Insurance: Mathematics and Economics 98 (2021) 92–105

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Cyber claim analysis using Generalized Pareto regression trees with applications to insurance

Sébastien Farkas, Olivier Lopez *, Maud Thomas

Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France

ARTICLE INFO

Article history:
Received January 2020
Received in revised form February 2021
Accepted 26 February 2021
Available online 17 March 2021

Keywords:
Cyber insurance
Extreme value analysis
Regression trees
Generalized Pareto distribution
Machine learning
Clustering

ABSTRACT

With the rise of the cyber insurance market, there is a need for better quantification of the economic impact of this risk and its rapid evolution. Due to the heterogeneity of cyber claims, evaluating the appropriate premium and/or the required amount of reserves is a difficult task. In this paper, we propose a method for cyber claim analysis based on regression trees to identify criteria for claim classification and evaluation. We particularly focus on severe/extreme claims, by combining a Generalized Pareto modeling – legitimate from Extreme Value Theory – and a regression tree approach. Coupled with an evaluation of the frequency, our procedure allows computations of central scenarios and of extreme loss quantiles for a cyber portfolio. Finally, the method is illustrated on a public database.

1. Introduction

Cyber risk is a natural consequence of the digital transformation. Digital technologies induce new vulnerabilities for economic actors, with a rapid evolution of practices, threats, and behaviors. With the increase of cyber threats, insurance contracts appear as fundamental tools to improve the resilience of society.

However while the cyber insurance market is growing fast (see for example the report of European Insurance and Occupational Pensions Authority (EIOPA), 2019), risk analysis faces a lack of consistent and reliable data in a context where the amount of claims is particularly volatile (see Matthews, 2019). Therefore, quantifying this emerging and evolving risk is a difficult task. In this paper, we propose to analyze cyber claims via regression trees in order to constitute clusters of cyber incidents. These clusters achieve a compromise between homogeneity and a sufficient size to allow a reliable statistical estimation of the risk. A particular attention is devoted to large claims, for which heavy tail distributions are fitted. The study of large claims raises the question of insurability of the risk, and the clustering technique we propose may help to separate types of incidents or circumstances according to whether they can be covered without endangering risk pooling. In the present work, we develop a regression tree methodology specifically adapted to the study of heavy-tailed distributions, and discuss its relevance to embrace the challenges of cyber risk quantification. A first contribution is methodological: by adapting the methodology of regression trees to extreme value regression purpose, we provide a flexible and still intelligible modeling tool, that can be valuable for a wide range of risks (not only in the field of cyber). On the other hand, we aim to discuss their practical behavior on a real cyber related database. In order to allow reproducibility of our work, we consider a publicly available database. Through this analysis, we show how the output of our method can be used for cyber insurance risk management, and to identify stylized facts useful for practitioners. Furthermore, public databases on cyber events are an important source in order to complement the information (usually poor due to the relative novelty of the risk) available for insurance companies. We hope that the detailed description of the path of our analysis can help to improve the integration of such public information in cyber risk quantification.

Topics recently addressed in cyber insurance are reviewed in Biener et al. (2015), Eling and Schnell (2016) and Marotta et al. (2017). However most of these approaches are performed from the point of view of a cyber analyst. For instance, Fahrenwaldt et al. (2018) study the topology of infected networks, and Insua et al. (2021) gather expert judgments using an Adversarial Risk Analysis. Eling and Loperfido (2017) and Edwards et al. (2016) developed more established insurance modeling methods illustrated on the Privacy Rights Clearinghouse (PRC) database (available for public download at https://privacyrights.org/data-breaches). PRC database has also been studied by Maillart and Sornette (2010). It gathers data breaches events for which a severity indication is given (through the volume of breached data), making it valuable for insurance applications. On the other hand,

* Corresponding author.
E-mail addresses: sebastien.farkas@sorbonne-universite.fr (S. Farkas), olivier.lopez@sorbonne-universite.fr (O. Lopez), maud.thomas@sorbonne-universite.fr (M. Thomas).

https://doi.org/10.1016/j.insmatheco.2021.02.009
0167-6687/

this database is not fed by an insurance portfolio but by various sources of information, each reporting heterogeneous types of claims. In particular, the exposure (that is the number of entities exposed to the risk in the scope of PRC organization) is blur.

In the present paper, we consider the same PRC database to illustrate our methodology, that can be easily extended to other types of data. The method we develop is adapted to detect such instabilities in this context of a database fed by sources of information which variety may disturb the evaluation of the risk. We especially focus on "extreme" events, that is events for which the severity of the claim is larger than a fixed (high) threshold, seeking to gain further insight on the impact of the characteristics of companies and of the circumstances on a cyber event. Therefore, relying on regression trees inference and extreme value theory, we introduce a statistical methodology that takes into account both the heterogeneity and the extreme features. In addition, we propose an insurance pricing and reserving framework based on assumptions on the exposure and on the costs of data breaches in order to take advantage of the PRC database within the realms of possibility.

Regression trees are good candidates to understand the origin of the heterogeneity, since they allow to perform regression and classification simultaneously. Since the pioneer works of Breiman et al. (1984) who introduced CART algorithm (Clustering And Regression Tree), regression trees have been used in many fields, including industry (see e.g. González et al., 2015), geology (see e.g. Rodriguez-Galiano et al., 2015), ecology (see e.g. De'ath and Fabricius, 2000), claim reserving (see e.g. Lopez et al., 2016). A nice feature of this approach is to introduce nonlinearities in the way the distribution is modeled, while furnishing an intelligible interpretation of the final classification of response variables. A further argument in favor of the use of regression trees is the simplicity of the algorithm: such models are fitted to the data via an iterative decomposition. The splitting criterion depends on the type of problems one wishes to investigate: the standard CART algorithm uses a quadratic loss since it aims at performing mean-regression. Alternative loss functions may be considered as in Chaudhuri and Loh (2002) in order to perform quantile regression or in Su et al. (2004) for log-likelihood loss for example. Loh (2011, 2014) provide detailed descriptions of regression trees procedures and a review of their variants. In the present paper, we use different types of splitting criteria, with a particular attention devoted to the tail of the distribution of the claim size, which describes the behavior of extreme events. We therefore use a Generalized Pareto distribution to approximate the tail of the distribution—which is at the core of the "Peaks Over Threshold" procedure in extreme value theory (see e.g. Pickands, 1975; Beirlant et al., 2004)—with parameters depending on the classes defined by the regression tree.

The rest of the paper is organized as follows. In Section 2, we give a short presentation of the PRC database, its advantages and its inconsistencies. The general description of regression trees and their adaptation to extreme value analysis is done in Section 3. These methodologies are applied to the PRC database in Section 4, leading to a model for the severity of claims. This model is combined with a frequency model in Section 5.2, in order to quantify the impact of this analysis on (virtual) insurance portfolios.

2. A public data breaches database

The Privacy Rights Clearinghouse (PRC) database is one of the few publicly available databases on cyber events which associates a quantification of the severity with a claim. This piece of information is crucial from an insurance perspective: evaluation of the risk associated with a policyholder requires to estimate the probability of being a victim of a cyber event (or the frequency of occurrence of such events), and to quantify the potential random loss. Regarding the severity, PRC database does not directly provide the loss associated with an event, but reports the number of records (that is the number of user accounts) affected by the breach. This number is correlated to the financial impact of the claim, which can be approximatively retrieved through a formula given in Jacobs (2014) which will be described later on in Section 5.1. We describe the database in Section 2.1. A focus on the sources feeding the database is done in Section 2.2. This short overview helps us to identify some characteristics and inconsistencies of cyber data summarized in Section 2.3, and will motivate the use of the methodology developed in the rest of the paper.

Table 1
List of the available variables in the PRC database.

PRC database	Variable
Victim data	Name of organization
	Sector of organization
	Geographic position of organization
Event data	Source of release
	Date of release
	Type of breach
	Number of affected records
	Description of the event

Table 2
Labels for activity sectors of victims in the PRC database.

BSF	Businesses - Financial and Insurance Services
BSO	Businesses - Other
BSR	Businesses - Retail/Merchant - Including Online Retail
EDU	Educational Institutions
GOV	Government & Military
MED	Healthcare, Medical Providers & Medical Insurance Services
NGO	Nonprofits

2.1. Description of the database

Privacy Rights Clearinghouse is a nonprofit organization founded in 1992 which aims at protecting US citizens privacy. Especially, PRC has maintained a chronology since 2005, listing companies that have been involved in data breaches affecting US citizens. This article is based on a download of this database made on January 23 2019, corresponding to 8860 cyber events on companies, mainly American companies. Among them, only 8298 events were kept for our analysis, since we eliminated duplicated and/or inconsistent events (e.g. information on the targeted company are sometimes not consistent).

The PRC database gathers information regarding each cyber event (its type, the number of records affected by the breach, a description of the event) and its victim (the targeted company name, its activities, its localization). These variables and their modalities are summarized in Tables 1 to 3. Additional statistics are shown in the supplementary material (Section 1).

