DAV/DGVFM
Jahrestagung
2025

Maud THOMAS, Laboratoire de Sciences Actuarielle et Financière
ISFA, Université Lyon 1

Extreme Risk Modeling: A Regression Tree Approach

joint work with S. Farkas, A. Heranval and O. Lopez

Jahrestagung, 29. April 2025

Extreme risks

Extreme claims

- Risk management
- Extreme event: some value exceeds a (high) threshold

- Lack of data and/or historical information
- Present some heterogeneity

Extreme claims

⇒ Evaluating the potential cost of extreme risks is a challenging task

Rare event vs extreme event?

- **Rare event** = probability of occurrence is very small
  → Think of a bimodal distribution
  → Time between two eruptions of a geyser, analysis of road traffic, water distribution to individual houses

- **Extreme event** = located in the distribution tail
  → Largest observations of a sample or those exceeding a certain threshold
  → Heatwaves, floods, windstorms, ...

Rare event vs extreme event?

- **Rare event** = probability of occurrence is very small
  $\rightarrow$ Think of a bimodal distribution
  $\rightarrow$ Time between two eruptions of a geyser, analysis of road traffic, water distribution to individual houses

- **Extreme event** = located in the distribution tail
  $\rightarrow$ Largest observations of a sample or those exceeding a certain threshold
  $\rightarrow$ Heatwaves, floods, windstorms, ...

To study extreme events is to study the largest observations or observations above a very high threshold

Some history

- In Delft (Netherlands) in 1953, a storm killed thousands of people and destroyed nearly 50,000 houses
- Government decision: build a dyke such that there is no more than one flood every 10,000 years
- BUT the available data only covers 100 years

Some history

- In Delft (Netherlands) in 1953, a storm killed thousands of people and destroyed nearly 50,000 houses
- Government decision: build a dyke such that there is no more than one flood every 10,000 years
- BUT the available data only covers 100 years

How to determine the height of the dike?

Some history

How to determine the height of the dike?

- **How to determine the height of the dike?**
  $\rightarrow$ Take the highest wave as a reference
- **What is the probability of overtaking the highest wave?**
  $\rightarrow$ Calculate the empirical frequency of past events

Some history

- **How to determine the height of the dike?**
  $\rightarrow$ Take the highest wave as a reference
- **What is the probability of overtaking the highest wave?**
  $\rightarrow$ Calculate the empirical frequency of past events

$\Rightarrow$ Consider that the worst has already happened 🤔

Some history

Goals of Extreme Value Theory

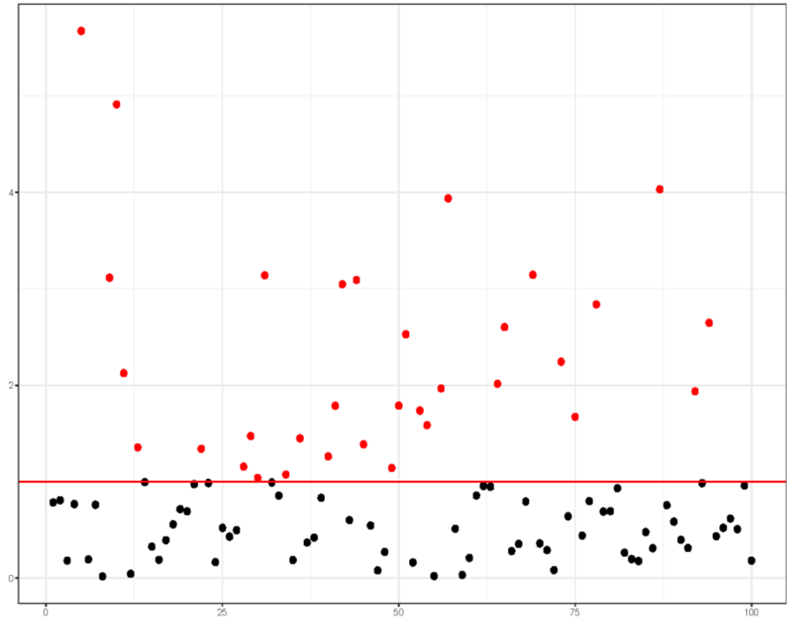
1. Estimate the probability of occurrence of an event that has not (yet) been observed
2. Estimate an extreme quantile

$\Rightarrow$ Inference outside the sample support

"Peaks-over-Threshold"method

- $Y_1, Y_2, \ldots$ a series of random variables i.i.d.
- Fix a (high) threshold u
- Extreme event = Y_i exceeds u

  $\rightarrow$ Given that $Y_i > u$, an excess is defined by $Z_i = Y_i - u$



"Peaks-over-Threshold"method

- $Y_1, Y_2, \ldots$ a series of random variables i.i.d.
- Fix a (high) threshold u
- Extreme event = Y_i exceeds u

  $\rightarrow$ Given that $Y_i > u$, an excess is defined by $Z_i = Y_i - u$
- Excess distribution $\overline{F}_u(z) = P[Y_1 - u > z \mid Y_1 > u] = \frac{\overline{F}(u+z)}{\overline{F}(u)},\ z > 0$

