La théorie des valeurs extrêmes et son application

Maud Thomas

ISUP, Sorbonne Université

Institut des ACTUAIRES

Plan de la présentation

1. Introduction
   - Qu'est-ce qu'un événement extrême ?
   - Pourquoi a-t-on besoin d'une théorie spécifique ?
   - Lien avec la gestion de risque
2. Lois limites des maxima
   - Lois de valeurs extrêmes généralisées
   - Domaines d'attraction
   - Quantiles extrêmes
   - Méthodes des maxima par blocs
3. Lois limites des excès
   - Lois de Pareto généralisées
   - Méthode « Peaks over Threshold »
   - Choix du seuil
   - Quantiles extrêmes
4. Généralisation aux suites non i.i.d.
5. Classification des sinistres extrêmes

Institut des ACTUAIRES

Événement extrême ?

Événement rare vs événement extrême ?

- **Événement rare** = probabilité d'occurrence est très petite
  → Penser à une loi bimodale
  → Temps entre deux éruptions d'un geyser, analyse du trafic routier, distribution en eau des habitations

- **Événement extrême** = situés dans la queue de distribution
  → Plus grandes observations d'un échantillon ou celles dépassant un certain seuil
  → Canicules, inondations, tempêtes de vent, ...

Institut des ACTUAIRES

Événement rare vs événement extrême ?

- **Événement rare** = probabilité d'occurrence est très petite
  → Penser à une loi bimodale
  → Temps entre deux éruptions d'un geyser, analyse du trafic routier, distribution en eau des habitations

- **Événement extrême** = situés dans la queue de distribution
  → Plus grandes observations d'un échantillon ou celles dépassant un certain seuil
  → Canicules, inondations, tempêtes de vent, ...

Étudier les événements extrêmes, c'est étudier le maximum des observations ou les données au-dessus d'un seuil très élevé

Institut des ACTUAIRES

Pourquoi la théorie des valeurs extrêmes ?

Un peu d'histoire

- A Delft (Pays-Bas) en 1953, une tempête a tué des milliers de personnes et détruit près de 50 000 habitations
- Décision du gouvernement : construire une digue telle qu'il n'y ait pas plus d'une inondation tous les 10 000 ans
- MAIS les données disponibles ne couvrent que 100 ans

Un peu d'histoire

- A Delft (Pays-Bas) en 1953, une tempête a tué des milliers de personnes et détruit près de 50 000 habitations
- Décision du gouvernement : construire une digue telle qu'il n'y ait pas plus d'une inondation tous les 10 000 ans
- MAIS les données disponibles ne couvrent que 100 ans

Un peu d'histoire

- **Comment déterminer la hauteur de la digue ?**
  → Prendre la vague la plus haute comme référence
- **Quelle est la probabilité de dépasser la vague la plus haute ?**
  → Calculer la fréquence des événements passés

Un peu d'histoire

- **Comment déterminer la hauteur de la digue ?**
  → Prendre la vague la plus haute comme référence
- **Quelle est la probabilité de dépasser la vague la plus haute ?**
  → Calculer la fréquence des événements passés

⇒ Considérer que le pire s'est déjà produit 🤔

Institut des ACTUAIRES

Un peu d'histoire

But de la théorie des valeurs extrêmes

1. Estimer la probabilité d'occurrence d'un événement qui n'a pas (encore) été observé
2. Estimer un quantile extrême

$\Rightarrow$ Inférence (estimation) en dehors du support de l'échantillon

Cadre

- $X_1, \ldots, X_n$ n variables aléatoires indépendantes et identiquement distribuées de fonction de répartition F inconnue
- $X_{(1)} \leqslant \ldots \leqslant X_{(n)}$ statistiques d'ordre associées

Deux questions

1. Pour $x \gg X_{(n)}$, pour $X^\star$ une nouvelle observation $\mathbb{P}\{X^\star > x\}$?
2. Estimer un quantile d'ordre $1 - 1/(xn)$ avec $x \gg 1$?

Approche classique

1. Fonction de répartition empirique

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leqslant x\}}$$

   → $\mathbb{P}\{X^\star > x\}$ estimée par 0

2. Quantile empirique $\widehat{F}_n^{-1}(p) = X_{(i)}$ pour $p \in]i/n, ((i+1)/n[$
   → Estimer $F^{-1}(1-\alpha)$ par $X_{(n)}$, le pire s'est déjà produit

Cadre

- $X_1, \ldots, X_n$ n variables aléatoires indépendantes et identiquement distribuées de fonction de répartition F inconnue
- $X_{(1)} \leqslant \ldots \leqslant X_{(n)}$ statistiques d'ordre associées

Deux questions

1. Pour $x \gg X_{(n)}$, pour $X^\star$ une nouvelle observation $\mathbb{P}\{X^\star > x\}$?
2. Estimer un quantile d'ordre $1 - 1/(xn)$ avec $x \gg 1$?

Deux points de vue en théorie des valeurs extrêmes

- Étude des lois limites du maximum d'un échantillon
- Étude des excès par rapport à un seuil u

→ Les deux sont équivalents : deux approches différentes pour étudier la queue de distribution
→ On peut aussi s'intéresser à la queue de distribution gauche

Gestion de risque et quantile extrême

Gestion du risque

Value-at-risk

Soit R la somme totale des provisions. Alors la Value-at-risk d'ordre β est définie comme

$$\mathrm{VaR}_\beta = \inf\{x : \mathbb{P}\{R \leqslant x\} \geqslant \beta\}$$

- Value-at-risk d'ordre 99.5% $\longleftrightarrow$ $\beta = 0.995$
- Calculer la VaR $\Leftrightarrow$ Calculer un quantile associé à la loi de R
- Valeur qui sera dépassée l'an prochain avec une probabilité de 0.005

Idée : proposer une loi pour R

Problème :

Ne permet pas de prendre en compte les événements extrêmes alors que la VaR est en fait un quantile extrême

Gestion du risque

Value-at-risk

Soit R la somme totale des provisions. Alors la Value-at-risk d'ordre β est définie comme

$$\mathrm{VaR}_\beta = \inf\{x : \mathbb{P}\{R \leqslant x\} \geqslant \beta\}$$

- Value-at-risk d'ordre 99.5% $\longleftrightarrow$ $\beta = 0.995$
- Calculer la VaR $\Leftrightarrow$ Calculer un quantile associé à la loi de R
- Valeur qui sera dépassée l'an prochain avec une probabilité de 0.005

Idée : proposer une loi pour R

Problème :

Ne permet pas de prendre en compte les événements extrêmes alors que la VaR est en fait un quantile extrême

$\Longrightarrow$ Théorie des valeurs extrêmes

Institut des ACTUAIRES

Lois limites des maxima

Étude de la loi du maximum

- $X_1, \ldots, X_n$ i.i.d. de fonction de répartition F
- Étudier le comportement de

$$X_{(n)} = \max(X_1, \ldots, X_n)$$

- Loi de $X_{(n)}$

$$\mathbb{P}\{X_{(n)} \leqslant z\} = F^n(z")$$

- Ce n'est pas très utile en pratique car F est inconnue
- Nous avons déjà vu que estimer F n'est pas une bonne solution
  $\Rightarrow$ **Idée :** Accepter que F est inconnue et chercher des modélisations pour F^n directement

- Loi limite de $X_{(n)}$

$$F^n(z) \xrightarrow[n\to\infty]{} \begin{cases} 0 & \text{si } z < z_F \\ 1 & \text{si } z \geqslant z_F \end{cases}$$

où $z_F = \sup\{z \in \mathbb{R}, F(z) \leqslant 1\}$ = point terminal de F

Étude de la loi du maximum

- $X_1, \ldots, X_n$ i.i.d. de fonction de répartition F
- Étudier le comportement de

$$X_{(n)} = \max(X_1, \ldots, X_n)$$

- Loi de $X_{(n)}$

$$\mathbb{P}\{X_{(n)} \leqslant z\} = F^n(z")$$

- Ce n'est pas très utile en pratique car F est inconnue
- Nous avons déjà vu que estimer F n'est pas une bonne solution

  $\Rightarrow$ **Idée :** Accepter que F est inconnue et chercher des modélisations pour F^n directement

- Loi limite de $X_{(n)}$

$$F^n(z) \xrightarrow[n\to\infty]{} \begin{cases} 0 & \text{si } z < z_F \\ 1 & \text{si } z \geqslant z_F \end{cases}$$

où $z_F = \sup\{z \in \mathbb{R}, F(z) \leqslant 1\}$ = point terminal de F

$\Rightarrow X_{(n)}$ tend en loi vers δ_{z_F} (loi dégénérée)

Analogie avec les sommes de v.a. i.i.d.

- Suppose $\mathbb{E}X_1 = m \in \mathbb{R}$ et $\operatorname{Var} X_1 = \sigma^2 \in]0, +\infty[$
- **Loi des grands nombres**

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p.s.} m$$

$\Rightarrow \overline{X}_n$ tend en loi vers δ_m (loi dégénérée)

$\rightarrow$ **Question :** Est-il possible de trouver deux suites $a_n > 0$ et b_n telles que

$$\frac{\overline{X}_n - b_n}{a_n}$$

converge en loi vers une loi non-dégénérée ? Si oui, quelle est cette loi limite ?