2.2. Multiple sources feeding the database

In this section, we focus on the variable "Source of release". The PRC organization gathers cyber events from different sources, which can be clustered into four groups:

- US Government Agencies on the federal level: in the healthcare domain, the Health Insurance Portability and Accountability Act (HIPAA) imposes a notification to the Secretary of the U.S. Department of Health and Human Services for each breach that affects 500 or more individuals, U.S. HHS department (2020b). Those notifications are reported online

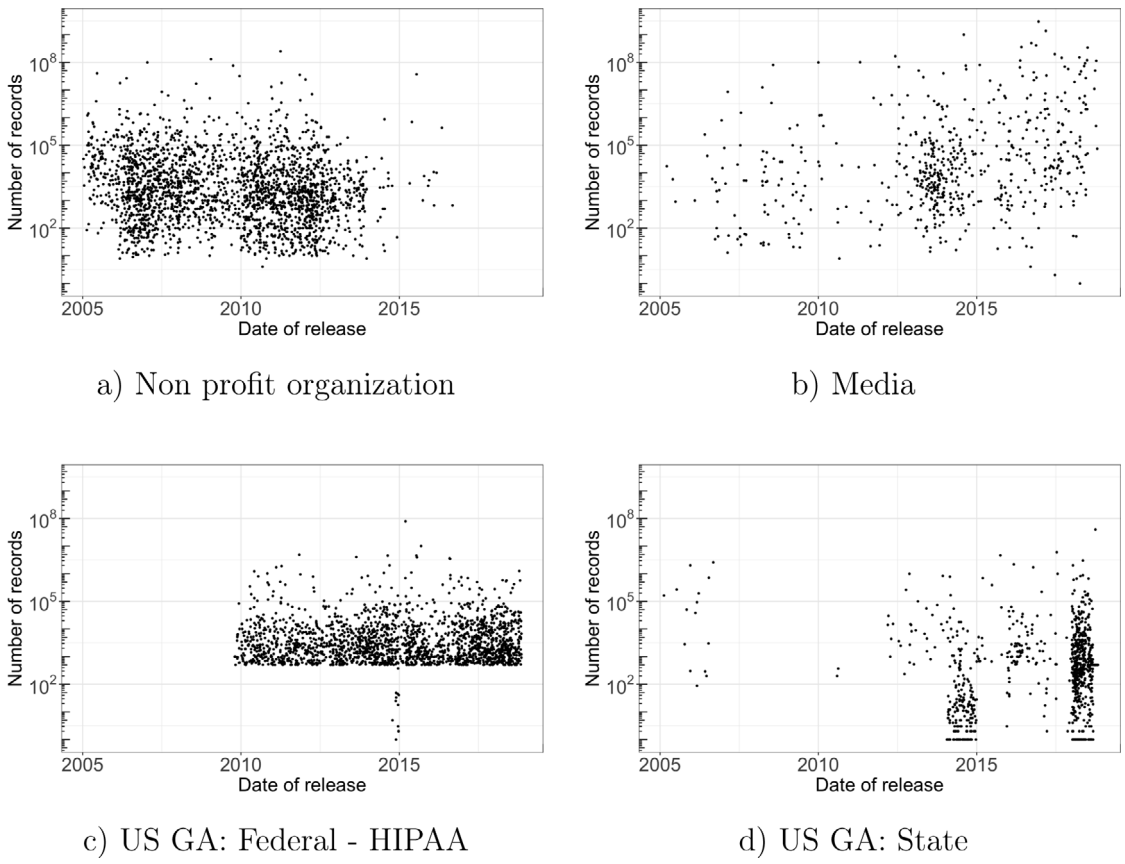

Fig. 1. Scatter plots of data breaches listed in the PRC database (the x-axis is the date of the release and the y-axis is the number of records) depending on the source of information.

Table 3
List of the types of data breaches as labeled in the PRC database.

CARD	Fraud involving debit and credit cards that is not accomplished via hacking
HACK	Hacked by outside party or infected by malware
INSD	Insider (someone with legitimate access intentionally breaches information)
PHYS	Includes paper documents that are lost, discarded or stolen (non electronic)
PORT	Lost, discarded or stolen laptop, PDA, smartphone, memory stick, CDs, hard drive, data tape, etc.
STAT	Stationary computer loss (lost, inappropriately accessed, discarded or stolen computer or server not designed for mobility)
DISC	Unintended disclosure (not involving hacking, intentional breach or physical loss)
UNKN	Unknown

with free access on the breach portal (U.S. HHS department, 2020a).

- US Government Agencies on the state level: since 2018, every state has a specific legislation related to data breaches. Differences have been studied by Privacy Rights Clearinghouse (2019). Particularly, there is no uniformity on the threshold (in terms of number of victims) above which a notification becomes mandatory. Some states publicly release notifications, which is the case of California through the online portal (State of California, 2020), but this is not systematic.
- Media: PRC organization monitors media to list data breaches that have received extensive media coverage.
- Non profit organizations: the PRC database includes the data breaches reported by other non profit organizations than PRC, for instance Databreaches.net (2020).

While merging different sources of notifications increases the scope of the PRC chronology, it also introduces some heterogeneity among the reported events, since each source reports a particular kind of claims. Additionally, the proportion of reported events from a given source fluctuates through time, as shown in Figs. 1 and 2.

2.3. Heterogeneity and inconsistencies in PRC database

The way the database has been fed has evolved over time. These changes have had an impact on our main objective, which is to analyze the severity of these events. Indeed one may for example guess that cyber claims that were exposed by media are more likely to be more "spectacular" (and hence more severe). This intuition will be confirmed by the quantitative results of Section 3.

Moreover, a short descriptive analysis of the severity variable ("number of records", see Table 4) shows that it is highly volatile. One can note an important difference between the median of the number of records (2000) and the empirical mean (1.821 millions) because the latter is mainly driven by extreme events (the largest having 3 billions of records). This important dispersion is expected, due to the extreme variety of situations considered in the database. This pleads for reducing this heterogeneity by introducing appropriate risk classes, and in which we could separate the sources of information if they appear to be correlated with the severity of the claim. To determine such classes, our procedure relies on regression trees which are described in Section 3. They

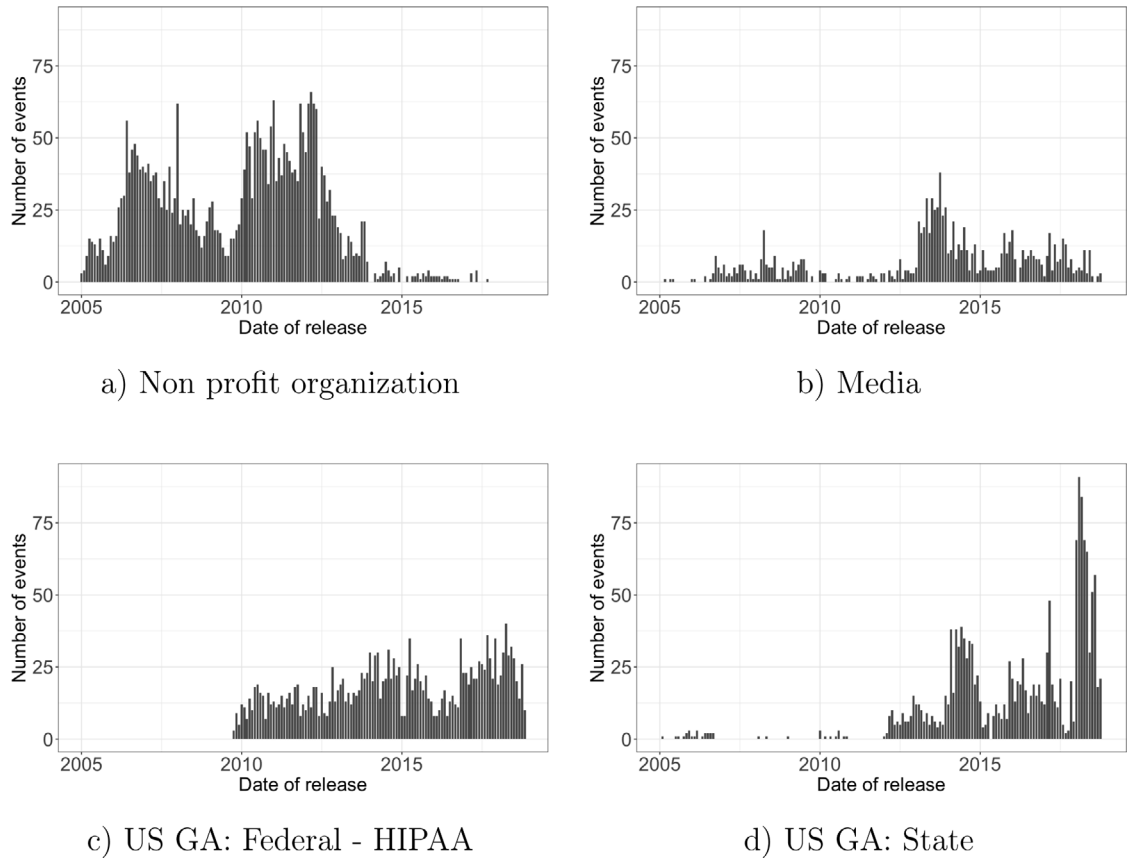

Fig. 2. Barplots of data breaches listed in the PRC database (the x-axis is the date of the release and the y-axis is the number of records) depending on the source of information.

Table 4
Descriptive statistics for the variable "Number of records" depending on the source of information (first column). q_α denotes the empirical α-quantile, that is such that α% of observations are smaller than q_α.

	Number	Mean	$q_{0.25}$	Median	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$	Max
Total	6160	1 821 682	597	2000	10 891	70 000	300 000	3 000 000 000
US GA: Federal - HIPAA	1949	84 358	981	2300	8009	28 440	75 016	78 800 000
US GA: State	888	89 377	20	4010	2403	18 000	63 825	40 000 000
Media	595	16 208 786	1400	11 266	137 193	4 420 000	41 029 090	3 000 000 000
Nonprofit organization	2309	422 623	380	2000	14 000	86 333	247 200	250 000 000
Unknown	419	853 736	958	2300	9154	30 194	61 863	191 000 000

offer the advantage to perform an automatic clustering, without any a priori on the covariates.

3. Regression trees and extreme value analysis

Regression trees are a convenient tool when one wants to simultaneously predict a response and filter heterogeneity by determining clusters among the data. In the sequel, Y denotes a response variable (a "cost" variable representing the severity of the claim), and $\mathbf{X} \in \mathbb{R}^d$ some covariates (the circumstances of the claim, the victim(s), the source which detected the event...). Our observation set is composed of i.i.d. replications $(Y_i, \mathbf{X}_i)_{1\leq i\leq n}$ of $(Y, \mathbf{X})$. Regression trees aim at determining "rules" to gather observations in risk classes depending on the values of their characteristics $\mathbf{X}_i$. Therefore they are particularly adapted to the situations where the variety of profiles of $\mathbf{X}_i$ induces some heterogeneity. The CART algorithm, used to compute the trees, is presented in Section 3.1. Depending on the purpose of regression trees (typically, in our situation, depending on whether we wish to investigate the center or the tail of the distribution), an appropriate loss function has to be defined in order to evaluate the quality of the tree and define splitting rules for the clustering part of the algorithm. Generalized Pareto regression trees, introduced in Section 3.2, rely on a splitting rule which is designed to focus on the tail of the distribution, due to key results in extreme value theory.

