# Tree-based methods and extreme value analysis for cyber insurance

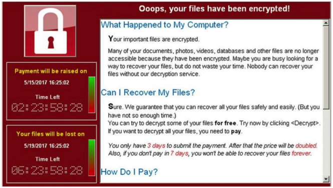
Olivier Lopez
(joint work with S. Farkas, A. Heranval, M. Thomas)
Sorbonne Université
Laboratoire de Probabilités Statistique et Modélisation UMR CNRS 8001

June 7th 2023

# Cyber-risk

- Cyber-risk: inappropriate use of numerical tools and information systems.
- A cyber incident can be voluntary (cyber attack) or not (accidents may happen).
- For hacking, hackers use vulnerabilities in information systems, from outside or from inside.
- Various types of attacks (ransomware, phishing, classic frauds...)
- Strike states, companies, people.

# Wannacry



Ooops, your files have been encrypted!

Payment will be raised on

5/15/2017 16:25:02

Time Left

02:23:58:28

Your files will be lost on

5/19/2017 16:25:02

Time Left

06:23:58:28

What Happened to My Computer?

Your important files are encrypted.

Many of your documents, photos, videos, databases and other files are no longer accessible because they have been encrypted. Maybe you are busy looking for a way to recover your files, but do not waste your time. Nobody can recover your files without our decryption service.

Can I Recover My Files?

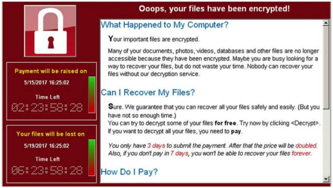
Sure. We guarantee that you can recover all your files safely and easily. (But you have not so enough time.)
You can try to decrypt some of your files **for free**. Try now by clicking <Decrypt>.
If you want to decrypt all your files, you need to **pay**.

*You only have 3 days to submit the payment. After that the price will be doubled.*
*Also, if you don't pay in 7 days, you won't be able to recover your files forever.*

How Do I Pay?

- Ransomware Wannacry : worldwide cyber attack in May 2017.
- Use the vulnerability "EternalBlue".
- Approximatively 200 000 infected computers across 150 countries over approximatively one week.
- Estimation of the cost : hundreds of million dollars, billions according to some estimations. (£100 millions for the NHS).

# Wannacry



Ooops, your files have been encrypted!

Payment will be raised on

5/15/2017 16:25:02

Time Left

02:23:58:28

Your files will be lost on

5/19/2017 16:25:02

Time Left

06:23:58:28

What Happened to My Computer?

Your important files are encrypted.

Many of your documents, photos, videos, databases and other files are no longer accessible because they have been encrypted. Maybe you are busy looking for a way to recover your files, but do not waste your time. Nobody can recover your files without our decryption service.

Can I Recover My Files?

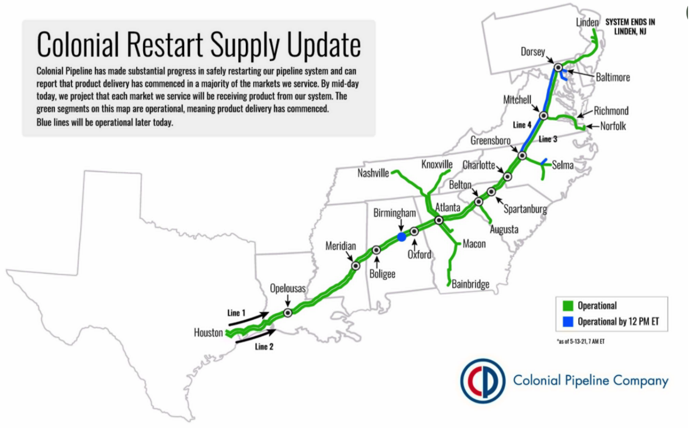
Sure. We guarantee that you can recover all your files safely and easily. (But you have not so enough time.)
You can try to decrypt some of your files **for free**. Try now by clicking <Decrypt>.
If you want to decrypt all your files, you need to **pay**.

*You only have 3 days to submit the payment. After that the price will be doubled.*
*Also, if you don't pay in 7 days, you won't be able to recover your files forever.*

How Do I Pay?

- Ransomware Wannacry : worldwide cyber attack in May 2017.
- Use the vulnerability "EternalBlue".
- Approximatively 200 000 infected computers across 150 countries over approximatively one week.
- Estimation of the cost : hundreds of million dollars, billions according to some estimations. (£100 millions for the NHS).
- NotPetya: June 2017, uses the same vulnerability, also huge losses.