$\rightarrow$ **Théorème Central Limite** $a_n = \sqrt{n\sigma^2}$ et $b_n = m$

$$\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Retour sur le maximum

Théorème [Fisher and Tippett, 1928, Gnedenko, 1943]

S'il existe deux suites $a_n > 0$ et b_n et une loi non-dégénérée G telles que

$$\mathbb{P}\left\{\frac{X_{(n)} - b_n}{a_n} \leqslant z\right\} = F^n(a_n z + b_n) \to G(z)$$

alors G est *nécessairement* du type

$$G_{\mu,\sigma,\gamma}(z) = \exp\left(-\left(1+\gamma\frac{z-\mu}{\sigma}\right)_+^{-1/\gamma}\right), z \in \mathbb{R}$$

- $z_+ = \max(z, 0)$: partie positive
- Si $\gamma = 0$, le membre de droite se lit $\exp(-\exp(-(z-\mu)/\sigma))$
- Si F vérifie les hypothèses ci-dessus, on dit que F appartient au max-domaine d'attraction de G_γ, $F \in \mathrm{DA}(\gamma)$

Institut des ACTUAIRES

Lois de valeurs extrêmes généralisées

- Famille des lois limites possibles $(G_{\mu,\sigma,\gamma})_{\mu,\sigma,\gamma}$ = Lois de valeurs extrêmes généralisées (GEV)
  - $\mu \in \mathbb{R}$ = paramètre de localisation
  - $\sigma > 0$ = paramètre d'échelle
  - $\gamma \in \mathbb{R}$ = **paramètre de forme**
    $\rightarrow$ reflète l'épaisseur de la queue de la distribution (cf prochain slide)

- Le théorème suggère l'utilisation de la famille des GEV pour modéliser les maximas de "longues suites"

Lois de valeurs extrêmes généralisées

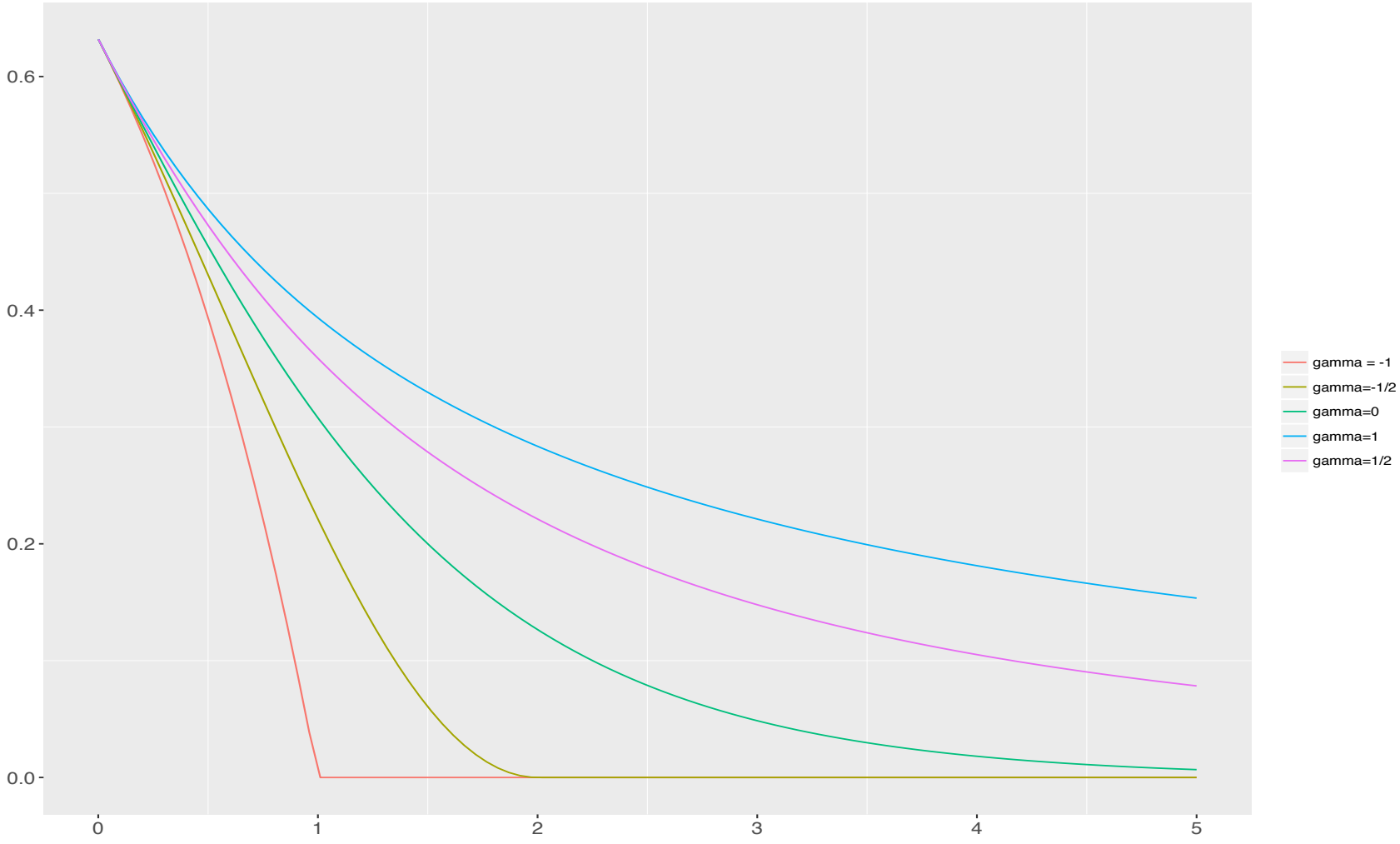


Figure – Fonctions de survie des lois de valeurs extrêmes généralisées.

Étude des domaines d'attraction

- Conditions d'appartenance à un domaine d'attraction ?
  → Condition nécessaire et suffisante = propriétés de variation régulière étendue sur F
- Si G_γ est max-stable, i.e. il existe $(a_n) > 0$ et b_n telles que

$$G_\gamma^n(a_n x + b_n) = G_\gamma(x)$$

  alors G_γ appartient à son propre domaine d'attraction

- 3 lois max-stable $\Leftrightarrow$ 3 domaines d'attraction
  1. Domaine de Fréchet ($\gamma > 0$) : domaine des lois à **queue épaisse**

$$1 - G_\gamma(x) \underset{+\infty}{\sim} \gamma^{-1/\gamma} x^{-1/\gamma}$$

     Les moments d'ordre supérieur à $1/\gamma$ n'existent pas.
     Loi de Fréchet

$$F(x) = \exp(-x^{-1/\gamma}) \mathbb{1}_{]0,\infty[}(x)$$

     Exemples : Lois de Cauchy, Pareto, log-gamma et Student

Étude des domaines d'attraction

- Conditions d'appartenance à un domaine d'attraction ?
  → Condition nécessaire et suffisante = propriétés de variation régulière étendue sur F
- Si G_γ est max-stable, i.e. il existe $(a_n) > 0$ et b_n telles que

$$G_\gamma^n(a_n x + b_n) = G_\gamma(x)$$

alors G_γ appartient à son propre domaine d'attraction

- 3 lois max-stable $\Leftrightarrow$ 3 domaines d'attraction
  2. Domaine de Gumbel($\gamma = 0$) : domaine des lois à **queue fine**

$$1 - G_0(x) \underset{+\infty}{\sim} \exp(-x)$$

Tous les moments existent
Loi de Gumbel

$$F(x) = \exp(-\exp(-x))$$

Exemples : Lois de gamma, exponentielle, normale, log-normale et...
Weibull !

Institut des ACTUAIRES

Étude des domaines d'attraction

- Conditions d'appartenance à un domaine d'attraction ?
  $\rightarrow$ Condition nécessaire et suffisante = propriétés de variation régulière étendue sur F
- Si G_γ est max-stable, i.e. il existe $(a_n) > 0$ et b_n telles que

$$G_\gamma^n(a_n x + b_n) = G_\gamma(x)$$

alors G_γ appartient à son propre domaine d'attraction

- 3 lois max-stable $\Leftrightarrow$ 3 domaines d'attraction
  3. Domaine de Weibull ($\gamma < 0$) : domaine des lois à **queue finie à droite**

$$1 - G_\gamma(\frac{1}{\gamma} - x) \underset{0}{\sim} (-\gamma)^{-1/\gamma} x^{-1/\gamma}$$

Tous les moments existent.
Loi de inverse-Weibull

$$F(x) = \exp(-(-x)^{-1/\gamma})\mathbb{1}_{]-\infty,0[}(x) + \mathbb{1}_{]0,\infty[}(x)$$

Exemples : Lois uniforme, Beta

Méthode des maxima par blocs

But

Construire une suite de maxima i.i.d. pour ajuster une GEV

- $X_1, X_2, \ldots, X_{nm}$ i.i.d.
- Ranger en m blocs de même taille n

$$\underbrace{X_1, \ldots, X_n}_{\text{Bloc } 1} \Big| \underbrace{X_{n+1}, \ldots, X_{2n}}_{\text{Bloc } 2} \Big| \cdots \Big| \underbrace{X_{(j-1)n+1}, \ldots, X_{jn}}_{\text{Bloc } j} \Big| \cdots \Big| \underbrace{X_{(m-1)n+1}, \ldots, X_{mn}}_{\text{Bloc } m}$$

- Pour chaque bloc j, $M_{n,j} = \max$ de $X_{(j-1)n+1}, \ldots, X_{jn}$
- $M_{n,1}, \ldots, M_{n,m}$ suite i.i.d. de maxima
- Ajustement d'une GEV sur $M_{n,1}, \ldots, M_{n,m}$

- Souvent, les blocs correspondent à une période de 1 an, n au nombre d'observations sur un an, les maxima par blocs aux maxima annuels

Institut des ACTUAIRES

Choix de la taille des blocs

- La méthode des maxima par blocs nécessite de choisir la taille des blocs
- Ce choix repose sur un compromis biais-variance
  → une taille trop petite ⇒ l'approximation asymptotique risque de ne pas être raisonnable (soit un grand biais)
  → une taille trop grande ⇒ trop peu de blocs, donc de maxima (soit une grande variance)
- Souvent, des considérations pragmatiques conduisent à prendre des blocs d'une durée d'un an
  → Par ex, seules les maxima annuels peuvent avoir été enregistrés
  → l'utilisation de blocs plus petits est impossible.