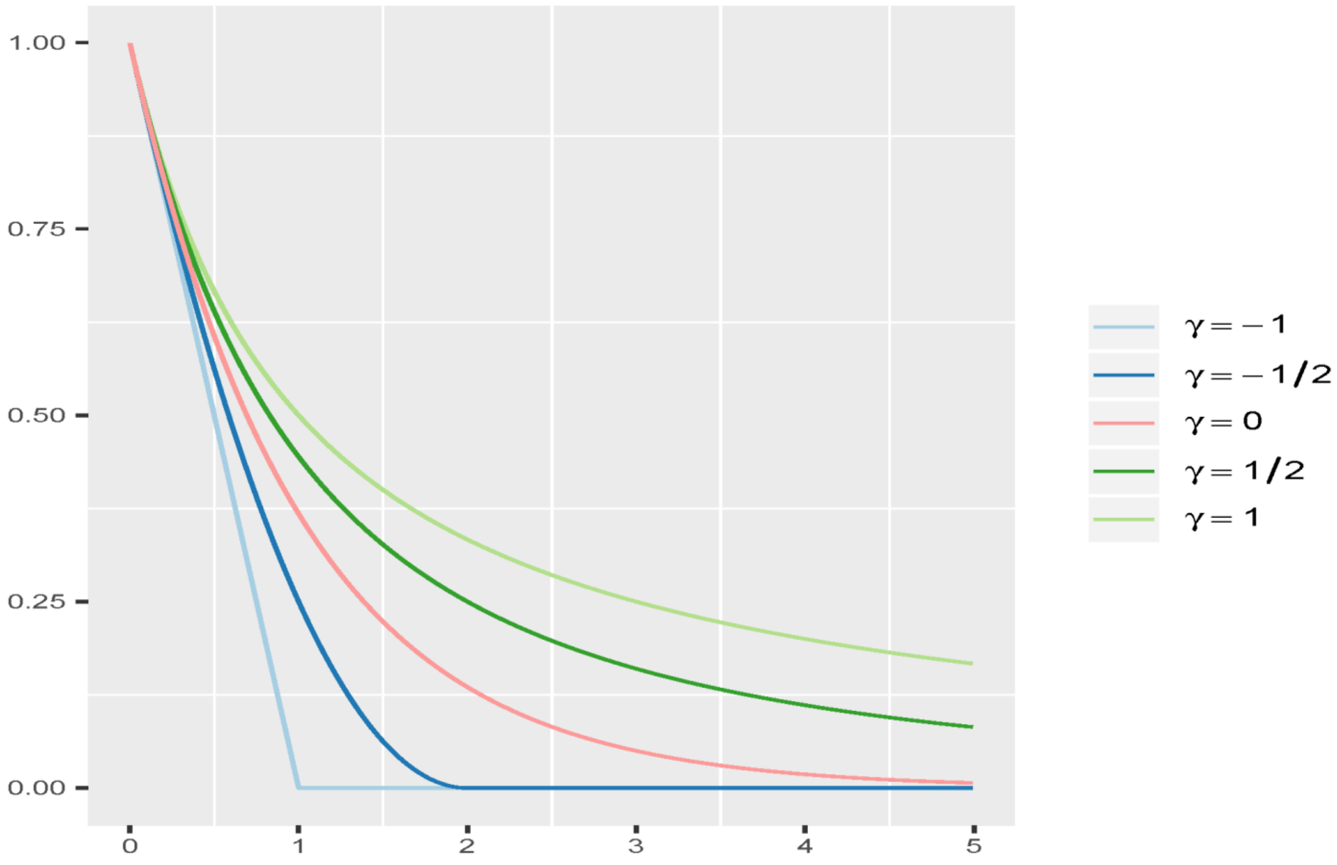
Balkema et de Haan (1974), Pickands (1975)

Under certain conditions, the distribution of excesses F_u converges, as $u \rightarrow \infty$, to a generalized Pareto distribution (GPD) whose distribution function is

$$H_{\sigma,\gamma}(z) = \begin{cases} 1 - \left(1 + \frac{\gamma}{\sigma} z\right)^{-1/\gamma} & \text{if } \gamma \neq 0 \\ 1 - \exp\left(-\frac{z}{\sigma}\right) & \text{if } \gamma = 0 \end{cases}$$

- Families of possible distributions for excesses = parametric family

Generalized Pareto distributions



3 domains of attraction

1. Fréchet domain ($\gamma > 0$): **heavy-tailed distributions**

$$1 - H_\gamma(z) \underset{+\infty}{\sim} \gamma^{-1/\gamma} z^{-1/\gamma}$$

   Examples: Cauchy, Log-gamma, Student

2. Gumbel domain ($\gamma = 0$): **thin tail distributions**

$$1 - H_0(z) \underset{+\infty}{\sim} \exp(-z)$$

   Examples: Gaussian, Gamma, Exponential

3. Weibull domain ($\gamma < 0$): **finite tail distributions**

$$1 - H_\gamma(z) = 0 \quad \text{for } x \geq -1/\gamma$$

   Examples: Uniform, Beta

Classification and Regression trees (CART)

Classification And Regression Trees (CART)

Regression tree (Breiman et al., 1984)