# Colonial Pipeline



## Colonial Restart Supply Update

Colonial Pipeline has made substantial progress in safely restarting our pipeline system and can report that product delivery has commenced in a majority of the markets we service. By mid-day today, we project that each market we service will be receiving product from our system. The green segments on this map are operational, meaning product delivery has commenced. Blue lines will be operational later today.

🟩 Operational

🟦 Operational by 12 PM ET

*as of 5-13-21, 7 AM ET

Colonial Pipeline Company

- 4.2% increase of WTI and Brent.
- "Double extorsion": ransomware attack combined with blackmail.
- Authors: the hacker group "Darkside" (Ransomware as a service).

# Some numbers (French market)

- The « Association pour le Management des Risques et des Assurances de l'Entreprise » (AMRAE) published its second version of the LUCY study in 2022.
- Used data from brokers in the French market.
- Loss ratios in 2021: 88%
- Previous Loss ratios: 167% in 2020, 84% in 2019.
- What is new in 2021:
  - Insurance capacities are smaller
  - Deductibles increase
  - Premium increases (+44,4% compared to an estimated growth of the market of 27,5%)
  - The coverage of large companies is diminishing (-4,4%)

# Report from the French National Direction of Treasury

- September 2022: report from the working group of the French National Direction of Treasury on cyber insurance:

  https://www.tresor.economie.gouv.fr/Articles/2022/09/07/

  remise-du-rapport-sur-le-developpement-de-l-assurance-du-risque-cyber
- Identifies some difficulties and levers to develop an efficient cyber insurance ecosystem.
- Among the needs mentioned by the report:
  - lack of data;
  - innovation in terms of models;
  - loss of mutualization, extreme claims, accumulation risk.

# Aim of this talk

- We propose a way to fit a model for the loss of a severe cyber event.
- This cyber event can be:
  - a catastrophic claim striking a single actor;
  - the aggregation of many claims occurring in a small amount of time (accumulation).
- Applications: pricing, reinsurance, selection of policyholders or of limits of guarantees...
- Possible connexions with some frequency models, see for example:

Hillairet, C., Lopez, O., d'Oultremont, L., Spoorenberg, B. (2022). "Cyber-contagion model with network structure applied to insurance," *Insurance: Mathematics and Economics*, vol. **107**, pp. 88-101.

# Outline

1 Introduction

2 Generalized Pareto regression trees
- Approximation of the tail via Generalized Pareto Distributions
- Generalized Pareto Regression Trees fitting
- Illustration in the case of data breaches

3 A Bayesian model
- Credibility theory
- Calibration of the prior

4 Conclusion

# Outline

1 Introduction

2 Generalized Pareto regression trees
- Approximation of the tail via Generalized Pareto Distributions
- Generalized Pareto Regression Trees fitting
- Illustration in the case of data breaches

3 A Bayesian model
- Credibility theory
- Calibration of the prior

4 Conclusion

# Notation and context

- In the following, $Y$ is a random variable used to model the loss linked to a cyber attack.
- We assume that $Y$ is heavy tail in the sense that

$$S_Y(t) = \mathbb{P}(Y \geq t) = \frac{l(t)}{t^{1/\gamma}},$$

  where $l$ is a slow-varying function, and $\gamma > 0$ is the tail index.
- $\mathbf{X} \in \mathbb{R}^d$ corresponds to covariates (quantitative and/or qualitative), like, for example:
  - type of attack (ransomware, Ddos...);
  - information on the victim (type of company, size, budget allocated to cyber security...);
  - consequences (business interruption, third party...)
- $(Y_1, \mathbf{X}_1, \cdots, Y_n, \mathbf{X}_n)$ = the sample of observations to calibrate the model.

# Extreme value theory

## Generalized Pareto Distribution

A random variable $Z$ with Generalized Pareto Distribution of parameters $(\gamma, \sigma)$ is characterized by its survival function

$$S_{\gamma,\sigma}(z) = \mathbb{P}(Z \geq z) = \begin{cases} \frac{1}{\left(1+\frac{z\gamma}{\sigma}\right)^{1/\gamma}} & , \quad \gamma \neq 0 \\ \exp\left(-\frac{z}{\sigma}\right) & , \quad \gamma = 0 \end{cases} .$$

# Extreme value theory

## Generalized Pareto Distribution

A random variable $Z$ with Generalized Pareto Distribution of parameters $(\gamma, \sigma)$ is characterized by its survival function

$$S_{\gamma,\sigma}(z) = \mathbb{P}(Z \geq z) = \begin{cases} \frac{1}{\left(1+\frac{z\gamma}{\sigma}\right)^{1/\gamma}} & , \quad \gamma \neq 0 \\ \exp\left(-\frac{z}{\sigma}\right) & , \quad \gamma = 0 \end{cases} .$$