Quantiles extrêmes

Quantiles extrêmes

- $\overline{F} = 1 - F$: fonction de survie
- **Quantile d'ordre** $1 - \alpha$

$$q(\alpha) = F^{\leftarrow}(1-\alpha) = \inf\{y\colon F(y) \geqslant 1-\alpha\} = \inf\{y\colon \overline{F}(y) \leqslant \alpha\}$$

$\rightarrow q(1/2) =$ médiane
$\rightarrow q(1/4) = 3^{\text{e}}$ quartile
$\rightarrow q(\alpha) = \text{VaR}_{1-\alpha}$

Quantiles extrêmes

Un quantile d'ordre extrême d'ordre $1 - \alpha_n$ est défini par

$$q(\alpha_n) = F^{\leftarrow}(1-\alpha_n) = \inf\{y\colon F(y) \geqslant 1-\alpha_n\} = \inf\{y\colon \overline{F}(y) \leqslant \alpha_n\}$$

avec $\alpha_n \to 0$ quand $n \to \infty$

- $\alpha_n \to 0 \Leftrightarrow$ l'information importante pour estimer un quantile extrême est contenue dans la queue de distribution

Estimation de quantiles extrêmes

Probabilité que le quantile soit plus grand que le max

$$\mathbb{P}\{X_{(n)} \leqslant q(\alpha_n)\} \underset{n\to\infty}{=} \exp(-n\alpha_n(1+o(1)))$$

- **1er cas : Si** $n\alpha_n \to \infty$**, alors** $\mathbb{P}\{X_{(n)} \leqslant q(\alpha_n)\} \to 0$
  → Estimer un quantile qui se trouve avec grande probabilité **dans** l'échantillon
  → Requiert une **interpolation** à l'intérieur de l'échantillon
  → Estimateur naturel $X_{(\lfloor n\alpha \rfloor)}$
  → Estimateur asymptotiquement normal

Estimation de quantiles extrêmes

Probabilité que le quantile soit plus grand que le max

$$\mathbb{P}\{X_{(n)} \leqslant q(\alpha_n)\} \underset{n\to\infty}{=} \exp(-n\alpha_n(1+o(1)))$$

- **2e cas : Si** $n\alpha_n \to 0$**, alors** $\mathbb{P}\{X_{(n)} \leqslant q(\alpha_n)\} \to 1$
  → Estimer un quantile qui se trouve avec grande probabilité **hors** l'échantillon
  → Inverser la fonction de répartition empirique n'est plus une solution
  → Requiert une **extrapolation** à l'extérieur de l'échantillon
- Estimer les quantiles d'une GEV

Quantiles d'une loi de valeurs extrêmes généralisées

- Il suffit d'inverser l'équation de la GEV

$$G_{\mu,\sigma,\gamma}(x) = \exp\left(-\left(1+\gamma\frac{x-\mu}{\sigma}\right)_+^{-1/\gamma}\right), x \in \mathbb{R}$$

Quantiles extrêmes

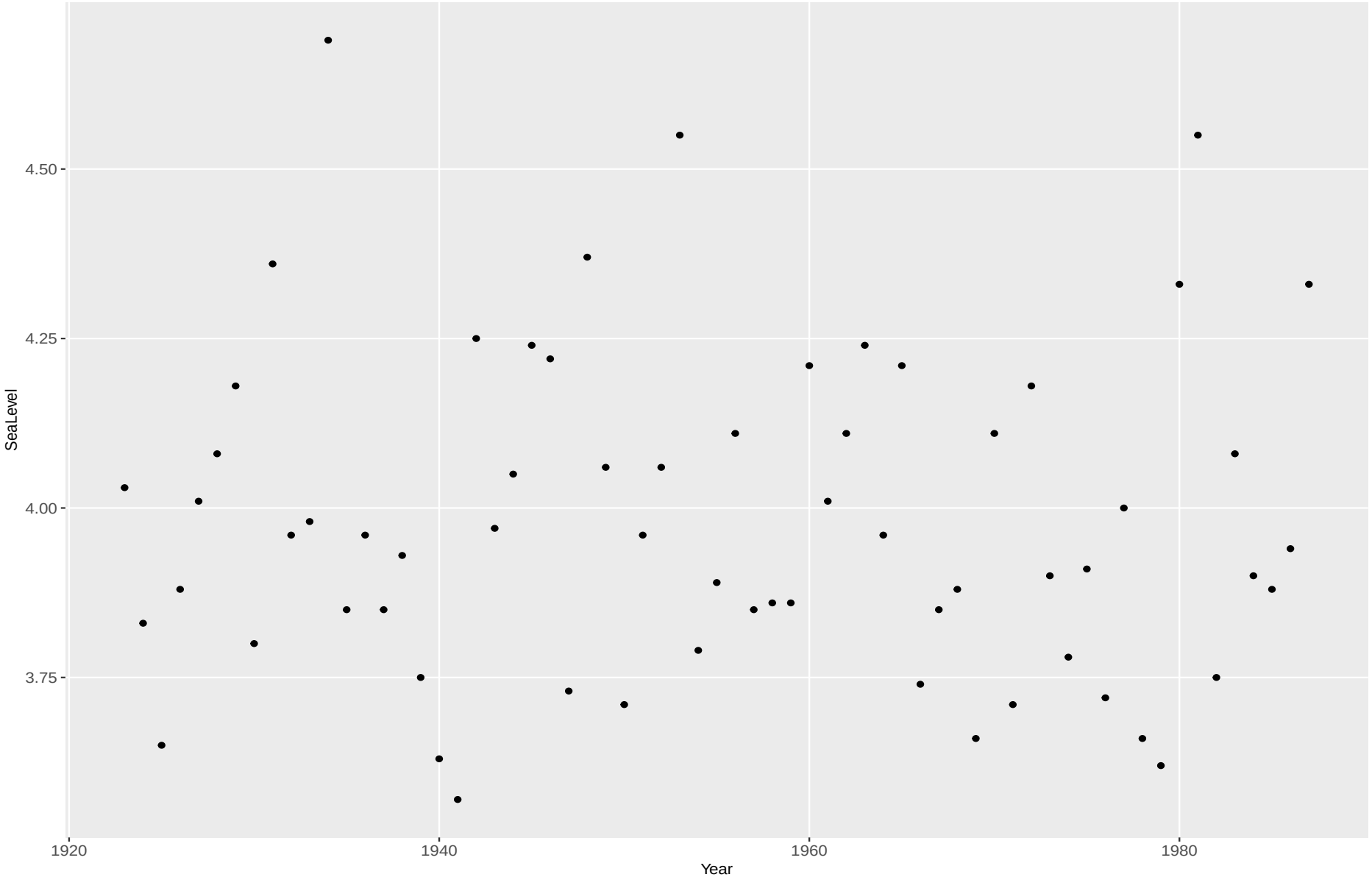
$$G_{\mu,\sigma,\gamma}(z_p) = 1-p \Leftrightarrow z_p = \begin{cases} \mu - \frac{\sigma}{\gamma}\left(1-(-\log(1-p))^{-\gamma}\right) & \text{pour } \gamma \neq 0 \\ \mu - \sigma\log\left(-\log(1-p)\right) & \text{pour } \gamma = 0 \end{cases}$$

- z_p est le niveau de retour associé à la période de retour $1/p$
- Pour les maxima annuels,
  $\rightarrow$ z_p devrait être dépassé en moyenne une fois tous les $1/p$ ans
  $\rightarrow$ z_p est dépassé par le maximum annuel pour une année donnée avec une probabilité p

Validation de modèle

- Les conditions d'appartenance à un DA ne peuvent pas être vérifiées en pratique
  $\rightarrow$ On applique la méthode des maxima par blocs et on vérifie l'adéquation à une GEV a posteriori
- QQPlot, return level plots
- Test du rapport de vraisemblance pour tester $H_0 : \gamma = 0$ contre $H_1 : \gamma \neq 0$

Exemple : Niveaux de la mer annuels maximaux à Port Pirie de 1923 à 1987 [Coles, 2001]



Exemple : Niveaux de la mer annuels maximaux à Port Pirie de 1923 à 1987 [Coles, 2001]

- Estimations de niveaux de retour
  → le niveau de la mer dépassera 4.30m avec probabilité $p = 0.1$ l'an prochain
  le niveau de la mer dépassera 4.30m en moyenne 1 fois toutes les 10 ans
  → le niveau de la mer dépassera 4.70m avec probabilité $p = 0.01$
  le niveau de la mer dépassera 4.70m en moyenne 1 fois toutes les 100 ans

- Probabilité de dépasser le maximum observé
  → Maximum observé = 4.69m
  → Le niveau de l'eau dépassera 4.69m avec probabilité 0.016 l'an prochain

Lois limites des excès

Étude de la loi des excès

- X v.a. de fonction de répartition F (fonction de survie $\overline{F} = 1 - F$)
- u seuil fixé
- Excès de X au dessus de u = v.a. $Y_u = X - u$ définie sur $\{X > u\}$

Loi d'excès

La loi de Y_u est la loi dont la fonction de survie est donnée par

$$\overline{F}_u(x) = \frac{\overline{F}(u+x)}{\overline{F}(u)}$$

- En pratique, F est inconnue, donc $\overline{F}_u$ aussi

Famille de lois limites des excès

Théorème [Balkema and de Haan, 1974, Pickands, 1975]

Si F appartient au max domaine d'attraction d'une GEV $G_{\mu,\sigma,\gamma}$, alors $\overline{F}_u$ peut être approchée par une loi de fonction de répartition donnée par

