Extreme claims analysis using Generalized Pareto Regression Trees

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

Maud Thomas

Sorbonne Université
Laboratoire de Probabilités, Statistique et Modélisation, UMR CNRS 8001

January 28, 2021

Actuarial modelling

- X characteristics of a policyholder
- N number of claims ($\mathbb{E}[N \mid X]$ =frequency)
- Y cost of a claim ($\mathbb{E}[Y \mid X]$ =severity)

Pricing principle = balance (in average) the cost of a policyholder and the commitments of the insurer

$$\pi(X) = E[N \mid X]E[Y \mid X]$$

- $\pi(X)$ = premium of the insurance contract of a policyholder with characteristics X
- Common assumption: Y and N are independent given X

Reserving = Need to estimate the whole conditional distribution of N and Y given X

Extreme claims

- Risk management
- Extreme event: some value exceeds a (high) threshold
- Lack of data and/or historical information
- Present some heterogeneity

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

Objectives of the presentation

Main goals

1. Study extreme claims
2. Gain further insight on their heterogeneity
3. Analyse the impact of characteristics on extreme claims

Focus on

- Tail of the distribution
- Severity of extreme claims

⇒ Two statistical tools :

1. Extreme value theory
2. Regression and classification trees

Statistical tools

Extreme Value theory

Extreme Value Theory

Goals of Extreme Value Theory

Goals of Extreme Value Theory

1. Estimate extreme quantiles
2. Estimate the occurrence probability of an event more extreme than previously observed

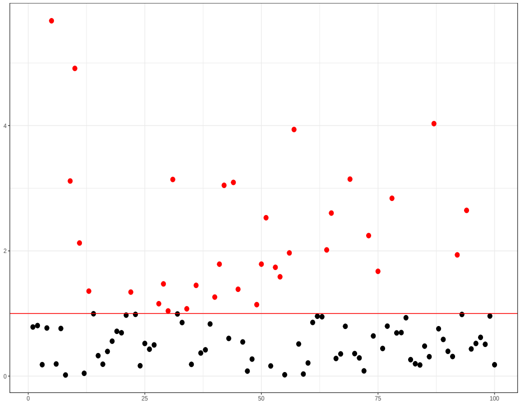
$\Rightarrow$ Inference outside of the range of the data

Extreme value theory

Peaks over threshold method

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

  → Given that $Y_i > u$, define the excess $X_i = Y_i - u$



Extreme value theory

Peaks over threshold method

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

  $\rightarrow$ Given that $Y_i > u$, define the excess $X_i = Y_i - u$

Balkema and de Haan (1974)

If there exist $(a_u) > 0$, (b_u) and a non-degenerated distribution function H such that,

$$\mathbb{P}[Y_i - u \geqslant a_u x + b_u \mid Y_i > u] \xrightarrow[u \to \infty]{d} 1 - H(x),$$

then H is necessarily of the form

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

- Possible limits of excesses = Parametric family of distributions

  $\hookrightarrow$ Generalized Pareto Distributions

Extreme value theory and regression models

- **Goal** : estimate $\gamma(X)$ where $\gamma(X)$ is the tail index of the distribution of $Y|X$.

- Existing methods :
  - Semi-parametric approaches
    - Exponenial regression model (Beirlant et al., 2003)
    - Smoothing splines (Chavez-Demoulin et al., 2015)

  - Non parametric approach (Beirlant and Goegebeur, 2004)
    - Local polynomial maximum likelihood
    - Only for continuous covariates

Statistical tools

CART algorithm

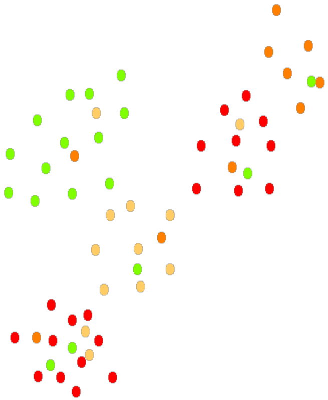
Classification And Regression Trees (CART)

Regression tree (Breiman et al., 1984)