**Approximation beyond a threshold:**
Pickands (1975): for a random variable $Y$, let $S_{u_n}(y) = \mathbb{P}(Y - u_n \geq y | Y \geq u_n)$, there exists $(\gamma, \sigma_n)$ such that

$$\lim_n |S_{u_n}(y) - S_{\gamma,\sigma_n}(y)| = 0,$$

where $u_n$ tends towards $\tau_S = \sup\{y : S(y) > 0\}$.

# Analysis on extreme cyber events

Edwards B., Hofmeyr S. and Forrest S. (2016), "Hype and heavy tails: A closer look at data breaches." *Journal of Cybersecurity*, vol. **2** (2057-2085), pp. 3-14.

Eling M. and Loperfido N. (2017), "Data breaches: Goodness of fit, pricing, and risk measurement." *Insurance: Mathematics and Economics*, vol. **75** (0167-6687), pp. 126-136.

Wheatley S., Maillart T. and Sornette D. (2016), "The extreme risk of personal data breaches and the erosion of privacy." *European Physical Journal B*, vol. **89** (1434-6036), pp. 7.

- In the particular case of data leaks, these authors show evidence that the distribution of the loss is heavy tail.

# GPD an insurability

- $Y =$ loss associated to a cyber claim.
- $\gamma > 0$ (heavy-tailed).
- If $Z = Y - u|Y \geq u$ is Generalized Pareto distributed with parameters $\gamma$ and $\sigma$,

$$E[Z|Z \geq 0] = \frac{\sigma}{1-\gamma},$$

  if $\gamma < 1$.
- If $\gamma \geq 1$, infinite expectation, "not insurable". The insurer :
  - canexclude the risk.
  - can introduce limits to guarantees (lower if $\gamma$ is high).

# Mixtures of GPD

- Consider that a population is a mix between to type of extreme behaviors.
- $Y = \delta Y_1 + (1-\delta) Y_2$, where $\delta$ is an (unobservable) Bernoulli distributed random variable, independent from $Y_1$ (tail index $\gamma_1$) and $Y_2$ (tail index $\gamma_2$).
- Then, the tail index of $Y$ is $\gamma = \max(\gamma_1, \gamma_2)$.
- Extends to more general mixtures.
- Consequence: if we do not manage to distinguish these two subpopulations, we will apply the worst case scenario.

# Mixtures of GPD

- Consider that a population is a mix between to type of extreme behaviors.
- $Y = \delta Y_1 + (1-\delta) Y_2$, where $\delta$ is an (unobservable) Bernoulli distributed random variable, independent from $Y_1$ (tail index $\gamma_1$) and $Y_2$ (tail index $\gamma_2$).
- Then, the tail index of $Y$ is $\gamma = \max(\gamma_1, \gamma_2)$.
- Extends to more general mixtures.
- Consequence: if we do not manage to distinguish these two subpopulations, we will apply the worst case scenario.
- Idea: use the risk factors $\mathbf{X} \in \mathbb{R}^d$ to identify different type of populations / claims that are associated with different values of $\gamma$.

# Classical way to deal with risk factors: Generalized Linear Model

## Generalized Linear Model

Let $Y$ denote the response variable, and $\mathbf{X} \in \mathbb{R}^d$ a set of covariates. In a GLM, one assumes that

$$g(E[Y|\mathbf{X}]) = \beta_0 + \beta^T \mathbf{X},$$

where

- $g$ is a monotonic (known) link function
- the distribution of $Y|\mathbf{X}$ belongs to a given exponential family of distributions.

# Why GLM does not seem a good idea

- Parametric model: relies on strong assumptions that may be far to be true in practice.
- Linearity (up to some known transformation) of the effects is a constraint.
- GLM does not allow to consider "extreme events".
- Targets the "central scenario" $E[Y|\mathbf{X}]$.

# Parametric methods

- We have risk factors $\mathbf{X} \in \mathbb{R}^d$ and we want to understand their impact on the tail index, say $\gamma(\mathbf{x})$ the tail index of the distribution of $Y|\mathbf{X} = \mathbf{x}$.

- First strategy: make a parametric assumption on $\gamma(\mathbf{x})$, for example

$$\gamma(\mathbf{x}) = f(\theta_0, \mathbf{x}),$$

  for some $f$ known, $\theta_0 \in \mathbb{R}^k$ unknown.

