Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance

Sébastien Farkas, Olivier Lopez, Maud Thomas

Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance

Sébastien FARKAS[1], Olivier LOPEZ[1], Maud THOMAS[1]

January 20, 2020

Abstract

With the rise of the cyber insurance market, there is a need of a better quantification of the economic impact of this new risk. Due to the relatively poor quality and consistency of databases on cyber events, and because of 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 based on regression trees to analyze cyber claims 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 the regression tree approach. Combined to an evaluation of the frequency, our procedure allows computations of central scenarios and extreme loss quantiles for a cyber portfolio. Finally, we illustrate on a public database.

Key words: Cyber insurance; Extreme value analysis; Regression Trees; Generalized Pareto Distribution.

Short title: Cyber-risk through GPD regression trees

[1] Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France, E-mails: sebastien.farkas@sorbonne-universite.fr, olivier.lopez@sorbonne-universite.fr, maud.thomas@sorbonne-universite.fr

1 Introduction

Cyber risk is a natural consequence of the digital transformation. Digital technologies induce new vulnerabilities for economic actors, with a fast 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 spectacularly growing, risk analysis faces a lack of consistent and reliable data in a context where the amount of claims is particularly volatile. 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 between type of incidents or circumstances that can or can not be covered without endangering risk pooling.

Topics recently addressed in cyber-insurance are reviewed in Biener et al. [2015], Eling and Schnell [2016] and [Marotta et al., 2017]. Although most of them approach this challenge 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. [2019] gather expert judgments using an Adversarial Risk Analysis. Eling and Loperfido [2017] and Forrest et al. [2016] developed more established insurance modeling methods illustrated on the Privacy Rights Clearinghouse (PRC) database (available for public download at `https://www.privacyrights.org/copyright`). 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 data breached), making it valuable for insurance applications. On the other hand, 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 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. Juárez, 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. Advocating for the use of regression trees is also 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 described 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 advantage 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 to a claim. This piece of information is crucial from an insurance perspective: evaluation the risk associated with a policyholder requires to estimate the probability of being 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 affected by the breach. This number is correlated to the financial impact of the claim, which can be approximatively retrieved by a formula given in Jacobs [2014] which will 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 that are summarized in Section 2.3, and will motivate the use of the methodology we develop in the rest of the paper.

2.1 Description of the database

Privacy Rights Clearinghouse is a nonprofit organization founded in 1992 which aims at protecting privacy for US citizens. Especially, PRC has been maintaining 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 Americans. Among them, only 8298 are kept for our analysis, since we eliminated duplicated and/or inconsistent events (e.g. information on the targeted company is 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 (targeted company name, its activities, its localization). These variables and their modalities are summarized in Tables 1 to 3. Additional statistics are shown in Section 7.3.

As already mentioned, the financial loss resulting from an event is absent from the database. However, the number of records is a key element to measure the severity of the event, at least for a significant number of cases (6160 observations out of 8802, on which we will perform our severity analysis). Section 5.1 below shows how a projected financial loss can be estimated from the number of records, accordingly to an approach that has been developed by previous authors (see e.g. Eling and Loperfido [2017]). Note that the

Table 1: List of the available variables of 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

Table 3: List of the types of data breaches as labelled 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

number of records should not be interpreted as the number of individuals affected by the breach, for example in the situation where a web-user possesses multiple accounts.

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 in 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, see [U.S. HHS department, b]. Those notifications are reported online with free access on the breach portal [U.S. HHS department, a].

- US Government Agencies on the state level: since 2018, every state has had a specific legislation related to data breaches. Differences have been studied by Maddie Ladner [2018]. 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, but this is not systematic.

- Media: PRC organization monitors media to list data breaches that leads to extensive press coverage.

- Non profit organizations: the PRC database includes the data breaches reported by other non profit organizations than PRC, for instance [Databreaches.net].

While merging different sources of notifications increases the scope of the PRC chronology, it also introduces 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 Figures 1 and 2.

This makes the estimation from this database of a frequency of occurrence particularly difficult (fundamental for insurance pricing). Indeed, there is no way to determine which part of the increase of the number of claims is caused by the evolution of the threat, and which is caused by the data collecting process. This could also have an effect on the analysis of the severity of the events, which is our main concern in this paper.

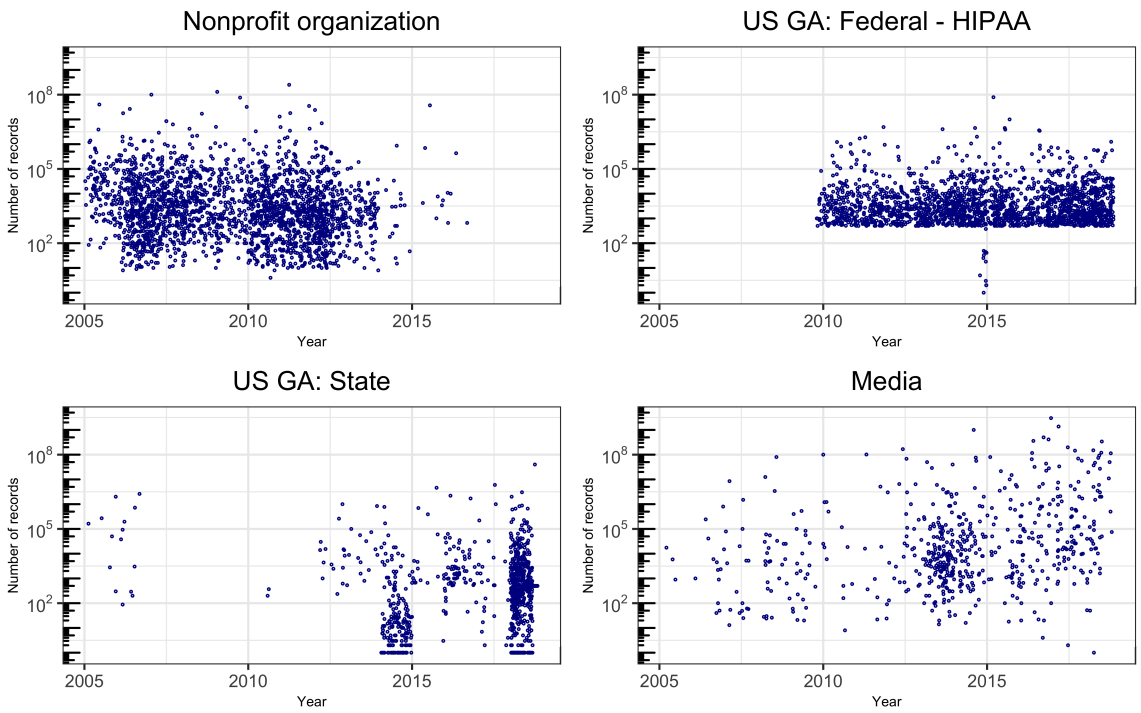

Figure 1: Data breaches listed in the PRC database through time and depending on the source of information.