3.1. Regression trees

Regression Trees are modeling tools that allow one to introduce modeling of (nonlinear) heterogeneity between the observations, by splitting them into classes on which different regression models are fitted. The aim is to retrieve a regression function $m^* = \arg\min_{m\in\mathcal{M}} E[\phi(Y, m(\mathbf{X}))]$, where, again, Y is our response variable (the severity of a cyber claim in our case), $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ is a set of covariates, $\mathcal{M}$ is a class of target functions on $\mathbb{R}^d$ and ϕ is a loss function that depends on the quantity we wish to estimate (see Section 3.1.2).

In the following, we will consider three different types of functions ϕ:

- the quadratic loss $\phi(y, m(\mathbf{x})) = (y - m(\mathbf{x}))^2$ corresponds to the situation where the objective is the conditional mean

$m^*(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$ and $\mathcal{M}$ is the set of functions of $\mathbf{x}$ with finite second order moment;
- the absolute loss $\phi(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$, where m^* is the conditional median;
- a log-likelihood loss $\phi(y, m(\mathbf{x})) = -\log f_{m(\mathbf{x})}(y)$, where $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ is a parametric family of densities. This corresponds to the case where one assumes that the conditional distribution of $Y|\mathbf{X} = \mathbf{x}$ belongs to the parametric family $\mathcal{F}$ for all $\mathbf{x}$, with parameter $m(\mathbf{x})$ depending on $\mathbf{x}$.

This split of the data is performed in an iterative way, by finding at each step an appropriate simple rule (that is a condition on the value of some covariate) to separate the data into two more homogeneous classes. The procedure includes two phases: a "growing" phase which corresponds to the CART algorithm, and a "pruning" step which consists in the extraction of a subtree from the decomposition obtained in the initial phase. Pruning can therefore be understood as a model selection procedure. In Section 3.1.1, we describe a general version of the CART algorithm, and explain in Section 3.1.2 how an estimation of a regression model can be deduced from a tree obtained in this first phase. The pruning step is then described in Section 3.1.3.

3.1.1. Growing step: construction of the maximal tree

The CART algorithm consists in determining iteratively a set of "rules" $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)}) \to R_j(\mathbf{x})$ to split the data, aiming at optimizing some objective function (also referred to as splitting criterion). More precisely, for each possible value of the covariates $\mathbf{x}$, $R_j(\mathbf{x}) = 1$ or 0 depending on whether some conditions are satisfied by $\mathbf{x}$, with $R_j(\mathbf{x})R_{j'}(\mathbf{x}) = 0$ for $j \neq j'$ and $\sum_j R_j(\mathbf{x}) = 1$. In case of regression trees, these partitioning rules have a particular structure, since they can be written as $R_j(\mathbf{x}) = \mathbf{1}_{\mathbf{x}_1 \leq \mathbf{x} < \mathbf{x}_2}$ for some $\mathbf{x}_1 \in \mathbb{R}^d$ and $\mathbf{x}_2 \in \mathbb{R}^d$, and the comparison symbols have to be understood as componentwise 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_j at step k generates two rules R_{j1} and R_{j2} (with $R_{j1}(\mathbf{x}) + R_{j2}(\mathbf{x}) = 0$ if $R_j(\mathbf{x}) = 0$) at step $k + 1$. The algorithm can be summarized as follows:

Step 1: $R_1(\mathbf{x}) = 1$ for all $\mathbf{x}$, and $n_1 = 1$ (corresponds to the root of the tree).

Stepk + 1 : Let $(R_1, \ldots, R_{n_k})$ denote the rules obtained at step k. For $j = 1, \ldots, n_k$,

- if all observations such that $R_j(\mathbf{X}_i) = 1$ have the same characteristics, then keep rule j as it is no longer possible to segment the population;
- else, rule R_j is replaced by two new rules R_{j1} and R_{j2} determined in the following way: for each component $X^{(l)}$ of $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$, define the best threshold $x^{(l)}_{j_\star}$ to split the data, such that $x^{(l)}_{j_\star} = \arg\min_{x^{(l)}} \Phi(R_j, x^{(l)})$, with

$$\begin{aligned}\Phi(R_j, x^{(l)}) = &\sum_{i=1}^n \phi(Y_i, \widehat{m}(R_j))R_j(\mathbf{x}) \\ &- \sum_{i=1}^n \phi(Y_i, m_{l-}(\mathbf{X}_i, R_j))\mathbf{1}_{X_i^{(l)} \leq x^{(l)}}R_j(\mathbf{x}) \\ &- \sum_{i=1}^n \phi(Y_i, m_{l+}(\mathbf{X}_i, R_j))\mathbf{1}_{X_i^{(l)} > x^{(l)}}R_j(\mathbf{x}),\end{aligned}$$

where

$$\widehat{m}(R_j) = \arg\min_{m \in \mathcal{M}} \sum_{i=1}^n \phi(Y_i, m(\mathbf{X}_i))R_j(\mathbf{X}_i),$$

$$m_{l-}(x, R_j) = \arg\min_{m \in \mathcal{M}} \sum_{i=1}^n \phi(Y_i, m(\mathbf{X}_i))\mathbf{1}_{X_i^{(l)} \leq x}R_j(\mathbf{X}_i),$$

$$m_{l+}(x, R_j) = \arg\min_{m \in \mathcal{M}} \sum_{i=1}^n \phi(Y_i, m(\mathbf{X}_i))\mathbf{1}_{X_i^{(l)} > x}R_j(\mathbf{X}_i).$$

Then, select the best component index to consider: $\widehat{l} = \arg\min_l \Phi(R_j, x^{(l)}_{j_\star})$.
Define the two new rules $R_{j1}(\mathbf{x}) = R_j(\mathbf{x})\mathbf{1}_{x^{(\widehat{l})} \leq x^{(\widehat{l})}_{j_\star}}$, and $R_{j2}(\mathbf{x}) = R_j(\mathbf{x})\mathbf{1}_{x^{(\widehat{l})} > x^{(\widehat{l})}_{j_\star}}$.

- Let n_{k+1} denote the new number of rules.

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

As it has already been mentioned, this algorithm has a binary tree structure. The list of rules $(R_j)_{1 \leq j \leq n_k}$ 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 $k + 1$.

In this version of the CART algorithm, all covariates are continuous 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_j 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 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.

3.1.2. From the tree to the regression function

From a set of rules $\mathcal{R} = (R_j)_{j=1,\ldots,s}$, an estimator $\widehat{m}^{\mathcal{R}}$ of the function m is given by

$$\widehat{m}^{\mathcal{R}}(\mathbf{x}) = \sum_{j=1}^s \widehat{m}(R_j)R_j(\mathbf{x}).$$

The final set of rules $\mathcal{R}^M$ obtained from the CART algorithm is called the maximal tree. This leads to a trivial estimator of m, since either the number of observations in a leaf is one, or all observations in this leaf have the same characteristics $\mathbf{x}$. The pruning step consists in extracting a subtree from the maximal tree, achieving a compromise between simplicity and good fit.

3.1.3. Selection of a subtree: pruning algorithm

For the pruning step, a standard way to proceed is to use a penalized approach to select the appropriate subtree (see Breiman et al., 1984; Gey and Nédélec, 2005). A subtree $\mathcal{S}$ of the maximal tree is associated with a set of rules $\mathcal{R}^{\mathcal{S}} = (R_1^{\mathcal{S}}, \ldots, R_{n_{\mathcal{S}}}^{\mathcal{S}})$ of cardinality $n_{\mathcal{S}}$. One then selects the subtree $\widehat{\mathcal{S}}(\alpha)$ that minimizes the criterion

$$C_\alpha(\mathcal{S}) = \sum_{i=1}^n \phi(Y_i, m^{\mathcal{R}^{\mathcal{S}}}(\mathbf{X}_i)) + \alpha n_{\mathcal{S}}, \tag{3.1}$$

among all subtrees of the maximal tree, where α is a positive constant. Hence, the trees with large numbers of leaves (i.e. of rules) are penalized compared to smaller ones. To determine this tree $\widehat{\mathcal{S}}(\alpha)$, it is not necessary to compute all the subtrees from the maximal tree. It suffices to determine, for all $K \geq 0$, the subtree $\mathcal{S}_K$ which minimizes the criterion (3.1) among all subtrees $\mathcal{S}$ with $n_{\mathcal{S}} = K$, and then to choose the tree $\mathcal{S}_K$ which minimizes the criterion with respect to K. From Breiman et al. (1984, p. 284–290), these $\mathcal{S}_K$ are easy to determine, since $\mathcal{S}_K$ is obtained by removing one leaf to $\mathcal{S}_{K+1}$.

The penalization constant α is chosen using a test sample or k-fold cross-validation. In the first case, data are split into two parts before growing the tree (a training dataset of size n and a test sample which is not used in computing the tree). In the second case, the dataset is randomly split into k parts which successively act as training or test sample.

Let $\widehat{\alpha}$ denote the penalization constant calibrated using the test sample or the k-fold cross-validation approach, our final estimator is then $\widehat{m}(\mathbf{x}) = m^{\widehat{S}(\widehat{\alpha})}(\mathbf{x})$.

3.2. Generalized Pareto regression trees for analyzing the tail of the distribution

Since the severity of cyber events is highly volatile, it seems necessary to develop a specific approach for the tail of distribution. In Section 3.2.1, we recall why Generalized Pareto (GP) distributions naturally appear in the analysis of heavy-tailed variables. This motivates our GP trees described in Section 3.2.2.

3.2.1. Peaks over threshold method for extreme value analysis

Extreme value analysis is the branch of statistics which has been developed and broadly used to handle extreme events, such as extreme floods, heat waves episodes or extreme financial losses (Katz et al., 2002; Embrechts et al., 1997). Given a series of independent and identically distributed observations $Y_1, Y_2, \dots$ 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 exceed 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

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

If $\overline{F}$ satisfies the following property

$$\lim_{t\to\infty} \frac{\overline{F}(ty)}{\overline{F}(y)} = y^{-1/\gamma}, \ \forall y > 0, \tag{3.2}$$

with $\gamma > 0$, then

$$\lim_{u\to\infty} \sup_{y>0} |\overline{F}_u(y) - \overline{H}_{\sigma_u,\gamma}(y)| = 0 \tag{3.3}$$

for some $\sigma_u > 0$ and $\overline{H}_{\sigma_u,\gamma}$ necessarily of the form

$$\overline{H}_{\sigma_u,\gamma}(y) = \left(1 + \gamma \frac{y}{\sigma_u}\right)^{-1/\gamma}, \ y > 0. \tag{3.4}$$

Here, $\sigma_u > 0$ is a scale parameter and $\gamma > 0$ is a shape parameter, which reflects the heaviness of the tail distribution. Especially, if $\gamma \in]0; 1[$, the expectation of Y is finite whereas if $\gamma \geq 1$ the expectation of Y is infinite. In our situation of highly volatile severity variables, the assumption $\gamma > 0$ is reasonable and supported by the empirical results of Maillart and Sornette (2010) (who even estimated $\gamma > 1$). The result from Balkema and De Haan (1974) states that, if the survival function of the normalized excesses above a high threshold u weakly converges towards a non-degenerate distribution, then the limit is a Generalized Pareto distribution (see also Pickands, 1975).