- Estimation can be performed by pseudo-maximum likelihood on the observations that exceed a certain threshold.

- Problem: which type of function $f$ should be chosen ?

# Nonparametric methods

- One assumes nothing but smoothness on $\gamma(\mathbf{x})$ (but is it realistic ?).
- Estimation relies on kernel smoothing, the simplest version is

$$(\hat{\gamma}(\mathbf{x}), \hat{\sigma}(\mathbf{x})) = \arg\max \sum_{i=1}^{n} K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right) \log f_{\gamma,\sigma}(Y_i - u)\mathbf{1}_{Y_i \geq u},$$

  where
  - $f_{\gamma,\sigma}$ is the density of a GPD with parameters $\gamma$ and $\sigma$
  - $u$ is a threshold
  - $K$ is a kernel function (integral $= 1$)
  - $h$ is a bandwidth close to zero
- Does not apply to discrete covariates.

# Clustering And Regression Trees (CART)

- Introduced by Breiman (1984). Many extensions: Su (2004), Hothorn (2006), Loh (2014),...
- Consider a random variable $Y$ and $\mathbf{X}$ some covariates.
- Regression trees:
  - combining clustering with regression (that is evaluation of the impact of covariates on a variable).
  - regression trees aim to estimate a function $m(x)$ (characterizing the distribution of $Y$ when $X = x$, for example $m(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$) by

$$\hat{m}(\mathbf{x}) = \sum_{j=1}^{K} m_j R_j(\mathbf{x}),$$

where $R_j$ are called a "rule," that is $R_j(\mathbf{x}) = 0$ or 1, and, for all $\mathbf{x}$, only one $R_j(\mathbf{x})$ is nonzero.

# Clustering And Regression Trees (CART)

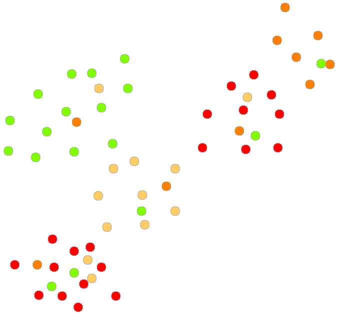
## Regression tree (Breiman et al., 1984)

$$m^* = \arg\min_{m \in \mathcal{M}} [\phi(Y, m(\mathbf{X}))],$$

- $Y$ 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 $\mathbb{R}^d$
- $\phi$ is a loss function that depends on the quantity we wish to estimate

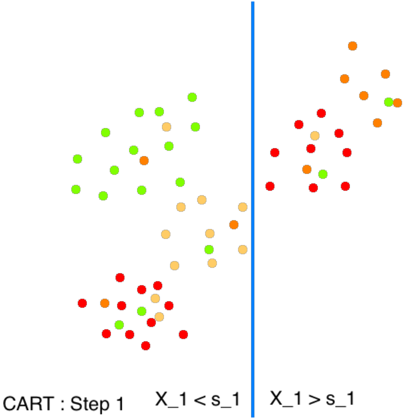
- if we take $\phi =$ quadratic loss, $m^*$ is the conditional expectation;
- if we take $\phi =$ absolute loss, $m^*$ is the conditional median;
- ...