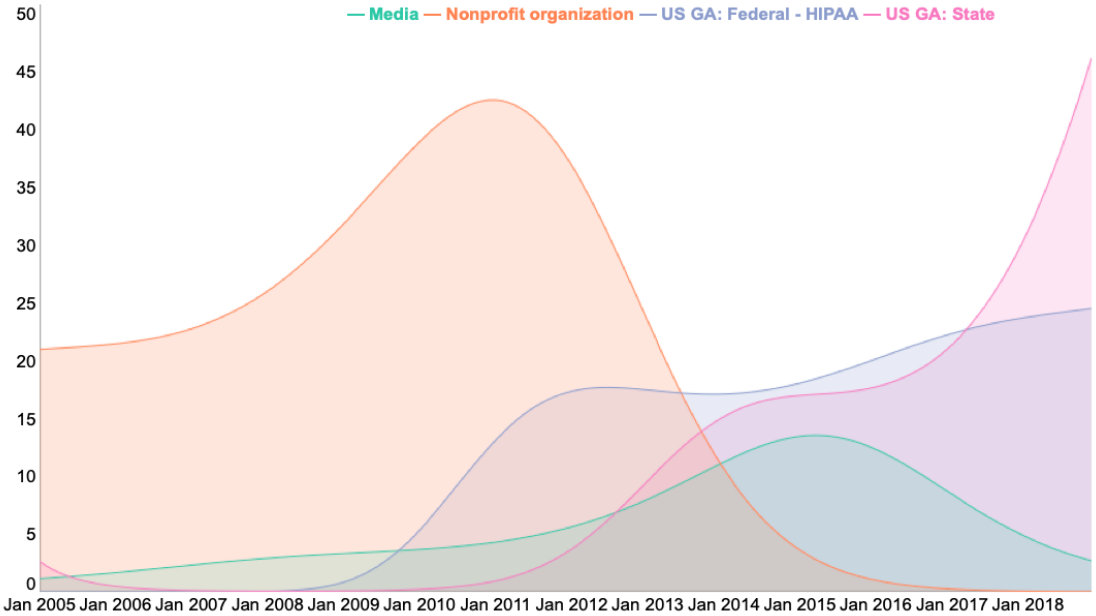

Figure 2: Evolution of the number of listed events by month depending on the source of information (smoothed curves).

2.3 Heterogeneity and inconsistencies in PRC database

As we already mentioned, the evolutions in the way different sources feed the database through time is of some concern in order to get a clear view on the frequency of cyber events. This evolution may also have impact on our main objective, which is analyzing the severity of these events. Indeed one may for example guess that cyber claims that were exposed by medias are more likely to be more "spectacular" (and hence more severe). In the same spirit, the fact that a legal source introduces a threshold under which the event is not necessarily reported creates a potential imbalance in the distribution. This intuition will be confirmed by our statistical modeling in Section 3.

Moreover, a short descriptive analysis of the severity variable ("number of records", see Table 4) shows that this variable is highly volatile: the severity of the worst data breach represents 27% of the total number of records affected by the totality of the data breaches. The severity of the top ten data breaches corresponds to 68% of the total severity and the severity of the top hundred data breaches to 97%. Furthermore, there is 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 risk classes which would be more homogeneous, and in which we could separate between the sources of information if they appear to be correlated with the severity of the claim. A few complementary statistics are reported in Table 20 in the Appendix section.

Table 4: Descriptive statistics for the variable "Number of records" depending on the source of information (first column). q_α denotes the empirical $\alpha-$quantile, that is such that $\alpha\%$ 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	2 000	10 891	70 000	300 000	3 000 000 000
US GA: Federal - HIPAA	1 949	84 358	981	2 300	8 009	28 440	75 015	78 800 000
US GA: State	888	89 377	20	4 010	2 403	18 000	63 826	40 000 000
Media	595	16 208 785	1 400	11 266	137 193	4 420 000	41 029 089	3 000 000 000
Nonprofit organization	2 309	422 623	380	2 000	14 000	86 333	247 200	250 000 000
Unknown	419	853 736	959	2 300	9 153	30 194	61 863	191 000 000

In view of performing this task, our idea is to rely on regression trees which are described in Section 3 below. They present the advantage to offer an automatic clustering,

without any a priori on the covariates present in the database and to confirm or infirm any intuition one may have on these characteristics (for example on the "Source" variable).

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 in 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, which are introduced in Section 3.2, are more promising tools to study 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 use three different 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 data into two more homogeneous classes. The procedure includes two phases: a "growing" phase through 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 $\sum_j R_j(\mathbf{x}) = 1$. 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).
Step k+1: Let $(R_1, ...R_{n_k})$ 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 $x_{j\star}^{(l)}$

to split the data, such that $x_{j\star}^{(l)} = \arg\max_{x^{(l)}} \Phi(R_j, x^{(l)})$, with

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

where

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

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

- 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 $\mathcal{R} = (R_j)_{j=1,\dots,s}$, 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 from the maximal tree a subtree that achieves 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 $\mathcal{R}^{\mathcal{S}} = (R_1^{\mathcal{S}}, \dots, R_{n_{\mathcal{S}}}^{\mathcal{S}})$ of cardinality $n_{\mathcal{S}}$. One then selects the subtree $\widehat{\mathcal{S}}(\alpha)$ that minimizes the criterion

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

among all subtrees of the maximal tree, where A is a positive constant. Hence, the trees with large numbers of leaves (i.e. of rules) are disadvantaged 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 data 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 a training or a 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 $\widehat{m}(\mathbf{x}) = m^{\widehat{\mathcal{S}}(\widehat{\alpha})}(\mathbf{x})$.

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 Distributions (GPD) naturally appear in the analysis of heavy-tailed variables. This motivates our GPD 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., 2013]. Given a series of independent and identically distributed observations $Y_1, Y_2, \ldots$ with an unknown survival function $\bar{F}$ (that is $\bar{F}(y) = P(Y_1 > y)$). 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

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

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. 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 toward a non-degenerate distribution, then the limit

distribution belongs to a parametric family called the Generalized Pareto distributions [see also Pickands, 1975].

In practice, the so-called Peaks over threshold method has been widely used since 1990 [see Davison and Smith, 1990]. It consists in choosing a high threshold u and fitting a Generalized Pareto distribution on the excesses above that threshold u. The estimation of the parameters σ and γ may be done by maximizing the Generalized Pareto 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.

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].

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 exponentially fast to 0. 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 a regression framework, for each value of the covariate $\mathbf{x}$, the conditional distribution of $Y|\mathbf{X} = \mathbf{x}$ is assumed to be heavy-tailed, but the parameters γ, σ (and the threshold u above which the GPD approximation seems satisfactory) depend on $\mathbf{x}$. More precisely, this means that equation (3.3) becomes

$$\lim_{u\to\infty} \sup_{y>0} |\overline{F}_u(y) - \overline{H}_{\sigma_u(\mathbf{x}),\gamma(\mathbf{x})}(y)| = 0 \tag{3.5}$$

To estimate the function $m(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$, we use a regression tree approach. The procedure of Section 3 is applied to the observations $(Y_i - u, \mathbf{X}_i)$ for which $Y_i \geq u$, using the Generalized Pareto log-likelihood as split function, that is

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

The threshold u is chosen large enough so that the Generalized Pareto approximation is correctly fitted to data (practical choice of this parameter will be discussed in Section 4.2, see also Remark 3.2 below). 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 et al., 1999] suppose a form of linearity. The more flexible nonparametric approaches, as in [Beirlant and Goegebeur, 2004] rely on smoothing techniques that require the covariate to be continuous. In [Chavez-Demoulin et al., 2015], the authors 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 GPD regression tree approach to their procedure in Section 4.4.

