables in the database and their characteristics. For the quantitative variables, Table a) shows the minimum, the first quartile, the median, the mean, the third quartile and the maximum, and for the categorical variables, Table b) the number of observations per category.

Variable	Min	1st Q	Median	Mean	3rd Q	Max
Cost (in euros)	100,093	199,287	477,943	6,066,835	1,941,047	380,487,161
Number of affected hydrological regions	1	1	2	4	4	35
Number of individual houses in flood risk area	0	5,874	20,692	92,477	71,094	4,097,075
Number of professional business premises in flood risk area	0	2,230	8,163	44,830	26,321	2,050,165

a)

Variable	Category	Number of observations
Meteorological regions	Center	60
	North West	85
	North	135
	North-East	87
	East	96
	South	209
	West	30
	South West	121
Seasons	Spring	272
	Summer	279
	Autumn	187
	Winter	85

b)

The tree obtained from GP regression procedure is shown in Figure 5 (the quantile-quantile plots of the GP fit in each leaf are shown in the supplementary material, Section C). The tree is composed of 6 leaves, with separations according to 3 criteria, the number of individual houses in flood risk area, the number of professional business premises in flood risk area, and the number of affected hydro-ecoregions. This seems consistent because the first two covariates represent the exposure to flooding but also the population density of the affected area, the third covariate captures the perimeter of the event. The most extreme case corresponds to the far right leaf, with a shape parameter of 0.92, it contains 7% of the events. It corresponds to an important number of affected individual houses and to a large area. Table 5 presents for each leaf the empirical median and mean of the costs and the theoretical median and mean of the corresponding GP distribution. Let us recall that for a GP distribution with a scale parameter σ and a shape parameter γ, the theoretical median is given by $\sigma(2^{\gamma}-1)/\gamma$ and the theoretical mean by $\sigma/(1-\gamma)$ for $\gamma < 1$ and ∞ for $\gamma \geq 1$. First of all, for every leaf, the median is much smaller than the mean suggesting that we are indeed dealing with extreme events. Then, the empirical and theoretical medians and means are of the same order for each leaf, and it appears that we have a good fit, especially for the median. To address the uncertainty concerning parameters estimation, we present in Table 6 the 95% confidence intervals for the shape parameter γ and in Table 7 the 95% confidence for scale parameter σ.

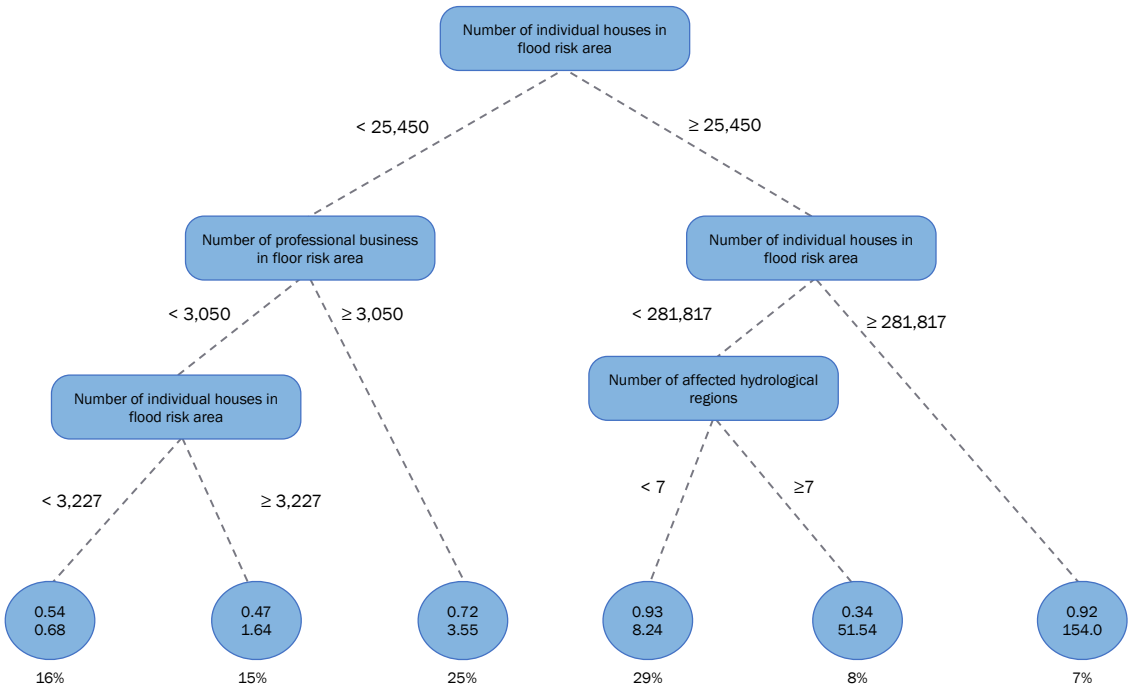

Figure 5: GP regression tree obtained for flooding events. For each leaf, the value of the shape parameter γ (first line) and the scale parameter σ at 10^{-5} (second line) are given. Percentage of observations affected to each leaf is mentioned.

Leaf	Shape parameter	Empirical Median	Theoretical Median	Empirical Mean	Theoretical Mean
1	0.54	161,694	157,697	239,923	249,456
2	0.47	226,196	234,764	399,274	410,387
3	0.72	455,663	419,978	1,439,087	1,390,099
4	0.93	950,181	902,387	4,144,876	11,877,446
5	0.34	4,215,647	4,140,879	7,982,445	8,009,145
6	0.92	15,555,487	15,090,137	52,203,995	281,103,859

Table 5: Empirical median and mean, and theoretical median and mean for each leaf (in euros).

Leaf	Shape parameter estimate	Lower CI	upper CI
1	0.54	0.27	0.82
2	0.47	0.21	0.73
3	0.72	0.50	0.95
4	0.93	0.67	1.19
5	0.34	0.03	0.67
6	0.92	0.38	1.46

Table 6: 95% confidence intervals for the shape parameter γ

5 Conclusion

In this paper, we investigated the consistency of Generalized Pareto regression trees, applied to extreme value regression. The results that we derive are non-asymptotic, and allow to justify the consistency of the pruning methodology used to select a proper subtree. Let us note that the conditions under which our results hold are relatively weak, in the sense that they hold even if the tail index γ is arbitrary close to zero (the special case $\gamma = 0$ is excluded) or large. Moreover, no regularity assumptions on the target parameters is required, due to the flexibility of the regression tree procedure.

Leaf	Scale parameter estimate	Lower CI	upper CI
1	0.68	0.47	0.90
2	1.64	1.15	2.14
3	3.55	2.66	4.44
4	8.24	6.06	10.42
5	51.54	31.36	71.53
6	154.0	69.51	238.33

Table 7: 95% confidence intervals for the scale parameter σ

Through the simulation study and the real data analysis, we investigated the practical performances of the methodology. The regression tree approach can be applied in various situations, and still provides interpretability of the results. On the other hand, regression trees may be unstable, since quite sensitive to some changes on the data that have been used to fit them. Hence, this work is a first step into the direction of studying other relied methodologies, like random forests [see for example Breiman et al., 1984] in this field of extreme value regression.

A Proofs

In this Section, we present in details the proof of the results presented throughout the paper. Concentration inequalities required to obtain the results are presented in Section A.1. These inequalities are used to obtain deviation bounds in Section A.2, which are the key ingredients of the proof of Theorem 1 (Section A.3), Corollary 8 (Section A.2), and Theorem 3 (Section A.5). Section B shows some results on covering numbers that are required to control the complexity of some classes of functions considered in the proofs. Some technical lemmas are gathered in Section C.

A.1 Concentration inequalities

The proofs of the main results are mostly based on concentration inequalities. The following inequality was proved initially Talagrand [1994], [see also Einmahl et al., 2005].

