Contributions à la théorie des valeurs extrêmes et à la gestion du risque

Maud Thomas

LPSM, Sorbonne Université

Plan

. Introduction à la théorie des valeurs extrêmes

1. Bornes non asymptotiques pour l'estimation de l'indice de valeurs extrêmes
   - Contributions
   - Estimateur des moments pondérés robuste

2. Méthodes statistiques et d'apprentissage statistique en gestion du risque
   - Contributions
   - Prédiction en temps réel de la survenue d'une épidémie de grippe extrême

3. Méthodes d'arbres de régression
   - Contributions
   - Arbres de régression Pareto généralisés

. Perspectives de recherche

But de la théorie des valeurs extrêmes

But de la théorie des valeurs extrêmes

1. Estimer la probabilité d'occurrence d'un événement plus extrême que les événements passés
2. Estimer un quantile extrême

⇒ Inférence en dehors du support de l'échantillon

Approche classique

- Basée sur la fonction de répartition empirique
- Considérer que le pire s'est déjà produit

Deux points de vue en théorie des valeurs extrêmes

- Étude des lois limites du maximum d'un échantillon
- Étude des lois limites des excès au-dessus d'un seuil élevé

Famille de lois limites du maximum

- $Y_1, \ldots, Y_n$ v.a. i.i.d. de loi F inconnue

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

$$\mathbb{P}\left\{\frac{\max Y_i - b_n}{a_n} \leq z\right\} \xrightarrow[n\to\infty]{} G(z)$$

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 domaine d'attraction de G_γ, $F \in \mathrm{DA}(\gamma)$
- Famille des lois limites des excès = Lois de valeurs extrêmes généralisées (GEV)

Famille de lois limites des excès

- Y v.a. de fonction de répartition F (fonction de survie $\overline{F} = 1 - F$)
- u seuil fixé
- Excès de Y au dessus de u = v.a. $Z_u = Y - u$ définie sur $\{Y > u\}$
- **Loi des excès :** $\overline{F}_u(z) = \overline{F}(u+z)/\overline{F}(u)$

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

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

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

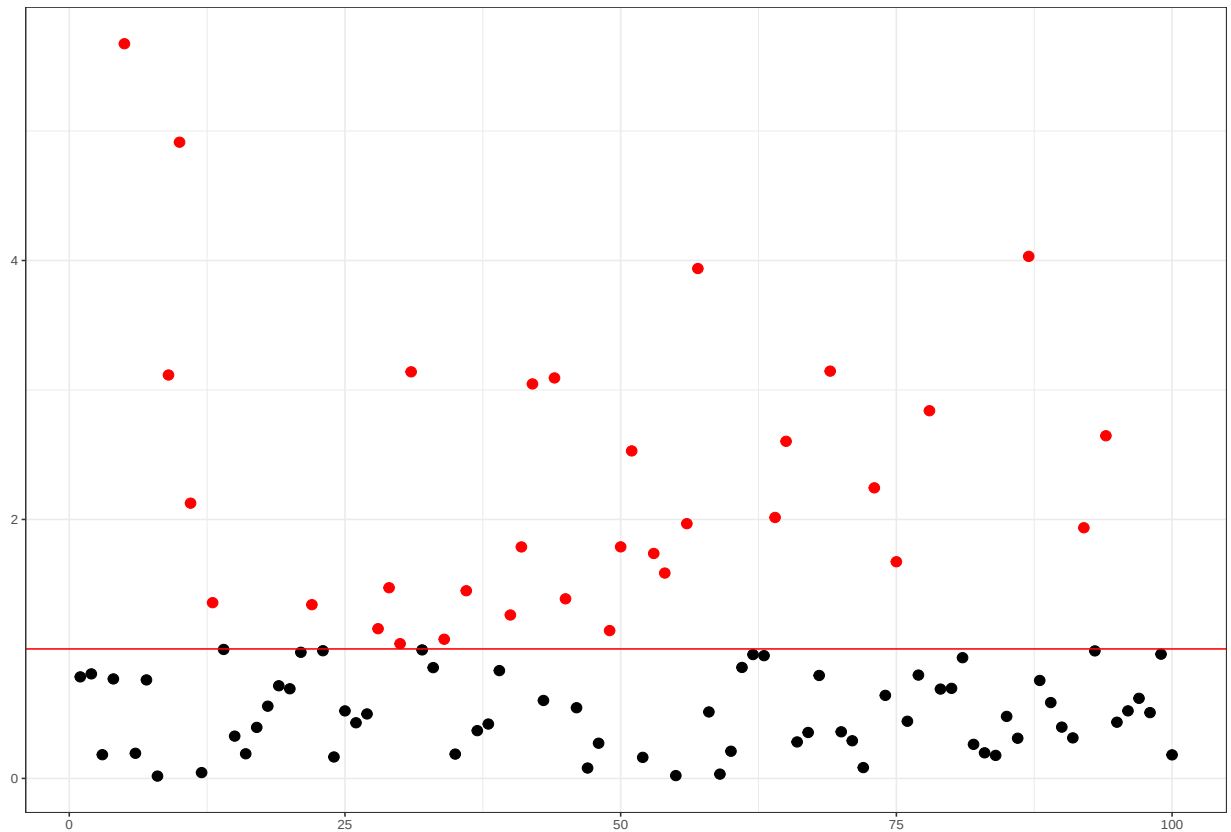
avec $\tilde{\sigma} = \sigma + \gamma(u - \mu)$

- Si $\gamma = 0$, le membre de droite se lit $1 - \exp(-z/\tilde{\sigma})$
- Famille des lois limites des excès = Lois de Pareto généralisées (GPD)

Méthode « Peaks over Threshold » (PoT)

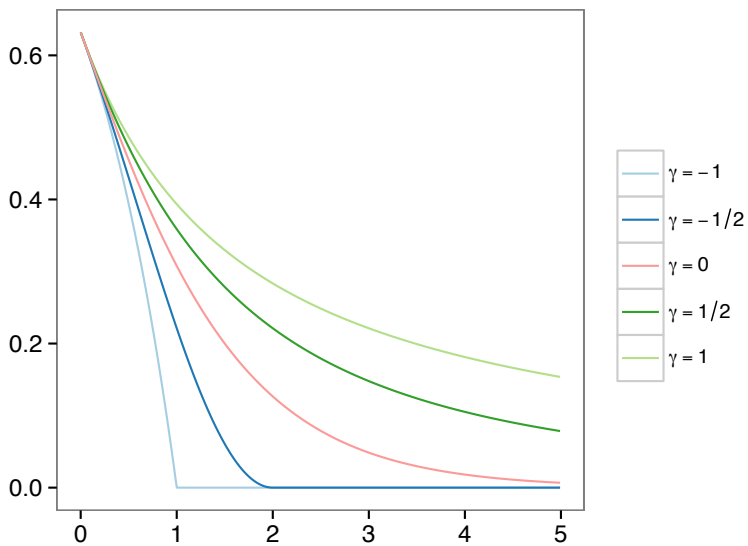
- $Y_1, Y_2, \ldots$ une suite de variables aléatoires i.i.d.
- Fixer un seuil (élevé) u
- Événement extrême = Y_i dépasse u