In practice, the so-called Peaks over Threshold (PoT) method has been widely used since 1990 (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 and γ may be done by maximizing the GP likelihood. The choice of the threshold u implies a balance between bias and variance. Too low a threshold is likely to violate the asymptotic basis of the model, leading to bias; too high a threshold will generate few excesses with which the model can be estimated, leading to high variance. The standard practice is to choose as low a threshold as possible, subject to the limit model providing a reasonable approximation.

Remark 3.1. Property (3.2) is called regular variation. When $\gamma > 0$, we say that $\overline{F}$ is heavy-tailed, meaning that its tail decreases polynomially. Usual distributions as Pareto, Cauchy and Student distributions satisfy this property. For more details, see De Haan and Ferreira (2007, Appendix B).

3.2.2. Generalized Pareto regression trees

When it comes to studying the severity of cyber claims, we expect to see a potential heterogeneity in the tail of the distribution. In order to improve the precision of our analysis, a natural idea is to study the impact of the circumstances of the claim and of the characteristics of the victim on the response variable. In our regression framework, for each value of the covariate $\mathbf{x}$, we assume the conditional distribution of $Y|\mathbf{X} = \mathbf{x}$ to be heavy-tailed, but the parameters γ, σ (and the threshold u above which the GP distribution approximation seems satisfactory) depend on $\mathbf{x}$. More precisely, this means that (3.2) becomes

$$\lim_{t\to\infty} \frac{\overline{F}(ty|\mathbf{x})}{\overline{F}(y|\mathbf{x})} = y^{-1/\gamma(\mathbf{x})}, \ \forall y > 0, \tag{3.5}$$

where $\overline{F}(y|\mathbf{x}) = \mathbb{P}(Y \geq y|\mathbf{X} = \mathbf{x})$ with $\gamma(\mathbf{x}) > 0$ for all $\mathbf{x}$, and (3.3) becomes

$$\lim_{u(\mathbf{x})\to\infty} \sup_{y>0} |\overline{F}_{u(\mathbf{x})}(y \mid \mathbf{x}) - \overline{H}_{\sigma_{u(\mathbf{x})}(\mathbf{x}),\gamma(\mathbf{x})}(y)| = 0. \tag{3.6}$$

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

The idea is then to apply the procedure of Section 3 to the observations $(Y_i - u(\mathbf{X}_i), \mathbf{X}_i)$ for which $Y_i \geq u(\mathbf{X}_i)$, using the Generalized Pareto log-likelihood as split function, that is

$$\phi(y, m(\mathbf{x})) = -\log(\sigma(\mathbf{x})) - \left(\frac{1}{\gamma(\mathbf{x})} + 1\right) \log\left(1 + \frac{y\gamma(\mathbf{x})}{\sigma(\mathbf{x})}\right),$$

where $m(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$ (we use the notation $\sigma(\mathbf{x}) = \sigma_{u(\mathbf{x})}(\mathbf{x})$ to simplify). The function $u(\mathbf{x})$ is an input of the procedure, and has to be taken so that the GP distribution fit seems appropriate for all considered values of covariates. The practical choice of this function is a delicate problem (see Section 4 in Beirlant and Goegebeur, 2004). To simplify, we consider in the following a fixed threshold $u(\mathbf{x}) = u$ for all values of covariates $\mathbf{x}$. The threshold u is chosen large enough so that the GP approximation is correctly fitted to the data (practical choice of this parameter will be discussed in Section 4.2, see also Remark 3.2). In the end, the leaves of the tree identify classes, each corresponding to different tail behaviors (that is with different values of $m(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$, the function m being constant on each leaf.

Compared to competing approaches in extreme value regression, the advantage of the procedure is to introduce discontinuities in the regression function while parametric approaches, like in Beirlant and Goegebeur (2003), suppose a form of linearity. More flexible nonparametric approaches, as in Beirlant and Goegebeur (2004), rely on smoothing techniques that require covariates to be continuous. Chavez-Demoulin et al. (2015) propose a semiparametric framework to separate the continuous covariates from the discrete ones. Smoothing splines are used to estimate nonparametrically the continuous part, while the influence of discrete covariates is captured by a parametric function. Due to the nice properties of this technique applied on operational risk data in Chavez-Demoulin et al. (2015), we compare the results of our GP regression tree approach to their procedure in Section 4.3.

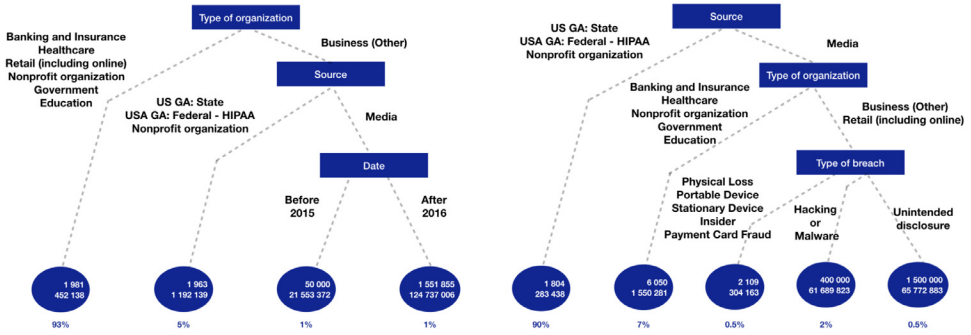

Fig. 3. Trees obtained from the CART algorithm based on the quadratic (left-hand side) and the absolute (right-hand side) losses. For each leaf, the value of the empirical median (first line) and mean (second line) are given. Percentage of observations affected to each leaf is mentioned.

Remark 3.2. As already stated, the conditional version of (3.4) used in extreme value regression leads to the introduction of a threshold function $u(\mathbf{x})$ that potentially depends on $\mathbf{x}$. A possibility would be to adapt the CART algorithm to select, at each step, a choice of threshold that could be different in each leaf. However, this complexifies considerably the technique, and we did not consider it.

4. PRC database analysis with regression trees

In this section, we apply the different variations of the regression tree approach of Section 3 to the response variable $Y =$ "Number of Records" in the PRC database. Let us note that, despite its name, this variable can be considered as continuous, since this number of records takes a wide range of values (see Table 4) with few ties (caused by rounded numbers). Section 4.1 describes regression tree analysis of the central part of the distribution, while the tail part is considered in Section 4.2, applying GP trees. Comparison with the fit of a GAM model as in Chavez-Demoulin et al. (2015) is shown in Section 4.3. Section 4.4 shows how our two regression tree approaches (one for the central part of the distribution, one for the tail) can be combined to provide a global analysis of the distribution. A discussion on the insurability of cyber risk—which, from a probabilistic point of view, is closely related to the value of the tail parameter γ—is done in Section 4.5.

4.1. Central part of the severity distribution

In order to estimate the conditional mean $E[Y|\mathbf{X} = \mathbf{x}]$, with a regression tree, the loss function ϕ has to be chosen as the quadratic loss $\phi(y, m(\mathbf{x})) = (y - m(\mathbf{x}))^2$. The conditional mean is particularly important in view of computing a pure premium in insurance (pure premium corresponds to estimating the expectation of the cost, which requires to estimate the frequency of occurrence and the mean value of a claim), but this indicator is not robust, due to its sensitivity to large observations. Let us also observe that this conditional expectation may even not be defined for some values of $\mathbf{X}$ since Y is heavy-tailed. Since the variable Y we study is highly volatile, investigating the conditional median of the distribution of $Y|\mathbf{X} = \mathbf{x}$ (that is $med(Y|\mathbf{X} = \mathbf{x}) = \inf\{y : F(y|\mathbf{x}) \geq 1/2\}$, where $F(y|\mathbf{x}) = P(Y \leq y|\mathbf{X} = \mathbf{x})$) may be more stable. Estimating the conditional median corresponds to the choice of the absolute loss as the loss function, that is $\phi(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$.

We fit regression trees using these two loss functions. These trees are computed using the `R` package `rpart` (see Therneau and Clinic, 2019), by using a user defined split function. The pruning step has been done thanks to a 10-fold cross validation used for error measurement and the selection of a proper subtree. The obtained trees are shown in Fig. 3.

The structure of the trees is different for the conditional median compared to the conditional expectation, although some similarities exist. For example, the category of victims "Business (Other)" seems generally associated with higher severity: for the mean tree, all events are gathered in the same leaf, except for those affecting this category of targets, which are associated with the largest predicted values. The picture is slightly different for the median tree: the highest predicted values are still linked with the "Business (Other)" category, but only under particular circumstances. In both cases, the Media source is generally associated with larger events.

The leaves of the trees determine clusters. If one wants to get a distribution for the claim severity, a distribution can be fitted on each leaf, see the supplementary material (Section 1.3) for more details.

Remark 4.1. Our procedure consists in first determining clusters (using regression trees with L^2 or L^1 loss), and then fitting log-normal distributions to each leaf. This last step is only required if one wishes to have a global model for the distribution of Y. One could directly use a log-normal log-likelihood as split criterion to obtain different clusters that should improve the log-normal fit. The reason for not choosing this path is because our purpose is essentially to understand which covariates drive the central part of the distribution (in order to compare it to the study of the tail, which is our main objective), but comparisons with a direct log-normal fit can be found in Section 2.1 of the supplementary material. The L^2 and L^1 trees are supposed to provide clusters that are based on the expectation or the median, with no particular assumption on the conditional distribution of Y in each leaf. In fact, fitting these trees can be done even in the case where all the leaves do not correspond to the same family of distribution (one may fit a gamma distribution in one leaf, a log-normal in another).