Proposition 4. *Let $(\mathbf{V}_i)_{1\leq i\leq n}$ denote i.i.d. replications of a random vector $\mathbf{V}$, and let $(\varepsilon_i)_{1\leq i\leq n}$ denote a vector of i.i.d. Rademacher variables (that is, $\mathbb{P}(\varepsilon_i=-1)=\mathbb{P}(\varepsilon_i=1)=1/2$) independent from $(\mathbf{V}_i)_{1\leq i\leq n}$. Let $\mathfrak{F}$ be a pointwise measurable class of functions bounded by a finite constant M_0. Then, for all t,*

$$\mathbb{P}\left(\sup_{\varphi\in\mathfrak{F}}\left\|\sum_{i=1}^{n}\{\varphi(\mathbf{V}_i)-\mathbb{E}[\varphi(\mathbf{V})]\}\right\|_\infty > A_1\left\{E\left[\sup_{\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}}\left\|\sum_{i=1}^{n}\varphi(\mathbf{V}_i)\varepsilon_i\right\|_\infty\right]+t\right\}\right)$$
$$\leq 2\left\{\exp\left(-\frac{A_2t^2}{nv_{\mathfrak{F}}}\right)+\exp\left(-\frac{A_2t}{M_0}\right)\right\},$$

with $v_{\mathfrak{F}}=\sup_{\varphi\in\mathfrak{F}}\mathrm{Var}(\|\varphi(\mathbf{V})\|_\infty)$, and where A_1 and A_2 are universal constants.

The difficulty in using Proposition 4 comes from the need to control the symmetrized quantity $\mathbb{E}\left[\sup_{\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}}\left\|\sum_{i=1}^{n}\varphi(\mathbf{V}_i)\varepsilon_i\right\|\right]$. Proposition 5 is due to Einmahl et al. [2005] and allows this control via some assumptions on the considered class of functions $\mathfrak{F}$.

We first need to introduce some notations regarding covering numbers of a class of functions. More details can be found for example in [van der Vaart, 1998, Chapter 2.6]. Let us consider a class of functions $\mathfrak{F}$ with envelope Φ (which means that for (almost) all v, $\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}$, $|f(v)|\leq\Phi(v)$). Then, for any probability measure $\mathbb{Q}$, introduce $N(\varepsilon,\mathfrak{F},\mathbb{Q})$ the minimum number of $L^2(\mathbb{Q})$ balls of radius ε to cover the class $\mathfrak{F}$. Then, define

$$\mathcal{N}_\Phi(\varepsilon,\mathfrak{F})=\sup_{\mathbb{Q}:\mathbb{Q}(\Phi^2)<\infty}N(\varepsilon(\mathbb{Q}(\Phi^2)^{1/2}),\mathfrak{F},\mathbb{Q}).$$

Proposition 5. *Let $\mathfrak{F}$ be a point-wise measurable class of functions bounded by M_0 with envelope Φ such that, for some constants $A_3,\alpha\geq 1$, and $0\leq\sqrt{v}\leq M_0$, we have*

(i) $\mathcal{N}_\Phi(\varepsilon, \mathfrak{F}) \leq A_3 \varepsilon^{-\alpha}$, *for* $0 < \varepsilon < 1$,

(ii) $\sup_{\varphi \in \mathfrak{F}} \mathbb{E}\left[\varphi(\mathbf{V})^2\right] \leq v$,

(iii) $M_0 \leq \frac{1}{4\alpha^{1/2}} \sqrt{nv/\log(A_4 M_0/\sqrt{v})}$, *with* $A_4 = \max(e, A_3^{1/\alpha})$.

Then, for some absolute constant A_5,

$$\mathbb{E}\left[\sup_{\varphi_{\boldsymbol{\theta}} \in \mathfrak{F}} \left\| \sum_{i=1}^n \varphi(\mathbf{V}_i)\varepsilon_i \right\|\right] \leq A_5 \sqrt{\alpha n v \log(A_4 M_0/\sqrt{v})}.$$

A.2 Deviation results

We first introduce some notations that will be used throughout Sections A.2 to B. In the following, $\varphi_{\boldsymbol{\theta}}$ is a function indexed by $\boldsymbol{\theta} = (\sigma, \gamma)^t$ denoting either $\phi(\cdot, \boldsymbol{\theta})$, $\partial_\sigma \phi(\cdot, \boldsymbol{\theta})$, or $\partial_\gamma \phi(\cdot, \boldsymbol{\theta})$.

We consider in the following the class of functions $\mathfrak{F}$ defined as

$$\mathfrak{F} = \{ y \mapsto \varphi_{\boldsymbol{\theta}}(y-u)\mathbf{1}_{y \geq u}\mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell},\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ u \in [u_{\min}; u_{\max}], \ell = 1, ..., K \}. \tag{10}$$

By Lemma 11, the functions $y \mapsto \partial_\sigma \phi(y-u, \boldsymbol{\theta})$ and $y \mapsto \partial_\gamma \phi(y-u, \boldsymbol{\theta})$ are uniformly bounded (eventually up to some multiplication by a constant) by $\Phi(y) = \log(1+wy)$, where $w = \gamma_{\max}/\sigma_{\min}$. On the other hand, $y \mapsto \phi(y-u, \boldsymbol{\theta})$ is bounded by $\log \sigma_n + \Phi(y) = O(\log(k_n)) + \Phi(y)$.

Next, for $\ell = 1, \ldots, K$, and $\boldsymbol{\theta} = (\sigma, \gamma)^t \in \boldsymbol{\Theta}$, let

$$L_n^\ell(\boldsymbol{\theta}, u) = \frac{1}{k_n} \sum_{i=1}^n \phi(Y_i - u, \boldsymbol{\theta}) \mathbf{1}_{Y_i > u} \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell},$$

be the (normalized) negative GP log-likelihood associated with the leaf ℓ of a tree T_K with set of K leaves $(\mathcal{T}_\ell)_{\ell=1,\ldots,K}$. Let $L^\ell(\boldsymbol{\theta}, u) = \mathbb{E}[L_n^\ell(\boldsymbol{\theta}, u)]$. The key results behind Theorems 1 and 3 relies on studying the deviations of the processes, indexed by $\boldsymbol{\theta}$, u and ℓ,

$$\begin{aligned} \mathcal{W}_0^\ell(\boldsymbol{\theta}, u) &= L_n^\ell(\boldsymbol{\theta}, u) - L^\ell(\boldsymbol{\theta}, u), \\ \mathcal{W}_1^\ell(\boldsymbol{\theta}, u) &= \nabla_{\boldsymbol{\theta}} L_n^\ell(\boldsymbol{\theta}, u) - \nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}, u). \end{aligned}$$

Let $M_n = \beta \log k_n \leq \beta a_1 \log(n)$ with $\beta > 0$ and $a_1 > 0$ (with a_1 defined in Assumption 1). We study the deviations of these processes by decomposing $\mathcal{W}_i^\ell(\boldsymbol{\theta}, u)$, for $i = 0, 1$, (which is a sum of i.i.d. observations) into two sums.

- the first one gathers observations smaller than some bound (more precisely, such that $\Phi(Y_i) \leq M_n$), which is considered in Theorem 6. Since these observations are bounded (even if this bound in fact depends on n and can tend to infinity when n grows), we can apply a concentration inequality such as the one of Section A.1. Let us stress that $\sup_{\varphi_{\boldsymbol{\theta}} \in \mathfrak{F}} \|\varphi_{\boldsymbol{\theta}}(y)\mathbf{1}_{\Phi(y) \leq M_n}\|_\infty \leq M_n$;
- in the second one (Theorem 7), we consider the observations larger than this bound, and control them through the fact that the function Φ has finite exponential moments (see Lemma 11).

Corollary 8, which provides deviation bounds for estimation errors in the leaves of the tree, is then a direct consequence.