  → Sachant que $Y_i > u$, un excès est défini par $Z_i = Y_i - u$

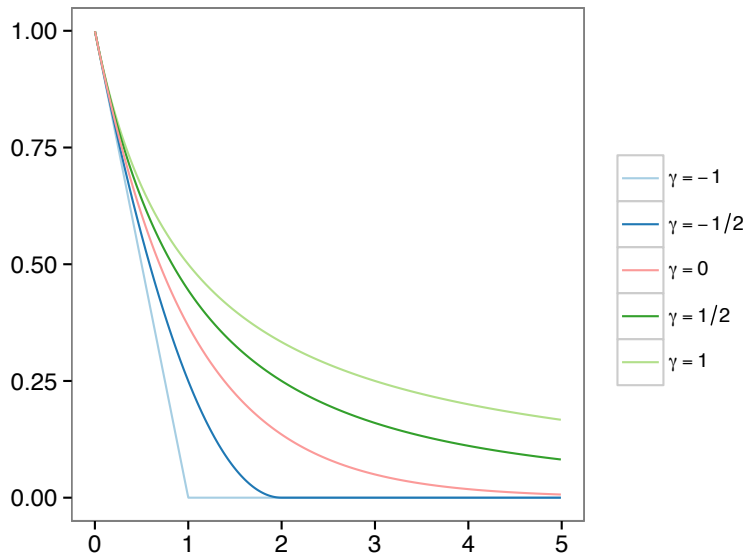


Domaines d'attraction

- Paramètre γ = paramètre de forme
  → reflète l'épaisseur de la queue de distribution
  → appelé l'indice de valeurs extrêmes
- 3 domaines d'attraction
  1. Domaine de Fréchet ($\gamma > 0$) : lois à queue lourde, décroissance polynonimale
  2. Domaine de Gumbel ($\gamma = 0$) : lois à queue fine, décroissance exponentielle
  3. Domaine de Weibull ($\gamma < 0$) : lois à queue finie à droite



Fonctions de survie des GEV



Fonctions de survie des GPD

Plan

. Introduction à la théorie des valeurs extrêmes

1. Bornes non asymptotiques pour l'estimation de l'indice de valeurs extrêmes
   - Contributions
   - Estimateur des moments pondérés robuste

2. Méthodes statistiques et d'apprentissage statistique en gestion du risque
   - Contributions
   - Prédiction en temps réel de la survenue d'une épidémie de grippe extrême

3. Méthodes d'arbres de régression
   - Contributions
   - Arbres de régression Pareto généralisés

. Perspectives de recherche

Bornes non asymptotiques pour l'estimation de γ

- γ reflète le comportement de la queue de distribution
  → Estimateurs = fonctions des plus grandes statistiques d'ordre
- Inférence basée sur un nombre d'observations restreint
- Or, la théorie repose essentiellement sur des résultats asymptotiques
- Élaboration des résultats non asymptotiques
  ⇒ garantir son usage pour des échantillons de taille finie

Contributions

1. Bornes non-asymptotiques de la variance et de la queue de probabilité des statistiques d'ordre d'un échantillon i.i.d.
   [Boucheron and Thomas, 2012]
2. Version adaptative de l'estimateur de Hill grâce à la méthode de Lepski et à partir d'inégalités de concentration
   [Boucheron and Thomas, 2015]
3. Nouvelle classe d'estimateurs des moments pondérés sous-gaussiens et robustes inspirée du cadre de *median-of-means*
   [Ben-Hamou, Naveau, and Thomas, 2023]

Bornes non asymptotiques pour l'estimation de γ

- γ reflète le comportement de la queue de distribution
  → Estimateurs = fonctions des plus grandes statistiques d'ordre
- Inférence basée sur un nombre d'observations restreint
- Or, la théorie repose essentiellement sur des résultats asymptotiques
- Élaboration des résultats non asymptotiques
  ⇒ garantir son usage pour des échantillons de taille finie

Contributions

1. Bornes non-asymptotiques de la variance et de la queue de distribution des statistiques d'ordre d'un échantillon i.i.d.
   [Boucheron and Thomas, 2012]
2. Version adaptative de l'estimateur de Hill grâce à la méthode de Lepski et à partir d'inégalités de concentration
   [Boucheron and Thomas, 2015]
3. Nouvelle classe d'estimateurs des moments pondérés sous-gaussiens et robustes inspirée du cadre de *median-of-means*
   [Ben-Hamou, Naveau, and Thomas, 2023]

Estimateurs des moments pondérés

- Les moments pondérés (PWM pour *Probability Weighted Moments*) d'une v.a. réelle X de loi F telle que $\mathbb{E}|X| < +\infty$ sont définis, pour tous entiers r, s, par

$$\mathbb{E}\left[XF(X)^r\overline{F}(X)^s\right]$$

- Utilisation motivée par les hydrologues et les statisticiens appliqués [Landwehr et al., 1979, Greenwood et al., 1979, Hosking and Wallis, 1987]
- Paramètres des lois GEV et GPD facilement exprimés comme des fonctions des PWM [de Haan and Ferreira, 2007]

  → Par ex., si $X \sim \mathrm{GEV}_\gamma$, $\gamma < 1$, alors

$$\frac{3\mathbb{E}[XG_\gamma^2(X)] - \mathbb{E}[X]}{2\mathbb{E}[XG_\gamma(X)] - \mathbb{E}[X]} = \frac{3^\gamma - 1}{2^\gamma - 1}$$

  → Méthode d'estimation rapide et efficace

Estimateurs des moments pondérés

- **Lien étroit entre les PWM et les statistiques d'ordre**

$$\theta_{k:m} := \mathbb{E}\left[X_{(k:m)}\right] = (m-k+1)\binom{m}{m-k+1}\mathbb{E}\left[XF(X)^{m-k}\overline{F}(X)^{k-1}\right], \text{ pour } 1 \leq k \leq m$$

$X_{(k:m)}$ est la k^{e} plus grande statistique d'ordre dans un échantillon de taille m

$\Rightarrow$ Les estimateurs des PWM peuvent être exprimés comme des fonctions des statistiques d'ordre.

- Estimateur classique de $\theta_{k:m}$ = combinaison linéaire des statistiques d'ordre

$$\frac{1}{\binom{n}{m}}\sum_{i=1}^{n}\binom{n-i}{m-k}\binom{i-1}{k-1}X_{(n-i+1:n)}$$

- Autre choix naturel pour estimer : utiliser l'estimateur sans biais

$$\frac{1}{\binom{n}{m}}\sum_{1\leq i_1<\cdots<i_m\leq n}\Psi_k(X_{i_1},\ldots,X_{i_m})$$

où $\Psi_k(X_{i_1},\ldots,X_{i_m})$ correspond à la k^{e} plus grande statistique d'ordre de $(X_{i_1},\ldots,X_{i_m})$.