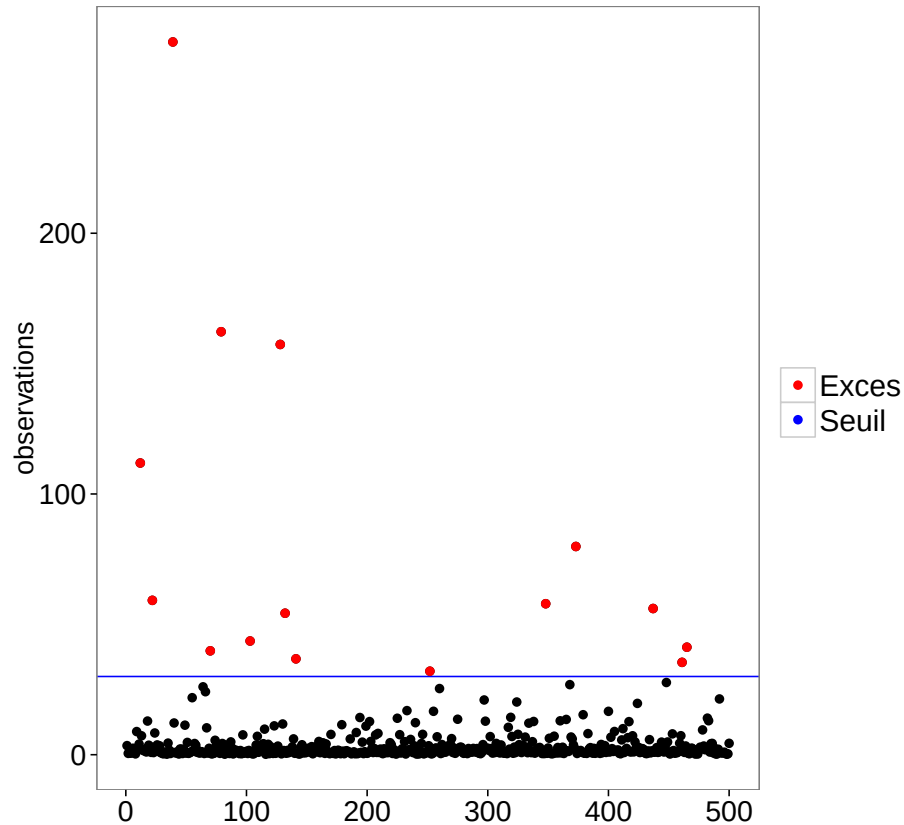
$$H_{\sigma,\gamma}(x) = 1 - \left(1 + \gamma \frac{x}{\tilde{\sigma}}\right)_+^{-1/\gamma} \quad x > 0$$

avec $\tilde{\sigma} = \sigma + \gamma(u - \mu)$

- Ici la convergence est obtenue lorsque $u \to \infty$
- Si $\gamma = 0$, le membre de droite se lit $1 - \exp(-x/\tilde{\sigma})$
- Si F vérifie les hypothèses ci-dessus, on dit que F appartient au domaine d'attraction des excès de $H_{\sigma,\gamma}$
- Domaine d'attraction des excès = max-domaine d'attraction
- Lois de Pareto généralisées (GPD) = famille des lois limites des excès

Méthode « Peaks over Threshold »

- Données brutes : observations $X_1, \ldots, X_n$ i.i.d.
- **Événement extrêmes** : observations ayant dépassé un certain seuil u précédemment choisi
  $\rightarrow X_j$ si $X_j > u$ for $j = 1, \ldots, k$
- **Excès** : $Z_j = X_j - u$ si $X_j > u$ pour $j = 1, \ldots, k$
  $\rightarrow$ les Z_j = v.a. i.i.d. dont la loi peut être approchée par une GPD

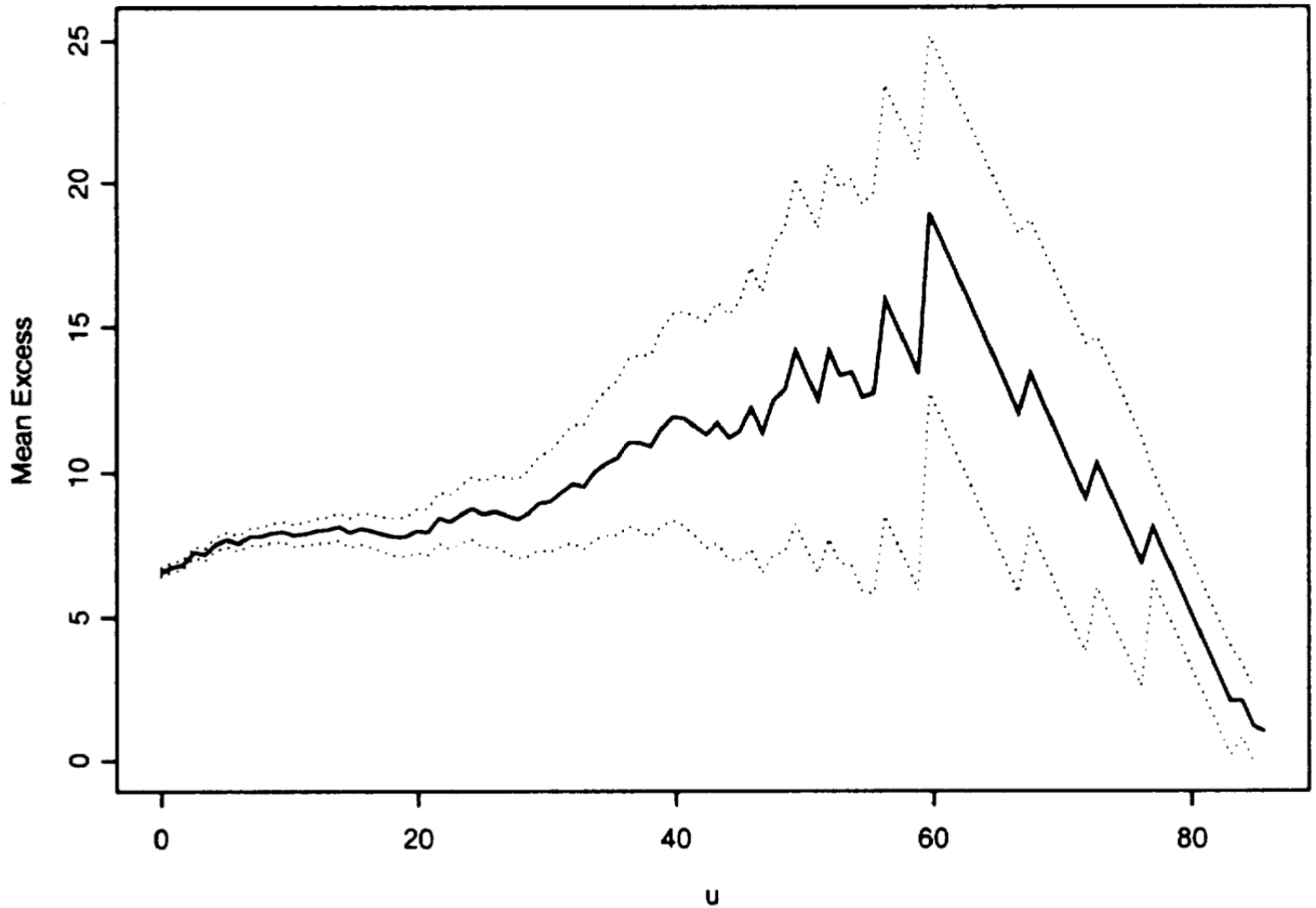


Choix du seuil

- Choix du seuil ↔ Choix de la taille du bloc
  → Compromis biais-variance
  → Un seuil trop bas est susceptible de transgresser le modèle asymptotique
  → Un seuil trop élevé génère peu d'excès avec lesquels le modèle peut être estimé, entraînant une variance élevée
- En pratique, on choisit le seuil le plus bas possible, sous réserve que le modèle asymptotique fournisse une approximation raisonnable
- Plusieurs méthodes
  1. Mean residual life plot (réalisée avant l'estimation du modèle)
  2. Évaluation de la stabilité des estimations des paramètres, à partir de l'ajustement des modèles pour une série de seuils différents.
  3. Parfois, le choix du seuil est guidé par les experts (vérification avec à ces deux techniques + qqplots)
- Il existe d'autres méthodes non-graphiques
  [Scarrott and MacDonald, 2012]

Exemple d'un mean residual life plot

Source : [Coles, 2001]



Quantiles d'une loi de Pareto généralisée

- Les quantiles extrêmes peuvent être également vus comme des quantiles d'une GPD
- Il suffit d'inverser l'équation de la GPD