# A few words about the algorithm

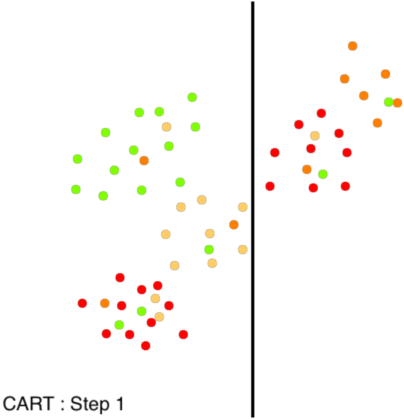


CART : Step 0

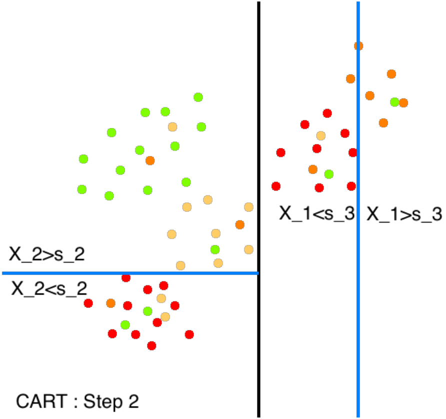
# A few words about the algorithm



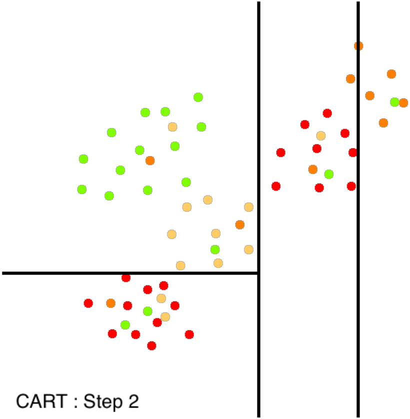
# A few words about the algorithm



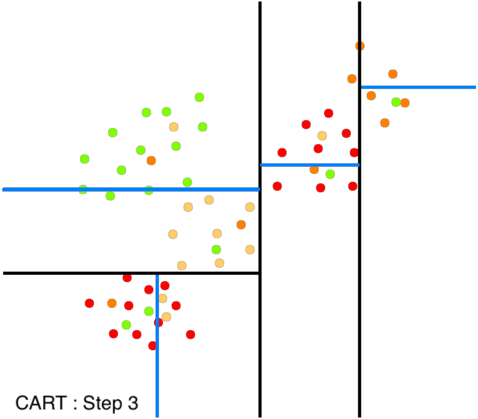
# A few words about the algorithm



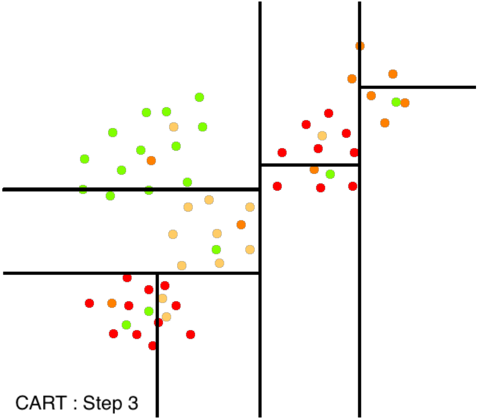
# A few words about the algorithm



# A few words about the algorithm



# A few words about the algorithm



# The splitting rule and loss functions

Loss functions considered:

- to analyze the "center of the distribution":
  - the quadratic loss $\phi(y, m(\mathbf{x})) = (y - m(\mathbf{x}))^2$,
  - the absolute loss $\phi(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$,
- to analyze the tail:
  - a log-likelihood loss $\phi(y, m(\mathbf{x})) = -\log f_{m(\mathbf{x})}(y)$, where $\mathcal{F} = \left\{f_\theta : \theta \in \Theta \subset \mathbb{R}^k\right\}$ is a parametric family of densities.
- Generalized Pareto log-likelihood as splitting criterion:

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

  where $m(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$.
- In this last case, this loss function is applied only to observations larger than some threshold.

# Pruning step: model selection

- Let $T_{\max}$ be the maximal tree obtained in the first phase and $K_{\max}$ the number of its leaves
- Pruning: consists in the extraction of a subtree from $T_{\max}$
- Penalized criterion ($n_T$ number of leaves of tree $T$)

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

- One can shows that the "best" tree with $K$ leaves can be extracted from the "best" tree with $K+1$ leaves, which makes the selection feasible through some kind of backward selection.
- $\alpha > 0$ is chosen by cross-validation or with test sample.

# Pruning step: model selection

- Let $T_{\max}$ be the maximal tree obtained in the first phase and $K_{\max}$ the number of its leaves
- Pruning: consists in the extraction of a subtree from $T_{\max}$
- Penalized criterion ($n_T$ number of leaves of tree $T$)

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

- One can shows that the "best" tree with $K$ leaves can be extracted from the "best" tree with $K+1$ leaves, which makes the selection feasible through some kind of backward selection.
- $\alpha > 0$ is chosen by cross-validation or with test sample.
- Denote $\widehat{T}_K$ the best tree with $K$ leaves according to this criterion, $T_K^*$ the best tree with $K$ leaves for the criterion $E[C_\alpha(T)]$.
- $\hat{T}$ the tree minimizing the penalized criterion, $\hat{K}$ its number of leaves.

# Some theory (short)

- Let $\|T - U\|_2^2 = \int (T(x) - U(x))^2 d\mathbb{P}(x)$.

