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

Sébastien Farkas, Olivier Lopez, Maud Thomas

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

Sébastien Farkas[1], Olivier Lopez[1], Maud Thomas[1]

May 2, 2019

Abstract

In this paper we propose a methodology to analyze the heterogeneity of cyber claim databases. This heterogeneity is caused by the evolution of the risk but also by the evolution in the quality of data and of sources of information through time. We consider a public database, already studied by Eling and Loperfido [2017], which is considered as a benchmark for cyber event analysis. Using regression trees, we investigate the heterogeneity of the reported cyber claims. A particular attention is devoted to the tail of the distribution, using a Generalized Pareto likelihood as splitting criterion in the regression trees. Combining this analysis with a model for the frequency of the claims, we develop a simple model for pricing and reserving in cyber insurance.

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

Evaluating the potential cost of cyber-risk is a challenging task due to the lack of consistent and reliable data on this topic. While the cyber-insurance market is constantly growing, the determination of an appropriate premium and the design of appropriate risk management strategies are particularly difficult due to the immaturity of the market and the poor historical information on a relatively new risk, as pointed by Biener et al. [2015] and Eling and Schnell [2016]. See also [Marotta et al., 2017] for a review of topics recently addressed in cyber-insurance. In Fahrenwaldt et al. [2018], the authors propose a pricing model of cyber contracts based on the topology on the infected networks. On the other hand, Eling and Loperfido [2017] recently proposed a pricing model based on the Privacy Rights Clearinghouse (PRC) database (available for public download at `https://www.privacyrights.org/copyright`) without requiring network data. Let us also mention Insua et al. [2019] who used Adversarial Risk Analysis to aggregate expert judgments. The statistical evaluation of the risk has also been studied by Maillart and Sornette [2010] and Forrest et al. [2016a], the authors focused mostly on the severity of the claim—that is on the amount of a claim once it occurred. On the other hand, few attention has been brought to the heterogeneity of the data used to calibrate these models. This heterogeneity—in terms of type of events and of reporting sources—undermines the reliability of the evaluation of the frequency of cyber events. It can also have impacts on the analysis of the distribution of the severity.

In the present paper, we propose to gain further insight on this heterogeneity by relying on regression trees inference for the severity of cyber claims. We also take a closer look at the analysis of the frequency with less deepness since the lack of public data on this issue is patent. This is typically due to the poor information one may have on the exposure - that is the number of entities exposed to the risk. The general objective is to analyze the impact of characteristics (victim information, source of the reporting,...) on cyber events. We especially focus on "extreme" events, that is events for which the severity of the claim is larger than a fixed (high) threshold. Based on this modeling, a pricing methodology is proposed, along with elements to perform reserving through the understanding of the potential impact of the most severe events.

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 responses. 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 propose a basic framework to model the loss of an insurance company, and the methodology we use to estimate the amount of a claim from public data. Section 3 is devoted to the general theoretical background of the statistical tools we use to consider the data (namely regression trees and extreme value analysis). Section 4 takes a closer look at the PRC data and shows the results of the models we consider. These models are then applied to virtual portfolios in Section 5.

2 Application to pricing of insurance contracts

2.1 Pricing model

From an insurance point of view, the finality of the quantification of cyber-risk essentially aims at determining a price for cyber-insurance contracts, and a reserve to face the potential future claims. A basic pricing approach consists in equalizing the expectation pay-off of both sides: the (potential) policyholder with characteristics $\mathbf{X}$ who pays a deterministic premium $\pi(\mathbf{X})$, and the insurer who will provide a random amount A as a compensation to the losses during some period of time. Computing the pure premium hence consists in

determining the function

$$\pi(\mathbf{x}) = E[A|\mathbf{X} = \mathbf{x}].$$

On the other hand, reserving consists in evaluating how to protect the insurer from deviations from this central scenario. Considering a portfolio with n policyholders with characteristics $(\mathbf{X}_i)_{1\leq i\leq n}$ and losses $(A_i)_{1\leq i\leq n}$, the total loss is $S = \sum_{i=1}^{n} A_i$, and the question is to determine an amount r such that $\mathbb{P}(S > r|\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_n = \mathbf{x}_n)$ is sufficiently small.

In both cases, the goal is to have the clearest possible vision of the distribution of $A|\mathbf{X} = \mathbf{x}$, considering that the components of the random vector $(A_i, \mathbf{X}_i)_{1\leq i\leq n}$ are independent and identically distributed (i.i.d.) in the second case. The difference between the two cases is that the pure premium computation only requires to focus on the center of this conditional distribution, while reserving requires to model the tail of the distribution.

Following a classical risk theory approach, the analysis of the distribution of $A|\mathbf{X} = \mathbf{x}$ can be decomposed into a "frequency" analysis and a "severity" analysis. Indeed, $A = \sum_{j=1}^{N} L_j$, where N is the number of claims registered by the policyholder—representing the frequency of the claims, and $(L_j)_{1\leq j\leq N}$ the list of successive amounts—representing the severity. In the following, we assume that N is independent of the claim amounts L_j, $j = 1, \ldots, N$. We further assume that, conditionally on $\mathbf{X}$, the claim amounts L_j are independent of N, and are i.i.d. with the same distribution as a random variable L. This simplification allows us to study separately the frequency variable N in Section 4.3, and the severity variable L in Section 4.4.1.

Since the database we consider contains different types of cyber events (hacking, malware...) as described further in Section 4.3.2, we will distinguish between types of claims by considering several counting processes $N_1, ...N_M$ corresponding to the M different types of events.