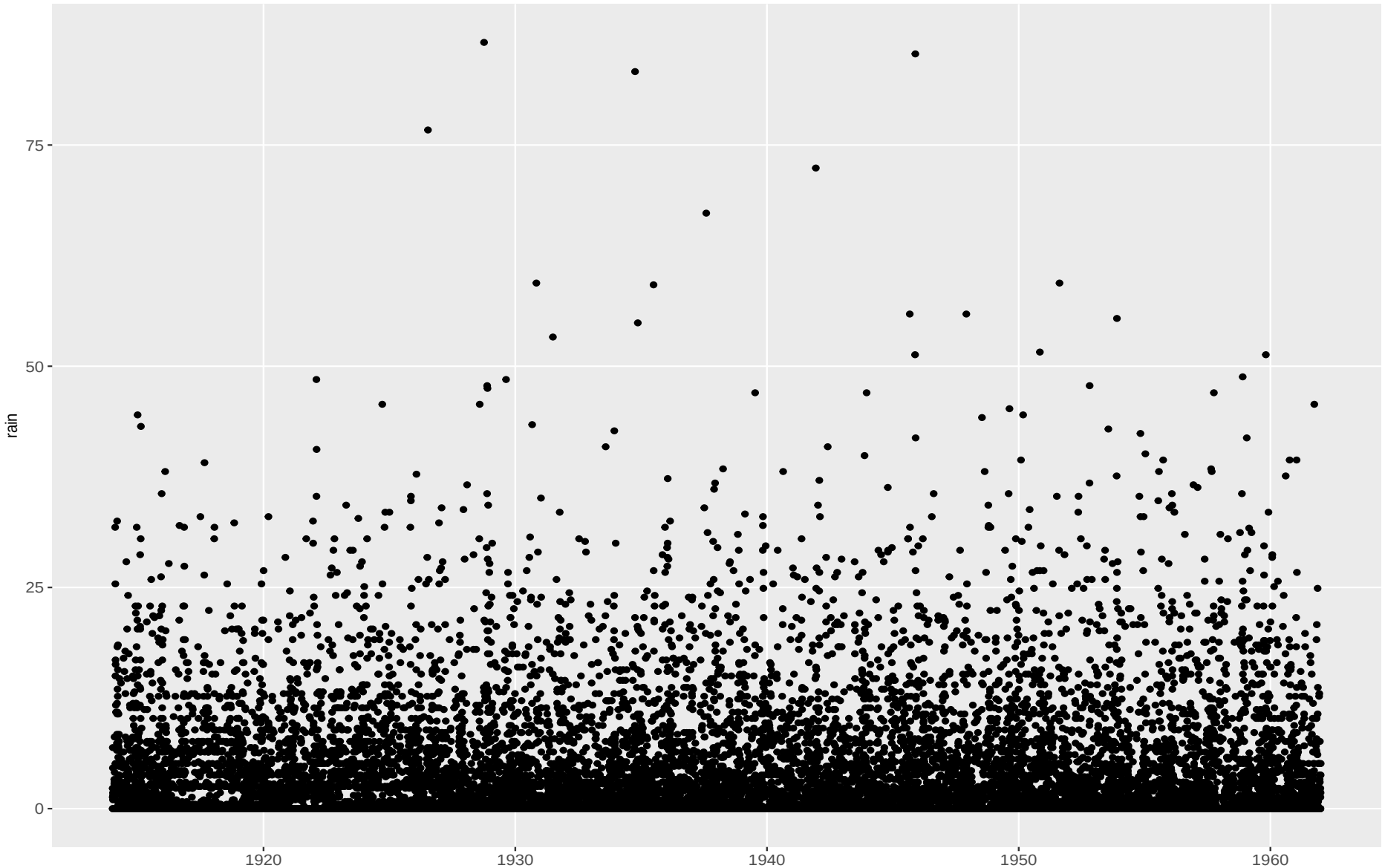
$$H_{\sigma,\gamma}(x) = 1 - \left(1 + \gamma\frac{x}{\sigma}\right)_+^{-1/\gamma} \quad x > u$$

Quantiles extrêmes

$$\mathbb{P}[X > x_m \mid X > u] = 1 - \frac{1}{m} \;\Leftrightarrow\; x_m = \begin{cases} u + \frac{\sigma}{\gamma}\left((m\zeta_u)^\gamma - 1\right) & \text{pour } \gamma \neq 0 \\ u + \sigma \log(m\zeta_u) & \text{pour } \gamma = 0 \end{cases}$$

- **Interprétation :** x_m est le seuil dépassé en moyenne toutes les m observations
  → Appelé **m-observation return level**

Exemple : Précipitations journalières à une station en Angleterre de 1914–1962 [Coles, 2001]



Exemple : Précipitations journalières à une station en Angleterre de 1914-1962 [Coles, 2001]

- Estimations de niveaux de retour
  → le niveau de précipitations dépassera 66mm avec probabilité $p = 0.1$ l'année prochaine
  le niveau de précipitations dépassera 66mm en moyenne 1 fois toutes les 10 ans
  → la quantité de précipitations dépassera 106mm avec probabilité $p = 0.01$
  la quantité de précipitations dépassera 106mm en moyenne 1 fois toutes les 100 ans

- Probabilité de dépasser le maximum observé
  → Maximum observé = 86.6mm
  → La quantité de précipitations dépassera 86.6mm avec probabilité 0.027 l'an prochain

Généralisation à des variables non i.i.d.

Modélisation des séries stationnaires

Maxima par blocs

- La réponse est particulièrement simple
- Pourvu que la dépendance à long terme à des niveaux extrêmes soit faible
  $\Rightarrow$ La loi des maxima par blocs est également une GEV
- Donc il est toujours approprié de modéliser la distribution du maximum annuel en utilisant une GEV
  $\rightarrow$ Mais les paramètres seront différents de ceux qui auraient été obtenus si la série était indépendante
  $\rightarrow$ Mais comme les paramètres doivent être estimés de toute façon, cela n'a pas d'importance.
- **Conclusion :** Pour la méthode des maxima par blocs, la dépendance des données peut être ignorée

Modélisation de suites stationnaires

Méthode Peaks over Threshold

- Tout comme la GEV reste un modèle approprié pour les maxima par blocs des séries stationnaires, des arguments similaires suggèrent que la GPD reste appropriée pour les excès de seuil.
- MAIS, les extrêmes peuvent avoir tendance à se regrouper dans une série stationnaire
  $\rightarrow$ Il faut adapter la méthode
- Pour les séries stationnaires, les arguments asymptotiques habituels impliquent que la loi marginale des dépassements d'un seuil élevé est une GPD.
- MAIS ils ne conduisent pas à une spécification de la loi conjointe des excès voisins

Modélisation de suites stationnaires

Méthode Peaks over Threshold

- Diverses suggestions ont été faites pour résoudre ce problème.
  → La plus largement adoptée = **declustering**
  1. utiliser une règle empirique pour définir les clusters de dépassements
  2. identifier l'excès maximal dans chaque cluster
  3. supposer que les maxima des clusters sont indépendants de loin GPD
  4. ajuster le GPD aux maxima des clusters
- La méthode est simple mais a ses limites
  → Les résultats peuvent être sensibles aux choix arbitraires faits dans la détermination des grappes.
- On peut dire qu'il y a un gaspillage d'informations en rejetant toutes les données sauf les maxima des clusters.

Suites non stationnaires

Exemples de modèles

- Par exemple, une tendance dans les données peut être apparente
  → Cela soulève des doutes sur la pertinence d'un modèle constant dans le temps
  → Supposons que le paramètre μ change linéairement au cours du cours, mais qu'à d'autres égards, la distribution reste inchangée.

$$M_t \sim \text{GEV}(\mu(t), \sigma, \gamma)$$

où $\mu(t) = \beta_0 + \beta_1 t$, avec β_0 et β_1 des paramètres à estimer.

- Des changements plus complexes de μ peuvent également être envisagés
  → par exemple, un modèle quadratique

$$\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2$$

  → par exemple, un modèle de rupture

$$\mu(t) = \begin{cases} \mu_1 & \text{pour } t \leqslant t_0 \\ \mu_2 & \text{pour } t > t_0 \end{cases}$$

Modèles

- La non-stationnaire peut également être exprimée en termes des autres paramètres de la GEV
  $\rightarrow$ par exemple,
  $$\sigma(t) = \exp(\beta_0 + \beta_1 t)$$
  où l'exponentielle assure la positivité de σ.
- γ est déjà difficile à estimer, il est donc généralement irréaliste d'essayer de le modéliser comme une fonction lisse du temps
- **Alternative** = spécifier un modèle avec des paramètres différents pour chaque saison.
- Structure commune à tous ces exemples
  $$\lambda(t) = h(X\beta)$$
  où λ désigne l'un des paramètres et h est une fonction spécifiée, appelée fonction de lien inverse

Choix de modèle

- Possibilité de modéliser n'importe quelle combinaison de paramètres du modèle EVT en tant que fonctions du temps ou de covariables
  → Grand catalogue de modèles parmi lesquels choisir
  → Sélectionner un modèle approprié devient une question importante
- **Principe de base** : parcimonie
  → Obtenir le modèle le plus simple qui explique le mieux possible la variation des données.
- Comparaison avec les tests AIC, BIC et tests de rapport de vraisemblance (pour les modèles emboîtés)

Théorie des valeurs extrêmes et assurabilité

Théorie des Valeurs Extrêmes et Assurabilité

- N= nombre de sinistres ($\mathbb{E}[N]$ =Fréquence)
- Y coût moyen d'un sinistre ($\mathbb{E}[Y]$ =Sévérité)

Tarification = équilibre (en moyenne) entre le coût d'un assuré et les engagements de l'assureur

$$\pi = E[N]E[Y]$$

- π = prime du contrat d'un assuré
- Hypothèse classique : Y et N sont indépendantes

- **Mutualisation** : si on dispose de n assurés identiques, la perte de l'assureur devrait être environ $n\pi$.

Théorie des Valeurs Extrêmes et Assurabilité

- $Y =$ coût d'un sinistre.
- $\gamma > 0$ (loi à queue lourde).
- Si $Z = Y - u | Y \geqslant u$ suit une GPD de paramètres γ et σ,

$$\mathbb{E}[Z|Z \geqslant 0] = \begin{cases} \frac{\sigma}{1-\gamma} & \text{si } \gamma < 1. \\ \infty & \text{si } \gamma \geqslant 1. \end{cases}$$

- Dans le cas $\gamma \geqslant 1$, le sinistre n'est "pas assurable". L'assureur :
  - peut exclure le risque.
  - peut introduire des limites de garanties (d'autant plus basses que γ est grand).

Classification des sinistres extrêmes

Théorie des valeurs extrêmes et régression

- Cadre de la régression
  - Considère une observation de caractéristiques X
  - Suppose que la loi de $Y \mid X$ est à queue lourde = la loi des excès $Z \mid X$ converge vers une GPD de paramètre $\gamma(X) > 0$ et $\sigma(X)$

$$H_{\sigma,\gamma}(z) = 1 - \left(1 + \frac{\gamma(X)}{\sigma(X)} z\right)^{-1/\gamma(X)}$$

- But : estimer $\gamma(X)$

- Espoir : pour certaines valeurs de X, $\gamma(X) < 1$.