## Consistency of the tree

Under some assumptions,

$$\mathbb{P}\left(\|T_K - T_K^*\|_2^2 \geq t\right) \leq 2\left\{\exp\left(-\frac{C_1 k_n t}{K[\log n]^2}\right) + \exp\left(-\frac{C_2 k_n t^{1/2}}{K^{1/2}\log n}\right)\right\} + \frac{C_3 K}{k_n t^{3/2}},$$

and

$$E\left[\|\widehat{T}_K - T_K^*\|_2^2\right] \leq C_4 \frac{K(\log n)^2 \log(n/k_n)}{k_n}.$$

# Consistency of pruning step

- Let $K_0$ denote the number of leaves of the "best" $T_K^*$ according to $E[C_\alpha(T)]$.

## Consistency of the pruning step

Under some assumptions,

$$E[\|\hat{T} - T_{K_0}^*\|_2^2] \leq C_4 \frac{K_0(\log n)^2 \log(n/k_n)}{k_n}.$$

- More details:

S. Farkas, A. Heranval, O. Lopez, M. Thomas, "Generalized Pareto Regression Trees for extreme events analysis" (2023) *Preprint* https://arxiv.org/abs/2112.10409.

# The case of data breaches (PRC database)

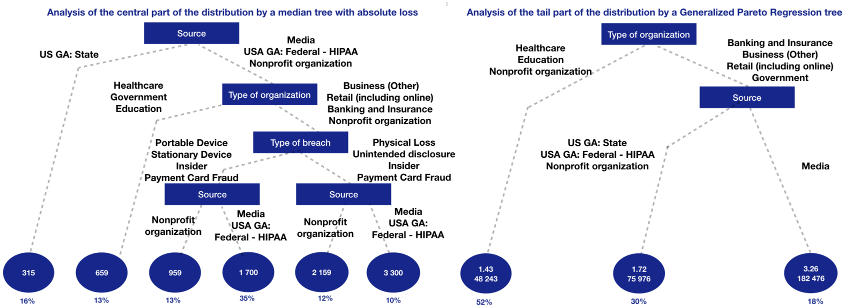
- Chronology of data breaches maintained by Privacy Rights Clearinghouse association (US) since 2005.
- $Y$ is here the "number of records" affected (gives an idea of the volume of data exposed by the event).
- Attempt to link the cost of the event to the number of records:

$$\log(Costs) \approx 9.59 + 0.57 \times \log(Records).$$

- This (very) rough formula is an updated version of:
  - a formula computed by Jacobs in the Cost of Data Breaches report 2014 (based on incidents in 2013 and 2014) from the Ponemon Institute;
  - with inclusion of more recent mega breaches.
- More details :

Farkas, S., Lopez, O., Thomas, M. (2021). "Cyber claim analysis using Generalized Pareto regression trees with applications to insurance," *Insurance: Mathematics and Economics*, vol. **98**, p. 92-105.

# GPD Regression Trees



Analysis of the central part of the distribution by a median tree with absolute loss

Source

US GA: State

Media
USA GA: Federal - HIPAA
Nonprofit organization

Type of organization

Healthcare
Government
Education

Business (Other)
Retail (including online)
Banking and Insurance
Nonprofit organization

Type of breach

Portable Device
Stationary Device
Insider
Payment Card Fraud

Physical Loss
Unintended disclosure
Insider
Payment Card Fraud

Source

Nonprofit organization

Media
USA GA: Federal - HIPAA

Source

Nonprofit organization

Media
USA GA: Federal - HIPAA

315 — 16%

659 — 13%

959 — 13%

1 700 — 35%

2 159 — 12%

3 300 — 10%

Analysis of the tail part of the distribution by a Generalized Pareto Regression tree

Type of organization

Healthcare
Education
Nonprofit organization

Banking and Insurance
Business (Other)
Retail (including online)
Government

Source

US GA: State
USA GA: Federal - HIPAA
Nonprofit organization

Media

1.43
48 243
52%

1.72
75 976
30%

3.26
182 476
18%

- Let us recall that $\gamma_L \approx 0.5\gamma_Y$.
- Corresponding tail index when it comes to the cost: (from left to right) 0.82, 0.98, 1.86.
- Tail index estimation if one does not separate claims into clusters: 2.16 (for the cost: 1.23).

# Remarks on regression trees

- The method identifies clusters of claims, and can help to draw a line between what can be insured and what can not.
- On the data: the variables that drive the central part of the distribution are not the same as the one that drive the tail.
- This database is an illustration, but the lack of data on cyber risk by insurance companies needs to be compensate by looking at public databases to build a "prior".
- In the paper: also a frequency analysis of the claims, and how it can be used to price (or compute reserves) for virtual portfolios.
- Extension: "black-box methods" (Random Forests, Gradient Boosting...)