Remark 3.2 *In extreme value regression, the conditional version of* (3.4) *leads to the introduction of a threshold u that potentially depends on $\mathbf{x}$ on the event $\mathbf{X} = \mathbf{x}$. 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. 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 GPD trees. Section 4.3 shows how these two approaches can be combined to provide a global analysis of the distribution. Comparison with the fit of a GAM model as in Chavez-Demoulin et al. [2015] is shown in Section 4.4. A discussion on the insurability of cyber-risk—which, from a probabilist 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. 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, 2018], 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 Figure 3.

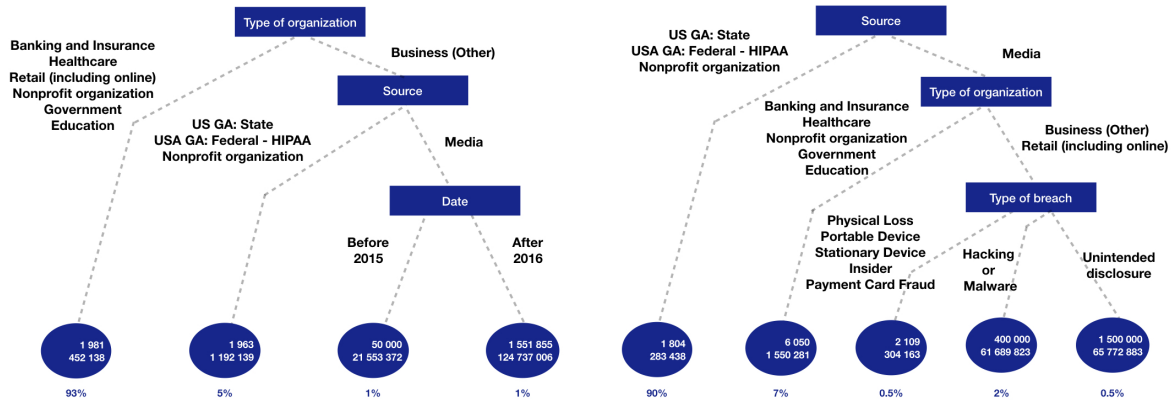

Figure 3: Trees obtained by the CART algorithm based on the quadratic (left-hand side) and the absolute (right-hand side) losses. For each leaf, the value of the empirical median (first line) and mean (second line) are given.

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 a 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.

Table 5 shows the estimation of the variable importance for quadratic and absolute trees. Variable importance is a common way to perform a ranking of the covariates in terms of their impact on the response, see Section 7.2.1. The picture is significantly different depending on the loss. The most important variable for the quadratic loss is the source, while it is only the fourth for the absolute loss.

Table 5: Variable importance of the regression trees obtained using the quadratic and absolute losses (in %).

	Source	Type of breach	Type of organization	Year
Quadratic loss	35	2	16	47
Absolute loss	6	56	27	11

The leaves of the trees determine clusters. If we want to get a distribution for the claim severity, one may fit a distribution on each leaf. In Table 6, we report the fitted parameters of a log-normal distribution on each leaf.

4.2 Tail part of the severity distribution

In view of applying the GPD regression tree approach of Section 3.2.2, our first task is to determine the threshold u above which the GPD approximation seems reasonable. This choice is made from the Hill plot [see Resnick, 2007, pp 85–89] for more details on Hill plots) in Figure 4. From the shape of the curve, we chose $u = 27\ 999$, 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.

Figure 5 shows the obtained GPD tree, and variable importance is evaluated in 8. The confidence intervals for the parameters estimates in each leaf are reported in Table 19 (see Section 7.3). Let us first note that the structure of the GPD tree is quite different from the ones obtained from the central part of the distribution. The values of the shape and scale parameters on each leaf have first to be compared to the values obtained if we fit a GPD 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

Table 6: Log normal estimated parameters on the leafs of the tree based on absolute loss. 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 trees from Figure 3.

Absolute	μ	σ
Leaf 1	7.56 [7.49;7.63] (0.14)	2.66 [2.61;2.71] (0.10)
Leaf 2	8.88 [8.58;9.18] (0.60)	3.11 [2.90;3.32] (0.42)
Leaf 3	8.19 [7.03;9.36] (2.33)	3.25 [2.43;4.08] (1.65)
Leaf 4	12.67 [11.86;13.48] (1.62)	4.19 [3.62;4.76] (1.14)
Leaf 5	13.47 [12.06;14.86] (2.8)	4.42 [3.44;5.40] (0.96)
Quadratic	μ	σ
Leaf 1	7.68 [7.61;7.75] (0.14)	2.75 [2.70;2.80] (0.10)
Leaf 2	7.88 [7.53;8.23] (0.70)	2.98 [2.73;3.22] (0.49)
Leaf 3	10.88 [9.99;11.77] (1.78)	3.84 [3.21;4.47] (1.26)
Leaf 4	13.68 [12.42;14.94] (2.52)	4.71 [3.82;5.60] (1.78)

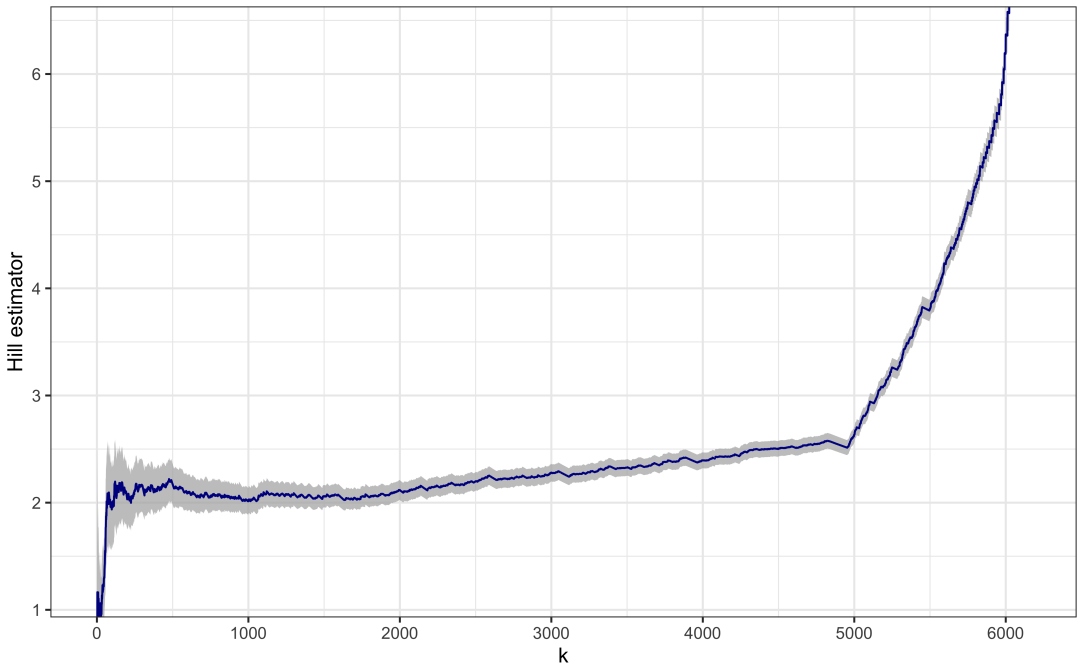

Figure 4: Hill plot for the number of records.