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

- Z is our response variable
- $\mathbf{X} \in \mathscr{X} \subset \mathbb{R}^d$ is a set of explanatory variables
- $\mathscr{F}$ is a class of target functions on $\mathbb{R}^d$
- ϕ is a loss function which depends on the quantity we wish to estimate

Loss function

- Quadratic loss $\rightarrow$ "Mean regression"(conditional mean)

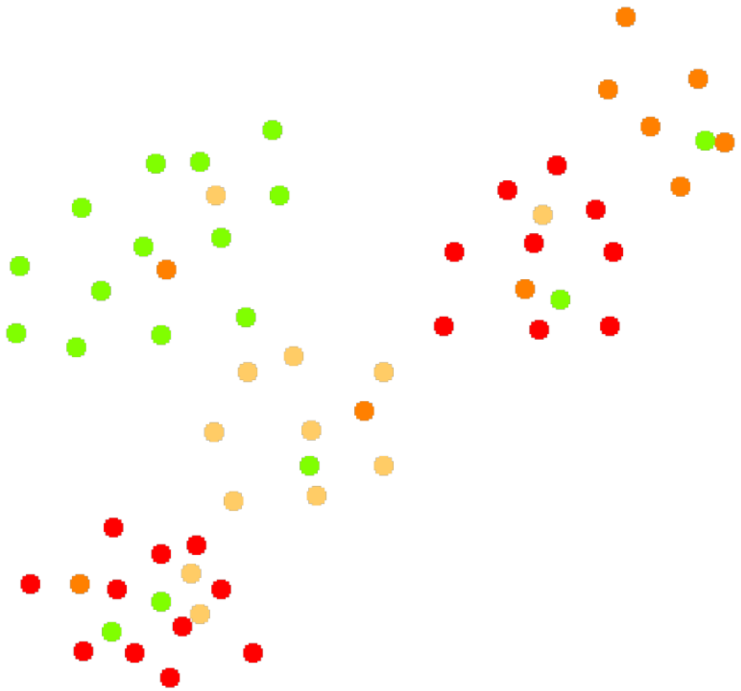
$$\varphi(z, \theta(\mathbf{x})) = (z - \theta(\mathbf{x}))^2$$

  $\hookrightarrow \theta^*(\mathbf{x}) = \mathbb{E}[Z \mid \mathbf{X} = \mathbf{x}]$
- Absolute loss $\rightarrow$ "Median regression"(conditional median)

$$\varphi(z, \theta(\mathbf{x})) = |z - \theta(\mathbf{x})|$$

  $\hookrightarrow \theta^*(\mathbf{x}) =$ conditional median
- Loss as negative log-likelihood $L_n(\boldsymbol{\theta})$
  $\hookrightarrow \theta^*(\mathbf{x}) = \arg\max_{\boldsymbol{\theta} \in \Theta} L_n(\boldsymbol{\theta})$