Estimateurs des moments pondérés

Obtenir des bornes non asymptotiques pour les estimateurs PWM sous des conditions de moments minimales

- Utilisation des outils de la théorie de la concentration
  → Les estimateurs des PWM sont liés aux statistiques d'ordre
  → Reprendre les techniques de Boucheron and Thomas [2012, 2015]
- MAIS celles-ci imposent l'existence de moments exponentiels [Boucheron et al., 2013]
  → Hypothèse non satisfaite par les lois à queues lourdes

- De plus, les difficultés liées à la présence de valeurs aberrantes peu abordées en théorie des valeurs extrêmes
  [Hubert et al., 2008, Dupuis and Victoria-Feser, 2006, Brazauskas and Serfling, 2006, Bhattacharya and Beirlant, 2019, Bhattacharya et al., 2019]
  → Résultats essentiellement asymptotiques

But

Proposer un estimateur PWM robuste et pour lequel les inégalités de concentration peuvent être obtenue sous l'existence d'un moment du second ordre

Un nouvel estimateur *median-of-means* de $\theta_{k:m}$

- Cadre *median-of-means* inspiré de Devroye et al. [2016], Lecué and Lerasle [2019], Joly and Lugosi [2016], Lecué and Lerasle [2020]

1. Choisir un niveau de significativité $\delta \in [e^{-\lfloor n/m \rfloor}, 1[$.
2. Diviser $X_1, \ldots, X_n$ en $K = \lceil \ln(1/\delta) \rceil$ blocs disjoints $B_1, \ldots, B_K$, chacun de taille

$$|B_j| \geq \left\lfloor \frac{n}{K} \right\rfloor \geq m.$$

3. Sur chaque bloc j, construire la U-statistique

$$\widehat{\theta}^{(j)}_{k:m} = \frac{1}{\binom{|B_j|}{m}} \sum_{A \subset B_j \ |A|=m} \Psi_k(X_A),$$

4. Calculer la médiane entre les blocs, soit

Estimateur PWM *median-of-means* de $\theta_{k:m}$

$$\widehat{\theta}_{k:m} = \text{median}\left(\widehat{\theta}^{(1)}_{k:m}, \ldots, \widehat{\theta}^{(K)}_{k:m}\right)$$

Inégalités sous-gaussiennes pour $\widehat{\theta}_{k:m}$

Propositions 1 et 2 [Ben-Hamou, Naveau, and Thomas, 2023]

Soit $X_1, \dots, X_n$ un échantillon i.i.d. à valeurs réelles, et un entier positif $m \in \{1, \dots, n\}$. Alors, pour tout $\delta \in [e^{-\lfloor n/m \rfloor}, 1)$, et $K = \lceil \ln(1/\delta) \rceil$

1. $$\mathbb{P}\left(|\widehat{\theta}_{k:m} - \theta_{k:m}| > 2e\sqrt{\frac{2m v_m \lceil \ln(1/\delta) \rceil}{n}} \right) \leq \delta \,,$$

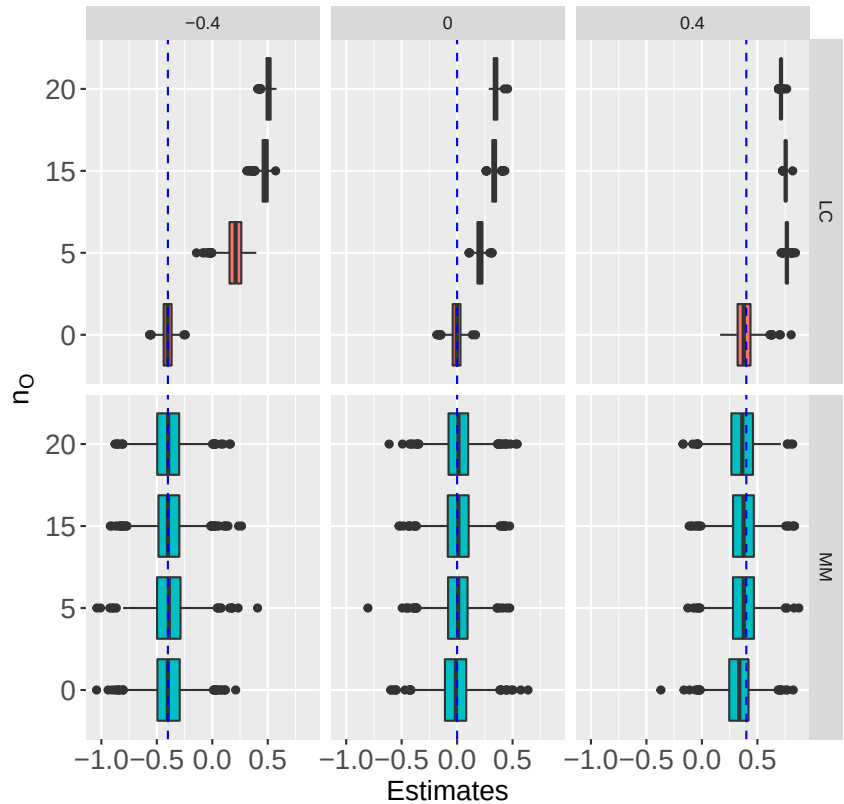
   avec $v_m = \mathrm{Var}[X_{(k:m)}]$
2. Si l'échantillon contient un nombre de valeurs aberrantes $\leq K/4$, alors

$$\mathbb{P}_{\text{Contamin}}\left(|\widehat{\theta}_{k:m} - \theta_{k:m}| > \frac{16e^2}{3\sqrt{3}}\sqrt{\frac{2m v_m \lceil \ln(1/\delta) \rceil}{n}} \right) \leq \delta \,,$$

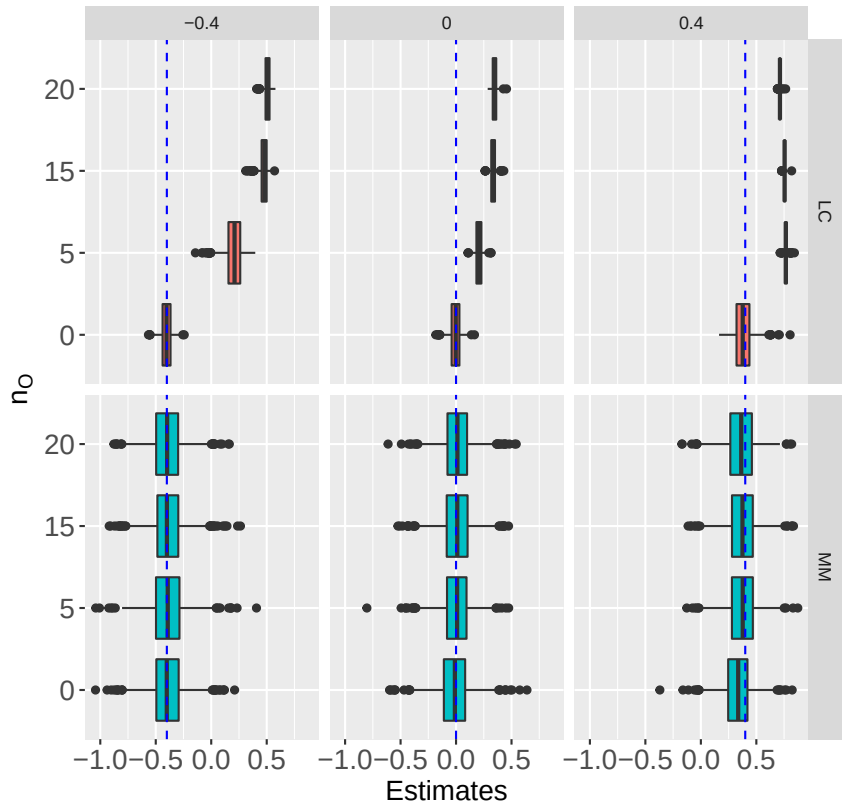
- On peut également montrer que les estimateurs PWM sont sous-gamma avec le bon facteur de variance.
- On en déduit des résultats analogues pour les estimateurs PWM des paramètres de la loi GEV [Ben-Hamou, Naveau, and Thomas, 2023, Proposition 3].