2.2 Loss quantification of a data breach

Obviously, a key element of cyber insurance evaluation is to obtain an estimation of the loss associated to an event. Public databases like the PRC one (described in details in Section 4) does not contain such an information. Nevertheless, the PRC database provides the number of records affected by a breach, which can be used to measure the severity of an event. In the sequel, the random variable Y denote the number of records breached during an event. This variable is expected to be strongly correlated with the loss variable L. Linking Y to L has been proposed by Romanosky [2016] who studied the cost

	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

Table 1: Data used to calibrate Formula (2.2): the costs of moderate breaches have been computed using Formula (2.1); the mega breaches are the only two communicated from CODB 2018.

of data breaches using a private database gathering cyber events and associated losses. However, the calibration they obtain requires to use information about the revenues of victim organizations: an information which is unavailable in the PRC database.

On the other hand, Jacobs [2014] analyzed data gathered by [Ponemon Institute LLC & IBM Security] for the publication of the 2013 and 2014 Cost of Data Breach (CODB) report. He proposed Formula (2.1) to deduce a cost L from a number of records Y,

$$\log(L) = 7.68 + 0.76 \log(Y). \tag{2.1}$$

A limit of this calibration is that, in 2014, the Ponemon Institute LLC had not yet observed the cost of "mega data breaches". This leads to a pessimistic evaluation of the cost of very larges breaches, compared to what has been observed in the next years following the study. However, as written in their 2018 CODB report, "for the first time in 2018, [they] attempt to measure the cost of a data breach involving more than one million compromised records, or what [they] refer to as a mega breach". Those recent observations can enrich the formula proposed by [Jacobs, 2014] based on data available in 2014 which only concerned data breaches with a number of records under 100 000, especially by taking into account the finding that the cost per record of a data breach seems to decrease significantly with the number of records. As an alternative to Formula (2.1), we propose to calibrate a second equation of the type

$$\log(L) = \alpha + \beta \log(\log(Y)), \tag{2.2}$$

where our estimation of α and β is based on two costs of moderate data breaches inferred from Formula (2.1) and the two costs of mega breaches contained in the 2018 Cost of Data Breach report summarized in Table 1. This leads to an estimation $\alpha = -1.998$ and $\beta = 7.503$.

This formula, based directly on 2018 CODB data and indirectly on 2014 CODB, seems to be a good trade off for the quantification of both moderate and mega breaches. We

Number of records	Costs inferred from Jacobs formula	Costs inferred from Formula (2.2)	Costs per records (Jacobs formula)	Costs per records (Formula (2.2))
10 000	2 373 458	2 329 378	237	233
50 000	8 064 897	7 798 660	161	156
100 000	13 657 827	12 426 702	137	124
1 000 000	78 592 594	48 803 702	79	49
50 000 000	1 536 734 440	316 874 975	31	6
100 000 000	2 602 445 366	422 540 274	26	4
1 000 000 000	14 975 509 984	1 022 505 107	15	1

Table 2: Comparison of Formula (2.1) and Formula (2.2) depending on the severity of the event (i.e. number of records).

compare the costs inferred from the Jacob formula (2.1) and the proposed formula (2.2) in Table 2. The practical comparison between the use of the formulas (2.1) and (2.2) is done in Section 5.

3 Regression Trees and extreme value analysis

In this section, we propose a general presentation of regression trees, and explain how they will be applied to the PRC database. The algorithm 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. Therefore, in Section 3.2, we introduce Generalized Pareto regression trees, motivated by theoretical results on extreme value theory.

3.1 Regression Trees

Regression Trees methods are designed to perform regression analysis and clustering simultaneously. They 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 Y is a response variable (the cost 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 has 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: obtention of the maximal tree

The CART algorithm consists in determining iteratively "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 $j = 1, ..., n_k$,

- 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 $\mathbf{X} = (X^{(1)}, ..., X^{(d)})$, define the best threshold $x_\star^{(l)}$

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

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

where

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

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

- 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 $\mathcal{R} = (R_j)_{j=1,\ldots,s}$, an estimator $\hat{m}^{\mathcal{R}}$ of the function m can be deduced, that is

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

where

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

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 $\mathcal{R}^{\mathcal{S}} = (R_1^{\mathcal{S}}, ...R_{n_{\mathcal{S}}}^{\mathcal{S}})$ of cardinality $n_{\mathcal{S}}$. One then selects the subtree $\hat{\mathcal{S}}(A)$ who minimizes the criterion

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

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 $\hat{\mathcal{S}}(A)$, it is not necessary to compute any subtree 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 A 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 $\hat{A}$ denote the penalization constant calibrated using the test sample or the $k-$fold cross-validation approach, our final estimator is then $\hat{m}(\mathbf{x}) = m^{\hat{\mathcal{S}}(\hat{A})}(\mathbf{x})$.

3.2 Analysis of the tail of the distribution through regression trees

3.2.1 Peaks over threshold method for extreme value analysis

Extreme value analysis is the branch of statistics which has been developped and broadly used to handle extreme events, such as extreme floods or heat waves episodes with extreme financial losses [Katz et al., 2002, Embrechts et al., 2013]. Given a series of observations $Y_1, Y_2, \ldots$ independent and identically distributed with an unknown survival function $\bar{F}$ (that is $\bar{F}(y) = \mathbb{P}(Y_1 > y)$). 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 $X_i > u$. The asymptotic behavior of extreme events is characterized by the distribution of the excesses which is given by

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

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 Balkema and De Haan [1974] have shown that there exist normalizing sequences $a(u) > 0$ and $b(u)$ and a non-degenerated distribution function H such that

$$\overline{F}_u(a(u)y + b(u)) \xrightarrow[n\to\infty]{d} \overline{H}_{\sigma,\gamma}(y),$$

with $\overline{H}_{\sigma,\gamma}$ necessarily of the form

$$\overline{H}_{\sigma,\gamma}(y) = \left(1 + \gamma\frac{y}{\sigma}\right)^{-1/\gamma},\ y > 0. \tag{3.3}$$

Here, $\sigma > 0$ is a scale parameter and $\gamma > 0$ is a shape parameter, which reflects the heaviness of the tail distribution. 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 distribution.

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].