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,n the two other leaves, representing 82% of the extreme events, are "lighter" (although still associated with a shape parameters 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 1.43, which is close to the estimate of the tail distribution provided by Maillart and Sornette [2010].

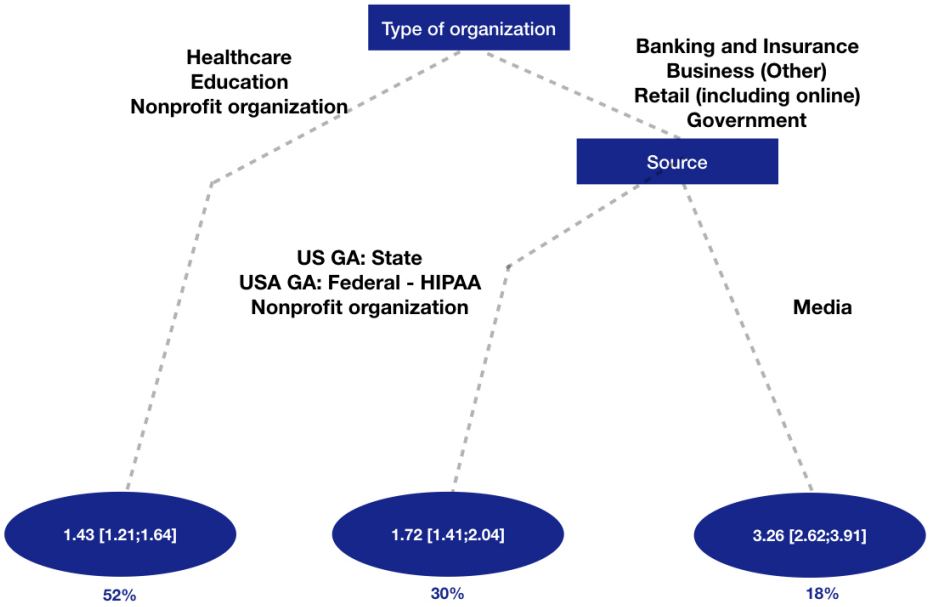

Figure 5: Tree obtained by the CART algorithm based on the Generalized Pareto log-likelihood splitting rule (fitted on the observations exceeding the threshold u). For each leaf, the estimates of γ and their 95% confidence intervals are given.

4.3 Global distribution analysis

The GPD tree of Figure 5 only provides an analysis of the distribution above a threshold u. If one wishes a global distribution, one must combine this approach to 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 Figure 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). The mean is indeed driven by the presence of "extreme" claims in each leaf. This invites to look at regression trees computed using the same method as in Section 4.1 (using quadratic loss or absolute loss) but only on observations smaller than the threshold u.

This leads to the regression trees of Figure 6. For both loss functions, we see that the gap between the median and the mean in each leaf has been drastically reduced. On the other hand, the trees have a different structure than the one obtained from the global set of observations in Figure 3, which shows that the presence of extreme values influences the obtained clusters. The structure of the absolute tree seems more stable than the quadratic one (in Figure 3, a leaf contains 93% of the observations while the repartition is more equilibrate in the tree of Figure 6, moreover the role of the variables Source and Type of organization are switched). This was expected, since the median (which is the target of the absolute tree, using the conditional median to determine its clusters) is a more robust indicator as the mean. This is why we prefer to use clusters based on the absolute tree in the application of Section 7.

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

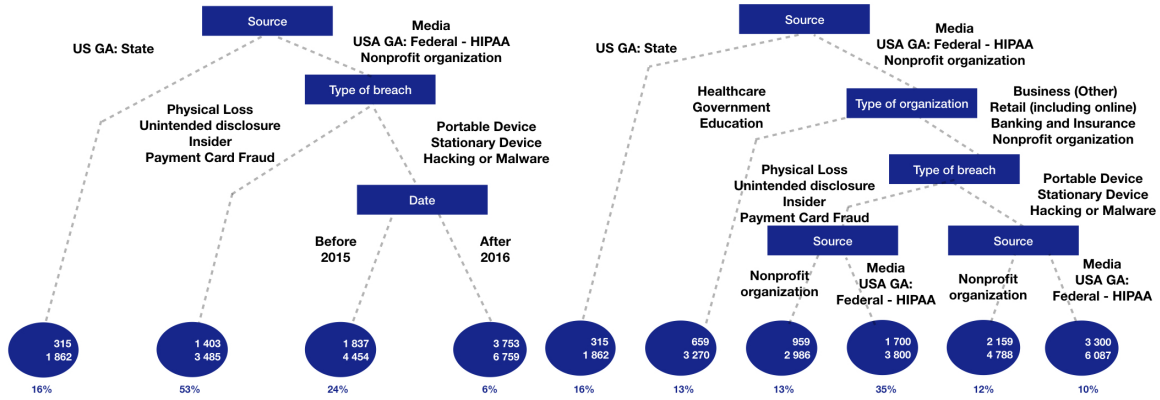

Figure 6: Trees obtained by the CART algorithm based on the quadratic (left-hand side) and the absolute (right-hand side) losses fitted on the observations such that the variable "Number of Records" is less than u.

To obtain the global distribution of the variable $Y =$ "Number of records", the combination of the results of the trees from Figures 6 and 5 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 7: Truncated log normal parameters estimated by the absolute loss tree based on central data. Signification of the parameters is similar to Table 6. Leaves are numerated from left to right according to the representation of the tree from Figure 6.

	Leaf 1	Leaf 2	Leaf 3
μ	5.62 [5.30;5.94] (0.64)	6.79 [6.53;7.04] (0.51)	6.95 [6.73;7.16] (0.43)
σ	3.37 [3.13;3.61] (0.49)	2.46 [2.26;2.65] (0.39)	2.19 [2.02;2.35] (0.33)
	Leaf 4	Leaf 5	Leaf 6
μ	7.64 [7.57 ;7.70] (0.13)	8.18 [7.84 ;8.53] (0.69)	8.70[8.36 ;9.05] (0.70)
σ	1.31 [1.26;1.36] (0.10)	2.26 [2.03;2.49] (0.45)	1.90 [1.68 ;2.12] (0.44)

- 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 Figure 6 (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 GPD tree of Figure 5;
- 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, let us mention that the variable importance for both trees involved in this scheme, which is reported in Table 8. Once again this Table shows the interest of separating the tail from the center of the distribution, since the variables which drive the tail are different (at least in term of hierarchy) from the ones driving the center.