4.2. Tail part of the severity distribution

In view of applying the GP regression tree approach of Section 3.2.2, our first task is to determine the threshold u above which the GP distribution approximation seems reasonable. This choice is made from the Hill plot (shown in the Appendix A, Fig. 6) (see Resnick, 2007, pp 85–89 for more details on Hill plots). From the shape of the curve, we chose $u = 27\ 999$ (which corresponds to a stabilization of the Hill plot) which leads to keep the 1000 highest observations (around 16% of the total number of breaches). Let us note that Hill plots are not designed for regression methods. In our context, as already pointed in Remark 3.2, one could look at thresholds depending on the covariates. See also Section 4 in Beirlant and Goegebeur (2004) who discussed this choice of thresholds in extreme value regression, and Section 4 of the supplementary material.

Table 5
Generalized Pareto parameters estimated by the Generalized Pareto regression Tree based on excesses and the 95% confidence intervals (given under brackets).

	Leaf 1	Leaf 2	Leaf 3
γ	1.43 [1.21;1.64]	1.72 [1.41;2.04]	3.26 [2.62;3.91]
$\sigma.10^{-5}$	0.36 [0.29;0.43]	0.76 [0.55;0.97]	1.82 [0.98;2.67]

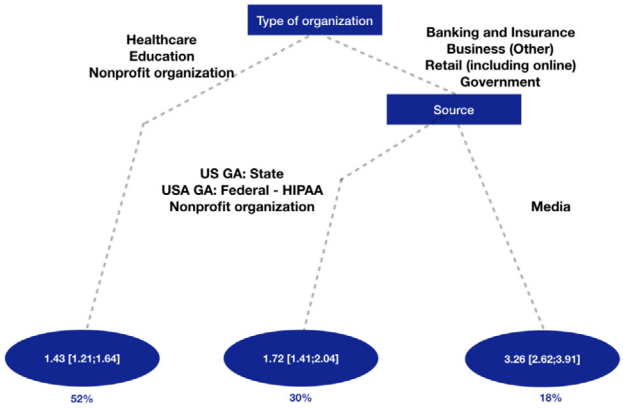

Fig. 4. Tree obtained from the CART algorithm based on the Generalized Pareto log-likelihood splitting rule (fitted on the observations exceeding the threshold u). For each leaf, the estimates of γ and their 95% confidence intervals are given.

Fig. 4 shows the obtained GP tree (fitted using the library `rpart` in R, with the appropriate user defined loss function), and variable importance is evaluated in Table 8 (more details on the computation of variable importance can be found in the supplementary material, Section 4). The confidence intervals for the parameters estimates in each leaf are reported in Table 5. Goodness-of-fit for the different leaves is shown through quantile–quantile plots in Appendix A.2, see Fig. 7. Let us first note that the structure of the GP tree is quite different from the ones obtained from the central part of the distribution. The estimated values of the shape and scale parameters on each leaf have first to be compared to the values obtained if we fit a GP distribution to the whole set of observations greater than u. In this case, maximum likelihood estimation leads to $\widehat{\sigma} = 48\,243$ (the 95% confidence interval is [40 685; 55 802]) and $\widehat{\gamma} = 2.16$ (the 95% confidence interval is [1.96; 2.36]). The worst case scenario, corresponding to the leaf with shape estimate 3.26, is even worse than this benchmark. Yet, the two other leaves, representing 82% of the extreme events, are "lighter" (although still associated with a shape parameter greater than 1, that is such that the expectation is not finite).

Moreover, let us observe that the major part of these events corresponds to a shape parameter equal to 1.43, which is close to the estimate of the tail distribution index provided by Maillart and Sornette (2010).

Remark 4.2. The value $\widehat{\gamma} = 2.16$ obtained from the whole sample implies that $E[Y] = \infty$. This indicates that the quadratic based regression method may not only lack robustness, but leads to ill-defined estimates (since the conditional expectation is not defined, at least for some leaves in the tree).

4.3. Comparison with generalized additive models

To compare the GP regression tree with competing extreme value regression approaches, we implemented the methodology developed by Chavez-Demoulin et al. (2015), that is using a Generalized Additive Model based on GP distributions for studying the tail (that is for $Y \geq u$). We will use the notation GAM GPD to refer to this technique. A short description of this technique is provided in the supplementary material (Section 3.1), along with estimates for the values of the model parameters.

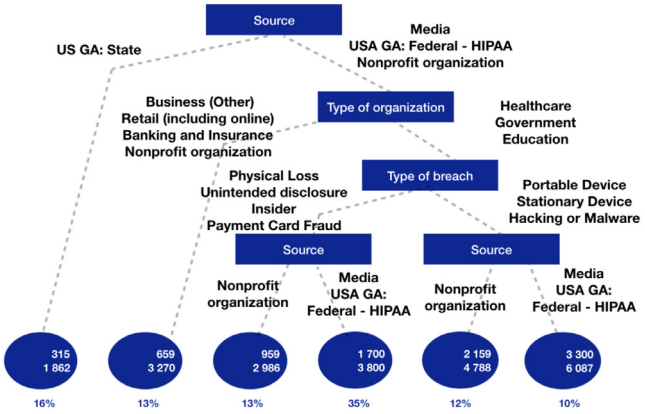

Fig. 5. Tree obtained from the CART algorithm based on the absolute loss fitted on the observations such that the variable "Number of Records" is less than u. For each leaf, the median (above) and the mean (below) are given.

Table 6 compares the fits of the GP tree with GAM GPD. Classical GP distribution fit (that is, using the PoT approach and without taking attention to the impact of the covariates) is also considered as a benchmark. We see that, in terms of log-likelihood and Akaike criterion (AIC), both regression techniques significantly improve this benchmark model, with a slightly better fit for the GP tree.

4.4. Global distribution analysis

The GP tree of Fig. 4 only provides an analysis of the distribution above a threshold u. If one wishes a global distribution, one must combine this approach with an analysis of the central part of the distribution. On the other hand, the analysis of Section 4.1 provides such a global analysis, but without taking the tail into account. Moreover, going back to the trees of Fig. 3, one can notice that, in each leaf, there is a significant difference between the value of the mean and the value of the median, as it is the case in the global set of observations (see Section 2.3). This invites us to look at a regression tree computed using the same method as in Section 4.1 (using absolute loss since less sensitive to large observations, the tree obtained via quadratic loss is shown in the supplementary material) but only on observations smaller than the threshold u. To summarize, observations are cut in two parts: observations with $Y \leq u$ are fitted using a regression tree based on absolute loss, while observations larger than u are fitted using the GP tree of Section 4.2.

This leads to the regression tree of Fig. 5. We see that the gap between the empirical median and the empirical mean in each leaf has been drastically reduced. On the other hand, the tree has a different structure than the one obtained from the global set of observations in Fig. 3, which shows that the presence of extreme values influences the obtained clusters.

A log-normal distribution (truncated by u) is fitted on the leaves of the absolute tree. The corresponding parameters are listed in Table 7.

To obtain the global distribution of the variable $Y =$ "Number of records", the combination of the results of the trees from Figs. 5 and 4 and Table 7 is done in the following way. We consider that the conditional distribution of Y is a mixture variable with same distribution as $\delta Z_1 + (1-\delta)Z_2$, where:

Table 6
Comparison of extreme value theory methodologies.

	Covariates used for σ	Covariates used for γ	LL	AIC
GP distribution	–	–	−2122	4249
GPD GAM	Organization and source	Date and organization	−2031	4098
GP tree	Type of organization and source	Type of organization and source	−2024	4072

Table 7
Truncated log normal parameters estimated by the absolute loss tree based on data below u. The parameter μ is the location parameter (expectation of the logarithm of the variable) and σ the scale parameter (standard deviation of the logarithm of the variable). Leaves are numerated from left to right according to the representation of the tree from Fig. 5. 95% intervals are given in brackets.

	Leaf 1	Leaf 2	Leaf 3
μ	5.62 [5.30;5.94]	6.79 [6.54;7.04]	6.95 [6.73;7.16]
σ	3.37 [3.13;3.61]	2.46 [2.26;2.65]	2.19 [2.02;2.35]
	Leaf 4	Leaf 5	Leaf 6
μ	7.64 [7.57;7.70]	8.20 [7.85;8.54]	8.72 [8.36;9.07]
σ	1.31 [1.26;1.36]	2.27 [2.04;2.49]	1.91 [1.69;2.13]

Table 8
Variable importance for the absolute tree of Fig. 5 and for the Generalized Pareto tree of Fig. 4 (in %).

	Source	Type of breach	Type of organization	Year
Central part tree	47	17	18	17
Tail part tree	35	–	48	17

- δ is a Bernoulli random variable independent from $\mathbf{X}$, and $p = P(\delta = 1)$ is the probability for an observation Y_i to be smaller than the threshold u;
- $Z_1|\mathbf{X} = \mathbf{x}$ has a distribution given by the absolute tree of Fig. 5 (where each leaf is associated with a truncated log-normal distribution determined by the parameters of Table 7);
- $Z_2|\mathbf{X} = \mathbf{x}$ has a distribution given by the GP tree of Fig. 4;
- δ is independent from (Z_1, Z_2) and Z_1 and Z_2 are independent conditionally to $\mathbf{X}$.

Let us recall that our estimate for p, in the PRC case, is the proportion of observations whose number of records is smaller than u, that is 0.84.

To complete this section, Table 8 reports the variable importance for both trees involved in this scheme. This confirms the relevance of separating the tail from the center of the distribution, since the variables driving the tail are different (at least in terms of hierarchy) from the ones driving the center.

4.5. Insurability of cyber risk

If we focus only on the tail of the distribution, the model fitted by the GP regression tree induces a mixture of three GP distributions for the unconditional distribution of Y. The advantage, compared to fitting a single GP distribution to all data larger than u, is that the tail index that the resulting shape index tends to be too pessimistic. Theoretically speaking, the tail index estimation of the global distribution should converge towards the worst tail index of the elements of the GP mixture. The GP tree technique presents the advantage to allow identification of some groups of claims that are still associated with a heavy tail behavior, but with more moderate consequences (in our example, all three leaves of the tree in Fig. 4 correspond to an infinite expectation, but let us recall that we are working with a proxy variable for the real amount of a claim). Hence we argue that using such techniques on more elaborate insurance databases can be a valuable tool to identify which types of cyber risks should be excluded from the policies (if the insurance company is unable to manage it), and potentially be used to reduce the premium if the insured population is associated with a lower risk.