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 maximising 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.

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, depending on the circonstances of the claim and on the characteristics of the victim. Several authors have proposed regression models to study the influence of some covariates on the parameters σ and γ in (3.3). Typically, these approaches are either parametric as in [Beirlant et al., 1999], or nonparametric using kernel smoothing (which supposes continuous covariates) as in [Beirlant and Goegebeur, 2004]. Alternatively, we propose to adapt the regression tree approach to study the tail of the distribution of the response variable Y.

Consider a threshold u. Based on observations such that $Y_i \geq u$, we fit a regression tree using the Generalized Pareto log-likelihood as splitting criterion, 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 $m(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$.

After completing the growth and the pruning phases as described in Section 3.1.1 and Section 3.1.3, the leaves give a decomposition of the data in subpopulations associated with a different tail behavior. On each of them, the estimated value of (σ, γ) gives the parameters of the proper Generalized Pareto distribution that better fits this subpopulation.

When adjusting a Generalized Pareto distribution to the tail of the distribution, the results are clearly threshold dependent. We will discuss in Section 4.4 a heuristic way to

determine such a threshold. Typically, we need to take a threshold small enough so that there is a sufficient number of observations beyond this threshold to fit a regression tree, and large enough so that the Generalized Pareto approximation is legitimate.

Remark 3.2 *In extreme value regression, the conditional version of* (3.3) *leads to the introduction of a threshold* u *that should depend 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. See also [Beirlant and Goegebeur, 2003] for a similar fixed threshold in Generalized Pareto regression.*

4 Analysis of cyber claims

In this section, we propose a detailed analysis of the cyber incidents gathered in the PRC database. The description of the database is done in Section 4.1—which focuses on the sources that feed the database—and in Section 4.2, where we give some elementary statistics on the variables available in PRC. The aim is then to calibrate a model, based on the database, describing cyber events. This is done in two steps, first by analyzing the frequency and the typology of the claims in Section 4.3, then by taking a closer look at the severity in Section 4.4 using regression trees. In this last part, we focus not only on the central part of the distribution but also on the tail.

4.1 Multiple sources feeding the database

Privacy Rights Clearinghouse (PRC) is a nonprofit organization founded in 1992 which aims at protecting privacy for US citizens by "empowering individuals and advocating for positive change". The PRC database is publicly available and this article is based on a download made on January 23 2019, corresponding to 8860 events. In order to raise awareness about privacy issues, PRC has been maintaining a chronology since 2005, listing companies that have been implicated in data breaches affecting US citizens. Unfortunately, although this chronology is not exhaustive, PRC organization is been trying to increase its scope by gathering notifications of data breaches from different sources of information which can be clustered in four groups:

- US Government Agencies on the federal level: a significant amount of the events is present due to the legal obligation to report data breaches (consequence of privacy

regulation). In the health 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]. One can note that there is no uniformity in the choice of a threshold (in terms of number of victims) to trigger the obligation to notify. Therefore, the requirement for an organization to report a data breach depends on its state. On the reporting procedure, some states use an online notification such as California, see State of California, but this is not systematic.

- Media: these events have been brought to the attention of the PRC organization due to their high visibility. They are typically more severe than the events from other sources of information.

- Non profit organizations: the PRC database includes the data breaches reported by non profit organizations such as [Databreaches.net].

The source of information is clearly identified in the database, and is available for each reported data breach. Although merging different sources of notifications increases the scope of the PRC chronology, it also introduces a new level of heterogeneity in data breaches events. Indeed, the source of the report gives an information on the typology of the event (for example, media reported events are usually more severe). On the other hand, let us also observe, as shown in Figure 4.1, that, through time, the proportion of events reported from one source compared to others strongly fluctuates (the global proportion of sources of reports is summarized in Table 3). Another visualization of this phenomenon is shown in Figure 2, representing the evolution of monthly numbers of reported events depending on the source. This constitutes an important issue at least when one tries to estimate the frequency of occurrence. Indeed, the increasing number of reported claims may be partially caused by an evolution of the risk, but also on an evolution on the way these claims are reported. For example, before the requirement to report from some government agencies, some claims may be absent from the database. In other words, Figure 4.1 tends to prove that the reported claims of the PRC database do

not refer to a stable population. We will develop this discussion on the exposure to the risk in Section 4.3.1.

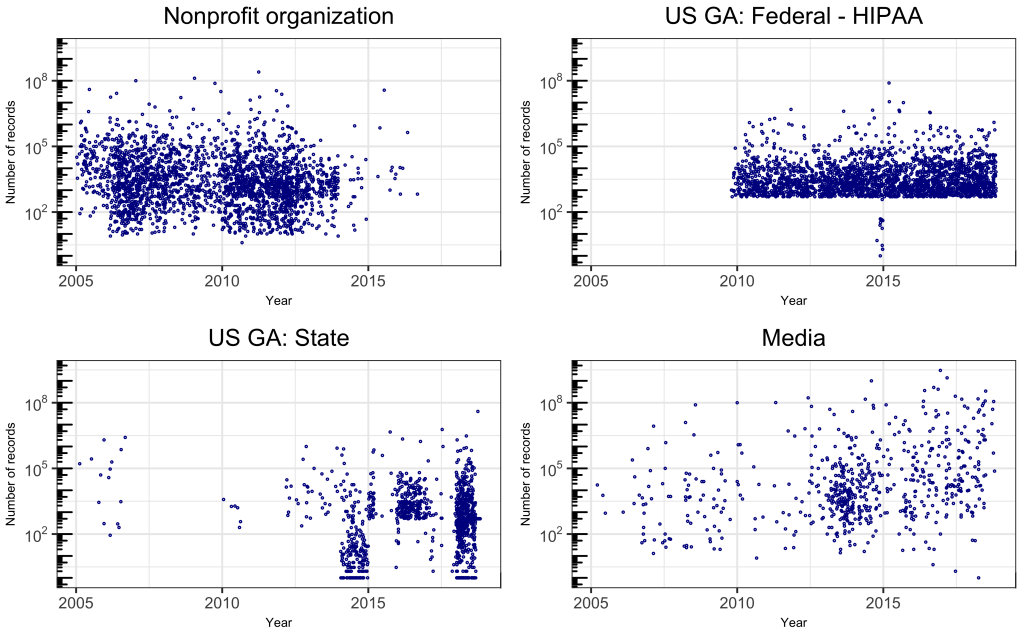

Figure 1: Time evolution of the reports of cyber events depending on the source of information.