Table 8: Variable importance for the absolute tree of Figure 6 and for the Generalized Pareto tree of Figure 5 (in %).

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

4.4 Comparison with General Additive Models

To compare the GPD 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 Section 7.1, along with estimates for the values of the model parameters.

Table 9 compares the fits of the GPD tree with GAM GPD. Classical GPD 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 GPD tree.

Table 9: Comparison of extreme value theory methodologies

	Covariates used for σ	Covariates used for γ	LL	AIC
GPD	-	-	-2122	4249
GPD GAM	Organization and Source	Date and Organization	-2031	4098
GPRT	Type of organization and Source	Type of organization and Source	-2014	4061

4.5 Insurability of cyber-risk

The model fitted by the GPD regression tree can be understood as a mixture of three GPD. The advantage, compared to fitting a single GPD distribution to all data, 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 GPD mixture. The GPD tree technique presents the advantage to allow identification of some groups of claims that are still associated with an heavy tail behavior, but with more moderate consequences (in our example, all three leaves of the tree of Figure 5 corresponds 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

In this section, we illustrate how the GPD regression trees can be used to project the result of a cyber insurance portfolio. We perform simulations on four portfolios of 1000 policies, where each portfolio is composed of policyholders coming from only of one of the following sectors of activities: BSF, BSO, BSR, or MED. The simulations use the different model 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 ;
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 $(Y_{i,j})_{1\leq i\leq n, 1\leq j\leq N_i}$ 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 & IBM Security] used Cost of Data Breach (CODB) report of 2013 and 2014. The formula is

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

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 10. 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,

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

Table 10: 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

The difference between the result of Formulas (5.1) and (5.2) is shown in Table 11. The results are relatively close for the most part of the events contained in the PRC database, but less pessimistic for the largest ones.

Clearly, we do not claim Formula (5.2) to be accurate to link the number of records to a financial loss. 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).

Table 11: Comparison of Formulas (5.1) and (5.2) depending on the severity of the event (i.e. number of records).

Number of records	Costs inferred from Formula 5.1	Costs inferred from Formula (5.2)	Costs per record Formula (5.1)	Costs per record Formula (5.2)
10 000	2 373 458	2 842 476	237	284
50 000	8 064 897	7 144 968	161	143
100 000	13 657 827	10 626 779	137	106
1 000 000	78 592 594	39 728 891	79	40
50 000 000	1 536 734 440	373 348 764	31	7
100 000 000	2 602 445 366	555 285 127	26	6
1 000 000 000	14 975 509 984	2 075 968 890	15	2

Remark 5.1 *The GPD regression tree of Figure 5 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 the most important, 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*

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

Hence, the shape parameter of $f(Y)$ is $\beta\gamma$. In (5.2)*, $\beta = 0.57$. Hence, the three leaves of the tree of Figure 5 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 Pareto 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 advocates for distinguishing tail behaviors depending on*

the covariates in order to identify more clearly which type of risks can be managed and which can not.

5.2 Frequency analysis

To provide an insurance pricing methodology, estimation of the 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 informations 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 Figure 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 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 as victim of breaches, 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 these multiple times breached companies 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). Table 12 summarizes descriptive count statistics for both portfolios.

Table 12: Number and percentage of companies having k breach events for $k = 1, \ldots, 12$ according to the PRC database and a matching with the NYSE quoted companies.

Number of events k	0	1	2	3	4	5	6	7	8	9	10	11	12	13
PRC Count	-	6782	362	103	55	14	12	5	2	4	1	1	0	1
PRC Percentage (in %)	-	92.4	4.9	1.4	0.7	0.2	0.2	0.1	0.0	0.1	0.0	0.0	0.0	0.0
NYSE Count	2615	120	24	10	6	3	1	1	0	2	0	0	0	0
NYSE Percentage (in %)	94.0	4.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

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

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

As shown in Table 13, the geometric GLM has a better fit than the Poisson one. The fitted parameters are shown in Table 14.

Table 13: Goodness of fit of GLM based on either truncated Poisson or truncated geometric distribution.

Models	Poisson	Geometric
PRC 1–truncated LL	-1361	-1278
PRC 1–truncated AIC	1375	1292
NYSE 0–truncated LL	-361	-319
NYSE 0–truncated AIC	369	327

Table 14: Fitted parameters for the Generalized Linear Model based on geometric distribution for the annual number of claims. The last two columns show an annual rate of occurrence (the rate has been computed considering that each entity has been exposed during 8 years: although the total time span of PRC is 14 years, no source of information is stably reported more than 8 years).

	$p^{(1)}_{PRC}$	$p^{(0)}_{NYSE}$	$E[N^{(1)}_{PRC}]$	$E[N^{(0)}_{NYSE}]$
BSF	0.92 [0.87;0.95] (0.08)	0.91 [0.86;0.94] (0.08)	9%	10%
BSO	0.90 [0.84;0.94] (0.10)	0.94 [0.90;0.97] (0.07)	11%	6%
BSR	0.93 [0.89;0.96] (0.07)	0.93 [0.87;0.96] (0.09)	7%	7%
EDU	0.89 [0.86;0.91] (0.05)	-	13%	-
GOV	0.96 [0.93;0.98] (0.05)	-	4%	-
MED	0.92 [0.90;0.93] (0.03)	0.94 [0.90;0.96] (0.06)	9%	7%
NGO	-	-	-	-
Unknown	0.95 [0.92;0.97] (0.05)	-	5%	-

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. In our model, we consider that, once an event has occurred, the type of event is determined by a multinomial random variable, which parameters only depend on the type of activity of the victim. Let S denote an indicator of the sector of activity, and M denote the type of breach. We can write

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

where $\beta_{s,0}$ corresponds to a reference category (here we took as reference category the incidents for which the type of organization is not known).

In full generality, this would lead to the estimation of a large number of coefficients, with to 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 $\beta_{s,m}$, see e.g. Tibshirani [1996]). This leads to the sparse matrix of coefficient of Table 15.

Table 15: Estimates of multinomial parameters $\widehat{\beta}_{s,m}$ depending on the sector and the type of breach.

$\widehat{\beta}_{s,m}$	CARD	DISC	HACK	INSD	PHYS	PORT	STAT
(Intercept)	-2.12	0.94	1.29	0.03	0.21	0.61	-0.96
BSO	-	-	0.90	-	-	-	-
MED	-	0.31	-	-	1.36	-	-
BSF	-	-	-	-	-	-	-
NGO	-	-	-	-	-	0.11	-
BSR	-	-0.23	0.33	-	-	-	-
EDU	-	0.04	-	-0.08	-	-	-
GOV	-	-	-0,40	-	-	-	-

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):

   - Quadratic tree: we use the tree obtained with the quadratic loss from Figure 3 to determine risk classes. The distribution of the claims in each leaf of the tree is considered as log-normal with the parameters of Table 6.
   - Absolute tree: same principle, using the second tree of Figure 3 and the log-normal parameters of Table 6.