- Méthodes :
  - Approches semi-paramétriques :
    - Exponenial regression model (Beirlant et al., 2003)
    - Smoothing splines (Chavez-Demoulin et al., 2015)

  - Approches non paramétriques (Beirlant and Goegebeur, 2004) :
    - Local polynomial maximum likelihood
    - Seulement pour les variables continues

Institut des ACTUAIRES

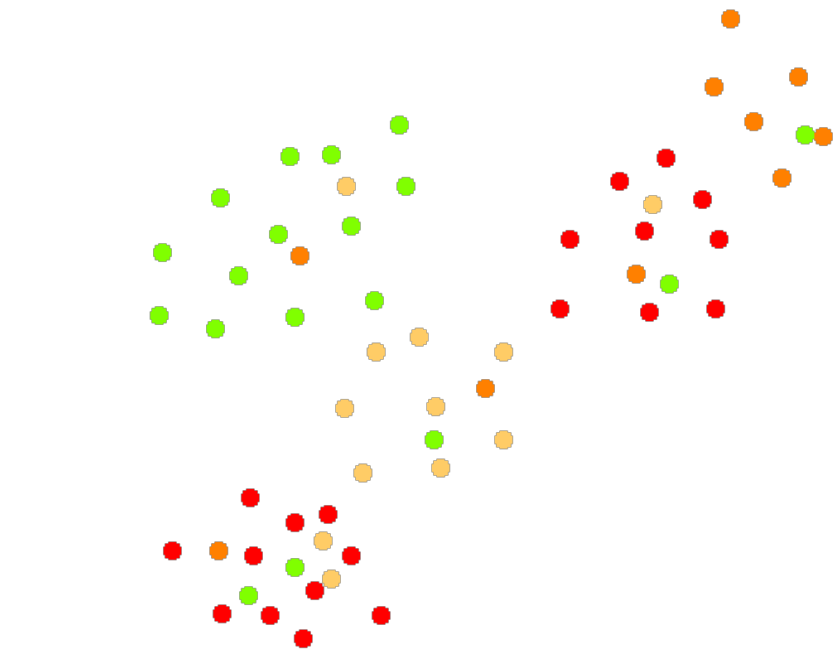
Classification And Regression Trees (CART)

Arbre de régression [Breiman et al., 1984]

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

- Z variable à expliquer (par ex. coût du sinistre)
- $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ ensemble de variables explicatives
- $\mathcal{F}$ classe de fonctions cibles sur $\mathbb{R}^d$
- ϕ fonction de perte qui dépend de la quantité que l'on souhaite estimer

Croissance de l'arbre



CART : Step 0

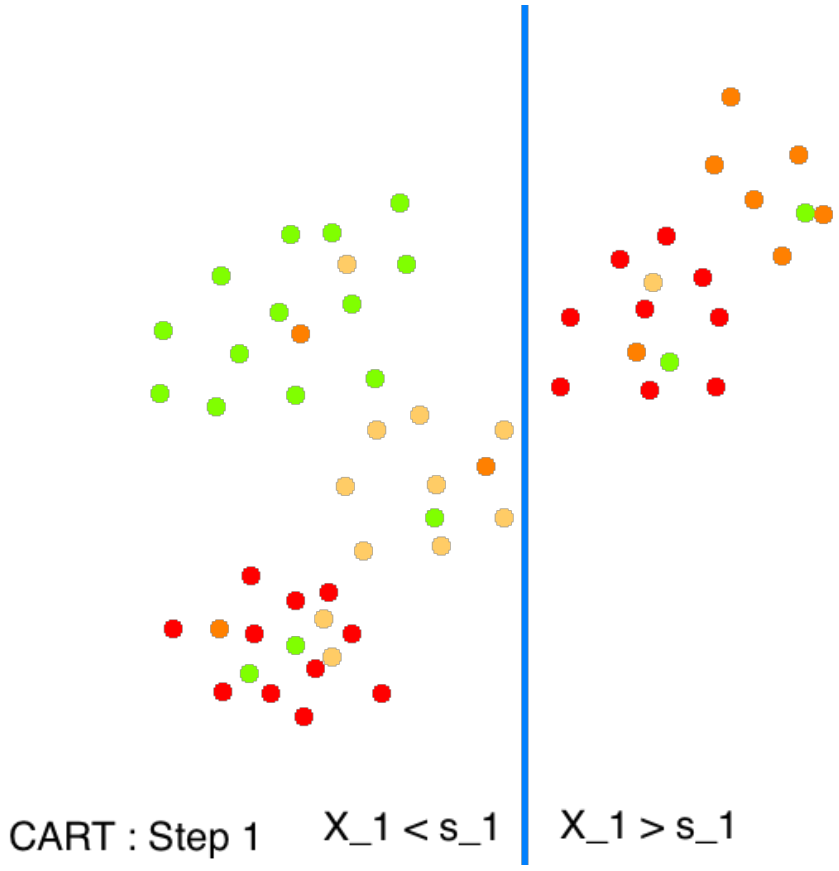
Croissance de l'arbre

Partitionnement

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

avec

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



Croissance de l'arbre

1. Pour chaque composante $X^{(\ell)}$, définir $x_\star^{(\ell)}$

$$x_\star^{(\ell)} = \arg\min_{x^{(\ell)}} \Phi(R_j, x^{(\ell)})$$

avec

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

et

$$\begin{aligned}\theta_{\ell-}(x, R_j) &= \arg\max_{\theta} \sum_{i=1}^{n} \phi(Y_i, \theta(\mathbf{X}_i)) \mathbf{1}_{X_i^{(\ell)} \leqslant x} R_j(\mathbf{X}_i), \\ \theta_{\ell+}(x, R_j) &= \arg\max_{\theta} \sum_{i=1}^{n} \phi(Y_i, \theta(\mathbf{X}_i)) \mathbf{1}_{X_i^{(\ell)} > x} R_j(\mathbf{X}_i).\end{aligned}$$