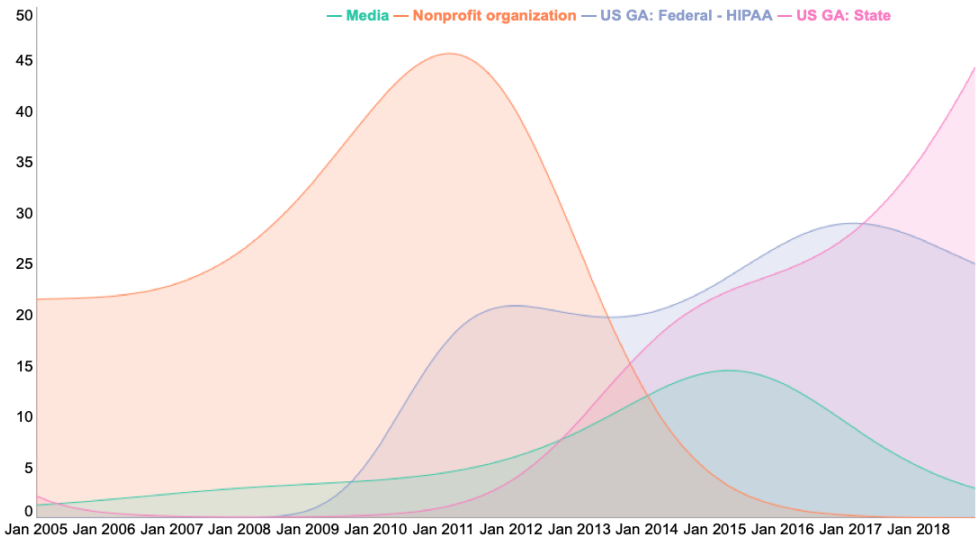

Figure 2: Evolution of monthly frequency of reporting depending on the source of information (smoothed curves).

Source	PRC: all events	PRC: events with known number of records	Events with records
Nonprofit organization	39%	37%	72%
US GA - HIPAA	27%	36%	99%
US GA - State	22%	17%	58%
Media	12%	10%	60%
Total	8860	6641	75%

Table 3: Relative weights of the different sources in PRC database: third column considers only events where the number of records is known by PRC; fourth column presents the proportion of events from a given source for which the number of records is known.

4.2 Description of the database

The PRC database gathers informations regarding each reported 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 different modalities are summarized in Tables 4 to 6.

The number of records is a key information for an actuarial analysis of the database. Indeed, although it does not quantify the financial loss generated by an event, it is an indicator in the PRC database of the severity of a data breach. The number of records is defined by the number of rows (not the number of cells) of the database concerned by a data breach. Although few rows may refer to the same person, for instance in case of multiple accounts, and that some rows may contain false information, for instance in case of a wrong filling, the number of records is often interpreted as the number of individuals affected by the breach.

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 4: List of the variables of 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 5: Description of sectors of organizations.

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

Table 6: Description of types of data breaches.

4.3 Number and typology of claims

We now analyze the process of occurrence of claims. In Section 4.3.1 we focus on the frequency of claims, that we modelize the number of events striking a company whatever its nature. Due to the lack of data, the typology of the claim, considered in Section 4.3.2, is assumed to be independent of the occurrence of a claim.

4.3.1 Frequency analysis

Estimation of the frequency of occurrence of cyber events based on the PRC database is not straightforward due to the difficulty to collect information on the exposure. Figures 4.1 and 2 highlight the significant evolution in the data collection process of the Privacy Right

Clearinghouse. 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. Therefore, the comment of Forrest et al. claiming that "there is no obvious reason why [the] size/frequency distributions [used in informal studies] should differ from PRC" (p.5 of [Forrest et al., 2016b]) may not seem adapted. This is not solely a question of unreported data breaches in PRC: Bisogni et al. [2017] claim that the majority of data breaches proves to be unreported. The exposure to the risk is not easy to track, since the population of potential victims that would report to PRC is not stable through time (or, at least not known in opposition to data coming from an insurance company which can have a clearer view on its exposure, for example).

More precisely, consider ν_i the random number of reported claims in a time period i. This number can be decomposed into $\nu_i = \sum_{j=1}^{w_i} n_j$, where n_j is the number of reported claims concerning the entity j and w_i is the number of entities constituting what could be referred to as the scope of the PRC database. By scope, we mean the set of potential victims on which PRC could obtain information if a claim would occur in the time period i. In a simple modeling where all the variables n_j are identically distributed, the expectation of ν_i is proportional to w_i. Since, from our observation of the database, w_i erratically evolves through time, there is a clear uncertainty about any modeling of the frequency of occurrence.

This is why, in this paper, we do not consider the model developed by Eling and Loperfido [2017], since it does not seem adapted to the poor quality of the PRC base regarding the frequency. The model from [Eling and Loperfido, 2017] has of course many advantages when dealing with a real insurance portfolio or with more consistant and reliable data: while we only focus on a Poisson modeling of the number of claims, Eling and Loperfido consider models with higher number of parameters, including a trend, and develop a goodness-of-fit procedure. We think that public data (and potentially with poor quality) will still constitute an important source of information for insurers until they get sufficient experience on the risk. Therefore, we develop in detail the methodology we used on the PRC as a way to treat such databases.