- Generalized Pareto Regression Tree: we use the combination of the trees of Figure 5 and 6, as described in Section 4.3. For the central part, log-normal distributions have been fitted using the parameters of Table 7.
- GAM GPD: for comparison, we considered the approach developed by Chavez-Demoulin et al. [2015], which is exposed in Section 7.1.

4. We use (5.2) to convert this number of records into a financial loss.

Results of these simulation procedures are summarized in Table 16. Let us first remark that, regarding the two approaches based on a single tree (quadratic or absolute), 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 Generalized Pareto 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.

Finally, we provide in Table 17 a short comparison between the use of Formula 5.2 and Jacobs Formula (5.1). Let us emphasize that this last formula (who differs from Formula (5.2) only through a slope coefficient) leads to much more pessimistic losses projections. This was expected, since we already identified in Section 5.1 that this formula probably was not adapted to mega breaches, but the sensitivity of the result to the choice of the model linking the number of records to the loss shows that there is still an important uncertainty around this projection. Hence a precise evaluation of the costs of cyber events—unavailable up to now due to lack of data—is essential to fill this gap.

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

Table 16: 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 100 simulations.

Modeling methodology			Clustering		GAM GPD		GPRT	
Source	Frequency	Organization	$q_{0.5}$	$q_{0.9}$	$q_{0.5}$	$q_{0.9}$	$q_{0.5}$	$q_{0.9}$
US GA State	NYSE	BSF	355	510	3 702	46 745	568	1 572
		BSO	206	315	1 084	27 543	295	1 219
		BSR	271	369	1 666	17 246	376	837
		MED	210	323	273	1 888	212	510
	PRC	BSF	297	416	3 032	56 506	434	1 355
		BSO	359	563	4 738	43 269	612	1 547
		BSR	233	383	1 625	15 898	310	809
		MED	300	444	445	2 669	313	624
Media	NYSE	BSF	1 111	1 722	3 182	81 847	5 275	77 200
		BSO	13 507	33 662	2 996	151 794	3 024	100 418
		BSR	13 710	51 561	2 604	40 015	2 438	35 792
		MED	628	1 128	556	2 361	240	626
	PRC	BSF	922	1 406	2 843	62 691	4 023	90 333
		BSO	24 010	73 423	9 407	161 508	7 016	104 425
		BSR	11 476	42 730	2 113	25 026	2 227	26 412
		MED	918	1 495	805	4 905	365	737

Table 17: Comparison of median and 0.9–quantile resulting of the GPRT methodology and depending on the data breaches cost formula (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 100 simulations.

Cost formula		Formula (5.2)		Formula (5.1)	
Source	Organization	$q_{0.5}$	$q_{0.9}$	$q_{0.5}$	$q_{0.9}$
US GA State	BSF	569	1 572	1 202	6 818
	BSO	295	1 219	562	6 344
	BSR	376	837	667	2 652
	MED	212	510	299	1 454
Media	BSF	5 275	77 200	34 135	1 723 780
	BSO	3 024	100 418	17 994	2 416 482
	BSR	2 438	35 792	13 169	593 715
	MED	240	626	315	1 809

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. 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. Finally, the results on our analysis based on GPD trees reveals that they 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 be interesting for designing appropriate risk management strategies for some type of claims at least. 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 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.

7 Appendix

7.1 Some elements on Generalized Additive Models and their combination with Generalized Pareto distributions

The core assumption of Generalized Additive Models (GAM) is to assume that a target regression function of l parameters $\theta(\mathbf{x}) = \left(\theta^k(\mathbf{x})\right)_{k\in\{1,\dots,l\}}$ is of the following form,

$$g^k(\theta^k(\mathbf{x})) = \sum_{i=1}^{p} h_i^k(\mathbf{X}_i)$$

where h_i^k is either a factor mapping function if $\mathbf{X}_i$ is a factor covariate (which is our case with PRC data), or a smooth function if $\mathbf{X}_i$ is continuous, and where g_k is a link function. h_i^k may be the null function invloving the non relevance of the covariates $\mathbf{X}_i$ while predicting θ^k according to the estimated Generalized Additive Model.

We first discuss the tail part analysis of the distribution using the approach introduced in Chavez-Demoulin et al. [2015], where the authors adapted GAM to extreme value regression. In details, we study $\theta_{\text{GAM GPD}}(\mathbf{x}) = (\nu(\mathbf{x}), \gamma(\mathbf{x}))$, which is a reparametrization of a GPD, with γ the shape parameter and $\nu = log\left((1+\gamma)\sigma\right)$. The link function used is the identity for both ν and γ. The form of the obtained model is

$$\nu(\mathbf{x}) = a_0 + a_1.t + \sum_{j=2}^{16} a_j \mathbf{S}_j, \text{ and } \gamma(\mathbf{x}) = \sum_{j=2}^{16} b_j \mathbf{S}_j,$$

where t is the difference between the year of the breach and the origin year 2005, and $(S_j)_{2\leq j\leq 16}$ are either 0 or 1 variables corresponding to the different modalities of the variables "Type of organtization," "Type of breach", and "Source of information." The coefficient a_0 corresponds to the remaining ordonal term when $t = 0$. The estimated coefficients are listed in Table 18.

Then, to complete the analysis of the excesses by the core of the distribution, we fitted GAM model combined with a truncated log-normal distribution. The density of a truncated log-normal distribution of parameters (μ, s) is defined for $x \in]0; 27999[$ by $g_{\mu,s}(x) = \frac{f_{\mu,s}(x)}{\int_{u=0}^{27999} f_{\mu,s}(u)du}$, where f is the density of the log-normal distribution that is given for $x \in]0; +\infty[$ by $f_{\mu,s}(x) = \frac{1}{xs\sqrt{2\pi}} \exp\left(-\frac{(\ln x-\mu)^2}{2s^2}\right)$. We consider $\theta_{\text{GAM LN}}(\mathbf{x}) = (\mu(\mathbf{x}), s(\mathbf{x}))$. For the parameter μ, the link function is identity whereas for s the log-link

function is used. The form of the resulting model is

$$\mu(\mathbf{x}) = c_0 + c_1.t + \sum_{j=2}^{16} c_j \mathbf{S}_j, \text{ and } s(\mathbf{x}) = \exp\left(d_0 + d_1.t + \sum_{j=2}^{16} d_j \mathbf{S}_j\right).$$

The coefficients c_0 and d_0 correspond to the remaining ordonal term when $t = 0$. The estimated coefficients are listed in Table 18.

Table 18: Fitted coefficients for the GAM LN and GAM GPD models for the variable "Number of records" (expressed in hundreds of thousands of lines for the GAM GPD). Model selection has been performed using the AIC criterion.