Illustration numérique de la robustesse

- Simulation de 1 000 échantillons de taille 200 de loi GEV_γ, avec $\gamma = -0.4, 0, 0.4$ contenant $n_{\mathscr{O}}$ valeurs aberrantes



γ



$q_\gamma(0.95)$

Plan

Méthodes d'apprentissage statistique en gestion du risque

- La théorie des valeurs extrêmes joue un rôle essentiel en gestion du risque
  → Développement de méthodes statistiques et d'apprentissage statistique
  → Problématique commune : prédiction d'événements rares
  → Classification binaire en présence de données déséquilibrées

Contributions

1. Méthodologie de prédiction du coût des événements de sécheresse en comparant différentes méthodes d'apprentissage statistique
   [Heranval, Lopez, and Thomas, 2022]
2. Détection d'anomalies dans les séries financières par une ACP combinée à un réseau de neurones
   [Crépey, Lehdili, Madhar, and Thomas, 2022]
3. Prédiction de la survenue d'événements extrêmes en santé publique
   [Thomas, Lemaitre, Wilson, Viboud, Yordanov, Wackernagel, and Carrat, 2016]
4. Prédiction en temps réel de la survenue d'une épidémie de grippe extrême
   [Thomas and Rootzén, 2022]
5. Étude asymptotique du comportement de produits d'assurance paramétrique dans le cadre de sinistres extrêmes
   [Lopez and Thomas, 2023]

Méthodes d'apprentissage statistique en gestion du risque

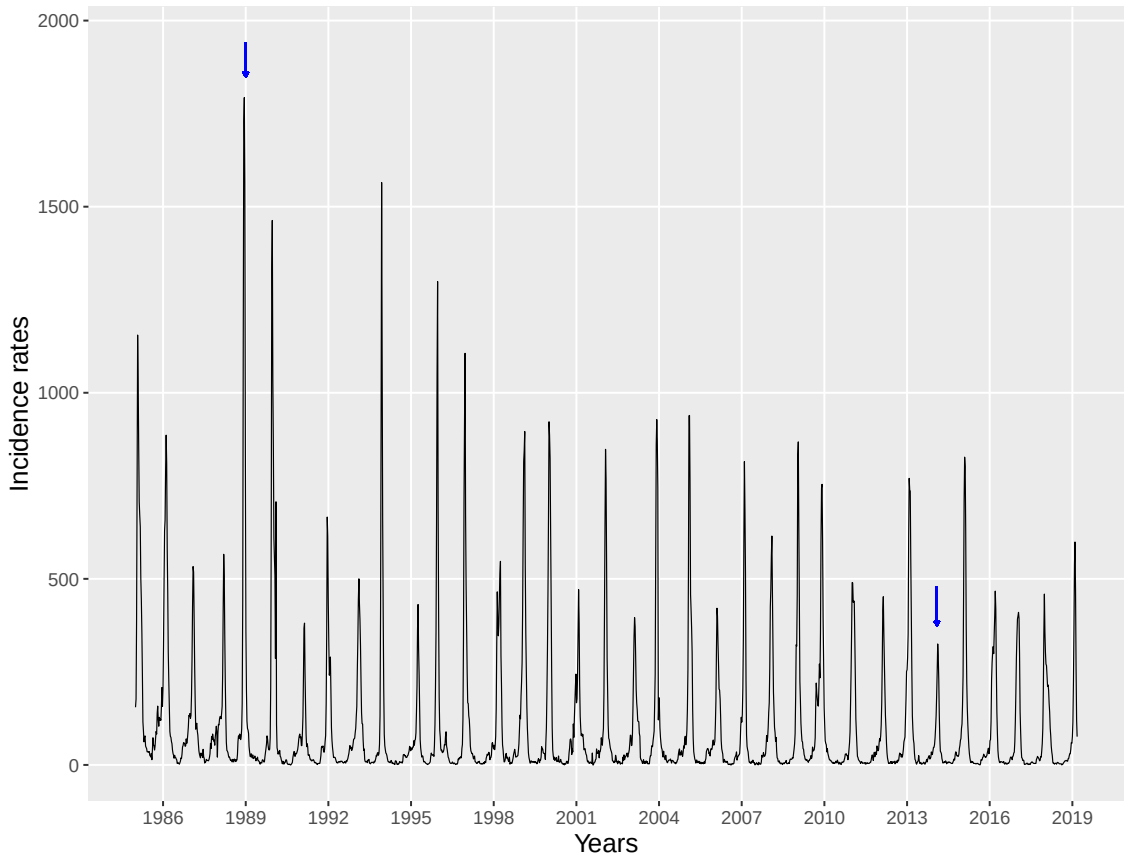
- La théorie des valeurs extrêmes joue un rôle essentiel en gestion du risque
  → Développement de méthodes statistiques et d'apprentissage statistique
  → Problématique commune : prédiction d'événements rares
  → Classification binaire en présence de données déséquilibrées

Contributions

1. Méthodologie de prédiction du coût des événements de sécheresse en comparant différentes méthodes d'apprentissage statistique
   [Heranval, Lopez, and Thomas, 2022]
2. Détection d'anomalies dans les séries financières par une ACP combinée à un réseau de neurones
   [Crépey, Lehdili, Madhar, and Thomas, 2022]
3. Prédiction de la survenue d'événements extrêmes en santé publique
   [Thomas, Lemaitre, Wilson, Viboud, Yordanov, Wackernagel, and Carrat, 2016]
4. Prédiction en temps réel de la survenue d'une épidémie de grippe extrême
   [Thomas and Rootzén, 2022]
5. Étude asymptotique du comportement de produits d'assurance paramétrique dans le cadre de sinistres extrêmes
   [Lopez and Thomas, 2023]