Theorem 6. *Let*

$$\underline{\mathcal{Z}}(M_n) = \sup_{\varphi_{\boldsymbol{\theta}} \in \mathfrak{F}} \left| \frac{1}{k_n} \sum_{i=1}^n \left(\varphi_{\boldsymbol{\theta}}(Y_i)\mathbf{1}_{\Phi(Y_i) \leq M_n} - \mathbb{E}\left[\varphi_{\boldsymbol{\theta}}(Y_i)\mathbf{1}_{\Phi(Y_i) \leq M_n}\right] \right) \right|.$$

If $k_n = O(n^{a_1})$ *with* $a_1 > 0$ *(Assumption 1), then, for* $t \geq \mathfrak{c}_1 (\log k_n)^{1/2} k_n^{-1/2}$,

$$\mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geq t\right) \leq 2\left(\exp\left(-\frac{C_1 k_n t^2}{\beta^2 (\log k_n)^2} \right) + \exp\left(-\frac{C_2 k_n t}{\beta \log k_n} \right) \right). \tag{11}$$

Proof. From Proposition 4,

$$\begin{aligned}\mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geq A_1\left\{\mathbb{E}\left[\sup_{\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}}\frac{1}{k_n}\left|\sum_{i=1}^{n}\varphi_{\boldsymbol{\theta}}(Y_i)\mathbf{1}_{\Phi(Y_i)\leq M_n}\varepsilon_i\right|\right]+t\right\}\right) \\ \leq \quad 2\left(\exp\left(-\frac{A_2k_n^2t^2}{nv_{\mathfrak{F}}}\right)+\exp\left(-\frac{A_2k_nt}{M_n}\right)\right),\end{aligned} \tag{12}$$

with $v_{\mathfrak{F}} = \sup_{\varphi\in\mathfrak{F}}\mathrm{Var}\left(|\varphi(Y)|\right)$. From Lemma 12, $v_{\mathfrak{F}} \leq M_n^2k_nn^{-1}$, which shows that the first exponential term on the right-hand side of (12) is smaller than

$$\exp\left(-\frac{A_2k_nt^2}{M_n^2}\right). \tag{13}$$

We can now apply Proposition 5 (combined with Lemma 10) to this class of functions with $v = M_n^2k_nn^{-1}$ and $M_0 = M_n$. Hence,

$$\mathbb{E}\left[\sup_{\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}}\frac{1}{k_n}\left|\sum_{i=1}^{n}\varphi_{\boldsymbol{\theta}}(Y_i)\mathbf{1}_{\Phi(Y_i)\leq M_n}\varepsilon_i\right|\right] \leq \frac{A_6}{k_n}\sqrt{nv\mathfrak{s}_n} = A_6\frac{\mathfrak{s}_n^{1/2}}{k_n^{1/2}},$$

where $A_6' > 0$ and $\mathfrak{s}_n = \log(\sigma_n^{\alpha}K^{4(d+1)(d+2)}n/k_n)$ ($\alpha > 0$ being defined in Lemma 10). From Assumption 1, we see that $\mathfrak{s}_n = O(\log(k_n))$ (let us recall that K is necessarily less than n). Whence, if $\mathfrak{c}_1 = 2A_1A_6'$, for $t \geq \mathfrak{c}_1\left\{\log\left(k_n\right)\right\}^{1/2}k_n^{-1/2}$,

$$\mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geq t\right) \leq \mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geq A_1\left\{\mathbb{E}\left[\sup_{\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}}\frac{1}{k_n}\left|\sum_{i=1}^{n}\varphi_{\boldsymbol{\theta}}(Y_i)\mathbf{1}_{\Phi(Y_i)\leq M_n}\varepsilon_i\right|\right]+\frac{t}{2A_1}\right\}\right).$$

Equation (11) follows from (12) and (13) with $C_1 = A_2A_1^{-2}/4$ and $C_2 = A_2A_1^{-1}/2$. □

Theorem 7. *Let*

$$\overline{\mathcal{Z}}(M_n) = \sup_{\varphi_{\boldsymbol{\theta}}\in\mathfrak{F}}\left|\frac{1}{k_n}\sum_{i=1}^{n}\left(f(Y_i)\mathbf{1}_{\Phi(Y_i)>M_n}\right)-\mathbb{E}\left[\varphi_{\boldsymbol{\theta}}(Y_i)\mathbf{1}_{\Phi(Y_i)>M_n}\right]\right|.$$

If $k_n = O(a_1)$ with $a_1 > 0$ (Assumption 1), then there exists $\rho_0 > 0$ (Lemma 11) such that for $\beta a_1 \geq 10/\rho_0$, and $t \geq \mathfrak{c}_2k_n^{-1/2}$,

$$\mathbb{P}\left(\overline{\mathcal{Z}}(M_n) \geq t\right) \leq \frac{C_3}{k_n^{5/2}t^3}. \tag{14}$$

Proof. Let $\beta' = \beta a_2$. $\overline{\mathcal{Z}}(M_n)$ is upper-bounded by

$$\frac{1}{k_n}\sum_{i=1}^{n}\left\{\Phi(Y_i)\mathbf{1}_{\Phi(Y_i)\geq M_n}\mathbf{1}_{Y_i\geq u_{\min}}+\mathbb{E}\left[\Phi(Y)\mathbf{1}_{\Phi(Y)\geq M_n}\mathbf{1}_{Y\geq u_{\min}}\right]\right\}.$$

A bound for $E_{1,n} = \mathbb{E}\left[\Phi(Y)\mathbf{1}_{\Phi(Y)\geq M_n}\mathbf{1}_{Y\geq u_{\min}}\right]$ is obtained from Lemma 13, and $nE_{1,n}/k_n \leq \mathfrak{e}_1k_n^{-1/2}$ if $\beta' \geq 2/\rho_0$.

Next, from Markov inequality,

$$\begin{aligned}t^3\mathbb{P}\left(\frac{1}{k_n}\sum_{i=1}^{n}\Phi(Y_i)\mathbf{1}_{\Phi(Y_i)\geq M_n}\mathbf{1}_{Y_i\geq u_{\min}}\geq t\right) \quad &\leq \quad \frac{nE_{3,n}}{k_n^3}+\frac{n(n-1)E_{2,n}E_{1,n}}{k_n^3} \\ &\quad +\frac{n(n-1)(n-2)E_{1,n}^3}{k_n^3}.\end{aligned}$$

From Lemma 13, we get

$$\begin{aligned}\frac{nE_{3,n}}{k_n^3} &\leq \frac{\mathfrak{e}_3n^{-(\rho_0\beta'/4-1/2)}}{k_n^{5/2}}, \\ \frac{n(n-1)E_{2,n}E_{1,n}}{k_n^3} &\leq \frac{\mathfrak{e}_2\mathfrak{e}_1n^{-(\rho_0\beta'/2-3/2)}}{k_n^{5/2}}, \\ \frac{n(n-1)(n-2)E_{1,n}^3}{k_n^3} &\leq \frac{\mathfrak{e}_1^3n^{-(\rho_0\beta'/4-5/2)}}{k_n^{5/2}}.\end{aligned}$$

Each of these terms is bounded by $\max(\mathfrak{e}_3, \mathfrak{e}_z\mathfrak{e}_1, \mathfrak{e}_1^3)k_n^{-5/2}$ for $\beta' \geq 10/\rho_0$. Thus, for $t \geq 2\mathfrak{e}_1 k_n^{-1/2}$ and $\beta' \geq 10/\rho_0$,

$$\begin{aligned}
&\mathbb{P}\left(\overline{\mathcal{Z}}_n \geq t\right) \\
&\quad \leq \quad \mathbb{P}\left(\frac{1}{k_n}\sum_{i=1}^{n}\Phi(Y_i)\mathbf{1}_{\Phi(Y_i)\geq M_n}\mathbf{1}_{Y_i\geq u_{\min}} \geq \frac{t}{2}\right) + \mathbb{P}\left(\mathbb{E}\left[\Phi(Y)\mathbf{1}_{\Phi(Y)\geq M_n}\mathbf{1}_{Y\geq u_{\min}}\right] \geq \frac{t}{2}\right) \\
&\quad \leq \quad \frac{8\max(\mathfrak{e}_3, \mathfrak{e}_z\mathfrak{e}_1, \mathfrak{e}_1^3)}{t^3 k_n^{5/2}}
\end{aligned}$$

□

We now apply these results to deduce deviation bounds on the estimators $\widehat{\boldsymbol{\theta}}_\ell$ in the leaves of the tree.