Our basic idea is to focus on what we consider to be a reliable part of the data. Let us first notice that among the 8860 events release in the database, only 7737 different companies are present, as summarized in Table 7.

Our main (strong) assumption is that, if a company has many reports in the database, this company is efficiently monitored by PRC. Hence we consider that it belongs to the

Number of reports k	1	2	3	4	5	6	7	8	9	10	11	12
Total : 7737	6929	634	103	41	12	6	7	3	1	0	0	1

Table 7: Number of companies having k reports in PRC database for $k = 1, \ldots, 12$ regardless the sector of activity.

Type of organization	$\hat{\lambda}(d)$
Unknown	0.27
Education	0.23
Banking and Insurance	0.15
Business (others)	0.13
Retail (including online)	0.13
Health	0.10
Government	0.07
Nonprofit organization	0.01

Table 8: Poisson rates $\hat{\lambda}(d)$ estimated on 1-truncated count data depending on the sector.

"stable" part of the scope of PRC. Due to the large number of companies with only 1 claim, we therefore only considered companies with more than 2 claims. Let N denote the number of claims that occur for a given company. We choose to model this random variable using a Poisson distribution with parameter depending on the sector of activity.

Let D denote the sector of activity of a given company, with $\lambda(d) = E[N|D = d]$ the corresponding Poisson parameter. We assume that $N|D = d$ is Poisson distributed with

$$\log E[N|D = d] = \alpha + \beta d,$$

that is a Poisson Generalized Linear Model with log link function. In terms of estimation, we take into account the fact that our data is 1-truncated, since our data consists of observations of N given that $N \geq 2$. The results of the estimation of $\hat{\lambda}(d)$ are shown in Table 8.

Figure 2 shows that none of the sources of information has been observed more than 8 years (except for the media source, but which constitutes a small proportion of the total data). Therefore, to compute annual rates of occurrence, we consider that each company has been under observation during only 8 years (and not during the whole 13 years of chronology reported in the database).

$\hat{p}_m(d)$	CARD	DISC	HACK	INSD	PHYS	PORT	STAT
BSF	0.03	0.17	0.3	0.14	0.09	0.23	0.04
BSO	0	0.11	0,62	0.06	0,06	0.13	0.02
BSR	0.06	0.12	0,5	0.12	0,06	0.11	0.03
EDU	0	0.3	0,36	0.03	0,08	0.17	0.06
GOV	0	0.3	0,2	0.11	0,14	0.23	0.03
MED	0	0.24	0,23	0.06	0,33	0.11	0.03
NGO	0	0.13	0,33	0.08	0,1	0.32	0/04

Table 9: Estimates of multinomial parameters $\hat{p}_m(d)$ depending on the sector and the type of breach.

4.3.2 Typology of claims

Until now, we have modeled N without taking into account the variety of situations contained in the variable "Type of breach". On the other hand, the severity can reasonably be thought as strongly dependent on the type of event. Let N_m denote the number of breaches of type m striking a company. We develop a compound Poisson approach, introducing a multinomial random variable Z taking its values in $\{1, ..., M\}$ (M denoting the number of types of breaches, here $M = 7$) and with parameters $p_m(d) = \mathbb{P}(Z = m|D = d)$. Then, considering i.i.d. copies $(Z_i)_{i\geq 1}$ of Z assumed to be independent of N, we consider that

$$N_m = \sum_{i=1}^{N} \mathbf{1}_{Z_i=m}.$$

Estimates of the parameters $p_m(d)$ are provided in Table 9.

4.4 Severity analysis

As we already mentioned, our only way to measure the severity of a cyber incident based on the PRC data is to use the number of records. This information is not available for all claims in the database, but only for 6641 events among the total of 8860. We make the assumption that, if the information about this variable is missing, this should be purely random and non related to the severity of the event.

Let us notice that the severity of data breaches is highly volatile. Indeed, 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%

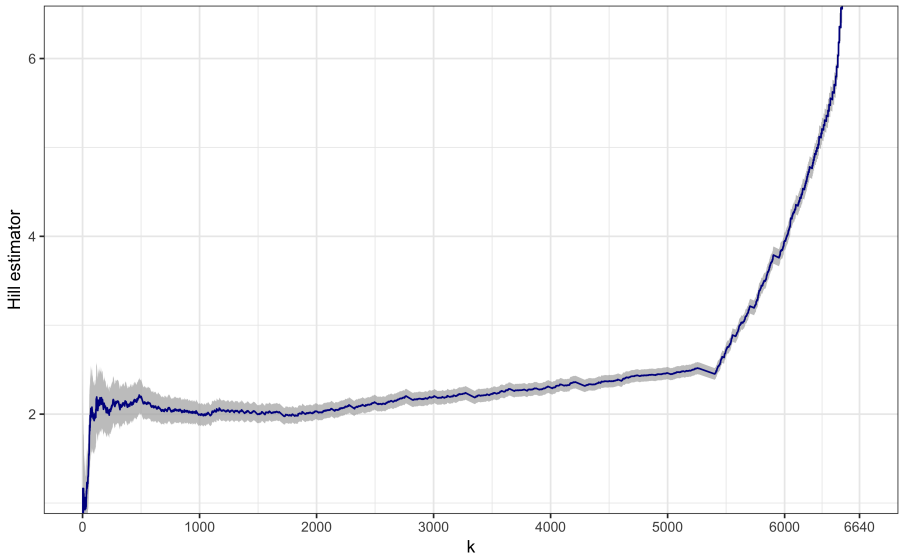

Figure 3: Hill plot for the number of records.

of the total severity and the severity of the top hundred data breaches to 97%. Therefore, the shape of the empirical distribution is highly skew. This motivates to separate the study of the center of the distribution from the right tail.

Another motivation for this separation is the important difference between the median of the number of records (2000) and the empirical mean (1.692 millions). This empirical mean is driven by extreme events (the largest having 3 billions of records).