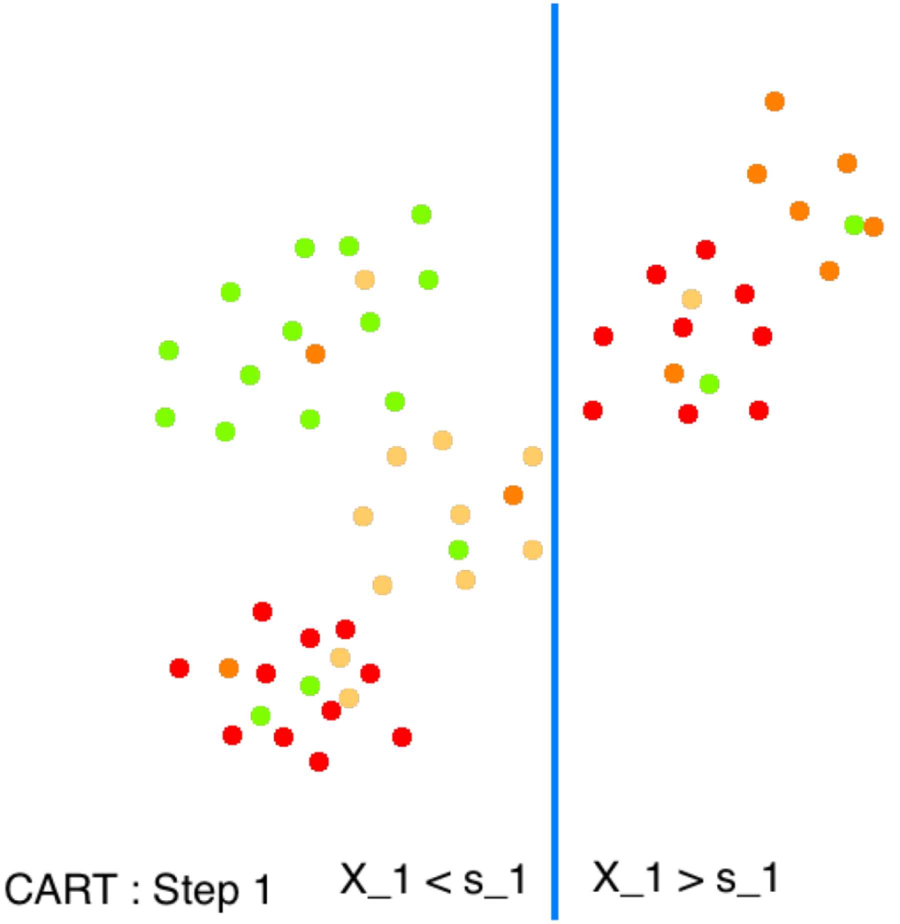
Step 1: growth of the tree

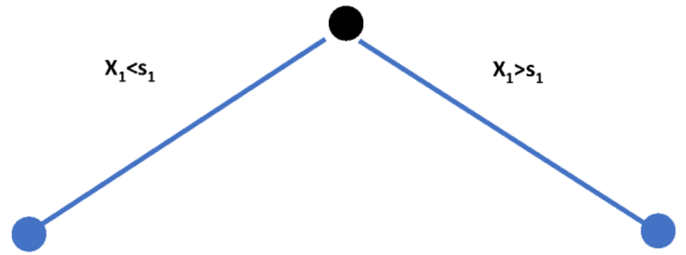


CART : Step 0

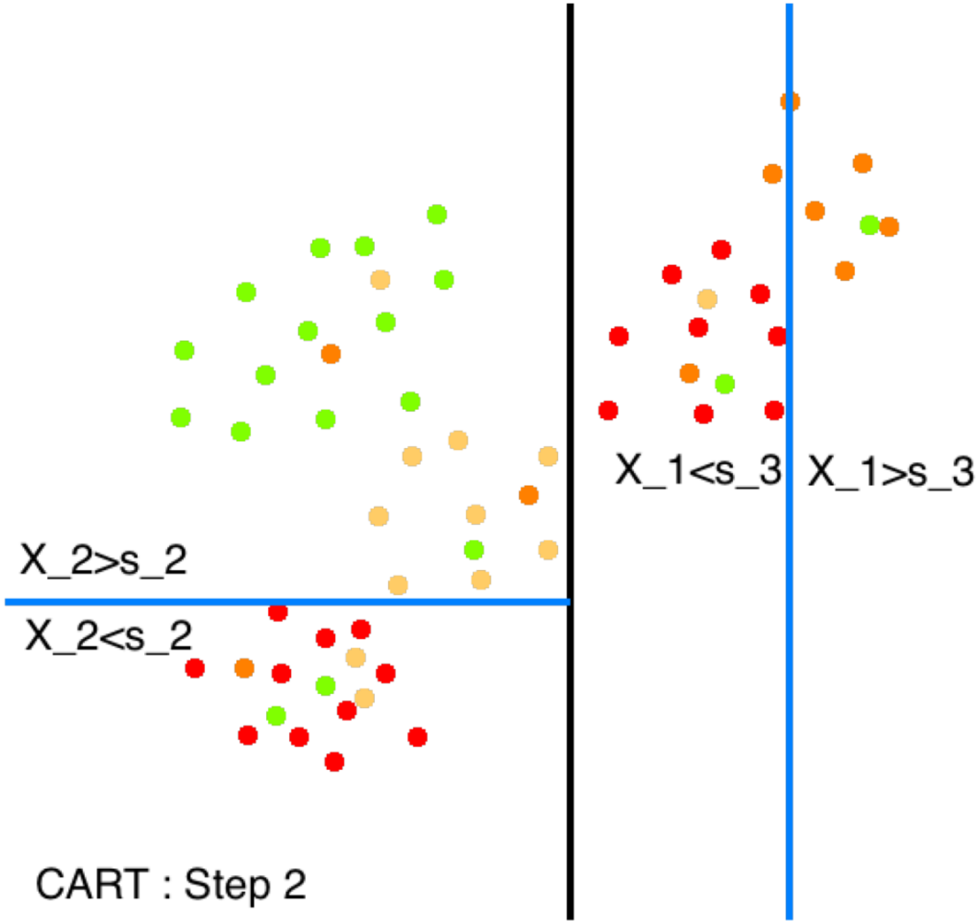
DAV Deutsche Aktuarvereinigung e.V. DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