5. An illustration on virtual cyber portfolios

The statistical approach performed in Section 4 is done using all the covariates present in the public PRC database. The aim is to achieve the best possible understanding of what drives the severity of cyber events. On the other hand, if one wishes to combine this analysis with an insurance perspective, an adaptation has to be made. It is the purpose of the present section to explain how this can be done. The question of coupling public database with intern information (from the history of the portfolio) is indeed fundamental in the context of cyber insurance, due to the lack of experience on the risk for many companies.

In this paper, we only address how to use the GP regression trees to project the result of a cyber insurance portfolio. In a real-life situation, this has to be combined with more reliable (but poorer) intern data. We perform simulations on four portfolios of 1000 policies, where each portfolio is composed of policyholders coming from only one of the following sectors of activities: BSF, BSO, BSR, or MED. The simulations use the different models we fitted on data. Nevertheless, the severity analysis we performed in Section 4 must be completed by three additional assumptions to produce an evaluation of the cost:

1. a transformation f that maps a number of records Y to a financial loss $f(Y)$;
2. a frequency analysis to model the occurrence of cyber claims, that is a distribution for N_i = number of incidents for the ith policyholder within 1 year;
3. once a claim has occurred, a probability distribution to determine the type of incident: indeed, since the type of breach has been seen to have a significant impact on the distribution of the claim size, we need to distinguish between these different categories of claims.

The total loss of the portfolio is then

$$S = \sum_{j=1}^{1000} \sum_{j=1}^{N_i} f(Y_{i,j}),$$

where $(Y_{i,j})_{1\le i\le n, 1\le j\le N_i}$ are the number of records for the claims of policyholder i (the number of records are supposed independent from N_i in this simple model). The distribution of S is then deduced from the points 1 to 3 above. In Sections 5.1 to 5.3 , we address successively each of these points. We then explain the simulation procedures we use to evaluate the total loss of each portfolio in Section 5.4.

5.1. Loss quantification of a data breach

Jacobs (2014) provided a model to transform a volume of data breach Y into a financial loss $L = f(Y)$. This model, which has also been used in Eling and Loperfido (2017), is based on data from Ponemon Institute LLC used Cost of Data Breach (CODB) reports of 2013 and 2014. The formula is

$$\log(L) = 7.68 + 0.76 \log(Y). \tag{5.1}$$

Table 9
Data breaches used to calibrate Formula (5.2): the costs of moderate breaches have been computed using Formula (5.1); the mega breaches are the only two communicated in CODB 2018.

	Moderate breaches		Mega breaches	
Number of records	10 000	100 000	1 000 000	50 000 000
Costs (in $)	2 373 458	13 657 827	39 490 000	350 000 000
Costs per record (in $)	237	137	39	7

A limit for this formula and analysis is that, in 2014, data gathered by the Ponemon Institute LLC was restricted. Indeed, the highest observed data breach had a size of 100 000 records, far from the highest one of the actual PRC database (which is 3 billions). Hence we propose to use a modified version of (5.1), using additional information contained in the 2018 CODB report, in which, "for the first time, [one] attempt[s] to measure the cost of a data breach involving more than one million compromised records, or what [one] refer[s] to as a mega breach".

Since only two costs of mega breaches are publicly available in the 2018 CODB report, we performed a rough fit of a linear relationship between $\log L$ and $\log Y$, based on four points detailed in Table 9. These four points are the two mega breaches, and two artificial points obtained, for moderate breaches, by the application of Formula (5.1). This presents the advantage to take Formula (5.1) into account and benefit from the fact that it has been calibrated on a large (non public) database, while using the additional information on mega breaches.

This leads to the following formula that will be used in our loss quantification,

$$\log(L) = 9.59 + 0.57 \log(Y). \tag{5.2}$$

The difference between the results of Formulas (5.1) and (5.2) is shown in the supplementary material (Section 3.2). The results are relatively close for the most part of the events contained in the PRC database, but for the largest ones, the difference becomes significant (with Formula (5.2) leading to smaller costs).

Clearly, we do not claim that Formula (5.2) is accurate for the association of a financial loss to the number of records. Our purpose is only to have a rough approximation of it. From the (public) data we have at our disposal, there is no way to pretend one is able to perform this evaluation with a good statistical precision. In practice, based on real loss data, the analysis that we provide can be seen as a rough benchmark that clearly needs to be improved by the use of more precise information.

Let us also note that Romanosky (2016) also studied the cost of data breaches using a private database gathering cyber events and associated losses. However, the obtained calibration requires information which is unavailable in the database used in this paper (but should be known from an insurance company when dealing with a real portfolio).

Remark 5.1. The GP regression tree of Fig. 4 has been done on the variable Y and not on the loss variable $f(Y)$. This choice has been done because we wanted to focus on the most reliable data, while Formula (5.2) is an approximation. However, the shape parameter of the GP distribution of $f(Y)$ can be easily deduced. Let us recall that this parameter is of most importance, since it gives us the decay of the survival function of $f(Y)$ (if this parameter is larger or equal to 1, $f(Y)$ has no expectation, and hence can be considered as "non-insurable" in a simplified vision of the problem). If $P(Y \geq y) \sim Cy^{-1/\gamma}$, where $\gamma > 0$ is the shape parameter of Y and C is a constant, considering $f(y) = \exp(\alpha + \beta \log y)$ leads to

$$P(f(Y) \geq z) = P\left(Y \geq \exp\left(\frac{\log z - \alpha}{\beta}\right)\right) \sim C \times \exp\left(-\frac{\alpha}{\beta\gamma}\right) z^{-\frac{1}{\beta\gamma}}.$$

Hence, the shape parameter of $f(Y)$ is $\beta\gamma$. In (5.2), $\beta = 0.57$. Hence, the three leaves of the tree of Fig. 4 have respective shape parameters 0.82, 0.98, 1.86. If we do not separate our claims into these three classes of risk, the shape parameters would have been $0.57 \times 2.16 = 1.23$. All of these numerical results should be taken carefully: the question of insurability is not so simple as determining if a GP shape parameter is smaller than one or not (and let us observe that, with Formula (5.1), all shapes parameters would have been greater than 1), but it still shows the importance to distinguish tail behaviors depending on the covariates in order to identify more clearly which type of risks can be managed and which cannot.

5.2. Frequency analysis

To provide an insurance pricing methodology, estimation of the annual frequency of claims is mandatory. The PRC database is not adequate to estimate this quantity rigorously. Nevertheless, we present here a possible way to roughly evaluate this frequency. This seems important for, at least, two reasons: (1) we want to provide an order of magnitude for the cost of cyber contracts; (2) even for an insurance company with a cyber portfolio, it is likely that frequency would be poorly estimated only based on internal historical data: since the risk is new, the number of reported claims would be too small to perform an accurate estimation. Hence, we believe that the combination of these information with external information – including public databases like PRC – is essential to improve the evaluation of the risk.

An important issue with the PRC database is the lack of knowledge of the exposure to the risk. Typically, it is impossible to know from such data which part of the increase of reported claims along time is caused by an evolution of the risk, and which is caused by an instability in the way the database is fed. This can be seen, for example, from Fig. 2. For example, the choice of PRC to stop gathering data breaches revealed by nonprofit organizations as from 2013 and a peak of data released by the media between 2015 and 2016 may be observed. Moreover, Bisogni et al. (2017) claim that the majority of data breaches proves to be unreported.

Hence, we propose two heuristics to derive a frequency analysis from the PRC database:

(H1) we restrain ourselves to companies listed in the PRC database that have been breached at least twice according to the PRC database. Since almost 90% of companies listed in PRC are reported only once, one may fear that the information about them is not completely reliable. On the other hand, a repeatedly reported company has more chances to have its major breaches exhaustively reported in the database. The frequency is estimated from companies that have been breached multiple times, considering that we are dealing with 1-truncated data.

(H2) we restrain ourselves to companies quoted on the New York Stock Exchange (NYSE) that have been breached at least once according to the PRC database. This idea has first been suggested by Wheatley et al. (2016). Here, 94% of companies of NYSE are absent from the PRC database. Assuming that no breach occurred for all of them seems unrealistic and would considerably lower the frequency: their absence is more likely due to the fact that these breaches have not been reported by the processes of PRC. If a company is associated with 0 claim, it is therefore not certain that this absence from PRC is really caused by the

absence of a breach, or by the fact that this entity was not in the scope of PRC. Hence, we consider that data from these companies is 0-truncated.

In the following, we consider two portfolios corresponding either to case (H1) (PRC portfolio) or case (H2) (NYSE portfolio). A summary of descriptive count statistics for both portfolios is given in the supplementary material (Section 1.2).

To model the number of claims striking a portfolio, we fit a Generalized Linear Model (GLM), considering the sector of activity as a covariate. For the PRC portfolio, we consider the sectors BSF, BSO, BSR, EDU, GOV and MED only, deliberately excluding the NGO sector because of lack of data on this category. The NYSE portfolio does not contain companies from sectors EDU, GOV and NGO. We consider two cases: a GLM based on a Poisson distribution, and one on a geometric distribution (for all $k \geq 0$, the probability that a geometric distribution is k is $p(1-p)^k$, where p is a parameter taking values in $(0, 1)$). More precisely, these two models can be written as

$$g(E[N|\mathbf{X}]) = \mathbf{X}\beta \text{ , with } \begin{cases} N \sim \mathcal{P}(\lambda) & \text{and} \quad g(x) = \log(x). \\ \quad \text{or} & \\ N \sim \mathcal{G}(p) & \text{and} \quad g(x) = \log\left(\frac{x}{1-x}\right). \end{cases} \tag{5.3}$$

On the PRC database, fitting indicators can be found in the supplementary material, Section 1.2, showing that the geometric GLM seems more adequate than the Poisson one.

5.3. Type of incident

The frequency of claims determined in Section 5.2 does not include the variety of cyber incidents: it is a global frequency, regardless the type of claims. If we want to simulate the impact on our insurance portfolio, we must simulate also a type of event once an event occurred. In our simulation scheme, the idea is to use a multinomial random variable to draw the type of event. We assume that the parameters only depend on the type of activity of the victim (which is the only variable available for the insurance company, among those present in the regression trees).