Corollary 8. *Under the assumptions of Theorems 6 and 7 and Assumption 2, for $t \geq \mathfrak{c}_3(\log k_n)^{1/2}k_n^{-1/2}$,*

$$\begin{aligned}
\mathbb{P}\left(\sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leq u\leq u_{\max}}} \|\widehat{\boldsymbol{\theta}}_\ell^K - \boldsymbol{\theta}_\ell^{*K}\|_\infty \geq t\right) \quad &\leq \quad 2\left(\exp\left(-\frac{C_4 k_n t^2}{\beta^2(\log k_n)^2}\right) + \exp\left(-\frac{C_5 k_n t}{\beta \log k_n}\right)\right) \\
&\quad + \frac{C_6}{k_n^{5/2}t^3}.
\end{aligned}$$

Proof. For $1 \leq \ell \leq K$ and $u_{\min} \leq u \leq u_{\max}$, let $\boldsymbol{\theta} = (s, \gamma)^t$ and, for $\ell = 1, \ldots, K$, $\boldsymbol{\theta}_\ell^{*K} = (s_\ell^{*K}(u), \gamma_\ell^{*K}(u))^t$, and let

$$\nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}, u) = \mathbb{E}\left[\begin{pmatrix} \partial_\sigma \phi(Y-u, \boldsymbol{\theta}) \\ \partial_\gamma \phi(Y-u, \boldsymbol{\theta}) \end{pmatrix} \mathbf{1}_{Y\geq u}\mathbf{1}_{\mathbf{X}\in\mathcal{T}_\ell}\right].$$

From Taylor series,

$$\nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}, u) = \mathbb{E}\left[H^\ell_{(\tilde{\sigma}_1,\gamma_1),(\sigma_1,\tilde{\gamma}_1),(\tilde{\sigma}_2,\gamma_2),(\sigma_2,\tilde{\gamma}_2)}(Y-u)\mathbf{1}_{\mathbf{X}\in\mathcal{T}_\ell}\right](\boldsymbol{\theta} - \boldsymbol{\theta}_\ell^{*K})^t,$$

for some parameters $\tilde{\sigma}_j$ (resp. $\tilde{\gamma}_j$) between σ and $\sigma_\ell^{*K}(u)$ (resp. γ and $\gamma_\ell^{*K}(u)$). From Assumption 2, we get, for all $\ell = 1, \ldots, K$,

$$\frac{n}{k_n}\|\nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}, u)\|_\infty \geq \mathfrak{C}_1\|\boldsymbol{\theta} - \boldsymbol{\theta}_\ell^{*K}(u)\|_\infty.$$

Hence, for all $\ell = 1, \ldots, K$,

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\theta}}_\ell^K - \boldsymbol{\theta}_\ell^{*K}\|_\infty \geq t\right) \leq \mathbb{P}\left(\frac{n}{k_n}\|\nabla_{\boldsymbol{\theta}} L^\ell(\widehat{\boldsymbol{\theta}}^K, u)\|_\infty \geq \mathfrak{C}_1 t\right).$$

Since for all $\ell = 1, \ldots, K$, $\nabla_{\boldsymbol{\theta}} L_n^\ell(\widehat{\boldsymbol{\theta}}^K) = 0$, $\mathcal{W}_1^\ell(\widehat{\boldsymbol{\theta}}^K(u), u) = -\frac{n}{k_n}\nabla_{\boldsymbol{\theta}} L^\ell(\widehat{\boldsymbol{\theta}}^K, u)$. Hence,

$$\mathbb{P}\left(\sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leq u\leq u_{\max}}} \|\widehat{\boldsymbol{\theta}}_\ell^K - \boldsymbol{\theta}_l^{*K}(u)\|_\infty \geq t\right) \leq \mathbb{P}\left(\sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leq u\leq u_{\max}}} \|\mathcal{W}_1^\ell(\widehat{\boldsymbol{\theta}}^K(u), u)\|_\infty \geq \mathfrak{C}_1 t\right),$$

and the right-hand side is bounded by

$$\mathbb{P}\left(\overline{\mathcal{Z}}(M_n) \geq \frac{\mathfrak{C}_1 t}{2}\right) + \mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geq \frac{\mathfrak{C}_1 t}{2}\right).$$

The result follows from Theorem 6 and 7. □

A.3 Proof of Theorem 1

The proof of the first part of Theorem 1 then consists in gathering the results on the leaves obtained in Corollary 8. Let $u_{\min} \leq u \leq u_{\max}$,

$$\|\widehat{T}_K - T_K^*\|_2^2 \leq \sum_{\ell=1}^{K} \|\widehat{\boldsymbol{\theta}}_\ell^K - \boldsymbol{\theta}_\ell^{*K}\|_\infty^2 \leq K \sup_{\ell=1,\ldots,K} \|\widehat{\boldsymbol{\theta}}_\ell^K - \boldsymbol{\theta}_\ell^{*K}\|_\infty^2.$$

Hence

$$\begin{aligned}
&\mathbb{P}\left(\sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K - T_K^*\|_2^2 \geq t \right) \\
&\leq \mathbb{P}\left(\sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leq u\leq u_{\max}}} \|\widehat{\boldsymbol{\theta}}_\ell^K - \boldsymbol{\theta}_\ell^{*K}\|_\infty \geq t^{1/2} K^{-1/2} \right).
\end{aligned}$$

The results follows from Corollary 8, and from the assumption on $K \leq K_{\max} = O(k_n^3)$ (Assumption 1).

To prove the second part of Theorem 1, write

$$\mathbb{E}\left[\sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K - T_K^*\|_2^2 \right] = \int_0^\infty \mathbb{P}(\sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K - T_K^*\|_2^2 \geq t) dt.$$

Let $t_n = c_1 K (\log k_n) k_n^{-1}$, then

$$\begin{aligned}
&\int_0^\infty \mathbb{P}(\sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K - T_K^*\|_2^2 \geq t) dt \\
&\leq \quad t_n + \int_{t_n}^\infty \mathbb{P}(\sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K - T_K^*\|_2^2 \geq t) dt.
\end{aligned}$$

We now use Theorem 1 to bound the integral on the right-hand side. Since $\int_0^\infty \exp(-at)dt = \frac{1}{a}$, $\int_0^\infty \exp(-a^{1/2}t^{1/2})dt = \frac{2}{a}$, and $\int_1^\infty t^{-3/2}dt = 2$, we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{u_{\min}\leq u\leq u_{\max}} \|\widehat{T}_K - T_K^*\|_2^2 \right] &\leq t_n + \frac{2K\beta^2(\log k_n)^2}{\mathcal{C}_1 k_n} + \frac{4K\beta^2(\log k_n)^2}{\mathcal{C}_2^2 k_n} + \frac{2\mathcal{C}_3 K}{k_n^{5/2}} \\
&\leq \frac{c_1 K \log k_n}{k_n} + \frac{2K\beta^2(\log k_n)^2}{\mathcal{C}_1 k_n} \\
&\quad + \frac{4K\beta^2(\log k_n)^2}{\mathcal{C}_2^2 k_n} + \frac{2\mathcal{C}_3 K}{k_n^{5/2}} \\
&\leq \frac{\mathcal{C}_4 K (\log k_n)^2}{k_n}.
\end{aligned}$$

A.4 Proof of Proposition 2

For all $\mathbf{x}$,

$$\|\boldsymbol{\theta}^*(\mathbf{x}) - \boldsymbol{\theta}_0(\mathbf{x})\|_\infty = \| \sum_{\ell=1}^{K_{\max}} (\boldsymbol{\theta}_\ell^* - \boldsymbol{\theta}_0(\mathbf{x}))\, \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} \|_\infty \leq \sum_{\ell=1}^{K_{\max}} \|\boldsymbol{\theta}_\ell^* - \boldsymbol{\theta}_0(\mathbf{x})\|_\infty \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}\,.$$

Now, from Taylor series, for $\ell = 1, \ldots, K$, conditionally on $\mathbf{X} \in \mathcal{T}_\ell$,

$$\nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}_0(\mathbf{X}), u) \quad = \quad \mathbb{E}\left[H^\ell_{(\tilde{\sigma}_1,\tilde{\gamma}_1),(\sigma_1,\tilde{\gamma}_1),(\tilde{\sigma}_2,\gamma_2),(\sigma_2,\tilde{\gamma}_2)}(Y-u) \mid \mathbf{X} \in \mathcal{T}_\ell \right] (\boldsymbol{\theta}_0(\mathbf{X}) - \boldsymbol{\theta}_\ell^*)^t,$$

for some parameters $\tilde{\sigma}_j$ (resp. $\tilde{\gamma}_j$) between $\sigma_0(\mathbf{X})$ and $\sigma_\ell^{*K}(u)$ (resp. $\gamma_0(\mathbf{X})$ and $\gamma_\ell^{*K}(u)$).