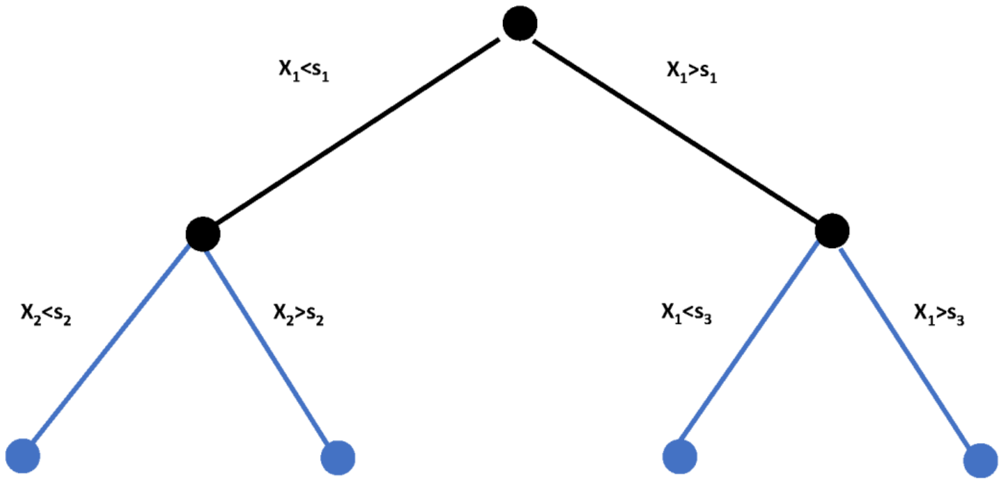
Step 1: growth of the tree



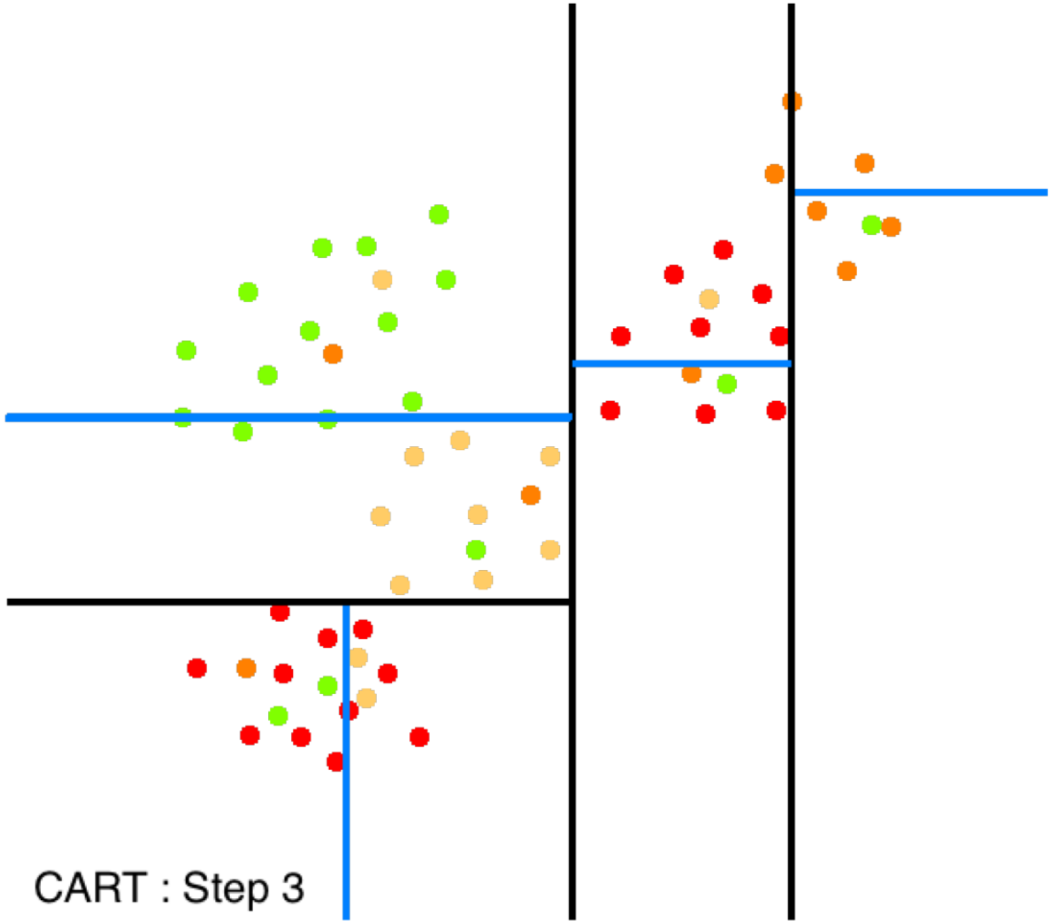


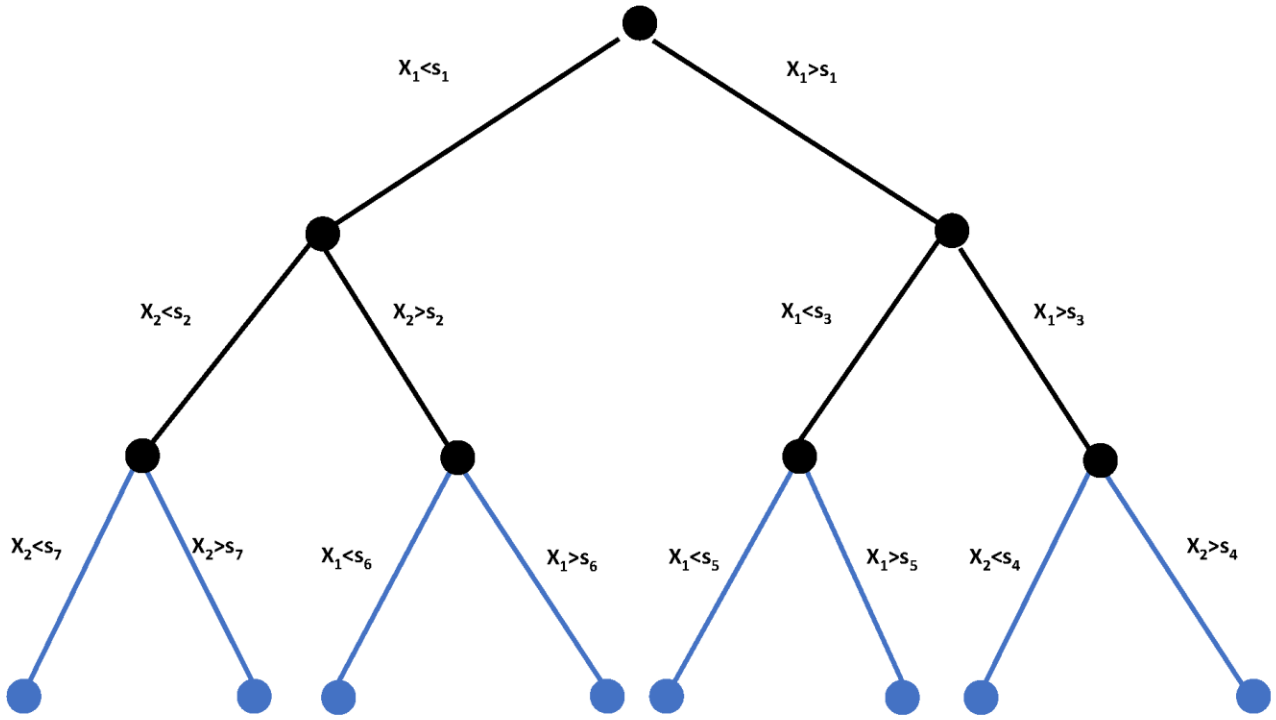
Step 1: growth of the tree





Step 1: growth of the tree





From the tree to an estimator

- Let $\widehat{T}_{\max}$ be the maximum tree obtained in the first phase and $K_{\max}$ be the number of its leaves.
- $\mathscr{T}_\ell$ ℓ-th leaf for $\ell = 1, \ldots, K_{\max}$
- **Estimator:** $\widehat{\theta}(\mathbf{X})$ **of the regression function** $\theta^*(\mathbf{X})$ given by

$$\widehat{\theta}(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\theta}_\ell \mathbf{1}_{\mathbf{x} \in \mathscr{T}_\ell}$$

$\rightarrow$ Piecewise constant function

Step:2 Pruning step (model selection)

Extract from $\widehat{T}_{\max}$ a subtree that achieves a compromise between simplicity and goodness-of-fit.

- Selected number of leaves

$$\widehat{K} = \arg \min_{K=1,\ldots,K_{\max}} \left\{ \frac{1}{n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \phi(Z_i, \widehat{\boldsymbol{\theta}}^K(\mathbf{X}_i)) \mathbf{1}_{\mathbf{X}_i \in \mathscr{T}_\ell} + \lambda K \right\},$$

$\lambda > 0$ chosen by cross-validation

- Selected tree : $\widehat{T} = \widehat{T}_{\widehat{K}}$

Loss function

- Quadratic loss $\rightarrow$ "Mean regression"(conditional mean)

$$\phi(z, \theta(\mathbf{x})) = (z - \theta(\mathbf{x}))^2$$