On the other hand, in the spirit of Section 3.2.1, we investigate the choice of a high threshold u after which a Generalized Pareto behavior is observed. The Hill plot [see Resnick, 2007, pp 85–89] in Figure 3 is a common graphical indicator to perform this choice in classical extreme value analysis. The choice of an appropriate threshold is done by looking at the stability of the Hill plot, leading us to take the value $u = 29156$. This choice corresponds to looking at the 1000 highest data breaches in the sample, standing for around 15% of the total number of breaches. Estimating shape and scale parameter of the Generalized Pareto distribution leads to $\hat{\sigma} = 5.14 \times 10^4$ and $\hat{\gamma} = 2.13$. This analysis is done without considering the potential impact of covariates on tail behavior, and has therefore to be compared with the GPD regression trees obtained in Section 4.4.2.

A first description of this heterogeneity is summarized in Table 10, indicating a need to develop a regression analysis, which is done in Section 4.4.1 for the central part of the distribution, and in Section 4.4.2 for the tail.

Variable	Modality	Number of data breaches	Median	Share of extreme values	Shape (lower bound 95%)	Shape (estimate)	Shape (upper bound 95%)
Source	Unknown	34	2894	21 %	0.44	3.46	6.48
	Media	641	10597	38 %	2.55	3.07	3.59
	US GA: State	1114	566	8 %	1.12	1.72	2.32
	US GA: Federal	2380	2267	9 %	1.15	1.48	1.80
	HIPAA						
	Nonprofit organization	2472	2000	18 %	1.29	1.53	1.76
Type of breach	CARD	32	300	12 %	-0.98	2.81	6.61
	STAT	184	3060	20 %	0.45	1.11	1.76
	INSD	377	651	12 %	1.14	2.07	2.99
	Unknown	636	589	8 %	1.16	1.93	2.71
	PORT	874	3575	21 %	1.02	1.37	1.72
	PHYS	1463	1726	7 %	0.89	1.37	1.85
	DISC	1469	1615	11 %	1.91	2.43	2.95
	HACK	1606	4605	26 %	2.19	2.54	2.88
Sector of organization	NGO	75	2000	16 %	0.32	1.92	3.53
	BSR	299	1000	27 %	1.94	2.75	3.56
	Unknown	376	500	6 %	0.21	1.25	2.29
	BSF	412	2005	24 %	1.63	2.29	2.95
	BSO	428	5830	36 %	2.62	3.30	3.99
	GOV	562	2837	23 %	1.12	1.58	2.03
	EDU	685	2400	17 %	0.65	1.01	1.38
	MED	3804	2039	10 %	1.14	1.38	1.62
Year	2005	117	16500	43 %	0.69	1.37	2.04
	2006	385	2000	21 %	0.90	1.49	2.08
	2007	340	3000	21 %	0.94	1.51	2.08
	2008	270	3800	23 %	1.03	1.66	2.28
	2009	193	2000	24 %	1.12	1.97	2.83
	2010	579	1907	13 %	0.97	1.56	2.14
	2011	585	1800	12 %	1.26	1.93	2.60
	2012	629	1593	11 %	1.09	1.73	2.36
	2013	589	2600	11 %	1.12	1.77	2.43
	2014	606	1204	14 %	1.52	2.19	2.87
	2015	377	2259	15 %	1.78	2.79	3.79
	2016	608	2500	15 %	2.05	2.85	3.66
	2017	449	2969	15 %	1.72	2.61	3.50
	2018	914	947	12 %	2.07	2.76	3.45

Table 10: Descriptive statistics for the different groups of the PRC database.

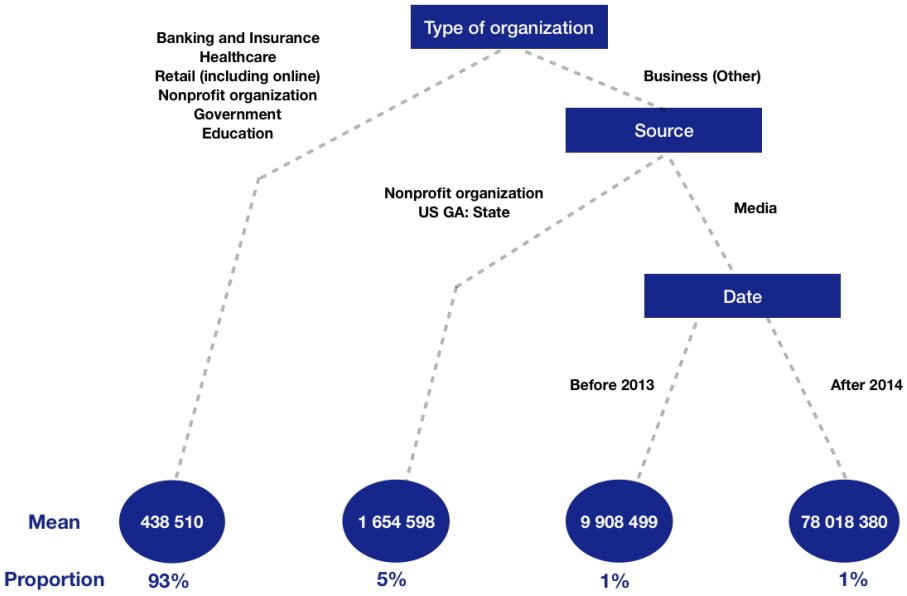

Figure 4: Tree obtained by the CART algorithm based on the quadratic loss.

4.4.1 Mean and median regression trees

To evaluate the center part of the distribution, we apply the Regression Tree procedure first using a square loss, that is, from the notations of Section 3.1, $\phi(y, m(\mathbf{x})) = (y - m(\mathbf{x}))^2$. The tree obtained in Figure 4 is then an estimator of $E[Y|\mathbf{X} = \mathbf{x}]$. The tree has been computed using the `R` package `rpart` [see Therneau and Clinic, 2018], decomposing randomly the database into a train set (67% of the data) on which the maximal tree is built, and a validation set (33% of the data) that is used for error measurement and the selection of a proper subtree from the pruning step.