Let S denote an indicator of the sector of activity, and M denote the type of breach. We can write

$$P(M = m|S = s) = \frac{e^{\beta_{s,0}+\beta_{s,m}}}{\sum_{m'} e^{\beta_{s,0}+\beta_{s,m'}}},$$

where $\beta_{s,0}$ corresponds to a reference category (here we took as reference category the incidents for which the type of organization is unknown).

In full generality, this would lead to the estimation of a large number of coefficients, with few data to calibrate them. To reduce the number of parameters, we used a LASSO dimension reduction technique (the log-likelihood is penalized using a L^1-penalty on the fitted coefficients $\beta_{s,m}$, with a parameter tuned through 10-fold cross validation, see e.g. Tibshirani (1996)). The matrix of fitted coefficients can be found in Section 1.2 of the supplementary material.

5.4. Results

We now show the impact of these models on our virtual portfolios. We recall that we consider four portfolios with 1000 policyholders, each composed of entities of a single category among BSF, BSR, EDU and MED. The losses of each portfolio are simulated according to the following procedure:

1. For each policyholder, we simulate a number of claims under the geometric model of Section 5.2.
2. For each claim, we determine which type of incident has caused the claim from the multinomial distribution of Section 5.3.
3. We simulate the number of records accordingly to four methodologies, assuming, in each case, that the distribution is the same as the one given by one single source of information (US GA State or Media):
   - Clustering: we use the tree obtained with the absolute loss from Fig. 3 to determine risk classes. The distribution of the claims in each leaf of the tree is considered as log-normal using the following set of parameters: $(7.56, 2.66)$ for leaf 1, $(8.88, 3.11)$ for leaf 2, $(8.19, 3.25)$ for leaf 3, $(12.67, 4.19)$ for leaf 4, $(13.47, 4.42)$ for leaf 5 (leaves numerated from left to right, first parameter (resp. second) is the expectation (standard deviation) of the logarithm of the log-normal variable).
   - GP regression tree: we use the combination of the trees of Figs. 4 and 5, as described in Section 4.4. For the central part, log-normal distributions are used, with the fitted parameters of Table 7.
   - GAM GPD: for comparison, we considered the approach developed by Chavez-Demoulin et al. (2015), which is exposed in detail in the supplementary material.
4. We use (5.2) to convert this number of records into a financial loss (a comparison with the use of (5.1) can be found in the supplementary material, Section 3.2).

Results of these simulation procedures are summarized in Table 10. Let us first remark that, regarding the clustering approach based on a single tree (built using absolute loss), the difference between the median quantile $q_{0.5}$ and $q_{0.9}$ is much smaller than for the two other approaches. This was expected, due to the use of a GP distribution to model the tail for the last two models. On the other hand, the order of magnitude of all tree-based methods is much smaller than for the GAM GPD approach, although all sectors generally keep the same ranking in terms of severity from one model to another.

It is also interesting to notice that, in our tree-based methods, separating the tail from the central part of the distribution pushes up the value of the median quantile of the loss (of course the push on the $q_{0.9}$ quantile was expected, because a specific model has been done on the tail of the distribution). Through this phenomenon, one can observe once again the benefit of separating "extreme" observations from the others: their presence in the sample distorts the fitting of the tree and of the log-normal distributions in the leaves, even though we chose a relatively stable procedure through the use of the absolute loss.

6. Conclusion

In this paper, we applied regression trees as a valuable tool for analyzing cyber claims. For reproducibility purpose, all models have been fitted on a public database, the PRC database. Although this database, widely used in the literature, presents serious drawbacks and inconsistencies as we discussed it intensively throughout the paper, the methodology can be easily extended to other private databases, and several conclusions we draw can be generalized. The first observation is the heterogeneity of cyber events in terms of severity. This is, of course, a well known fact. However the regression tree approaches allow a clarification and a quantification of some characteristics that create this heterogeneity. For example, some sectors of activity (Healthcare, Education, Nonprofit organization) seem to have significantly

Table 10
Comparison of median and 0.9-quantile depending on the methodology used (through columns) and the additional hypothesis regarding the source of information and the frequency portfolio (through lines). Quantities are given in million of dollars and have been obtained after 10 000 simulations.

Modeling methodology			Clustering		GAM GPD		GP tree	
Source	Frequency	Organization	$q_{0.5}$	$q_{0.9}$	$q_{0.5}$	$q_{0.9}$	$q_{0.5}$	$q_{0.9}$
US GA State	(H1)	BSF	286	424	2561	58 764	433	1 320
		BSO	363	522	4806	156 055	572	1 755
		BSR	235	358	1851	41 563	351	1 074
		MED	305	447	497	2446	284	574
	(H2)	BSF	342	498	3426	75 515	532	1 558
		BSO	202	317	1445	41 952	292	916
		BSR	244	374	2077	46 042	365	1 140
		MED	223	345	332	1 695	203	434
Media	(H1)	BSF	884	1 491	3651	82 832	3455	78 978
		BSO	23 686	62 795	6857	223 796	5602	123 629
		BSR	12 236	38 421	2695	58 175	2511	51 070
		MED	942	1 556	744	3 744	348	648
	(H2)	BSF	1 056	1 747	5110	105 458	4604	106 816
		BSO	11 860	37 837	2086	60 200	1900	40 166
		BSR	13 069	40 128	3005	66 493	2698	55 113
		MED	683	1 204	508	2 478	252	475

lighter tail than the others (see Fig. 4). Moreover, it appears that the central part of the distribution does not behave like the tail—in the sense that the impact of the covariates on this right tail does not seem to be identical to what we can observe on the core of the distribution. Among the categories of targeted organizations which are associated with the lightest tail, one can observe that Healthcare and Education are mainly affected to the right-hand side of the tree describing the central part of the distribution (see Fig. 5), meaning that the severity of claims striking them is, in average, higher. This shows the importance of a separate analysis of "typical" claims, and "extreme" ones. This dissemblance between what drives the center and what drives the tail of the distribution is not specific to cyber, but is probably reinforced by the various profiles of cyber criminals (home-made attacks versus larger scale criminal organizations). Finally, the results on our analysis based on GP trees reveal that there may be a significant operational impact if we pay attention to clustering types of "extreme" claims.

We want to emphasize this last point: our analysis tends to acknowledge that a classical peaks over threshold approach (that is ignoring the influence of covariates on the shape parameter) leads to considering the whole tail of the distribution as too heavy. On the other hand, identifying some clusters for extreme events could at least be interesting for designing appropriate risk management strategies for some type of claims. Our purpose is not to draw a clear line between which criterion should be used to exclude or not some type of claims from the perimeter of insurance contracts, our data are not accurate enough to elaborate precise recommendations. Nevertheless we strongly advocate for developing such regression approaches to better understand and manage extreme claims.

Regarding estimation of the frequency, the approach we took is very approximative due to the lack of consistency of data. Nevertheless, this analysis seemed to us essential in order to show how a whole insurance pricing and reserving methodology can be developed. Moreover, due to the relative novelty of the risk, the information gathered by insurance companies are sufficiently recent to take advantage on additional sources of (public) data. Hence we believe that a promising field of research is to find a proper way for companies to combine internal data and these external sources, provided that a rigorous statistical analysis has first identified and corrected their biases.

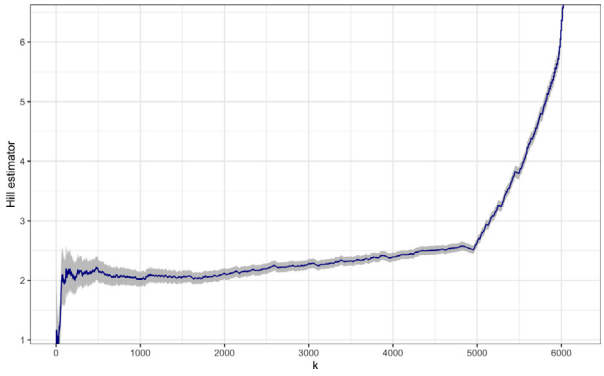

Fig. 6. Hill plot for the number of records.

Acknowledgment

The authors acknowledge funding from the project *Cyber Risk Insurance: actuarial modeling*, Joint Research Initiative under the aegis of Risk Foundation, France, with partnership of AXA, AXA GRM, ENSAE and Sorbonne Université.

R codes: The code is made publicly available at https://bitbucket.org/sebastien_farkas/cyber_claim_analysis_gpd_regression_trees/

Appendix A

A.1. Hill plot

Fig. 6 shows the Hill plot for the number of records (see Resnick, 2007, pp 85–89 for more details on Hill plots). From the shape of the curve, we chose $u = 27\,999$ (which corresponds to a stabilization of the Hill plot) which leads to keep the 1000 highest observations (around 16% of the total number of breaches).

A.2. Goodness of fit for GP tree and comparison tests

Fig. 7 gathers quantile–quantile plots corresponding to each leaf of our final GP tree of Fig. 4. After fitting the GP tree of Fig. 4, we check that the three clusters can be considered dissimilar enough so that they cannot be grouped into a single one (which