$\hookrightarrow \theta^*(\mathbf{x}) = \mathbb{E}[Z \mid \mathbf{X} = \mathbf{x}]$

- Absolute loss $\rightarrow$ "Median regression"(conditional median)

$$\phi(z, \theta(\mathbf{x})) = |z - \theta(\mathbf{x})|$$

$\hookrightarrow \theta^*(\mathbf{x}) =$ conditional median

- Loss as $\phi =$ negative log-likelihood , here GPD

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

$\rightarrow \boldsymbol{\theta}^*(\mathbf{x}) = (\sigma^*(\mathbf{x}), \gamma^*(\mathbf{x}))$

Extreme value theory and regression trees

- Regression framework
  - Consider an observation Y of characteristics $\mathbf{X}$.
  - Assume that the distribution of $Y \mid \mathbf{X} = \mathbf{x}$ is $\mathrm{RV}_{-1/\gamma_0(\mathbf{x})}$
  - Choice of a threshold $u(\mathbf{X})$.
  - Distribution of excesses $Z = Y - u(\mathbf{X}) \mid (\mathbf{X}, Y \geq u(\mathbf{X}))$ converges towards a GPD of parameters $\sigma_0(\mathbf{X})$ and $\gamma_0(\mathbf{X}) > 0$

$$H_{\sigma_0(\mathbf{x}),\gamma_0(\mathbf{x})}(z) = 1 - \left(1 + \frac{\gamma_0(\mathbf{X})}{\sigma_0(\mathbf{X})} z\right)^{-1/\gamma_0(\mathbf{X})}$$

- Purpose: to estimate $\theta_0(\mathbf{x}) = (\sigma_0(\mathbf{x}), \gamma_0(\mathbf{X}))$

- Reminder (PoT method): select the observations $Y_i \geq u(\mathbf{X}_i)$.
- Here we assume $u(\mathbf{x}) = u \in [u_{\min}, u_{\max}]$
- k_n: average number of observations Y_i above u
- Application of the CART algorithm with negative log-likelihood GPD to Z_i.