Thus, under Assumption 2,

$$\begin{aligned}
&\|\boldsymbol{\theta}_0(\mathbf{X}) - \boldsymbol{\theta}^*_\ell\|_\infty \\
&\quad\leq \quad \frac{1}{\mathfrak{C}_1}\|\nabla_{\boldsymbol{\theta}} L^\ell(\boldsymbol{\theta}_0(\mathbf{X}), u)\|_\infty \\
&\quad\leq \quad \frac{1}{\mathfrak{C}_1}\frac{k_n}{n}\max\left(|\mathbb{E}\left[\partial_\sigma\phi(Z, \boldsymbol{\theta}_0(\mathbf{X}))\mid \mathbf{X}\in\mathcal{T}_\ell\right]|, |\mathbb{E}\left[\partial_\gamma\phi(Z, \boldsymbol{\theta}_0(\mathbf{X}))\mid \mathbf{X}\in\mathcal{T}_\ell\right]|\right),
\end{aligned}$$

where Z is a random variable distributed according to the distribution F_u defined in Section 2.1 with $\sigma_0(\mathbf{X}) = u\gamma_0(\mathbf{X})$ and with

$$\begin{aligned}
\mathbb{E}\left[\partial_\sigma\phi(Z, \boldsymbol{\theta}_0(\mathbf{X}))\mid \mathbf{X}\in\mathcal{T}_\ell\right] &= -\frac{1}{u\gamma_0(\mathbf{X})} + \frac{1}{u^2\gamma_0(\mathbf{X})}\left(1+\frac{1}{\gamma_0(\mathbf{X})}\right)\mathbb{E}\left[\frac{Z}{1+Z/u}\mid \mathbf{X}\in\mathcal{T}_\ell\right] \\
\mathbb{E}\left[\partial_\gamma\phi(Z, \boldsymbol{\theta}_0(\mathbf{X}))\mid \mathbf{X}\in\mathcal{T}_\ell\right] &= -\frac{1}{\gamma_0(\mathbf{X})^2}\mathbb{E}\left[\log(1+Z/u)\mid \mathbf{X}\in\mathcal{T}_\ell\right] \\
&\quad + \frac{1}{u\gamma_0(\mathbf{x})}\left(1+\frac{1}{\gamma_0(\mathbf{X})}\right)\mathbb{E}\left[\frac{Z}{1+Z/u}\mid \mathbf{X}\in\mathcal{T}_\ell\right].
\end{aligned}$$

Under Assumption 3, we have

$$\overline{F}_u(z) = \left(1+\frac{z}{u}\right)^{-1/\gamma_0(\mathbf{X})}\left\{1 + c\psi(u)\int_1^{1+z/u} v^{\rho-1}\mathrm{d}v + o(\psi(u))\right\}.$$

$$\begin{aligned}
\mathbb{E}\left[\frac{Z}{1+Z/u}\mid \mathbf{X}\in\mathcal{T}_\ell\right] &= \int_0^u \overline{F}_u\left(\frac{t}{1-t/u}\right)\mathrm{d}t \\
&= \frac{u}{1+1/\gamma_0(\mathbf{X})}\left(1+\frac{c\psi(u)}{1+1/\gamma_0(\mathbf{X})-\rho}+o(\psi(u))\right) \\
&\leq u\left(1+c\gamma_0(\mathbf{X})\psi(u)+o(\psi(u))\right)
\end{aligned}$$

and then

$$\begin{aligned}
\mathbb{E}\left[\log(1+Z/u)\mid \mathbf{X}\in\mathcal{T}_\ell\right] &= \int_0^u \mathbb{P}\left[Z\geq u(\mathrm{e}^t-1)\mid \mathbf{X}\in\mathcal{T}_\ell\right]\mathrm{d}t \\
&= \gamma_0(\mathbf{X})\left(1+\frac{c\psi(u)}{1/\gamma_0(\mathbf{X})-\rho}+o(\psi(u))\right) \\
&\leq \gamma_0(\mathbf{X})\left(1+c\gamma_0(\mathbf{X})\psi(\mathbf{X})(u)+o(\psi(u))\right).
\end{aligned}$$

Consequently,

$$|\mathbb{E}\left[\partial_\sigma\phi(Z, \boldsymbol{\theta}_0(\mathbf{X}))\mid \mathbf{X}\in\mathcal{T}_\ell\right]| \leq \frac{1}{\gamma_{\min}}\left(1+\frac{1}{u}\left(1+\frac{1}{\gamma_{\min}}\right)\right)\left(1+c\gamma_0(\mathbf{X})\psi(u)+o(\psi(u))\right)$$

and

$$|\mathbb{E}\left[\partial_\gamma\phi(Z, \boldsymbol{\theta}_0(\mathbf{X}))\mid \mathbf{X}\in\mathcal{T}_\ell\right]| \leq \frac{1}{\gamma_{\min}}\left(1+\frac{1}{\gamma_{\min}}+\frac{\gamma_{\max}}{\gamma_{\min}}\right)\left(1+c\gamma_0(\mathbf{X})\psi(u)+o(\psi(u))\right).$$

Hence, conditionally on $\mathbf{X}\in\mathcal{T}_\ell$,

$$\|\boldsymbol{\theta}_0(\mathbf{X}) - \boldsymbol{\theta}^*_\ell\|_\infty \quad\leq\quad \mathfrak{C}_2(u)\frac{k_n}{n}\left(1+c\gamma_{\max}\psi(u)+o(\psi(u))\right),$$

where $\mathfrak{C}_2(u) = \frac{1}{\mathfrak{C}_1}\frac{1}{\gamma_{\min}}\max\left(1+\frac{1}{u}+\frac{1}{u\gamma_{\min}}, 1+\frac{1}{\gamma_{\min}}+\frac{\gamma_{\max}}{\gamma_{\min}}\right)$.

Finally, for all $\mathbf{x}$,

$$\begin{aligned}
\|\boldsymbol{\theta}^*(\mathbf{x}) - \boldsymbol{\theta}_0(\mathbf{x})\|_\infty &\leq \sum_{\ell=1}^{K_{\max}}\|\boldsymbol{\theta}^*_\ell - \boldsymbol{\theta}_0(\mathbf{x})\|_\infty \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} \\
&\leq \mathfrak{C}_2(u)\frac{k_n}{n}\left(1+c\gamma_{\max}\psi(u)+o(\psi(u))\right)\sum_{\ell=1}^{K_{\max}}\mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} \\
&\leq \mathfrak{C}_2(u)\frac{k_n}{n}\left(1+c\gamma_{\max}\psi(u)+o(\psi(u))\right).
\end{aligned}$$

A.5 Proof of Theorem 3

First, let us introduce some notations that are needed in the proof.

Define the log-likelihood $L_n(T_K, u)$ associated with a tree T_K with K leaves $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$ and with parameters $\boldsymbol{\theta}(u) = \left(\boldsymbol{\theta}_\ell^K(u)\right)_{\ell=1,\dots,K}$

$$L_n(T_K, u) = \sum_{\ell=1}^{K} L_n^\ell(\boldsymbol{\theta}_\ell^K, u) = \frac{1}{k_n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \phi(Y_i - u, \boldsymbol{\theta}_\ell^K) \mathbf{1}_{Y_i > u} \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell} \,,$$

and $L(T_K, u) = \mathbb{E}[L_n(T_K, u)]$. Finally, for two trees T and T', $\Delta L_n(T, T') = L_n(T, u) - L_n(T', u)$ and similarly, $\Delta L(T, S) = L(T, u) - L(T', u)$.

The following lemma will be needed to prove Theorem 3.