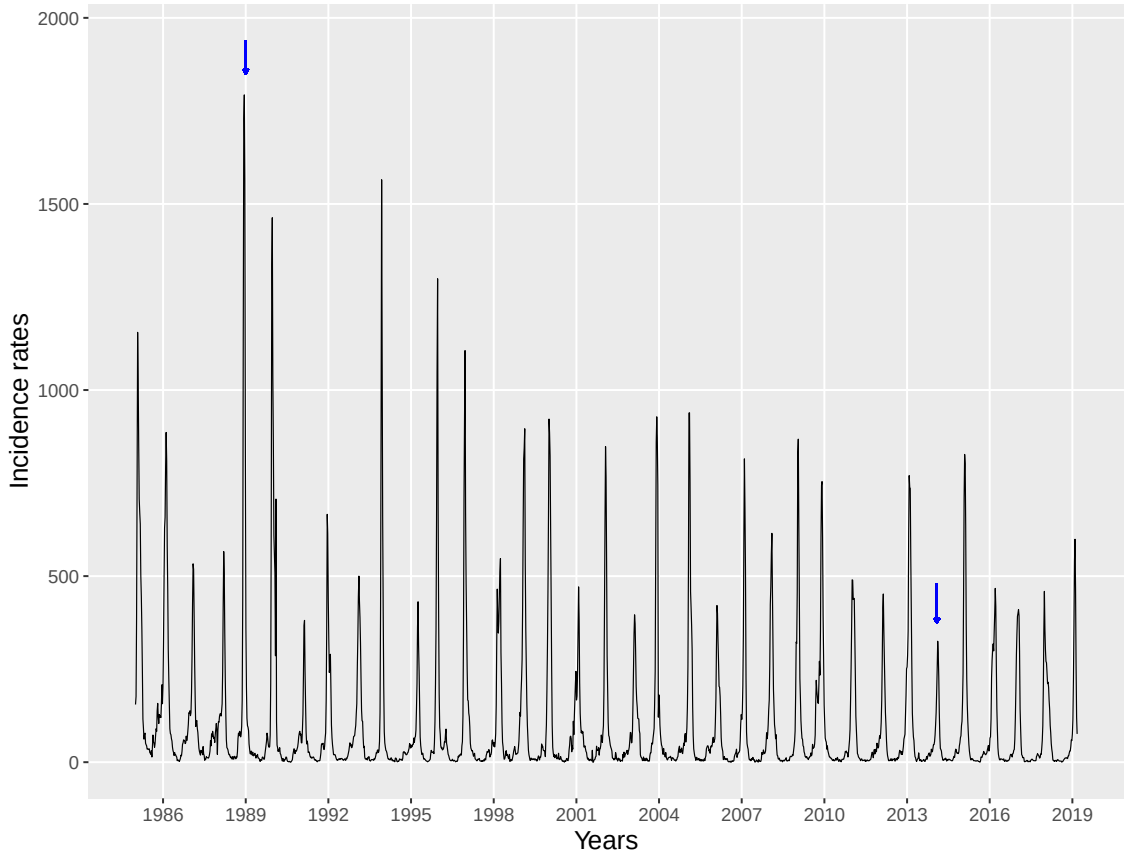
Syndromes grippaux (ILI pour *Influenza like illness*)

- Fièvre supérieure à 39°C, douleurs musculaires et symptômes respiratoires
- Bonne approximation de l'incidence de la grippe
- Réseau Sentinelles : taux d'incidence hebdomadaire des ILI pour 100 000 en France entre janvier 1985 et février 2019 [Réseau Sentinelles, 2019]
  → 35 épidémies, dont 33 extrêmes



Objectif

- Prédire 2019 à partir des données de 1985 et 2018



Objectif

Prédire en temps réel des taux d'incidence élevés des ILI pour une épidémie en cours

Cadre mathématique

Notations

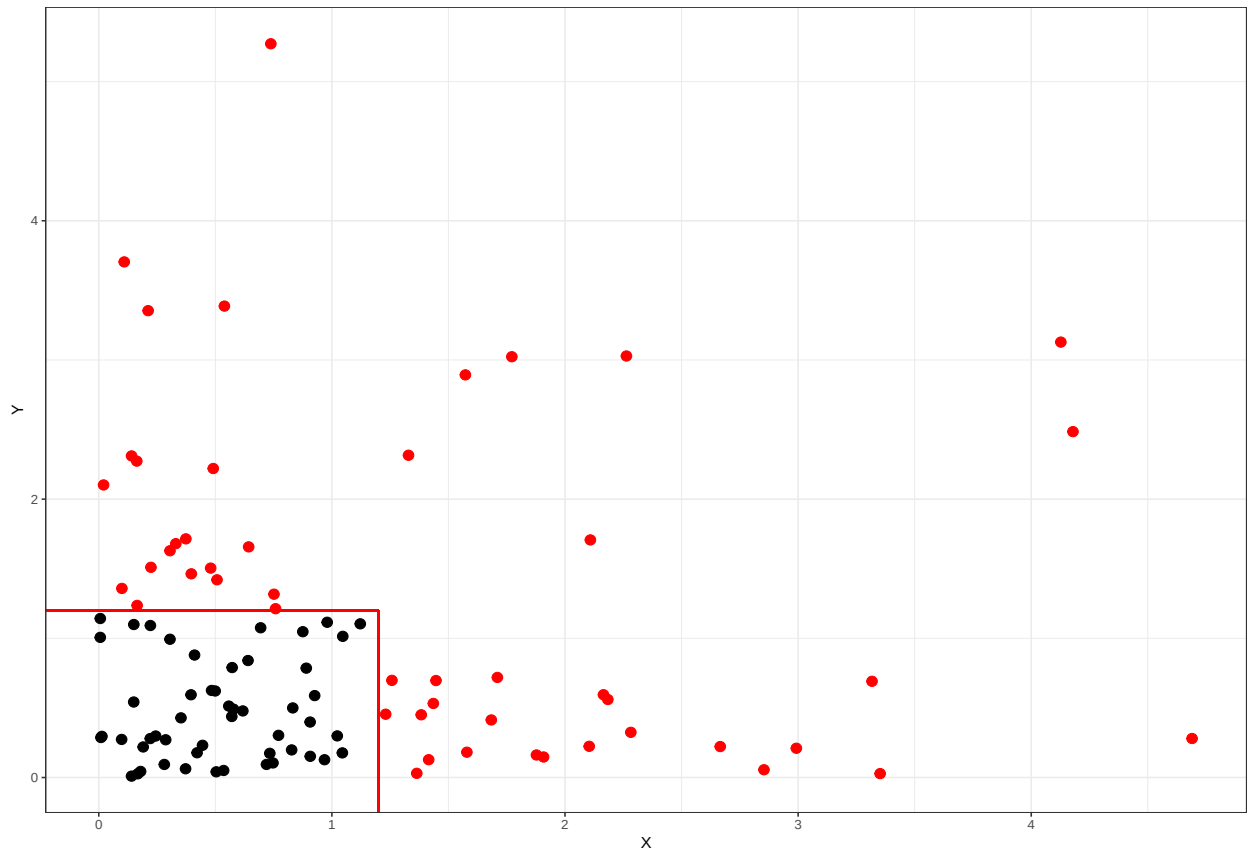
- Y_1 = taux d'incidence de la première semaine de l'épidémie
- Y_2 = taux d'incidence de la deuxième semaine de l'épidémie
- Y_3 = taux d'incidence de la troisième semaine de l'épidémie
- Ainsi de suite jusqu'à la fin de l'épidémie