Real-life applications: cyber risk and flood risk

Cyber-risk

- Cyber risk: inappropriate use of digital tools and information systems.
- A cyber incident may be intentional (cyber attack) or unintentional (accidental).
- In the case of hacking, hackers use vulnerabilities in information systems, but also human vulnerabilities in companies.
- Different types of attacks (ransomware, phishing, classic frauds using the digital vector...)
- States, industries, individuals can be affected

Privacy Rights Clearinghouse

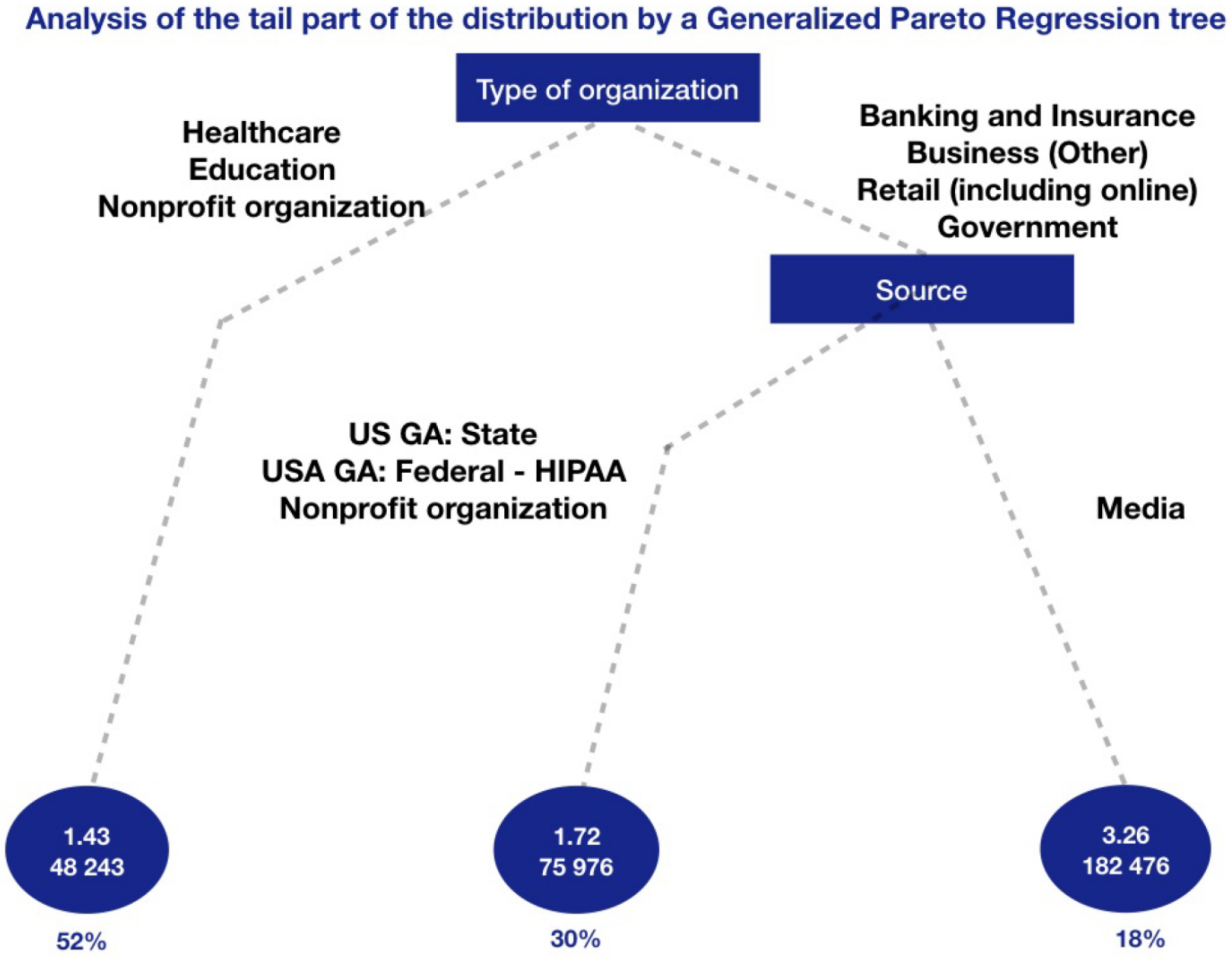
- Founded in 1992
- Public
- Reference for academic work on the analysis of cyber events related to data breach
- Aim: to raise awareness of privacy issues.
- Chronology of data breaches maintained since 2005.
- Contains information on events from multiple sources:
  - US government agencies (Federal level–HIPAA).
  - US government agencies (State level)
  - Media
  - Other organizations
- 8860 events

PRC database: variables

- Exposition variables : information about the victim (organization).
- Event variables : information about the data breach

Exposition variables	Name of the organization Type of organization Localization of the organization
Event variables	Source Date Type of breach Number of affected records Description of the event

Tail of the distribution: GP tree



Natural catastrophes

CatNat Regime

- Natural disasters insured by a private-public partnership
- Covers a large number of risks: floods, landslides, drought, earthquakes,...