From the obtained tree, we can observe that the first splitting variable is the type of organization, with essentially highest predicted costs for the category "Business (others)", which is probably the most heterogeneous due to the uncertainty about the activities of the corresponding companies. Another interesting issue is the importance of the source in the analysis: this variable only appears at one node, but it shows that the heterogeneity of the sources feeding the database has not only impact on the estimation of the frequency, but also on the estimation of the severity. On the other hand, most of the events (93%) are gathered in the same leaf, leading to a first impression that, regarding the conditional mean, the heterogeneity of the claims is not so obvious apart from very large ones.

Estimation of $E[Y|\mathbf{X} = \mathbf{x}]$ is of course crucial in order to perform pure premium pricing, but the estimation of the conditional mean is known to lack robustness, since the

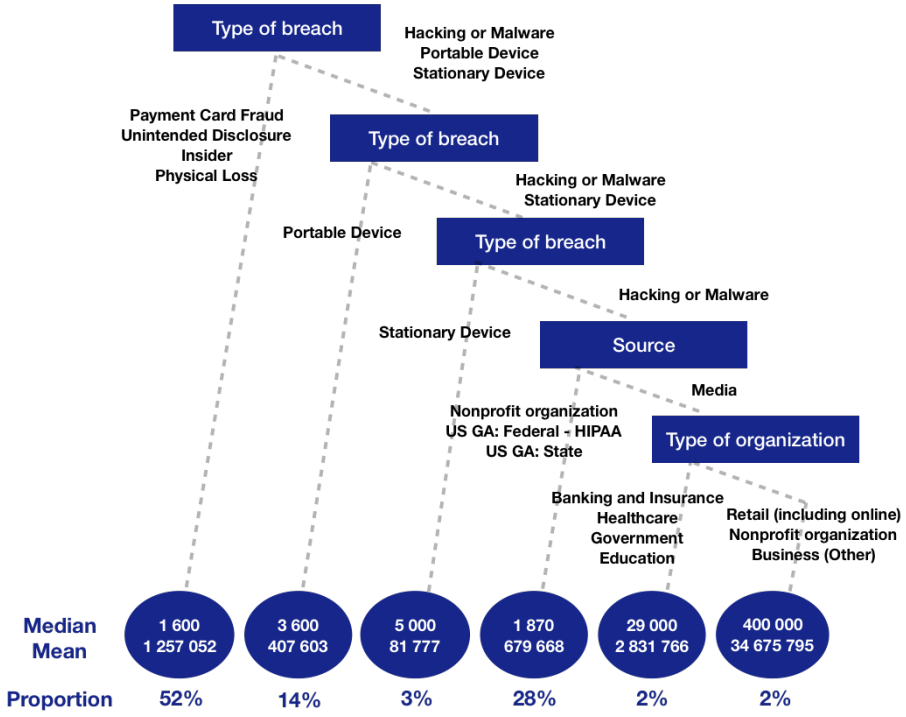

Figure 5: Tree obtained by the CART algorithm based on the median splitting rule.

quadratic loss may be influenced by too large observations. Since the variable Y is highly volatile, we challenge the tree of Figure 4 by computing a median regression tree, that is with loss function $\phi(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$, providing an estimator of the median of the distribution of $Y|\mathbf{X} = \mathbf{x}$. This leads to the tree presented in Figure 5.

We see that the structure of the obtained tree is quite different from Figure 4, which seems to indicate that the largest observations strongly impact the estimation of the conditional expectation. The obtained classes (that is leaves of the median tree) are associated with smaller values of the median than of the conditional expectation associated with leaves of the mean tree. Moreover, the situation in the category "Business (others)" is different from the tree of Figure 4, since the high predicted number is caused essentially by a particular shape of cyber incident (a malicious event caused by hacking or malware, and reported by a media source). The variable "Type of breach" also appears in the decomposition, which was not the case in the tree of Figure 4.

Figure 4 also presents the mean of the events that are gathered in each leaf. Indeed, a possible compromise between the two trees of Figure 4 and 5 is to use the decomposition obtained using the more robust criterion of the median, but still estimating an expectation

in each group. The impact of such an approach on the evaluation of the risk of virtual portfolios is considered in Section 5.

4.4.2 GPD Regression Tree

Let us recall that the GPD regression tree is built from the 1000 largest observations in terms of records. The tree has been computed decomposing randomly the extreme observations into a train set (67% of the data) on which the maximal tree is built, and a validation set (33% of the data) that is used for error measurement and the selection of a proper subtree from the pruning step. The obtained tree is shown in Figure 6.

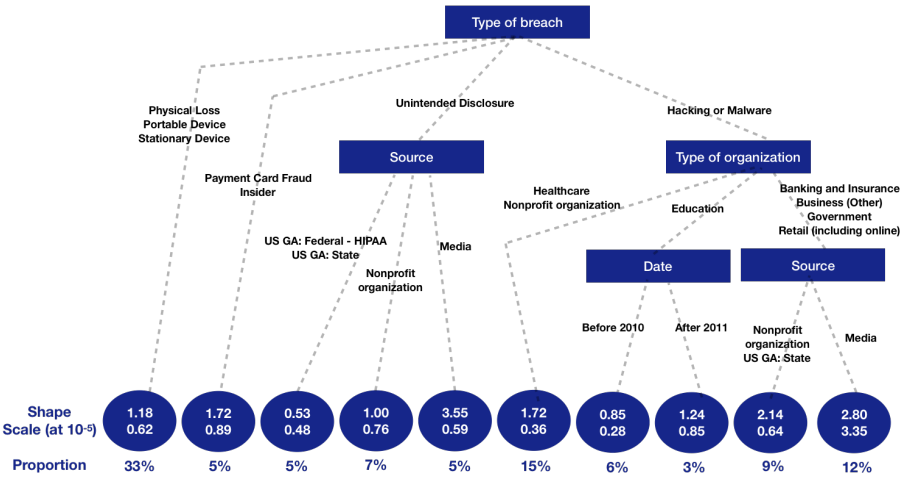

Figure 6: Tree obtained by the CART algorithm based on the Generalized Pareto likelihood splitting rule.