Lemma 9. *Let $\mathfrak{D} = \inf_u \inf_{K<K^*} \Delta L(T^*, T_K^*)$ and $u \in [u_{\min}, u_{\max}]$ fixed. Suppose that there exists a constant $c_2 > 0$ such that the penalization constant λ satisfies*

$$c_2 \{\log k_n\}^{1/2} k_n^{-1/2} \leq \lambda \leq (\mathfrak{D} - 2c_2 \{\log(k_n)\}^{1/2} k_n^{-1/2}) k_n^{-1},$$

then, under Assumptions 1 and 2, for $K > K^*$,

$$\begin{aligned} \mathbb{P}(\widehat{K} = K) \quad \leq \quad & 2 \left(\exp\left(- \frac{C_1 k_n \lambda^2 (K - K^*)^2}{\beta^2 (\log k_n)^2} \right) + \exp\left(- \frac{C_2 k_n \lambda (K - K^*))}{\beta \log k_n} \right) \right) \\ & + \frac{C_3}{k_n^{5/2} \lambda^3 (K - K^*)^3}, \end{aligned}$$

and, for $K < K^*$,

$$\begin{aligned} \mathbb{P}(\widehat{K} = K) \quad \leq \quad & 4 \exp\left(- \frac{C_1 k_n \{\mathfrak{D} - \lambda(K^* - K)\}^2}{\beta^2 (\log k_n)^2} \right) \\ & + 4 \exp\left(- \frac{C_2 k_n \{\mathfrak{D} - \lambda(K^* - K)\}}{\beta \log k_n} \right) \\ & + \frac{2 C_3}{k_n^{5/2} \{\mathfrak{D} - \lambda(K^* - K)\}^3}. \end{aligned}$$

Proof. Let $u \in [u_{\min}, u_{\max}]$ fixed. If $\widehat{K} = K$, this means that

$$\Delta L_n(T_K, T_{K^*}) := L_n(T_K, u) - L_n(T_{K^*}, u) > \lambda(K - K^*).$$

Decompose

$$\begin{aligned} \Delta L_n(T_K, T_{K_0}) \quad = \quad & \{L_n(T_K, u) - L_n(T_K^*, u)\} + \{L_n(T_K^*, u) - L_n(T^*, u)\} \\ & + \{L_n(T^*, u) - L_n(T_{K^*}, u)\}. \end{aligned}$$

Since $L_n(T^*, u) - L_n(T_{K^*}, u) < 0$,

$$\Delta L_n(T_K, T_{K^*}) \leq \{L_n(T_K, u) - L_n(T_K^*, u)\} + \{L_n(T_K^*, u) - L_n(T^*, u)\}.$$

For $K > K^*$, $T_K^* = T^*$, hence,

$$\begin{aligned} \mathbb{P}(\widehat{K} = K) \quad \leq \quad & \mathbb{P}\left(\Delta L_n(T_K, T_K^*) > \lambda(K - K^*)\right) \\ \leq \quad & \mathbb{P}\left(|\Delta L_n(T_K, T_K^*) - \Delta L(T_K, T_K^*)| > \lambda(K - K^*)\right). \end{aligned}$$

For $K > K^*$, a bound is then obtained from Theorems 6 and 7 if $\lambda(K - K^*) \geq c_1 \{\log(k_n)\}^{1/2} k_n^{-1/2}$, that is $\lambda \geq c_1 \{\log k_n\}^{1/2} k_n^{-1/2}$.

Now, for $K < K^*$,

$$\begin{aligned} \Delta L_n(T_K^*, T^*) \quad \leq \quad & |\Delta L_n(T_K^*, T^*) - \Delta L(T_K^*, T^*)| + \Delta L(T_K^*, T^*) \\ \leq \quad & |\Delta L_n(T^*, T_K^*) - \Delta L(T^*, T_K^*)| - \mathfrak{D}(K^*, K). \end{aligned}$$

where $\mathfrak{D} = \inf_{K<K^*, u\in[u_{\min},u_{\max}]} \mathfrak{D}(K^*, K)$, Hence,

$$\begin{aligned}
&\mathbb{P}(\widehat{K} = K) \\
&\leq \mathbb{P}\left(\Delta L_n(T_K, T_K^*) \geq \frac{\mathfrak{D} - \lambda(K^* - K)}{2}\right) \\
&\quad + \mathbb{P}\left(|\Delta L_n(T^*, T_K^*) - \Delta L(T^*, T_K^*)| \geq \frac{\mathfrak{D} - \lambda(K^* - K)}{2}\right) \\
&\leq \mathbb{P}\left(|\Delta L_n(T_K, T_K^*) - \Delta L(T_K, T_K^*)| \geq \frac{\mathfrak{D} - \lambda(K^* - K)}{2}\right) \\
&\quad + \mathbb{P}\left(|\Delta L_n(T^*, T_K^*) - \Delta L(T^*, T_K^*)| \geq \frac{\mathfrak{D} - \lambda(K^* - K)}{2}\right).
\end{aligned}$$

These two probabilities can be bounded using Theorems 6 and 7 provided that, for all $K < K^*$,

$$\frac{\mathfrak{D} - \lambda(K^* - K)}{2} \geq \mathfrak{c}_1\{\log(k_n)\}^{1/2} k_n^{-1/2},$$

that is,

$$\lambda \leq \mathfrak{D} - 2\mathfrak{c}_1\{\log(k_n)\}^{1/2} k_n^{-1/2}.$$

□

We are now ready to prove Theorem 3. Let $u \in [u_{\min}, u_{\max}]$ fixed.

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{T} - T^*\|_2^2\right] &= \sum_{K=1}^{K_{\max}} \mathbb{E}\left[\|T_K - T^*\|_2^2 \mathbf{1}_{\widehat{K}=K}\right] \\
&\leq \mathbb{E}\left[\|T_{K^*} - T^*\|_2^2\right] + \sum_{K=1, K\neq K^*}^{K_{\max}} K\mathbb{P}(\widehat{K} = K) \\
&\quad + \sum_{K=1, K\neq K^*}^{K_{\max}} \mathbb{E}\left[\|T_K - T^*\|_2^2 \mathbf{1}_{\|T_K - T^*\|_2^2 > K} \mathbf{1}_{\widehat{K}=K}\right] \\
&\leq \mathbb{E}\left[\|T_{K^*} - T^*\|_2^2\right] + \sum_{K=1}^{K^*-1} K\mathbb{P}(\widehat{K} = K) \\
&\quad + \sum_{K=K^*+1}^{K_{\max}} K\mathbb{P}(\widehat{K} = K) \\
&\quad + 2\sum_{K=1, K\neq K^*}^{K_{\max}} \mathbb{E}\left[\|T_K - T_K^*\|_2^2 \mathbf{1}_{\|T_K - T^*\|_2^2 > K}\right] \\
&\quad + 2\sum_{K=1, K\neq K^*}^{K_{\max}} \mathbb{P}(\widehat{K} = K)\|T^* - T_K^*\|_2^2.
\end{aligned}$$

Firstly, from Theorem 1,

$$\begin{aligned}
&\mathbb{E}\left[\|T_K - T_K^*\|_2^2 \mathbf{1}_{\|T_K - T^*\|_2^2 > K}\right] \\
&= K\mathbb{P}\left(\|T_K - T_K^*\|_2^2 > K\right) + \int_K^\infty \mathbb{P}\left(\|T_K - T_K^*\|_2^2 > t\right) \mathrm{d}t \\
&\leq 2K\left(1 + \frac{\beta^2(\log k_n)^2}{\mathcal{C}_1 k_n}\right) \exp\left(-\frac{\mathcal{C}_1 k_n}{\beta^2(\log k_n)^2}\right) \\
&\quad + 2K\left(1 + \frac{2\beta(\log k_n)}{\mathcal{C}_2 k_n} + \frac{2\beta^2(\log k_n)^2}{\mathcal{C}_2^2 k_n^2}\right) \exp\left(-\frac{\mathcal{C}_2 k_n}{\beta(\log k_n)}\right) + \frac{2\mathcal{C}_3 K^{1/2}}{k_n^{5/2}}.
\end{aligned}$$

Secondly, recall that

$$\|T_K^* - T^*\|_2^2 = \int \|\boldsymbol{\theta}^{*K}(\mathbf{x}) - \boldsymbol{\theta}^*(\mathbf{x})\|_\infty^2 \mathrm{d}P_{\mathbf{X}}(\mathbf{x}) \leq K_{\max} \sum_{\ell=1}^{K_{\max}} \mu(\mathcal{T}_\ell) \|\boldsymbol{\theta}_\ell^{*K} - \boldsymbol{\theta}_\ell^*\|_\infty^2 ,$$

where $\mu(\mathcal{T}_\ell) = \mathbb{P}(\mathbf{X} \in \mathcal{T}_\ell)$. Following the same idea as in the proof of Proposition 2, from Taylor series , under Assumptions 2 and 3,

$$\|\boldsymbol{\theta}_\ell^{*K} - \boldsymbol{\theta}_\ell^*\|_\infty^2 \quad \leq \quad \mathfrak{C}_2^2(u) \frac{k_n^2}{n^2} \left(1 + c\gamma_{\max}\psi(u) + o(\psi(u))\right)^2 .$$

Hence,

$$\begin{aligned} \|T_K^* - T^*\|_2^2 &\leq \mathfrak{C}_2^2(u) \frac{k_n^2}{n^2} (1 + c\gamma_{\max}\psi(u) + o(\psi(u)))^2 \sum_{\ell=1}^{K_{\max}} \mathbf{1}_{\mathbf{X}\in\mathcal{T}_\ell} \\ &\leq \mathfrak{C}_3(u) \frac{k_n^2}{n^2} . \end{aligned}$$

Finally,

$$\mathbb{E}\left[\|\widehat{T} - T^*\|_2^2\right] \quad \leq \quad \frac{\mathcal{C}_5 K^* (\log k_n)^2}{k_n},$$

for some constant $\mathcal{C}_5$. .