- **Notre objectif** : Sachant que l'on a observé Y_1 et Y_2, prédire que Y_3 dépasse un niveau (très) élevé v_3, soit

$$\mathbb{P}\left[Y_3 \geq v_3 \mid Y_1 = y_1, Y_2 = y_2\right]$$

Lois de Pareto multivariées

- $\mathbf{Y} = (Y_1, Y_2, Y_3)$ observations
- Choisir des seuils $\mathbf{u} = (u_1, u_2, u_3)$ pour chaque composante
- Événement extrême = AU MOINS un des Y_j a dépassé son seuil u_j (noté $\mathbf{Y} \not\leq \mathbf{u}$)



Lois de Pareto multivariées

- $\mathbf{Y} = (Y_1, Y_2, Y_3)$ observations
- Choisir des seuils $\mathbf{u} = (u_1, u_2, u_3)$ pour chaque composante
- Événement extrême = AU MOINS un des Y_j a dépassé son seuil u_j (noté $\mathbf{Y} \not\leq \mathbf{u}$)

Théorie asymptotique multivariée

- Lorsque $\boldsymbol{u} \to \infty$,
- $\mathbf{Z} = \mathbf{Y} - \mathbf{u} \mid \mathbf{Y} \not\leq \boldsymbol{u}$ converge vers une loi de Pareto généralisée multivariée (MGPD)
- Paramètre d'échelle $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$
- Paramètre de forme $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$

PAS de famille paramétrique de lois limites

$\Rightarrow$ Sélection de modèles à l'aide des résultats de Rootzén et al. [2018], Kiriliouk et al. [2019]

$\Rightarrow$ Calcul de probabilités conditionnelles à partir du modèle sélectionné

Évaluation de la prédiction en dehors du support de l'échantillon

- Simuler 1 500 jeux de données contenant 33 épidémies extrêmes à partir du modèle MGPD pour le taux d'incidence de la semaine 3
- Pour chaque jeu de données, ajuster le modèle MGPD sélectionnée sur les 32 premières épidémies
- Prédire la 33^{e} épidémie avec le modèle ajusté

- Valeurs de v_3 basé sur le maximum observé (= 1 729)

Multiple	0.5	0.75	0.95	1
v_3	864	1 297	1 643	1 729

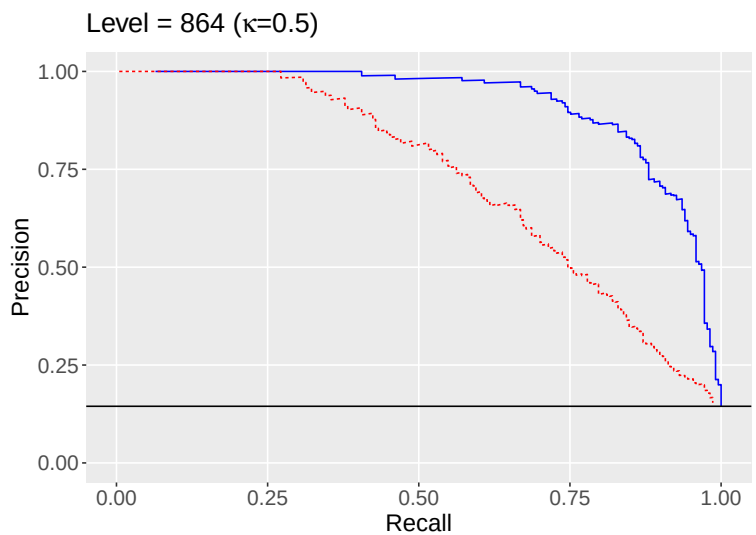
Évaluation de la prédiction en dehors du support de l'échantillon

Courbes de précision-rappel sont plus informatives que les courbes ROC

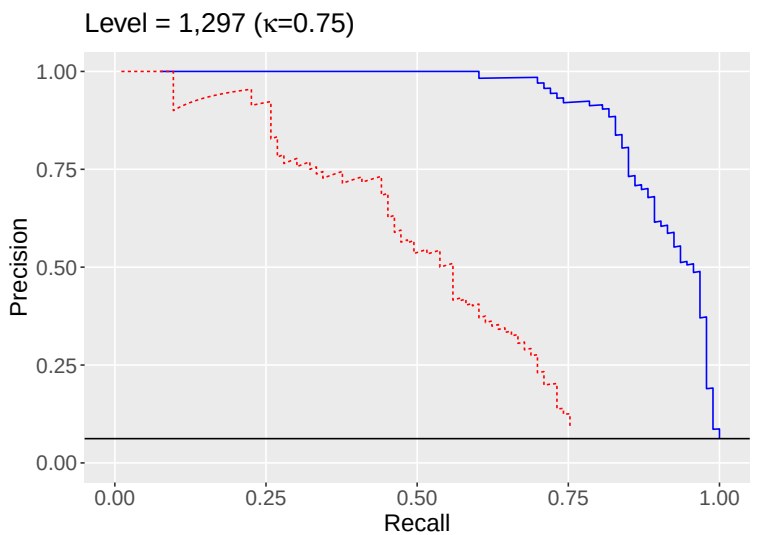
[Saito and Rehmsmeier, 2015]

$$\text{Precision}(p_c) = \frac{\text{vrai positifs}}{\text{vrai positifs} + \text{faux positifs}} \quad \text{et} \quad \text{Rappel}(p_c) = \frac{\text{vrai positifs}}{\text{vrai positifs} + \text{faux négatifs}}$$

p_c = probabilité critique variant entre 0 et 1.



$v_3 = 864$

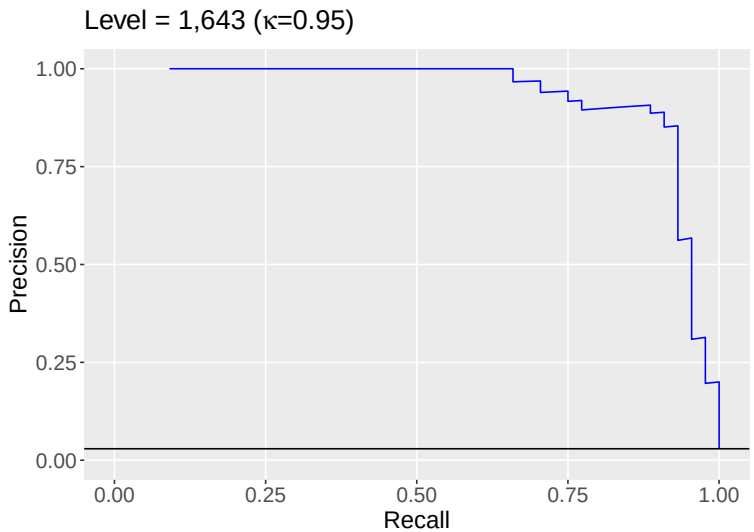


$v_3 = 1\,297$

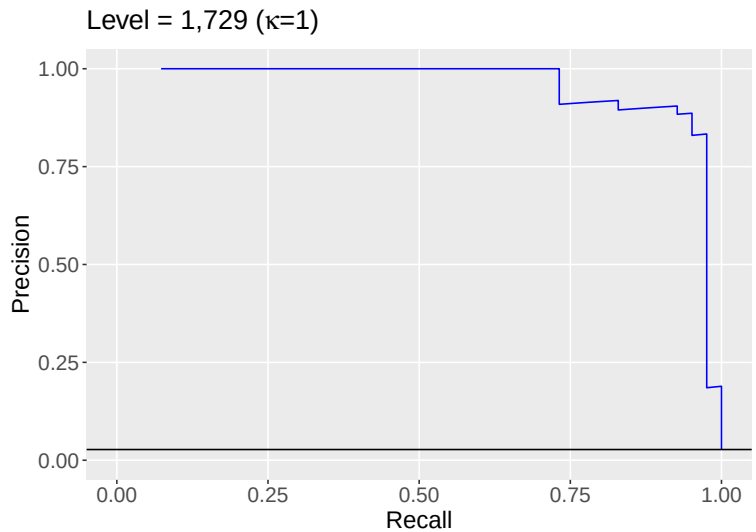
Évaluation de la prédiction en dehors du support de l'échantillon