	GAM GPD		GAM	
Covariates	a_j (for ν)	b_j (for γ)	c_j (for μ)	d_j (for s)
Intercept	9.42	-	5	-0.02
Year	0.11	-	-0.02	-0.04
BSO	0.26	1.76	-0.17	0.19
BSR	0.03	2.10	-0.25	-0.88
EDU	-1.54	0.37	-0.07	0.60
GOV	-0.59	1.55	0.01	0.54
MED	-1.05	0.71	-0.09	0.69
NGO	-0.39	0.72	-0.09	0.17
DISC	-	-	-0.18	0.29
HACK	-	-	0.05	1.04
INSD	-	-	-0.13	-0.68
PHYS	-	-	-0.18	0.25
PORT	-	-	-0.20	1.39
STAT	-	-	-0.21	1.11
Source NGO	-	-0.97	-0.36	-1.92
Source US GA HIPAA	-	-0.8	-0.82	-1.21
Source US GA State	-	-0.63	0.25	-3.16

7.2 Missing values handling: surrogate rules

The growing process of a tree detailed in Section 3.1.1 need to be slightly adapted to handle missing values. Indeed, at the step k + 1, the completeness of the data is required at two stages:

- for $l = 1, \ldots, d$, when finding the best split $x_{j\star}^{(l)}$ for the component $X^{(l)}$ of $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$, it is necessary to know all the observations of the component l.

- then, once the best component index $\widehat{l}$ is determined, the definition of the two new rules $R_{j1}(\mathbf{x})$ and $R_{j2}(\mathbf{x})$ are only relevant for observations where $\mathbf{x}^{(\widehat{l})}$ is not missing.

The process is therefore customized as follows:

- for $l = 1, \ldots, d$, the best split $x_\star^{(l)}$ for the component $X^{(l)}$ of $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$ is computed regarding only available observations of the component l.
- then, once the best component index $\widehat{l}$ is determined, the definition of the rules (R_{j1}, R_{j2}) is extending by rules $(\tilde{R}_{j1}, \tilde{R}_{j2})$ to be able to assign observations where $\mathbf{x}^{(\widehat{l})}$ is missing.

The add-on rules aim at mimicking the established rules by minimizing the classification error function Δ defining as follows:

$$\Delta : \Big((R_{j1}, R_{j2}), (\tilde{R}_{j1}, \tilde{R}_{j2})\Big) \longmapsto \frac{\sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{(\widehat{l})} \neq \mathrm{NA}} \Big(R_{j1}(\mathbf{X}_i)\tilde{R}_{j2}(\mathbf{X}_i) + R_{j2}(\mathbf{X}_i)\tilde{R}_{j1}(\mathbf{X}_i)\Big)}{\sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{(\widehat{l})} \neq \mathrm{NA}} \left(R_{j1}(\mathbf{X}_i) + R_{j2}(\mathbf{X}_i)\right)}.$$

One possible continuation of the rules is to point all the missing observations toward the more abundant rule, by introducing $(\tilde{R}_{j1}^{maj}, \tilde{R}_{j2}^{maj})$ as follows:

$$\begin{cases} (\tilde{R}_{j1}^{maj}, \tilde{R}_{j2}^{maj}) &= (R_j(\mathbf{x}), 0) \text{ if } \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{(\widehat{l})} \neq \mathrm{NA}} R_{j1}(\mathbf{X}_i) \geq \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{(\widehat{l})} \neq \mathrm{NA}} R_{j2}(\mathbf{X}_i), \\ & \text{or} \\ (\tilde{R}_{j1}^{maj}, \tilde{R}_{j2}^{maj}) &= (0, R_j(\mathbf{x})) \text{ if } \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{(\widehat{l})} \neq \mathrm{NA}} R_{j1}(\mathbf{X}_i) < \sum_{i=1}^n \mathbf{1}_{\mathbf{X}_i^{(\widehat{l})} \neq \mathrm{NA}} R_{j2}(\mathbf{X}_i), \end{cases}$$

leading to the error $\Delta_j^{\mathrm{maj}} = \Delta\left((R_{j1}, R_{j2}), (\tilde{R}_{j1}^{maj}, \tilde{R}_{j2}^{maj})\right)$.

The surrogate rules are then defined for $s \in \{1, \ldots, d\} \setminus \widehat{l}$, as all the couple of rules $(\tilde{R}_{j1}^s, \tilde{R}_{j2}^s)$ given by $\tilde{R}_{j1}^s = R_j(\mathbf{x})\mathbf{1}_{\mathbf{x}^{(s)} \leq x_{j\diamond}^{(s)}}$, and by $\tilde{R}_{j2}^s = R_j(\mathbf{x})\mathbf{1}_{\mathbf{x}^{(s)} > x_{j\diamond}^{(s)}}$, that meet the criterion $\Delta\left((R_{j1}, R_{j2}), (\tilde{R}_{j1}^s, \tilde{R}_{j2}^s)\right) < \Delta_j^{\mathrm{maj}}$, with $x_{j\diamond}^{(s)}$ defined as:

$$x_{j\diamond}^{(s)} = \arg\min_{x^{(s)}} \Delta\left((R_{j1}, R_{j2}), (R_j(\mathbf{x})\mathbf{1}_{\mathbf{x}^{(s)} \leq x^{(s)}}, R_j(\mathbf{x})\mathbf{1}_{\mathbf{x}^{(s)} > x^{(s)}})\right).$$

Letting Λ_j be the set of $s \in \{1, \ldots, d\} \setminus \widehat{l}$ that lead to a surrogate rule and writing Δ_j^s the resulting error for all $s \in \Lambda_j$, surrogate rules are used in descending order of Δ_j^s to allocate observations that may have missing values on possible few components. Then, if necessary and by default, rules $(\tilde{R}_{j1}^{maj}, \tilde{R}_{j2}^{maj})$ are used.

7.2.1 Analysis of a tree: variable importance

An attractive output of a tree is the variable importance. It aims to quantifying the involvement of each covariate in the decrease of the loss resulting from the tree. It is based on the improvement made at each step of the growing process, which is given by $\Phi(R_j, x_{j\star}^{(\widehat{l})})$ if R_j is replaced by two new rules thanks to the covariate $\widehat{l}$ and 0 otherwise. It also takes into account the potential components used in the surrogates rules. To be precise, the importance of each component s implicated in a surrogate rules is defined as the improvement of the principal rules $\Phi(R_j, x_{j\star}^{(\widehat{l})})$ weighted by $w_j^s = (\Delta_j^s - \Delta_j^{\text{maj}})/(1 - \Delta_j^{\text{maj}})$. Thus, the variable importance vector is obtained by normalizing the vector $I = (i_l)_{l\in 1,\dots,d}$ defined by:

$$i_l = \sum_{j=1}^{n_k-1} \left(\Phi(R_j, x_{j\star}^{(\widehat{l})}) \mathbf{1}_{l=\widehat{l}} + \Phi(R_j, x_{j\star}^{(\widehat{l})}) w_j^s \mathbf{1}_{l\in\Lambda_j} \right).$$

Remark 7.1 *To compute the variable importance, the search of all the surrogates rules have to be done regardless of the presence of missing values.*

7.3 Additional statistics

In this section, we give a few more descriptive statistics on the PRC database (Table 20). The confidence intervals for the parameters estimated on the leaves of the GPD tree of Figure 5 are shown in Table 19.