B Covering numbers

Lemma 10. *Following the notations of the proof of Theorem 6, the class of functions $\mathfrak{F}$ satisfies*

$$\mathcal{N}_\Phi(\varepsilon, \mathfrak{F}) \leq \frac{\mathfrak{C}_4 K^{4(d+1)(d+2)} \|\Phi\|_2^{\alpha_1} \sigma_n^\alpha}{\varepsilon^\alpha},$$

for some constants $\mathfrak{C}_4 > 0$ and $\alpha > 0$ (not depending on n nor K).

Proof. Let

$$\begin{aligned} g_{\boldsymbol{\theta}}(z) &= -\frac{1}{\sigma} + \left(\frac{1}{\gamma} + 1\right) \frac{\gamma z}{\sigma^2(1 + \frac{z\gamma}{\sigma})}, \\ h_{\boldsymbol{\theta}}(z) &= -\frac{1}{\gamma^2} \log\left(1 + \frac{z\gamma}{\sigma}\right) + \frac{\left(\frac{1}{\gamma} + 1\right) z}{\sigma + z\gamma}, \end{aligned}$$

for $z > 0$. For $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ in $\mathcal{S} \times \Gamma$, we have (from a straightforward Taylor expansion),

$$|g_{\boldsymbol{\theta}}(y - u) - g_{\boldsymbol{\theta}'}(y - u)| \leq C|\gamma - \gamma'| + C'|\sigma - \sigma'|,$$

for some constants C and C'. More precisely, one can take

$$\begin{aligned} C &= \frac{6}{\gamma_{\min}^2 \sigma_{\min}}, \\ C' &= \frac{1}{\sigma_{\min}^2} \left(1 + 3\left\{1 + \frac{1}{\gamma_{\min}}\right\}\right). \end{aligned}$$

Next, observe that

$$|g_{\boldsymbol{\theta}'}(y - u) - g_{\boldsymbol{\theta}'}(y - u')| \leq C''|u - u'|,$$

where $C'' = 4\gamma_{\max}^2/[\gamma_{\min}\sigma^3]$. Which leads to

$$|g_{\boldsymbol{\theta}}(y - u) - g_{\boldsymbol{\theta}'}(y - u')| \leq C_g \max(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty, |u - u'|),$$

for some constant $C_g > 0$. Similarly,

$$|h_{\boldsymbol{\theta}}(y-u) - h_{\boldsymbol{\theta}'}(y-u)| \leq C_1(4+\log(1+wy))|\gamma-\gamma'| + C_2|\sigma-\sigma'|,$$

Next,

$$|h_{\boldsymbol{\theta}'}(y-u) - h_{\boldsymbol{\theta}'}(y-u')| \leq C_7|u-u'|,$$

where $C_7 = 5/(\gamma_{\min}\sigma_{\min})$, leading to, for some $C_h > 0$,

$$|h_{\boldsymbol{\theta}}(y-u) - h_{\boldsymbol{\theta}'}(y-u')| \leq C_h \max(\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_\infty, |u-u'|).$$

On the other hand,

$$|\phi(y-u,\boldsymbol{\theta}) - \phi(y-u,\boldsymbol{\theta}')| \leq \frac{1}{\gamma_{\min}^2}(2+\log(1+wy))|\gamma-\gamma'| + \frac{3}{\gamma_{\min}\sigma_{\min}}|\sigma-\sigma'|,$$

and

$$|\phi(y-u,\boldsymbol{\theta}') - \phi(y-u',\boldsymbol{\theta}')| \leq \frac{1}{\sigma_{\min}}|u-u'|.$$

Define $\mathfrak{F}_1 = \{g_{\boldsymbol{\theta}}(\cdot - u) : \boldsymbol{\theta} \in \mathcal{S}\times\Gamma, u \in [u_{\min}, u_{\max}]\}$, $\mathfrak{F}_2 = \{h_{\boldsymbol{\theta}}(\cdot - u) : \boldsymbol{\theta} \in \mathcal{S}\times\Gamma, u \in [u_{\min}, u_{\max}]\}$, and $\mathfrak{F}_3 = \{\phi(\cdot - u, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{S}\times\Gamma, u \in [u_{\min}, u_{\max}]\}$. From [van der Vaart, 1998, Example 19.7], we get, for $i = 1, ..., 3$,

$$N(\varepsilon, \mathfrak{F}_i) \leq \varphi_i \|\Phi\|_2^{\alpha_1} \sigma_n^{\alpha_1} \varepsilon^{-\alpha_1},$$

for some $\alpha > 0$ and constants φ_i.

On the other hand, let

$$\mathfrak{F}_4 = \{\mathbf{x} \mapsto \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} : \ell = 1, \ldots, K\},$$

and

$$\mathfrak{F}_5 = \{y \mapsto \mathbf{1}_{y>u} : u \in \mathcal{U}\}.$$

From Lemma 4 in [Lopez et al., 2016], we have $N(\varepsilon, \mathfrak{F}_4) \leq m^k K^{\alpha_2} \varepsilon^{-\alpha_2}$, where $\alpha_2 = 4(d+1)(d+2)$, and where k is the number of discrete components taking at most m modalities. On the other hand, from Example 19.6 in [van der Vaart, 1998], $N(\varepsilon, \mathfrak{F}_5) \leq 2\varepsilon^{-2}$.

From [Einmahl et al., 2005, Lemma A.1], we get, for $i = 1, \ldots, 3$,

$$N(\varepsilon, \mathfrak{F}_i\mathfrak{F}_4\mathfrak{F}_5) \leq \frac{4m^k K^{\alpha_2} \max(C_g, C_h) \|\Phi\|_2^{\alpha_1} \sigma_n^{\alpha_1}}{\varepsilon^{\alpha_1+\alpha_2+\alpha_3}}.$$

Multiplying $\mathfrak{F}_i\mathfrak{F}_4\mathfrak{F}_5$ by a single indicator function $\mathbf{1}_{\Phi(Y_i)\leq M_n}$ does not change the covering number, and the result follows. □

C Technical Lemmas

Lemma 11. *1. The derivatives of the functions $y \to \phi(y-u,\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ are uniformly bounded by*

$$\Phi(y) = C(1+\log(1+wy)),$$

where C is a constant (not depending on n), and $w = \gamma_{\max}/\sigma_{\min}$.

2. There exists a certain $\rho_0 > 0$ such that

$$m_{\rho_0} := \mathbb{E}\left[\exp(\rho_0\Phi(Y))\right] < \infty.$$

Proof. To proof point 1, it is sufficient to derive the GP likelihood and see that they can be upper-bounded by Φ.

Now, for point 2, note that for all $\mathbf{x}$, $\gamma(\mathbf{x}) \geq \gamma_{\min} > 0$, Y is heavy-tailed random variable, then $\log(Y)$, and thus $\Phi(Y)$, is a light-tailed random variable. Thus $\Phi(Y)$ has finite exponential moments. □

Lemma 12. *With $v_{\mathfrak{F}}$ defined in Proposition 4,*

$$v_{\mathfrak{F}} \leq \frac{M_n^2 k_n}{n}.$$

Proof. We have

$$\begin{aligned} v_{\mathfrak{F}} &\leq \mathbb{E}\left[\Phi(Y)^2 \mathbf{1}_{Y \geq u_{\min}} \mathbf{1}_{\Phi(Y) \leq M_n}\right] \\ &\leq M_n^2 \mathbb{P}(Y \geq u_{\min}) = \frac{M_n^2 k_n}{n}. \end{aligned}$$

□

Lemma 13. *Define, for* $j = 1, 2, 3$,

$$E_{j,n} = \mathbb{E}\left[\Phi(Y)^j \mathbf{1}_{\Phi(Y) \geq M_n} \mathbf{1}_{Y \geq u_{\min}}\right].$$

Under the assumptions of Theorem 7,

$$E_{j,n} \leq \frac{\mathfrak{e}_j k_n^{1/2}}{n^{1/2} n^{\rho_0 \beta a_2/4}}.$$

Proof. Applying twice Cauchy-Schwarz inequality leads to

$$E_{j,n} \leq \mathbb{P}(Y \geq u_{\min})^{1/2} \mathbb{E}[\Phi(Y)^{2j} \mathbf{1}_{\Phi(Y) \geq M_n}]^{1/2} \leq \frac{k_n^{1/2}}{n^{1/2}} \mathbb{E}[\Phi(Y)^{4j}]^{1/4} \mathbb{P}(\Phi(Y) \geq M_n)^{1/4}.$$

Next, from Chernoff inequality,

$$\mathbb{P}(\Phi(Y) \geq M_n) \leq \exp(-\rho_0 M_n) \mathbb{E}[\exp(\rho_0 \Phi(Y))] \leq \frac{m_{\rho_0}}{n^{\rho_0 \beta a_2}}.$$

□

R codes: Tthe `R` codes are publicly available at `https://github.com/antoine-heranval/Generalized-Pareto-Regression-T`

Ethical Approval and Consent to participate All the authors approve and consent to participate.