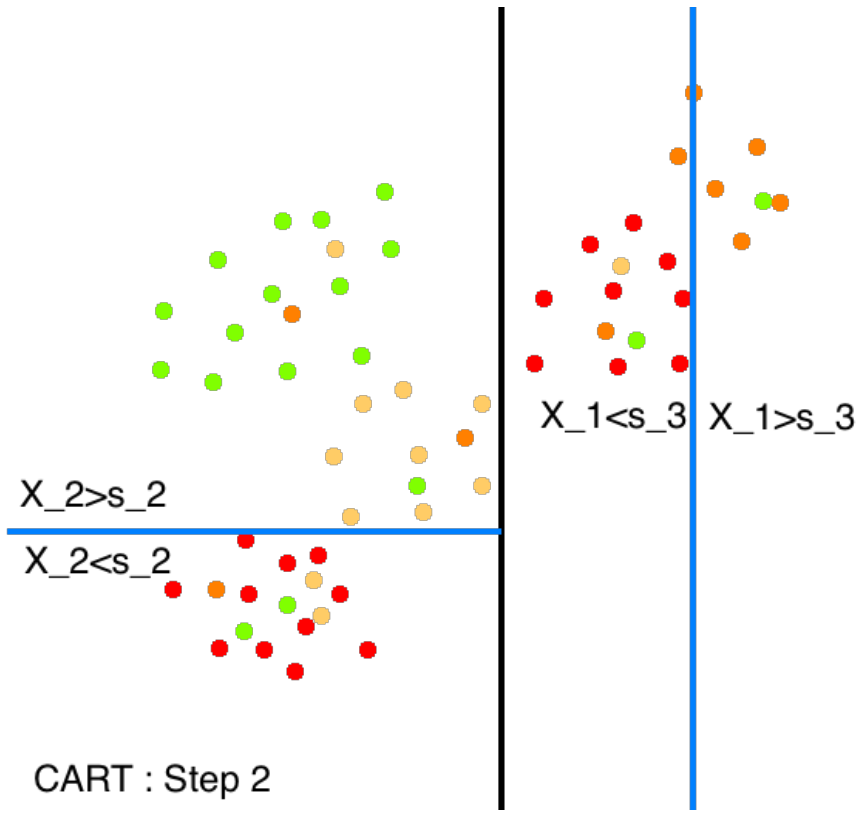
Croissance de l'arbre

2. Choisir le meilleur indice

$$\hat{\ell} = \arg\min_{\ell} \Phi(R_j, x_\star^{(\ell)})$$

3. Définir

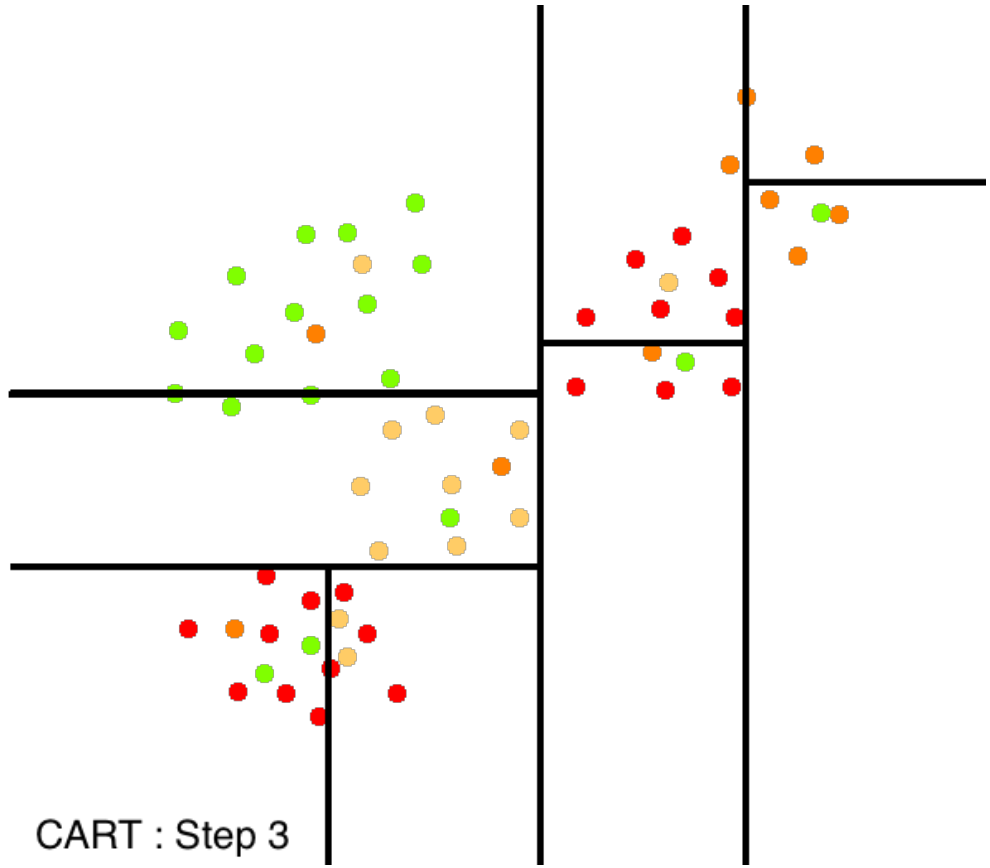
$$R_{j_1}(\mathbf{x}) = R_j(\mathbf{x}) \mathbb{1}_{x^{(\hat{\imath})} \leqslant x_\star^{(\hat{\ell})}} \quad \text{et} \quad R_{j_2}(\mathbf{x}) = R_j(\mathbf{x}) \mathbb{1}_{x^{(\hat{\imath})} > x_\star^{(\hat{\ell})}}$$



Croissance de l'arbre

Estimateur de la fonction de régression $\widehat{\theta}(\mathbf{X})$ est donné par

$$\widehat{\theta}(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\theta}(R_j) R_j(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\theta}_\ell \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell} \quad \text{où } \mathcal{T}_\ell = \text{ensemble des feuilles obtenues}$$



Fonction de perte

- Perte quadratique → "Mean regression" (espérance conditionnelle)

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

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

- Perte absolue → "Median regression" (médiane conditionnelle)

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

$\hookrightarrow \theta^*(\mathbf{x}) =$ médiane conditionnelle

- Perte liée à la log-vraisemblance négative, ici GPD

$$\phi(z, \theta(\mathbf{x})) = \log(\sigma) + \left(\frac{1}{\gamma(\mathbf{x})} + 1\right) \log\left(1 + \frac{\gamma(\mathbf{x})z}{\sigma(\mathbf{x})}\right), \quad z > 0$$

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

Fonction de perte

- Perte liée à la log-vraisemblance négative, ici GPD

$$\phi(z, \theta(\mathbf{x})) = \log(\sigma) + \left(\frac{1}{\gamma(\mathbf{x})} + 1\right) \log\left(1 + \frac{\gamma(\mathbf{x}) z}{\sigma(\mathbf{x})}\right), \quad z > 0$$

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

Application de la méthode PoT :

- sélectionner les observations $Y_i \geqslant u(\mathbf{X}_i)$, ici suppose $u(\mathbf{x}) = u \in [u_{\min}, u_{\max}]$
- k_n : nombre moyen de $Y_i \geqslant u$
- Appliquer la procédure CART aux excès $Z_i = Y_i - u > 0$

Élagage : sélection de modèle

- Soit $T_{\max}$ l'arbre maximal obtenu dans la première phase et $K_{\max}$ le nombre de ses feuilles
- Consiste à extraire de $T_{\max}$ un sous-arbre.
- Critère pénalisé (K nombre de feuilles de l'arbre T_K)

$$\frac{1}{k_n}\sum_{\ell=1}^{K}\sum_{i=1}^{n}\phi(Y_i-u,\widehat{\theta}^K(\mathbf{X}_i))\mathbf{1}_{Y_i>u}\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell}+\lambda K,$$

- $\lambda > 0$ est choisi par cross-validation
- On désigne par $\widehat{T}_K$ le meilleur arbre à K feuilles d'après ce critère.
- $\widehat{T}$ l'arbre minimisant le critère pénalisé, $\widehat{K}$ son nombre de feuilles ($\widehat{T} = \widehat{T}_{\widehat{K}}$)

Applications réelles : risque-cyber et risque inondation

Privacy Rights Clearinghouse (nonprofit association)

- Fondée en 1992
- Publique
- Référence pour les travaux académiques d'analyse des événements cyber liés à la fuite de données
- But : faire prendre conscience des problématiques de respect de la vie privée.
- Chronologie des fuites de données maintenue depuis 2005.
- Contient des informations sur des événements de multiples sources :
  - Agences gouvernementales US (Federal level–HIPAA).
  - Agences gouvernementales US (State level)
  - Média
  - Autres organisations
- 8860 événements

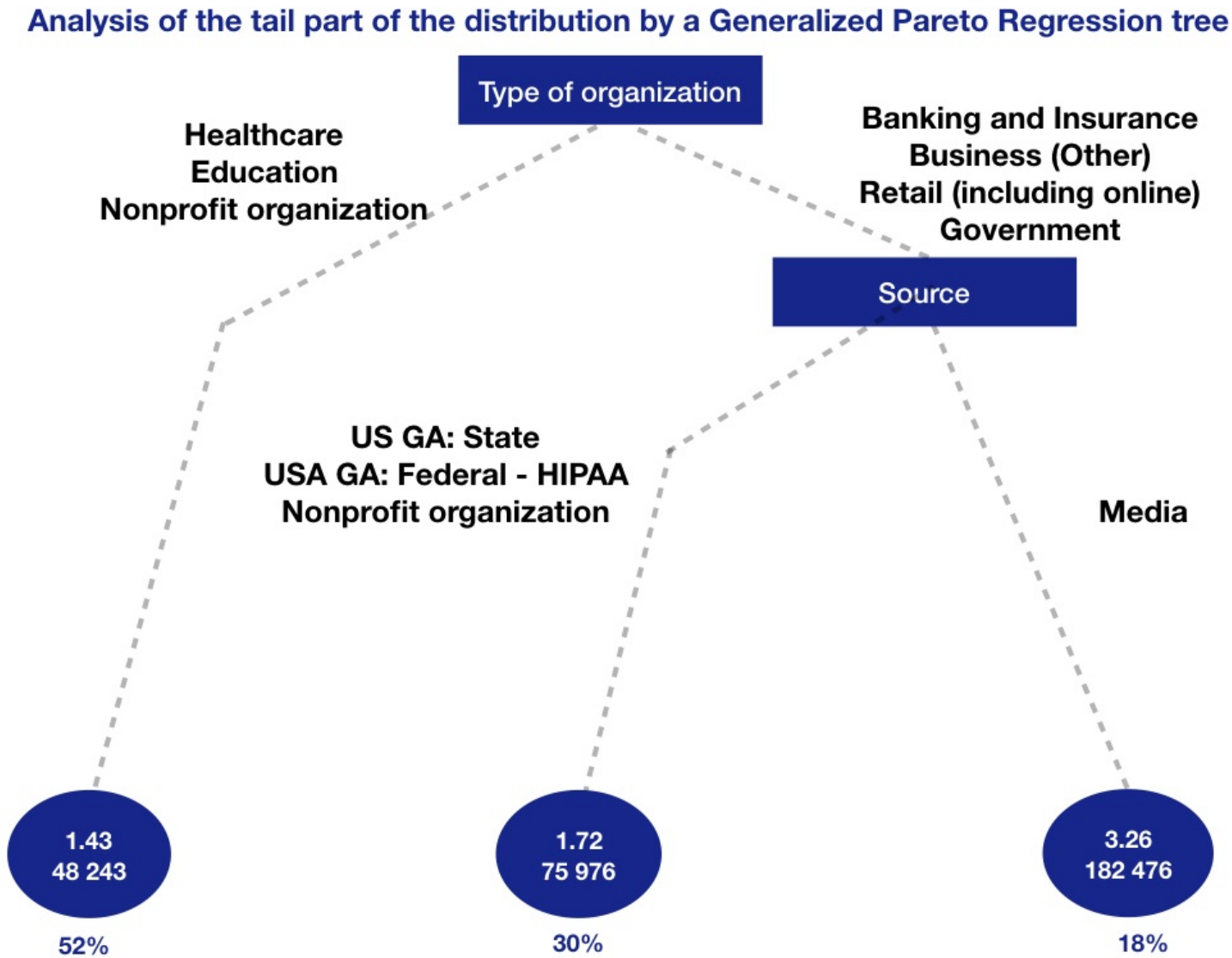
La base de données PRC : variables

- Variables liées à l'exposition : informations sur la victime.
- Variables liées à l'événement : informations sur la fuite de données.