Courbes de précision-rappel sont plus informatives que les courbes ROC

[Saito and Rehmsmeier, 2015]



$v_3 = 1\,643$



$v_3 = 1\,729$

Plan

Méthodes d'arbres de régression

- Les modèles de régression permettent de comprendre quels facteurs ont un impact sur les risques étudiées
  → Arbres de régression permettent de regrouper les données en sous-groupe ayant un comportement moyen similaire
  → Non adapté à l'étude des risques extrêmes
  → Nécessité de comprendre les facteurs de risque qui pèsent sur la queue de distribution

Contributions

1. Méthodologie pour identifier les facteurs responsables du déclin de la grippe pendant la pandémie COVID-19
   [Bonacina, Boëlle, Colizza, Lopez, Thomas, and Poletto, 2023]
2. Adaptation de l'algorithme CART pour l'analyse des événements cyber extrêmes : arbres de régression Pareto généralisés
   [Farkas, Lopez, and Thomas, 2021b]
3. Développement de résultats non asymptotiques pour assurer la consistance des arbres de régression de Pareto généralisés
   [Farkas, Heranval, Lopez, and Thomas, 2021a]

Méthodes d'arbres de régression

- Les modèles de régression permettent de comprendre quels facteurs ont un impact sur les risques étudiées
  → Arbres de régression permettent de regrouper les données en sous-groupe ayant un comportement moyen similaire
  → Non adapté à l'étude des risques extrêmes
  → Nécessité de comprendre les facteurs de risque qui pèsent sur la queue de distribution

Contributions

1. Méthodologie pour identifier les facteurs responsables du déclin de la grippe pendant la pandémie COVID-19
   [Bonacina, Boëlle, Colizza, Lopez, Thomas, and Poletto, 2023]
2. Adaptation de l'algorithme CART pour l'analyse des événements cyber extrêmes : arbres de régression Pareto généralisés
   [Farkas, Lopez, and Thomas, 2021b]
3. Développement de résultats non asymptotiques pour assurer la consistance des arbres de régression de Pareto généralisés
   [Farkas, Heranval, Lopez, and Thomas, 2021a]

Classification And Regression Trees (CART)

Arbres de classification et de régression [Breiman et al., 1984]

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

- Z est une variable à expliquer
- $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ est un ensemble de variables explicatives
- $\boldsymbol{\Theta} \subset \mathbb{R}^k$ représente l'espace des paramètres
- ϕ est une fonction de perte qui dépend de la quantité que l'on souhaite estimer

Fonction de perte

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

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

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

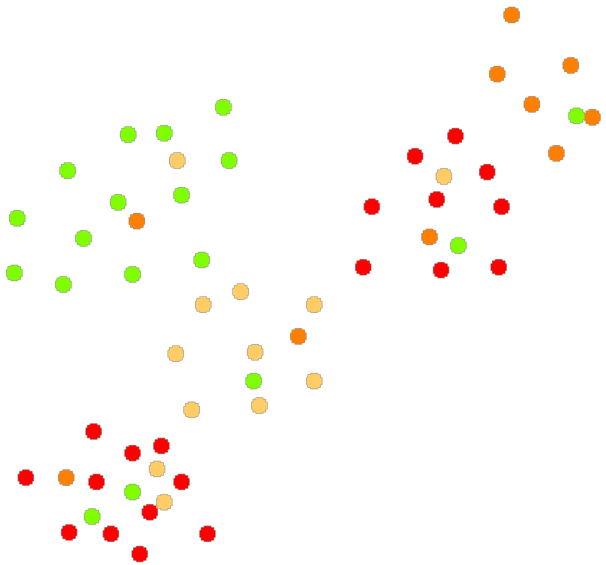
- Perte absolue → "Median regression" (médiane conditionnelle)

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

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

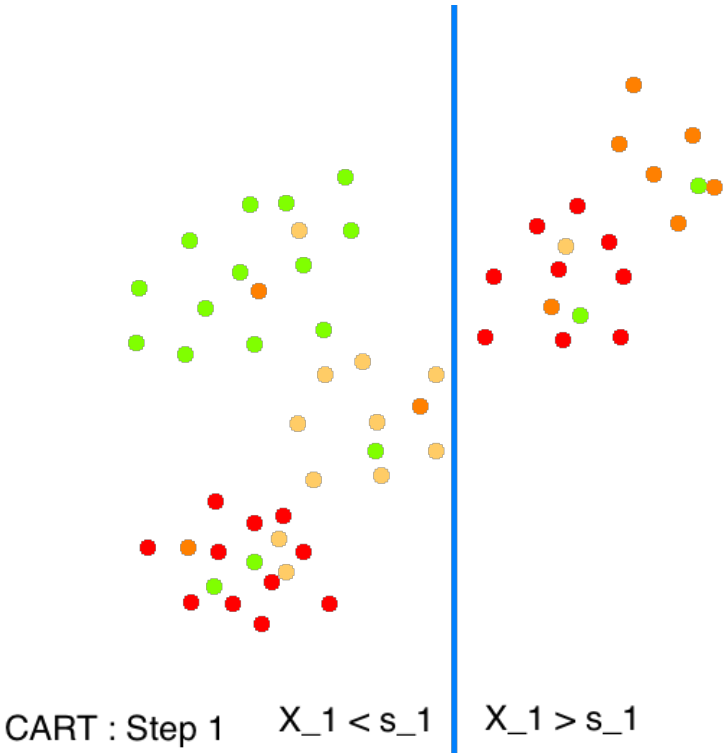
- Perte liée à une log-vraisemblance négative,
  → $\theta^*(\mathbf{x})$ = paramètres de la loi.