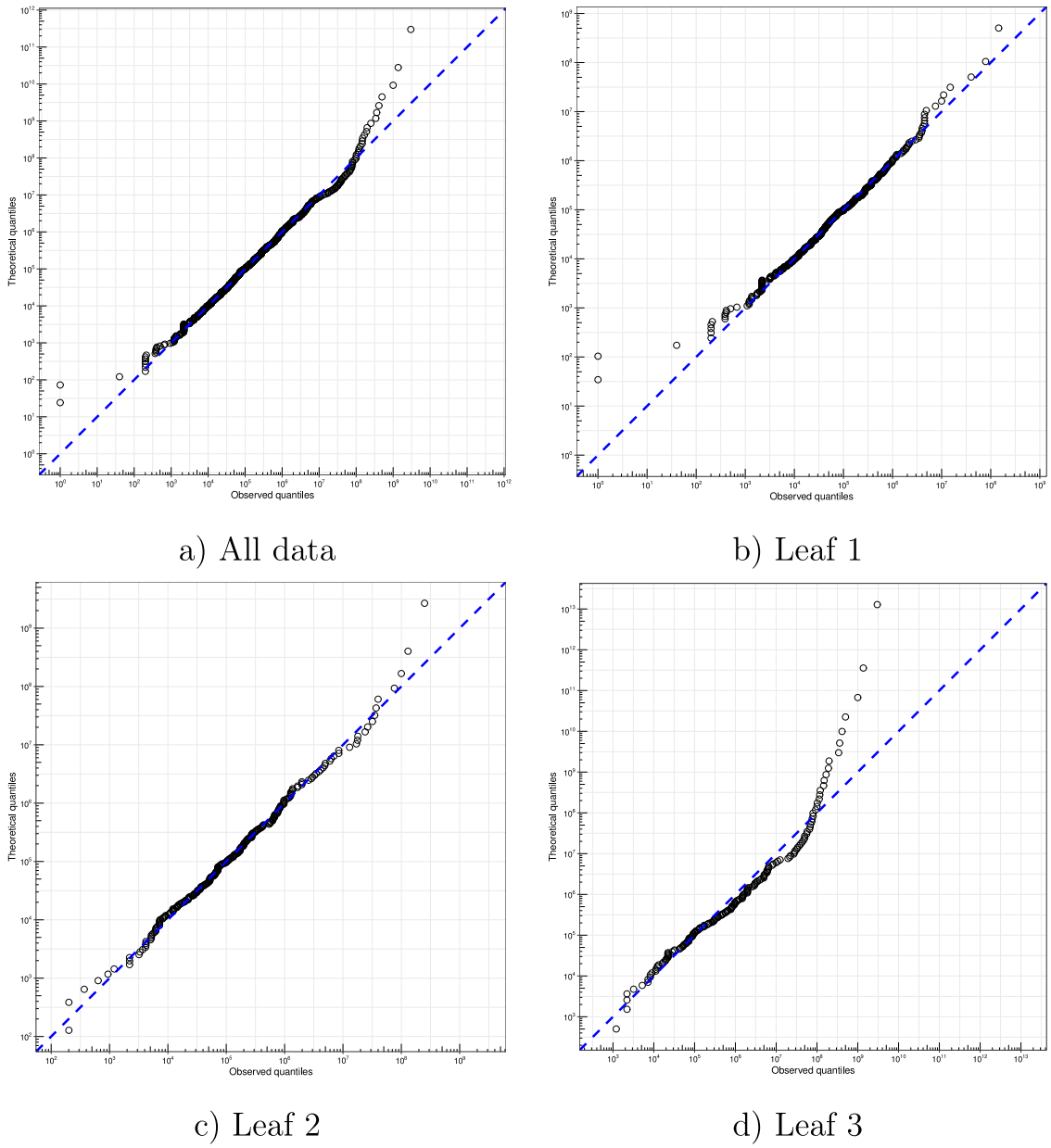

Fig. 7. Quantile plots of Generalized Pareto distribution fits on all observations exceeding the threshold u (Figure (a)) and on samples of each of the 3 leaves of the Generalized Pareto tree of Fig. 4 (Figures (b), (c) and (d)).

Table 11
Statistics and p-values of the two sample Kolmogorov–Smirnov tests computing on samples of the leaves of the Generalized Pareto tree, two by two.

Leaf of the first sample	Leaf of the second sample	KS statistic	KS p-value
1	2	0.22	1.94×10^{-8}
1	3	0.44	$<2.2 \times 10^{-16}$
2	3	0.32	8.07×10^{-11}

would considerably simplify the study by performing standard extreme value analysis methodologies, i.e. without taking covariates into account). A Kolmogorov–Smirnov test (see Section 6.9 in Lehmann and Romano, 2006) has been used to compare the empirical distribution of each couple of leaves. The p-values are given in Table 11. They suggest a rejection of the null hypothesis. We also considered a likelihood ratio test (see Section 12.4.4 in Lehmann and Romano, 2006) which uses the particular structure of GP distribution. This consists in computing the difference between the log-likelihood obtained from the tree to the log-likelihood obtained when a single GP distribution is fitted to the whole set of observations. The value of this test statistic is 169.8, leading to a p-value lower than 2.2×10^{-16}, which once again suggests a significant improvement of the fit.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.insmatheco.2021.02.009.

References

Balkema, A.A., De Haan, L., 1974. Residual life time at great age. Ann. Probab. 2 (5), 792–804.
Beirlant, J., Goegebeur, Y., 2003. Regression with response distributions of Pareto-type. Comput. Statist. Data Anal. 42 (4), 595–619.

Beirlant, J., Goegebeur, Y., 2004. Local polynomial maximum likelihood estimation for Pareto-type distributions. J. Multivariate Anal. 89 (1), 97–118.
Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.L., 2004. Statistics of Extremes: Theory and Applications. John Wiley & Sons Ltd., Chischester.
Biener, C., Eling, M., Wirfs, J.H., 2015. Insurability of cyber risk: An empirical analysis. Geneva Pap. Risk Insur. Issues Pract. 40 (1), 131–158.
Bisogni, F., Asghari, H., Van Eeten, M.J., 2017. Estimating the size of the iceberg from its tip: An investigation into unreported data breach notifications. In: Proceedings of 16th Annual Workshop on the Economics of Information Security 2017.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. In: Wadsworth Statistics/Probability Series, Wadsworth Advanced Books and Software, Belmont, CA.
Chaudhuri, P., Loh, W.-Y., 2002. Nonparametric estimation of conditional quantiles using quantile regression trees. Bernoulli 8 (5), 561–576.
Chavez-Demoulin, V., Embrechts, P., Hofert, M., 2015. An extreme value approach for modeling operational risk losses depending on covariates. J. Risk Insurance 83 (3), 735–776.
Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer Verlag, London.
Databreaches.net. (n.d.), 2020. Databreaches reporting. (https://www.databreaches.net/about/).
Davison, A.C., Smith, R.L., 1990. Models for exceedances over high thresholds. J. R. Stat. Soc. Ser. B Stat. Methodol. 52 (3), 393–425.
De Haan, L., Ferreira, A., 2007. Extreme Value Theory: An Introduction. Springer, New York.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology 81 (11), 3178–3192.
Edwards, B., Hofmeyr, S., Forrest, S., 2016. Hype and heavy tails: A closer look at data breaches. J. Cybersecur. 2 (1), 3–14.
Eling, M., Loperfido, N., 2017. Data breaches: Goodness of fit, pricing, and risk measurement. Insurance Math. Econom. 75, 126–136.
Eling, M., Schnell, W., 2016. What do we know about cyber risk and cyber risk insurance?. J. Risk Finance 17 (5), 474–491.
Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modelling Extremal Events. In: Applications of Mathematics, vol. 33, Springer-Verlag, Berlin.
European Insurance and Occupational Pensions Authority (EIOPA), 2019. Cyber risk for insurers - challenges and opportunities. Retrieved from https://www.eiopa.europa.eu/sites/default/files/publications/reports/eiopa_cyber_risk_for_insurers_sept2019.pdf.
Fahrenwaldt, M.A., Weber, S., Weske, K., 2018. Pricing of cyber insurance contracts in a network model. Astin Bull. 48 (3), 1175–1218.
Gey, S., Nédélec, E., 2005. Model selection for CART regression trees. IEEE Trans. Inf. Theory 51 (2), 658–670.
González, C., Mira-McWilliams, J., Juárez, I., 2015. Important variable assessment and electricity price forecasting based on regression tree models: classification and regression trees, Bagging and Random Forests. IET Gener. Transm. Distrib. 9 (11), 1120–1128.
Insua, D.R., Vieira, A.C., Rubio, J.A., Pieters, W., Labunets, K., Rasines, D.G., 2021. An adversarial risk analysis framework for cybersecurity. Risk Anal. 41 (1), 16–36.
Jacobs, J., 2014. Analyzing ponemon cost of data breach. Retrieved from http://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/.
Katz, R.W., Parlange, M.B., Naveau, P., 2002. Statistics of extremes in hydrology. Adv. Water Resour. 25 (8–12), 1287–1304.
Lehmann, E.L., Romano, J.P., 2006. Testing Statistical Hypotheses. Springer, New York.
Loh, W.-Y., 2011. Classification and regression trees. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1 (1), 14–23.
Loh, W.-Y., 2014. Fifty years of classification and regression trees. Internat. Statist. Rev. 82 (3), 329–348.
Lopez, O., Milhaud, X., Thérond, P.-E., 2016. Tree-based censored regression with applications in insurance. Electron. J. Stat. 10 (2), 2685–2716.
Maillart, T., Sornette, D., 2010. Heavy-tailed distribution of cyber-risks. Eur. Phys. J. B 75 (3), 357–364.
Marotta, A., Martinelli, F., Nanni, S., Orlando, A., Yautsiukhin, A., 2017. Cyber-insurance survey. Comp. Sci. Rev. 24, 35–61.
Matthews, D., 2019. Report on the Cybersecurity Insurance and Identity Theft Coverage Supplement. NAIC, Retrieved from https://content.naic.org/sites/default/files/inline-files/Cyber_Supplement_2019_Report_Final_1.pdf.
Pickands, J., 1975. Statistical inference using extreme order statistics. Ann. Statist. 3 (1), 119–131.
Ponemon Institute LLC, 2018. Cost of a data breach study: global overview. (https://www.ibm.com/downloads/cas/AEJYBPWA).
Privacy Rights Clearinghouse. 2019. Retrieved from https://privacyrights.org/data-breaches.
Resnick, S.I., 2007. Heavy-TAil Phenomena: Probabilistic and Statistical Modeling. Springer, New York.
Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818.
Romanosky, S., 2016. Examining the costs and causes of cyber incidents. J. Cybersecur. 2 (2), 121–135.
State of California. (n.d.), 2020. California list of Data Security Breaches. Retrieved from https://oag.ca.gov/privacy/databreach/list.
Su, X., Wang, M., Fan, J., 2004. Maximum likelihood regression trees. J. Comput. Graph. Statist. 13 (3), 586–598.
Therneau, T., Clinic, M., 2019. User written splitting functions for RPART. Retrieved from https://cran.r-project.org/web/packages/rpart/vignettes/usercode.pdf.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
U.S. HHS department. (n.d.-a), 2020a. Retrieved from https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf.
U.S. HHS department. (n.d.-b), 2020b. Retrieved from https://www.hhs.gov/hipaa/for-professionals/breach-notification/index.html.
Wheatley, S., Maillart, T., Sornette, D., 2016. The extreme risk of personal data breaches and the erosion of privacy. Eur. Phys. J. B 89 (7).