Variables d'exposition	Nom de l'organisation Type d'organisation Localisation de l'organisation
Variables de l'événement	Source Date Type de brèches Nombre de lignes touchées Description de l'événement

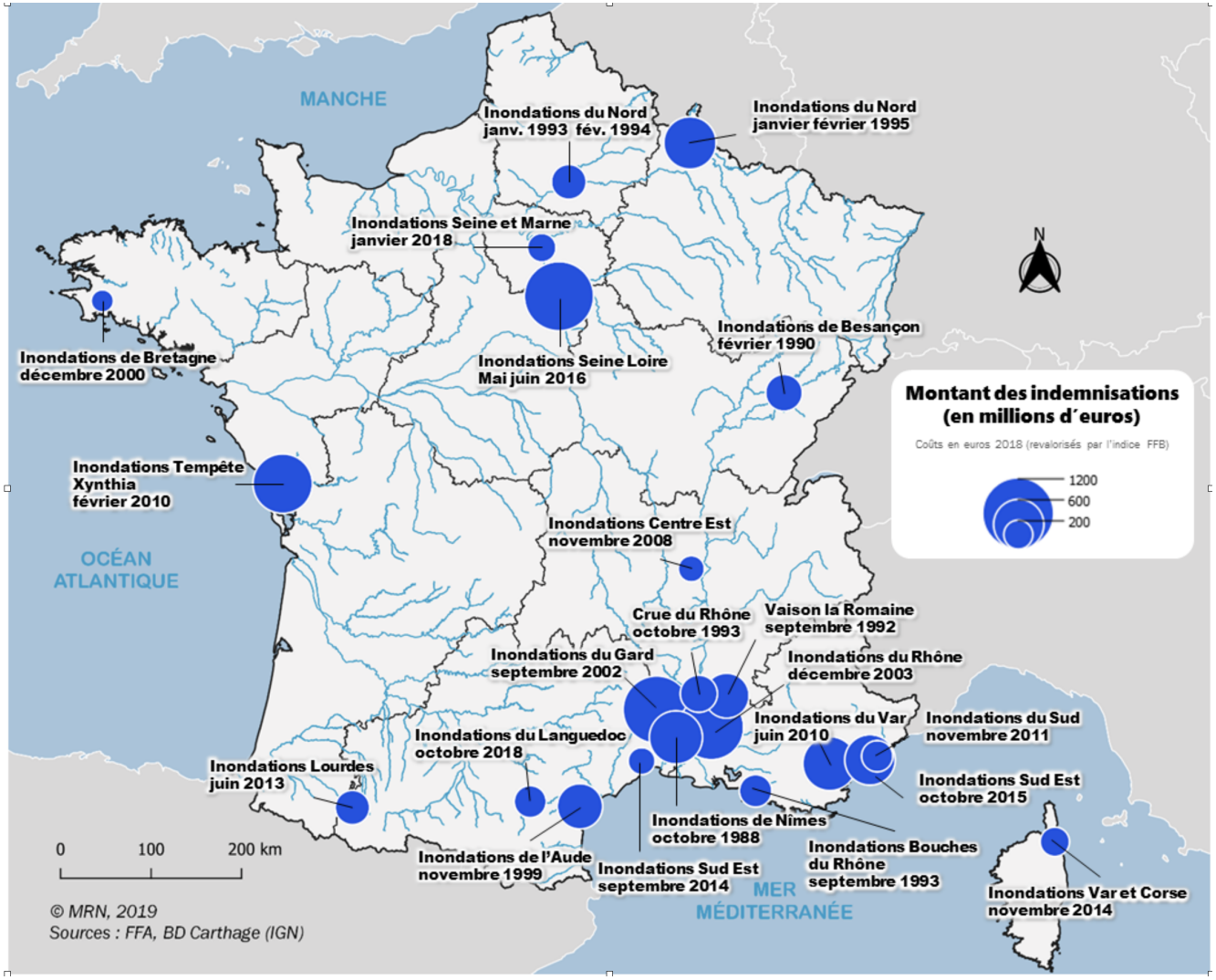
Application à la base de données PRC

- On considère les observations telles que le nombre de lignes est au-dessus de u =27 000



Les inondations en France

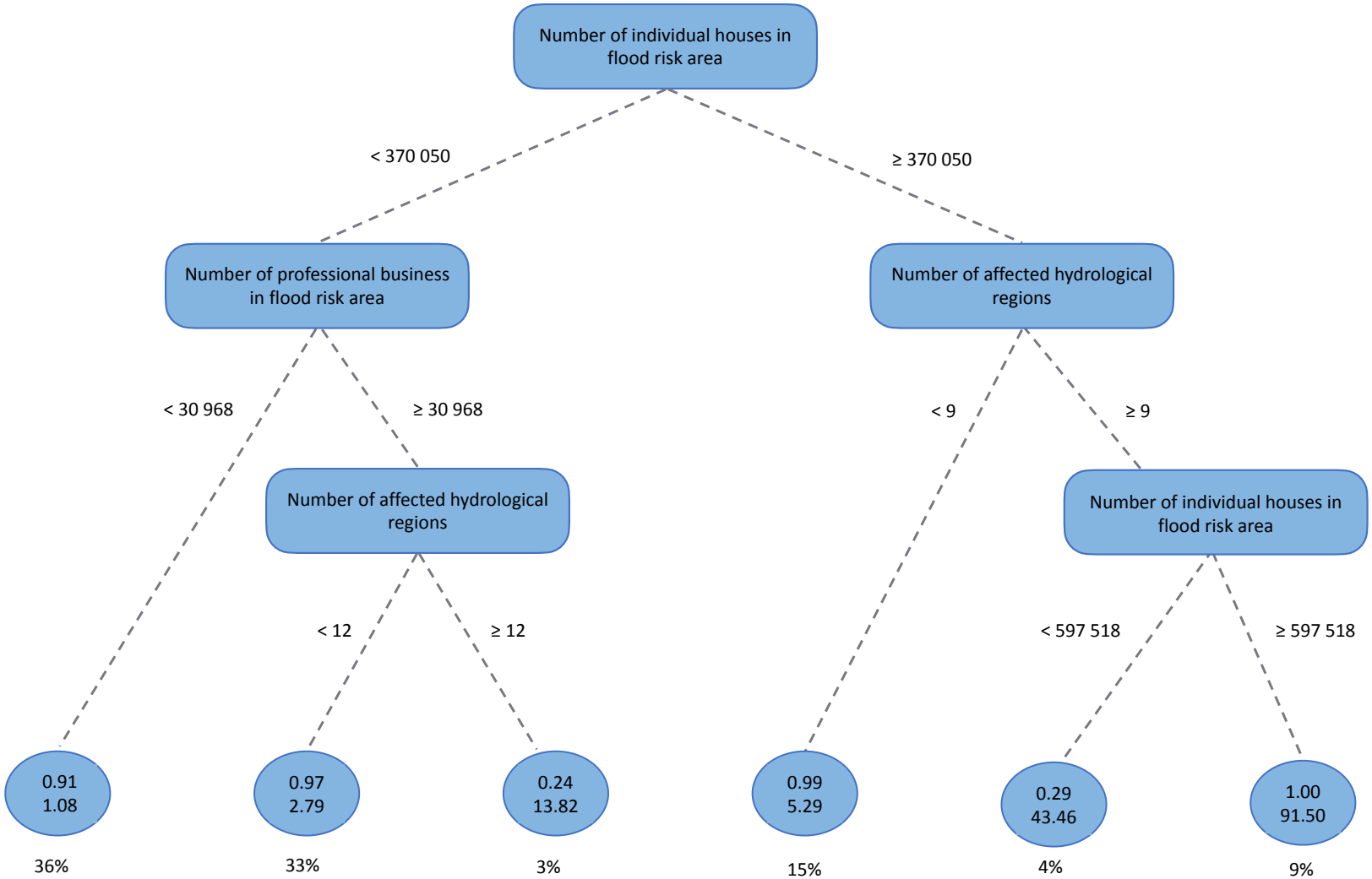
Les 22 plus grandes inondations en France depuis l'instauration du régime CatNat



Application aux événements inondations

- Base de données SILECC
  - Partenariat avec la MRN
  - Constituée des sinistres des plus grandes compagnies d'assurance (70% du marché français)
  - 700 000 sinistres de 1990 à 2019 dont **3 147 événements inondations**
- Covariables (disponibles peu de temps après l'occurrence de l'événement)
  - la région météorologique
  - la saison
  - le type d'inondations
  - le nombre de régions hydrologiques touchées
  - le nombre de maisons de maisons individuelles
  - le nombre de locaux professionnels dans la zone inondable
- Le seuil u a été choisi égal à 100 000€, ce qui correspond à 1 083 événements

Application aux événements inondations



Conclusion

- Classification des comportements extrêmes via des méthodes d'arbres :
  - permet de considérer des nonlinéarités dans cette dépendance
  - adapté aux variables $\mathbf{X}$ qui sont discrètes comme continues (intéressant notamment pour l'étude des comportements car de nombreuses variables sont qualitatives)
  - permet une classification
  - mais peut parfois être instable
- Des extensions naturelles, moins intelligibles mais plus précises (gradient boosting, forêts aléatoires) peuvent être utilisées
- De façon générale, outil d'aide à la décision pour tracer la ligne entre ce qui est « assurable »et ce qui ne l'est pas.

- Autre travail sur les sinistres extrêmes

  [Lopez and Thomas, 2023] *Parametric insurance for extreme risks : the challenge of properly covering severe claims*

Take Home message

- Les événements extrêmes nécessitent une attention particulière
- Ils sont par définition rares mais leurs conséquences peuvent être dramatiques
- Théorie des valeurs extrêmes = extrapoler au delà du support de l'échantillon
- Quantiles extrêmes (type Value at Risk) doivent être estimés à l'aide de méthodes de théorie des valeurs extrêmes
- Deux méthodes principales (en univarié)
  1. Méthode des maxima par blocs
  2. Méthode Peaks over Threshold
- Outil d'aide à la décision pour tracer la ligne entre ce qui est « assurable » et ce qui ne l'est pas (cadre régression)

Institut des ACTUAIRES

Merci de votre attention

« Il est impossible que l'improbable n'arrive jamais », Gumbel (1958)

Institut des ACTUAIRES

Références I

Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, pages 792–804.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer London.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pages 180–190. Cambridge University Press.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of mathematics*, pages 423–453.

Lopez, O. and Thomas, M. (2023). Parametric insurance for extreme risks : the challenge of properly covering severe claims.

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

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical journal*, 10(1) :33–60.