# Outline

1 Introduction

2 Generalized Pareto regression trees

- Approximation of the tail via Generalized Pareto Distributions
- Generalized Pareto Regression Trees fitting
- Illustration in the case of data breaches

3 A Bayesian model

- Credibility theory
- Calibration of the prior

4 Conclusion

# A Bayesian credibility model

- Let us consider that we have an history of claim losses $(Y_1, \cdots, Y_N)$ on a given segment.
- By segment we mean either:
  - a policyholder or a class of policyholders;
  - a type of cyber event;
  - a type of cyber event on a class of policyholders...
- Let us assume that :
  - there is a hidden factor $\theta$ such that $(Y_i)_{1 \leq i \leq N}$ are independent, identically distributed conditionally on $\theta$;
  - $Y_1 | \theta = t \sim \mathcal{E}(t)$;
  - prior distribution: $\theta \sim \Gamma(r, \lambda)$.

# Posterior distribution

- Closed formula for the posterior distribution of $\theta | Y_1, \cdots, Y_N$ :

$$\theta | Y_1, \cdots, Y_N \sim \Gamma\left(r+n, \lambda + \sum_{i=1}^{N} Y_i\right).$$

- Let us assume that we want to compute a pure premium:

$$\begin{aligned}\pi(Y_1, \cdots, Y_N) &= E\left[Y_{N+1} | Y_1, \cdots, Y_N\right] \\ &= E\left[E\left[Y_{N+1} | \theta, Y_1 n \cdots, Y_N\right] | Y_1, \cdots, Y_N\right] \\ &= E\left[\frac{1}{\theta} | Y_1, \cdots, Y_N\right] \\ &= c_N(r) \frac{\sum_{i=1}^{N} Y_i}{N} + (1 - c_N(r)) \frac{\lambda}{r-1},\end{aligned}$$

for $r > 1$, with

$$c_N(r) = \frac{N}{r+N-1}.$$

# Distribution of $Y$

- Write

$$\begin{aligned}\mathbb{P}(Y \geq y) = E\left[\mathbb{P}(Y \geq y|\theta)\right] &= \int_0^{\infty} \exp(-ty) p_{r,\lambda}(t)dt \\ &= \left(\frac{\lambda}{\lambda + y}\right)^r.\end{aligned}$$

- Consequence: the distribution of $Y$ is a GPD with parameters

$$\begin{aligned}\gamma &= \frac{1}{r}, \\ \sigma &= \frac{\lambda}{r}.\end{aligned}$$

- The expectation of $Y$ (if finite) is

$$E[Y] = \frac{\lambda}{r-1}.$$

# How to calibrate the prior distribution?

- We have two types of information:
  - a collective database, on which we have elements on claims, with $(Z_1, \mathbf{X}_1, \cdots, Z_n, \mathbf{X}_n)$ where $Z_i =$ loss, $\mathbf{X}_i =$ covariates;
  - we have individual information $(Y_1, \cdots, Y_N)$ as in the hidden factor model we considered.
- The collective database is assumed to be i.i.d. with same distribution as $Y$.
- Examples:
  - the collective database is an external database (provided by cybersecurity experts, national statistics...);
  - the collective database corresponds to the data of the whole portfolio, while $(Y_1, \cdots, Y_N)$ concern a single policyholder.
- (In the last case, the independence assumption does not perfectly hold).

# Application (analogy)

- Approach developed in the field of natural disasters (collaboration with "Mission Risques Naturels").
- Use case:
  - a natural disaster occurs (a flood), with some characteristics;
  - the tree based model is used to fit a Generalized Pareto distribution whose parameters are adapted to the nature of the event;
  - we deduce the corresponding values $r$ and $\lambda$ of the prior distribution;
  - we have individual data on the area that is stroke (usually small amount of information) $(Y_1, \cdots, Y_N)$ that we combine with the prior to predict the loss.

# Application (analogy)

- Approach developed in the field of natural disasters (collaboration with "Mission Risques Naturels").
- Use case:
  - a natural disaster occurs (a flood), with some characteristics;
  - the tree based model is used to fit a Generalized Pareto distribution whose parameters are adapted to the nature of the event;
  - we deduce the corresponding values $r$ and $\lambda$ of the prior distribution;
  - we have individual data on the area that is stroke (usually small amount of information) $(Y_1, \cdots, Y_N)$ that we combine with the prior to predict the loss.
- Useful to evaluate the amount of the loss soon after an event, or to study scenarios.