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

- Y is a response variable (the cost of a cyber claim in our case)
- $\mathbf{X} \in \mathscr{X} \subset \mathbb{R}^d$ is a set of covariates
- $\mathscr{M}$ is a class of target functions on $\mathbb{R}^d$
- ϕ is a loss function that depends on the quantity we wish to estimate

Growing phase



CART : Step 0

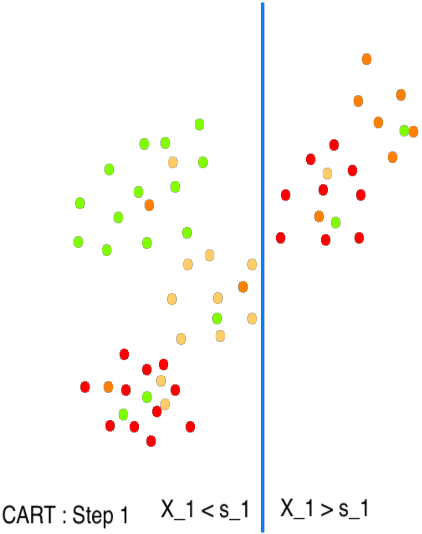
Growing phase

Splitting rules

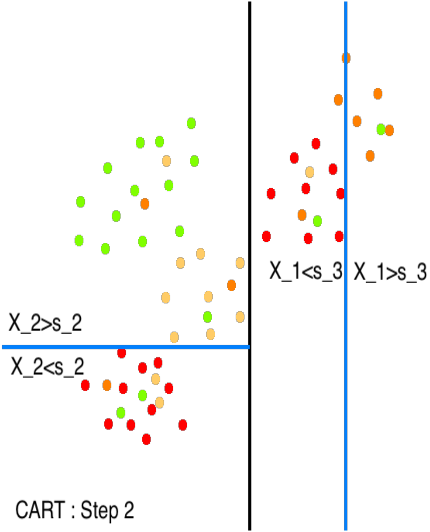
$$\mathbf{x} = (x^{(1)}, \ldots, x^{(d)}) \longrightarrow R_j(\mathbf{x})$$

with

$$\begin{cases} R_j(\mathbf{x}) & = 0 \text{ ou } 1 \\ R_j(\mathbf{x}) R_{j'}(\mathbf{x}) & = 0 \text{ for } j \neq j' \\ \sum_j R_j(\mathbf{x}) & = 1 \end{cases}$$



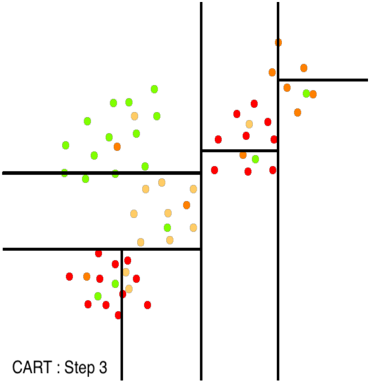
Growing phase



Growing phase

Regression estimator $\hat{m}^{\mathscr{R}}(\mathbf{x})$ of m^* given by

$$\hat{m}^{\mathscr{R}}(\mathbf{x}) = \sum_{j=1}^{s} \hat{m}(R_j) R_j(\mathbf{x}) \quad \text{where} \quad \hat{m}(R_j) = \arg\min_{m \in \mathscr{M}} \sum_{i=1}^{n} \phi(Y_i, \mathbf{X}_i) R_j(\mathbf{X}_i)$$



The splitting rule and loss functions

- Quadratic loss → Mean regression

$$\phi(y, m(\mathbf{x})) = (y - m(\mathbf{x}))^2$$

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

- Absolute loss → Median regression

$$\phi(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$$

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

- Log-likelihood loss, here GPD

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

$\hookrightarrow m^*(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$

The splitting rule and loss functions

- Quadratic loss → Mean regression

$$\phi(y, m(\mathbf{x})) = (y - m(\mathbf{x}))^2$$

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

- Absolute loss → Median regression

$$\phi(y, m(\mathbf{x})) = |y - m(\mathbf{x})|$$

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

- Log-likelihood loss, here GPD

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

$\hookrightarrow m^*(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$

Pruning step: model selection

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

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

- $\alpha > 0$ is chosen by cross-validation
- 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 $E[C_\alpha(T)]$.
- $\widehat{T}$ the tree minimizing the penalized criterion, $\widehat{K}$ its number of leaves.
- k_n = nombre d'observations au-dessus du seuil u