Table 19: Generalized Pareto parameters estimated by the Generalized Pareto Regression Tree based on excesses and the 95% confidence intervals.

	Leaf 1	Leaf 2	Leaf 3
γ	1.43 [1.21;1.64] (0.42)	1.72 [1.41;2.04] (0.63)	3.26 [2.62;3.91] (1.29)
$\sigma.10^{-5}$	0.36 [0.29;0.43] (0.14)	0.76 [0.55;0.97] (0.41)	1.82 [0.98;2.67] (1.68)

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

Table 20: Descriptive statistics for the PRC database. Shape estimates denote the estimation of parameter γ on this sub-population.

Covariate	Modality	Number	Median	Share of extremes	Shape estimates
Total	-	6160	2000	16 %	2,16 [1,96;2,36]
Source	Media	595	11266	40 %	3,14 [2,60;3,68]
	Nonprofit organization	2309	2000	19 %	1,52 [1,29;1,76]
	US GA: Federal - HIPAA	1949	2300	10 %	1,53 [1,17;1,89]
	US GA: State	888	409,5	8 %	1,81 [1,13;2,50]
	NA	419	2300	11 %	2,28 [1,36;3,21]
Breach	CARD	32	300	12 %	2,75 [-0,97;6,46]
	DISC	1353	1600	12 %	2,48 [1,95;3,02]
	HACK	1467	4700	28 %	2,50 [2,16;2,84]
	INSD	359	566	12 %	1,98 [1,08;2,87]
	PHYS	1248	1700	7 %	1,38 [0,87;1,90]
	PORT	769	4000	23 %	1,39 [1,03;1,75]
	STAT	152	3561,5	22 %	1,11 [0,42;1,80]
	NA	780	949,5	9 %	2,28 [1,52;3,04]
Sector	BSF	404	1936	24 %	2,21 [1,56;2,87]
	BSO	404	5750	37 %	3,17 [2,49;3,84]
	BSR	284	870	25 %	2,80 [1,94;3,67]
	EDU	680	2400	17 %	0,99 [0,64;1,34]
	GOV	540	2773	23 %	1,61 [1,14;2,08]
	MED	3331	2078	11 %	1,42 [1,17;1,67]
	NGO	70	2200	19 %	2,04 [0,3;3,72]
	NA	447	500	11 %	2,79 [1,66;3,92]
Year	2005	117	16500	43 %	1,30 [0,66;1,94]
	2006	381	2089	22 %	1,49 [0,91;2,06]
	2007	334	3000	22 %	1,52 [0,95;2,09]
	2008	269	4000	24 %	1,63 [1,02; 2,24]
	2009	192	2000	25 %	2,00 [1,15;2,86]
	2010	540	2000	14 %	1,51 [0,93;2,08]
	2011	548	1807	13 %	1,89 [1,23;2,55]
	2012	584	1595	12 %	1,86 [1,17;2,54]
	2013	536	2711,5	12 %	1,95 [1,22;2,67]
	2014	581	1251	14 %	2,17 [1,51;2,84]
	2015	333	2208	16 %	2,84 [1,77;3,91]
	2016	456	2554	19 %	3,21 [2,27;4,15]
	2017	411	3000	16 %	2,66 [1,76;3,57]
	2018	878	991,5	12 %	2,75 [2,07;3,44]

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

References

A. A. Balkema and L. De Haan. Residual life time at great age. *The Annals of probability*, pages 792–804, 1974.

J. Beirlant and Y. Goegebeur. Local polynomial maximum likelihood estimation for pareto-type distributions. *Journal of Multivariate Analysis*, 89(1):97–118, 2004.

J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys. Tail index estimation and an exponential regression model. *Extremes*, 2(2):177–200, Jun 1999.

J. Beirlant, Y. Goegebeur, J. Segers, J. Teugels, D. De Waal, and C. Ferro. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2004.

C. Biener, M. Eling, and J. H. Wirfs. Insurability of cyber risk: An empirical analysis. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 40(1):131–158, 2015.

F. Bisogni, H. Asghari, and M. J. Van Eeten. 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*, 2017.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

P. Chaudhuri and W.-Y. Loh. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 8(5):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, Feb. 2015. doi: 10.1111/jori.12059. URL `https://doi.org/10.1111/jori.12059`.

Databreaches.net. Databreaches reporting. `https://www.databreaches.net/about/`.

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.

L. De Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

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.

M. Eling and N. Loperfido. Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance: Mathematics and Economics*, 75:126–136, 2017.

M. Eling and W. Schnell. What do we know about cyber risk and cyber risk insurance? *The Journal of Risk Finance*, 17(5):474–491, 2016.

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

M. A. Fahrenwaldt, S. Weber, and K. Weske. Pricing of cyber insurance contracts in a network model, 2018.

S. Forrest, S. Hofmeyr, and B. Edwards. Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, 2(1):3–14, 2016.

S. Gey and E. Nédélec. Model Selection for CART Regression Trees. *IEEE Transactions on Information Theory*, 51(2):658–670, 2005.

D. R. Insua, A. C. Vieira, J. A. Rubio, W. Pieters, K. Labunets, and D. G. Rasines. An adversarial risk analysis framework for cybersecurity. *CoRR*, abs/1903.07727, 2019.

J. Jacobs. Analyzing ponemon cost of data breach. *Data Driven Security*, 11, 2014.

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*, 9:1120–1128(8), 2015.

R. W. Katz, M. B. Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in water resources*, 25(8-12):1287–1304, 2002.

W.-Y. Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

W.-Y. Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014.

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.

P. Maddie Ladner. Data breach notification in the united states and territories, 2018. `https://www.privacyrights.org/sites/default/files/Data%20Breach%20Notification%20in%20the%20United%20States%20and%20Territories_0.pdf`.

T. Maillart and D. Sornette. Heavy-tailed distribution of cyber-risks. *The European Physical Journal B*, 75(3):357–364, 2010.

A. Marotta, F. Martinelli, S. Nanni, A. Orlando, and A. Yautsiukhin. Cyber-insurance survey. *Computer Science Review*, 24:35–61, 2017.

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

Ponemon Institute LLC & IBM Security. 2018 cost of a data breach study: Global overview.

S. I. Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

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.

S. Romanosky. Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, 2(2):121–135, 08 2016.

State of California. California list of breaches. `https://oag.ca.gov/privacy/databreach/list`.

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

T. Therneau and M. Clinic. User written splitting functions for rpart. 02 2018.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x.

U.S. HHS department. HSS breach portal, a. https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf.

U.S. HHS department. HIPAA breach notification index, b. https://www.hhs.gov/hipaa/for-professionals/breach-notification/index.html.

S. Wheatley, T. Maillart, and D. Sornette. The extreme risk of personal data breaches and the erosion of privacy. *The European Physical Journal B*, 89(1), Jan 2016. ISSN 1434-6036. doi: 10.1140/epjb/e2015-60754-4. URL http://dx.doi.org/10.1140/epjb/e2015-60754-4.