# Adaptation to cyber (forthcoming)

- We define a cyber attack using some characteristics $\mathbf{X}$ available in the database used for calibration:
    - what is the modus operandi ? (ransomware, Ddos, double extorsion,...)
    - what are the characteristics of the target ? (sector of activity, size,...)
- We fit a GP tree to a database of events and get $\gamma(\mathbf{X})$ and $\sigma(\mathbf{X})$.

# Adaptation to cyber (forthcoming)

- We define a cyber attack using some characteristics $\mathbf{X}$ available in the database used for calibration:
    - what is the modus operandi ? (ransomware, Ddos, double extorsion,...)
    - what are the characteristics of the target ? (sector of activity, size,...)
- We fit a GP tree to a database of events and get $\gamma(\mathbf{X})$ and $\sigma(\mathbf{X})$.

# Adaptation to cyber (forthcoming)

- We define a cyber attack using some characteristics $\mathbf{X}$ available in the database used for calibration:
  - what is the modus operandi ? (ransomware, Ddos, double extorsion,...)
  - what are the characteristics of the target ? (sector of activity, size,...)
- We fit a GP tree to a database of events and get $\gamma(\mathbf{X})$ and $\sigma(\mathbf{X})$.
- We consider historical data (if available) on the target:

# Adaptation to cyber (forthcoming)

- We define a cyber attack using some characteristics $\mathbf{X}$ available in the database used for calibration:
  - what is the modus operandi ? (ransomware, Ddos, double extorsion,...)
  - what are the characteristics of the target ? (sector of activity, size,...)
- We fit a GP tree to a database of events and get $\gamma(\mathbf{X})$ and $\sigma(\mathbf{X})$.
- We consider historical data (if available) on the target:
  - if no history of claims, we keep the rough estimation using $(\gamma(\mathbf{X}), \sigma(\mathbf{X}))$ and get a distribution of the claim size.

# Adaptation to cyber (forthcoming)

- We define a cyber attack using some characteristics $\mathbf{X}$ available in the database used for calibration:
  - what is the modus operandi ? (ransomware, Ddos, double extorsion,...)
  - what are the characteristics of the target ? (sector of activity, size,...)
- We fit a GP tree to a database of events and get $\gamma(\mathbf{X})$ and $\sigma(\mathbf{X})$.
- We consider historical data (if available) on the target:
  - if no history of claims, we keep the rough estimation using $(\gamma(\mathbf{X}), \sigma(\mathbf{X}))$ and get a distribution of the claim size.
  - if previous claims are available, we use the posterior distribution from the Bayesian model.

# Adaptation to cyber (forthcoming)

- We define a cyber attack using some characteristics $\mathbf{X}$ available in the database used for calibration:
  - what is the modus operandi ? (ransomware, Ddos, double extorsion,...)
  - what are the characteristics of the target ? (sector of activity, size,...)
- We fit a GP tree to a database of events and get $\gamma(\mathbf{X})$ and $\sigma(\mathbf{X})$.
- We consider historical data (if available) on the target:
  - if no history of claims, we keep the rough estimation using $(\gamma(\mathbf{X}), \sigma(\mathbf{X}))$ and get a distribution of the claim size.
  - if previous claims are available, we use the posterior distribution from the Bayesian model.
- Note:
  - when we say "the target", this can be a single company, but it can be a generic category (more precise than the characterization used to fit the GPD)
  - prevention and evolution of the risk are taken into account through the covariates $\mathbf{X}$.

# Outline

1 Introduction

2 Generalized Pareto regression trees
- Approximation of the tail via Generalized Pareto Distributions
- Generalized Pareto Regression Trees fitting
- Illustration in the case of data breaches

3 A Bayesian model
- Credibility theory
- Calibration of the prior

4 Conclusion

# Limits and extension

- Misspecification: how does the model work if it is misspecified?
- Extension: adding expert judgment to improve the model.
- Here, we used a GP regression tree, but any extreme value regression method could be used instead (blackbox or not).

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

# Data...

- Analyzing the tail of the distribution requires a significant amount of data.
- Reliable individual data on losses:
  - very scarce!
  - public data: some data on data breaches, but with no precise indication about the cost;
  - from police, justice and related entities: not easy to track the total amount of the prejudice, but some elements for particular type of cyber claims can be obtained.
- Data coming from insurance portfolios are usually more precise, but:
  - one must pay attention that the real (total) loss is usually unknown from the insurer (only the loss corresponding to what is covered by the policy is);
  - need to wait for the stabilization of the claim.

Thank you for your attention !

To know more about our research on cyber risk, visit the web site of the Joint Research Initiative

https://sites.google.com/view/cyber-actuarial/home?authuser=0