Consent for publication All the authors consent for publication.

Human and Animal Ethics Not applicable.

Availability of supporting data Since the data were provided by a private partnership with the Mission Risques Naturels, the data are not publicly available.

Competing interests The authors have no competing interests.

Funding Not applicable.

Authors' contributions All the authors wrote the main manuscript text, and the supplementary material All the authors prepared the all figures and tables All the authors reviewed the manuscript.

Acknowledgments The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project).

References

Catastrophe naturelle, assurance et prévention. Technical report, Mission Risques Naturels, 2016. URL `https://www.mrn.asso.fr/wp-content/uploads/2019/03/190603_mrn_guidecatnat_15x21cm_ecran.pdf`.

D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974. doi: https://doi.org/10.1080/00401706.1974.10489157.

M. Allouche, S. Girard, and E. Gobet. Estimation of extreme quantiles from heavy-tailed distributions with neural networks. working paper or preprint, 2022. URL `https://hal.science/hal-03751980`.

A. A. Balkema and L. de Haan. Residual life time at great age. *The Annals of probability*, pages 792–804, 1974. doi: https://doi.org/10.1214/aop/1176996548.

A. M. Barlow, E. Mackay, E. Eastoe, and P. Jonathan. A penalised piecewise-linear model for non-stationary extreme value analysis of peaks over threshold. *Ocean Engineering*, 267:113265, 2023.

J. Beirlant and Y. Goegebeur. Local polynomial maximum likelihood estimation for Pareto-type distributions. *Journal of Multivariate Analysis*, 89(1):97–118, 2004. doi: https://doi.org/10.1016/S0047-259X(03)00125-8.

J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of extremes: Theory and Applications*. John Wiley & Sons, 2004. ISBN 978-0-471-97647-9.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

J. Carreau and M. Vrac. Stochastic downscaling of precipitation with neural network conditional mixture models. *Water Resources Research*, 47(10), 2011.

A. Charpentier, L. Barry, and M. R. James. Insurance against natural catastrophes: balancing actuarial fairness and social solidarity. *The Geneva Papers on Risk and Insurance - Issues and Practice*, May 2021. ISSN 1018-5895, 1468-0440. doi: https://doi.org/10.1057/s41288-021-00233-7.

P. Chaudhuri. Asymptotic consistency of median regression trees. *Journal of statistical planning and inference*, 91(2):229–238, 2000. doi: https://doi.org/10.1016/S0378-3758(00)00180-4.

P. Chaudhuri and W.-Y. Loh. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, pages 561–576, 2002.

V. Chavez-Demoulin, P. Embrechts, and M. Hofert. An extreme value approach for modeling operational risk losses depending on covariates. *Journal of Risk and Insurance*, 83(3):735–776, 2015. doi: https://doi.org/10.1111/jori.12059.

V. Chernozhukov. Extremal quantile regression. *The Annals of Statistics*, 33(2):806–839, 2005.

S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer London, 2001.

A. C. Davison and R. L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):393–425, 1990. doi: https://doi.org/10.1111/j.2517-6161.1990.tb01796.x.

G. De'ath and K. E. Fabricius. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, 81(11):3178–3192, 2000. doi: https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

U. Einmahl, D. M. Mason, et al. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005. doi: https://doi.org/10.1214/009053605000000129.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

S. Farkas, O. Lopez, and M. Thomas. Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, 98:92–105, 2021. doi: https://doi.org/10.1016/j.insmatheco.2021.02.009.

L. Gardes and G. Stupfler. An integrated functional weissman estimator for conditional extreme quantiles. *REVSTAT-Statistical Journal*, 17(1):109–144, 2019.

S. Gey and E. Nedelec. Model selection for cart regression trees. *IEEE Transactions on Information Theory*, 51(2):658–670, 2005. doi: https://doi.org/10.1109/TIT.2004.840903.

N. Gnecco, E. M. Terefe, and S. Engelke. Extremal random forests. *arXiv preprint arXiv:2201.12865*, 2022.

C. González, J. Mira-McWilliams, and I. Juárez. Important variable assessment and electricity price forecasting based on regression tree models: Classification and regression trees, Bagging and Random Forests. *IET Generation, Transmission Distribution*, 9(11):1120–1128, 2015. doi: https://doi.org/10.1049/iet-gtd.2014.0655.

W. K. Huang, D. W. Nychka, and H. Zhang. Estimating precipitation extremes using the log-histospline. *Environmetrics*, 30(4):e2543, 2019.

R. W. Katz, M. B. Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in water resources*, 25(8-12):1287–1304, 2002. doi: https://doi.org/10.1016/S0309-1708(02)00056-8.

W.-Y. Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011. doi: https://doi.org/10.1002/widm.8.

W.-Y. Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3): 329–348, 2014. doi: https://doi.org/10.1111/insr.12016.

O. Lopez, X. Milhaud, and P.-E. Thérond. Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics*, 10(2):2685–2716, 2016. doi: https://doi.org/10.1214/16-EJS1189.

O. C. Pasche and S. Engelke. Neural networks for extreme quantile regression with an application to forecasting of flood risk. *arXiv preprint arXiv:2208.07590*, 2022.

J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131, 1975.

J. Richards and R. Huser. A unifying partially-interpretable framework for neural network-based extreme quantile regression. *arXiv preprint arXiv:2208.07581*, 2022.

T. Rietsch, P. Naveau, N. Gilardi, and A. Guillou. Network design for heavy rainfall analysis. *Journal of Geophysical Research: Atmospheres*, 118(23):13–075, 2013.

V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71:804–818, 2015. doi: https://doi.org/10.1016/j.oregeorev.2015.01.001.

E. Ross, S. Sam, D. Randell, G. Feld, and P. Jonathan. Estimating surge in extreme north sea storms. *Ocean Engineering*, 154:430–444, 2018.

C. Scarrott and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical journal*, 10(1):33–60, 2012.

R. L. Smith. Threshold methods for sample extremes. In *Statistical extremes and applications*, pages 621–638. Springer, 1984.

R. L. Smith. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, pages 367–377, 1989.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

X. Su, M. Wang, and J. Fan. Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, 13(3):586–598, 2004. doi: https://doi.org/10.1198/106186004X2165.

M. Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

P. Tencaliec, A.-C. Favre, P. Naveau, C. Prieur, and G. Nicolet. Flexible semiparametric generalized pareto modeling of the entire range of rainfall amount. *Environmetrics*, 31(2):e2582, 2020.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998.

J. Velthoen, J.-J. Cai, G. Jongbloed, and M. Schmeits. Improving precipitation forecasts using extreme quantile regression. *Extremes*, 22(4):599–622, 2019. doi: https://doi.org/10.1007/s10687-019-00355-1.

J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke. Gradient boosting for extreme quantile regression. *arXiv preprint arXiv:2103.00808*, 2021.

H. J. Wang, D. Li, and X. He. Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association*, 107(500):1453–1464, 2012. doi: https://doi.org/10.1080/01621459.2012.716382.

B. D. Youngman. Generalized additive models for exceedances of high thresholds with an application to return level estimation for us wind gusts. *Journal of the American Statistical Association*, 114(528): 1865–1879, 2019.