As we can see, we observe significant differences in the tail behaviors depending on the cases. Let us observe that the worst case corresponds to a shape parameter of 3.55, far from the shape parameter 2.14 obtained by our preliminary analysis on the whole set of 1000 largest observations (that is without clustering). This can be explained by the fact that the media source seems to deviate from the rest of the population, since this source is characterized by larger events. Clearly, one may fear the instability of the calibrated model, since the structure of regression trees is known to be sensitive to the introduction of new data, and this model should be monitored on a regular basis.

5 An application on virtual portfolios

In this section, we try to measure the impact of the previous analysis on virtual portfolios. Considering a given number of policies ($n_p = 20000$ in the following), we consider different configurations for the composition of the virtual portfolios in terms of types of insured companies, which are summarized in Table 11.

VP	q_{BSF}	q_{BSO}	q_{BSR}	q_{MED}
$1st$	25 %	25 %	25 %	25 %
$2nd$	40 %	20 %	20 %	20 %
$3rd$	33 %	33 %	33 %	0 %
$4th$	20 %	30 %	30 %	20 %

Table 11: Proportions of sectors of the companies in virtual portfolios. q_{x} stands for the proportion of companies from sector x.

Compared to the PRC database, we only consider 4 types of companies. Then we simulate number of claims (see Section 4.3.1), type of claims (see Section 4.3.2). Our models for the severity require the knowledge of the type of reporting source. We assume that this source is government agencies at a federal level, since regulation considerations seem to make this source more reliable.

We then try to evaluate a "central scenario" for each portfolio. This is done thanks to the linearity of the expectation by using two different severity models. First by using the mean regression tree of Figure 4: conditionally on the numbers and the typologies of the claims, we look at the corresponding leaf of the tree and report the mean value for the number of records. The second way consists in using the clustering obtained through the median tree, and then predicting the amount using the estimated mean in each leaf (see Section 4.4.1). In each case, the link between the number of records Y and the amount L is done first using (2.1), then (2.2).

We also want to evaluate the weight of the tail of the distribution. To this aim, we use the GPD tree of Section 4.4.2, see Figure 6. We focus in the 15% highest part of the distribution i.e. on claims in excess of the threshold u used in the calibration of our GPD distributions. This is done by simulating the size of the claims (10 000 simulations by portfolio) based on the distribution of the GPD trees (and augmented by the threshold u), by computing an associated loss and by retaining only the amount of the claim in excess of the loss associated with a claim of size u. The results are gathered in Table 12.

VP	Expected loss per policy				Share of loss in excess per policy			
	Mean tree		Median tree		GPD tree / Mean tree		GPD tree / Median tree	
	Jacobs formula	Formula (2.2)	Jacobs formula	Formula (2.2)	Jacobs formula	Formula (2.2)	Jacobs formula	Formula (2.2)
1^{st}	74 210 855	15 353 511	8 751 594	5 551 376	5 182 %	20.9 %	43 941 %	57.9 %
2^{nd}	60 640 433	13 214 110	9 055 042	5 743 248	6 674 %	23.9 %	44 694 %	55.0 %
3^{rd}	97 518 899	19 422 361	9 024 747	5 776 944	7 352 %	21.1 %	79 441 %	70.9 %
4^{th}	87 987 095	17 643 633	8 681 808	5 523 106	5 169 %	19.5 %	52 388 %	62.4 %

Table 12: Evaluation of the mean loss for the virtual portfolios using mean-based and median-based trees; comparison of the cost of the 10% upper part of the distribution with the central scenario of the mean tree and median tree.

As expected, the results highlight the influence of the composition of portfolios on an expected loss and a loss in excess for a (re)insurer point of view. It also enhances the limit of the Jacobs formula concerning mega data breaches that seems to overestimate the costs. We see that even with the less pessimist Formula (2.2), the weight of the cost of the largest claim is heavy compared to the central scenario.

6 Conclusion

In this paper, we have taken a closer look at the heterogeneity of cyber events in order to improve their evaluation and anticipation for insurance purpose. The main difficulty is clearly the lack of data: the public PRC database is very rich in terms of reported breaches, but has not been designed for insurance purpose. Therefore, we have tried to emphasize the difficulty to estimate the frequency of claims using such data with such an uncertainty on the exposure. This uncertainty is not expected from databases provided by insurance companies, since they are aware of the number of policies they have, and have informations on their characteristics. On the other hand, the study of databases such as PRC is still important (at least for combining it with insurance data) since it is one of the few databases gathering such a high number of cyber events, in a context where the experience of insurers is quite recent. From a methodological point of view, our main contribution is the analysis of the severity using regression trees. It raises the difficulty to estimate a conditional mean due to the high volatility in the severity variable, therefore the combination of a median regression tree to determine cluster, and a mean evaluation inside these clusters seems to us a reasonable solution. Regarding the tail of the distribution, our study via GPD trees highlights an heterogeneous severity. Understanding its heterogeneity is important since we see, from the empirical study on virtual portfolios

of Section 12, that it can have strong impacts on the final result depending on the structure of the portfolio.

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.

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. Regression with response distributions of pareto-type. *Computational statistics & data analysis*, 42(4):595–619, 2003.

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.

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, 2016a.

S. Forrest, S. Hofmeyr, and B. Edwards. Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, 2, 2016b.

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.

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`.