Consistency of the algorithm

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

Consistency of the tree

Under some assumptions,

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

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,

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

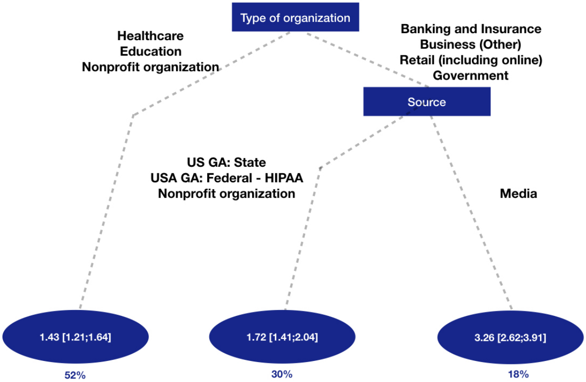
Application to real data: cyber-claims

(Farkas et al, 2020)

- Privacy Rights Clearinghouse (nonprofit association)
- Founded in 1992
- Publicly available
- Benchmark for Cyber event analysis
- Aim at raising awareness about privacy issues.
- Chronology of data breaches maintained from 2005.
- Gathering events information from multiple sources:
  - US Government Agencies (Federal level–HIPAA): Health domain, obligation to declare any breach that affects more than 500 individuals
  - US Government Agencies (State level): since 2018, each state has a specific legislation related to data breaches
  - Media
  - Non profit organizations.
- Focus on the Tail of the distribution
  - Consider only the number of affected records above 27 000
  - Fit a GPD CART

Application to real data: cyber-claims

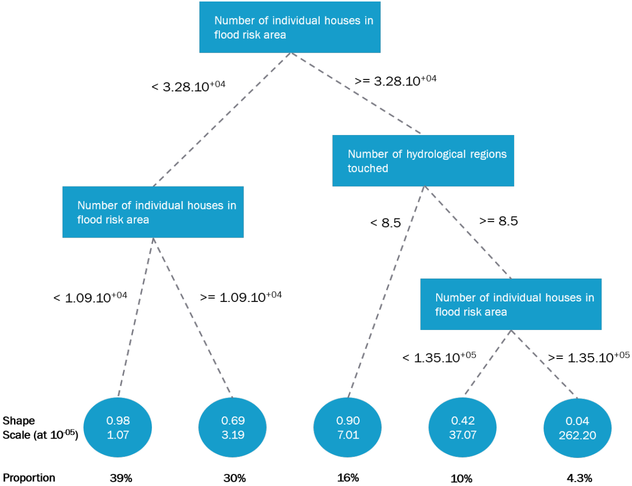
Farkas et al, 2020



Application to real data: cost prediction of floods in France

- Goal = improve the cost prediction of an event of floods, shortly after its occurrence in France
- In collaboration with MRN (Mission risques naturels) and partnership with Fédération Française des Assurances
- Access to a large volume of events: all events of floods that have been identified as "CAT NAT" over the past 20 years in France
  → Events built by the MRN from claims reported by insurance companies
- Database fed by 13 contributors including major French insurance companies
- Including 70% of the total amount paid for non life insurance
- 31 000 events
- Focus on the Tail of the distribution
  - Consider only the events with a cost larger than $u = 1e5$
  - Fit a GPD CART

Application to real data: floods



Conclusion

- Propose a methodology to study extreme claims by taking into account
  - heterogeneity,
  - impact of the covariates
  - evolution through time
- Give theoritical guarantees

- Advantage: interpretation.
- Drawbacks: the robustness of the tree structure and the estimator.

- Future works: consider random forest

- Corresponding article:
  S. Farkas, O. Lopez and M. Thomas. Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance pricing and reserving, Preprint
  `https://hal.archives-ouvertes.fr/hal-02118080v2/document`