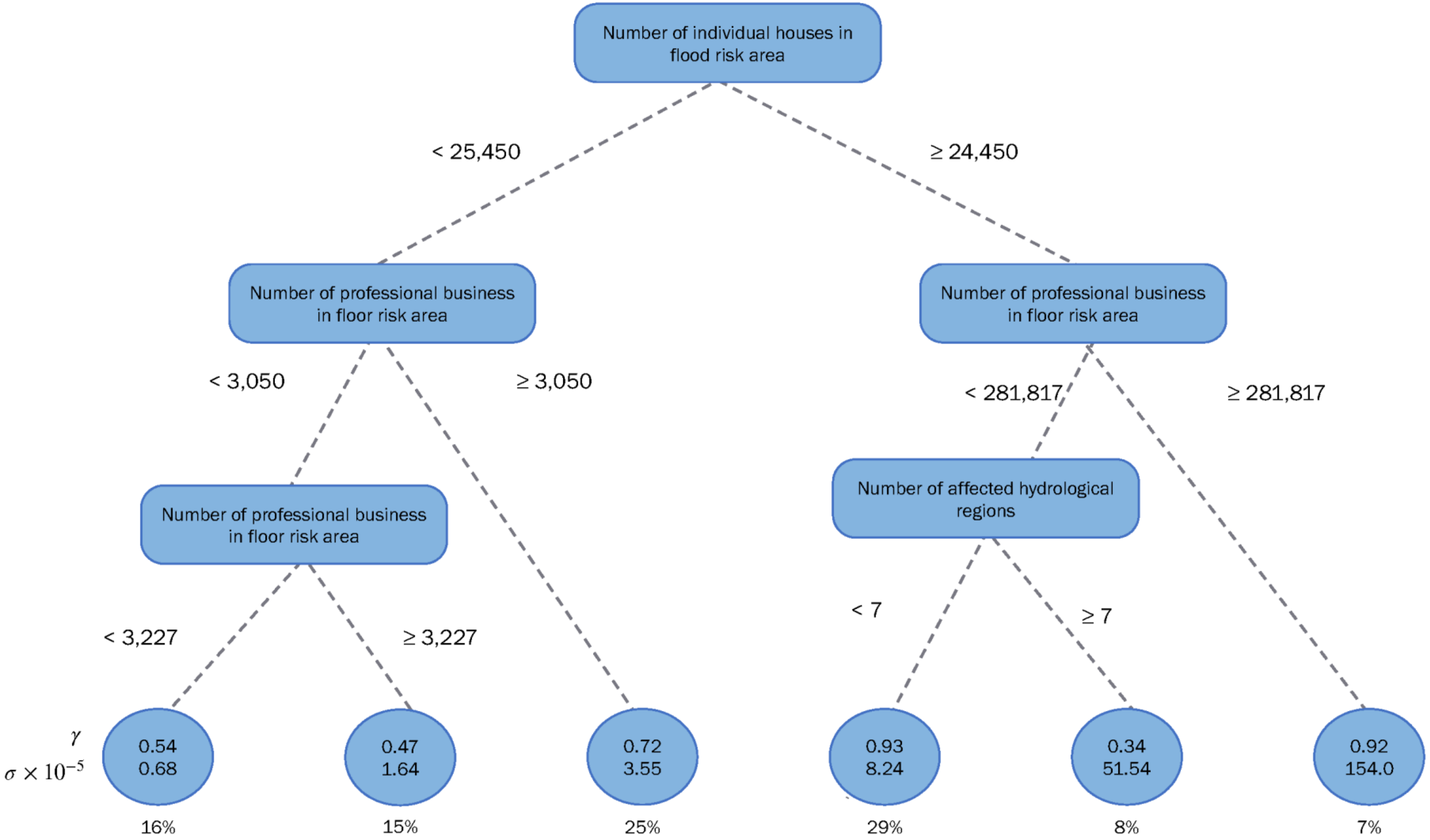
- A few figures
  - between 1988-2013:
    $\rightarrow$ 48.3M€ were paid by insurance companies for natural disasters
    $\rightarrow$ 431 claims per year
    $\rightarrow$ 1.9M per year
  - between 2014 - 2019:
    $\rightarrow$ 92M€ should be paid out by insurance companies for natural disasters
    $\rightarrow$ 114% on the cost of floods, +39% on the cost of storms
- **Specificity**: to receive compensation, the insured city must apply to a committee to be recognized as a state of Natural Disaster

$\rightarrow$ Is it possible to predict its cost shortly after its occurrence and for the entire French market?

Application to flooding events

- SILECC database
  - Partnership with MRN
  - Consists of claims from the largest insurance companies (70% of the French market)
  - 700 000 claims from 1990 to 2019 including 3 147 flood events
- Covariates (available shortly after the event)
  - the weather region
  - the season
  - the type of flooding
  - the number of affected hydrological regions
  - the number of individual houses
  - the number of business premises in the flood zone
- The threshold u was chosen equal to 100 000€, which corresponds to 1 083 events

Application to flooding events



Conclusion

- Classification of extreme behaviour via tree methods:
  - allows to consider nonlinearities in this dependence
  - adapted to **X** variables which are discrete as well as continuous (interesting in particular for the study of behaviours because many variables are qualitative)
  - allows a classification
  - but can sometimes be unstable
- Natural extensions, less intelligible but more precise (gradient boosting, random forests) can be used
- In general, a decision support tool to draw the line between what is ïnsurableänd what is not.

Vielen Dank für Ihre Aufmerksamkeit.

Maud THOMAS, Laboratoire de Sciences Actuarielle et Financière
ISFA, Université Lyon 1