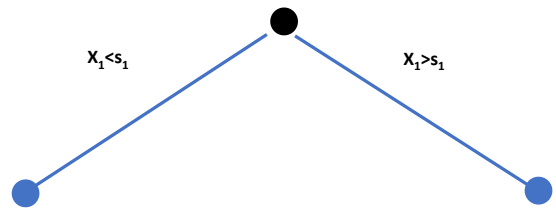
Construction de l'arbre



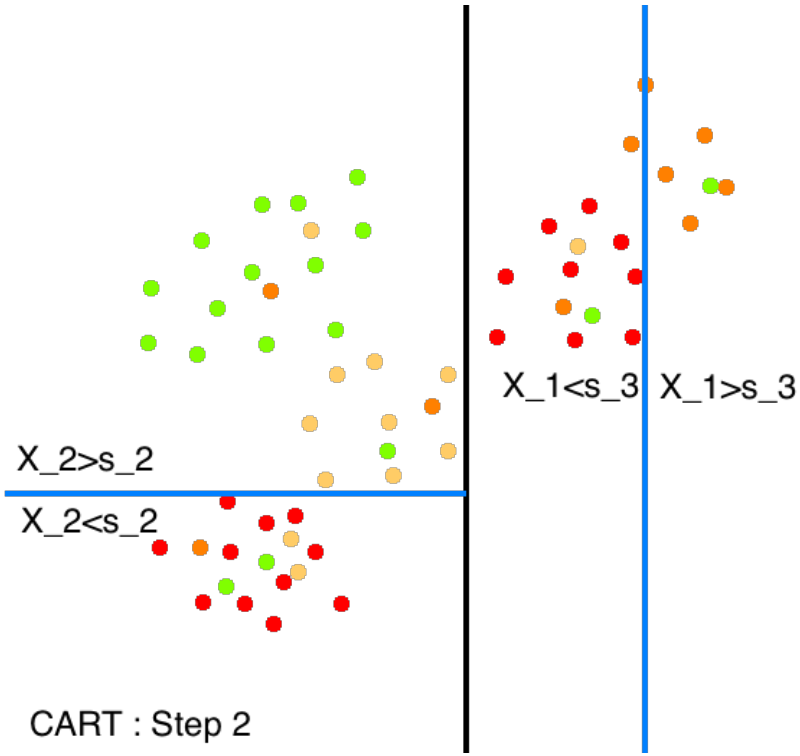
CART : Step 0

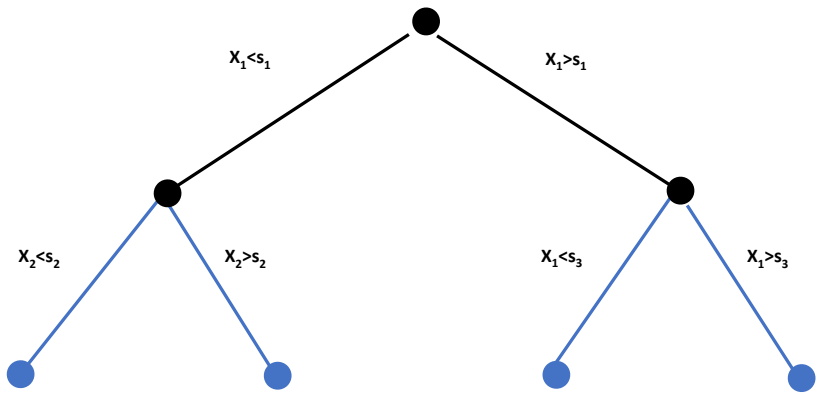
Construction de l'arbre



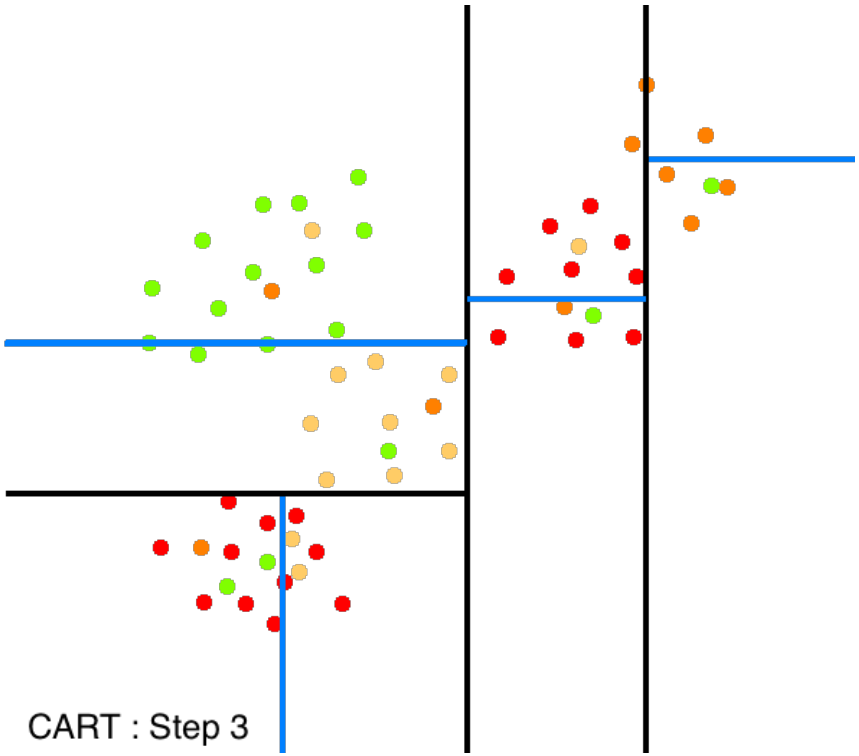


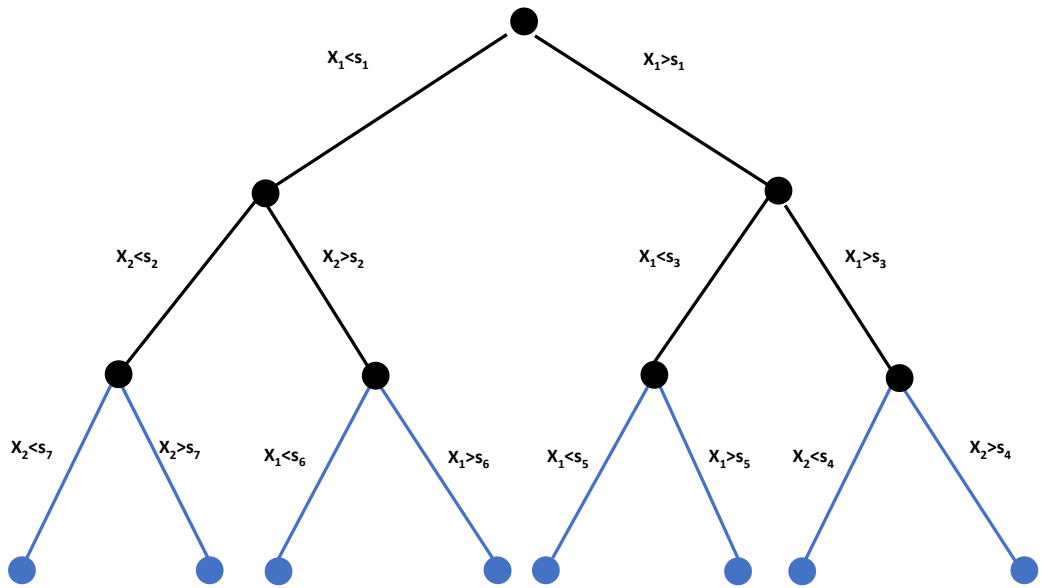
Construction de l'arbre





Construction de l'arbre





De l'arbre à l'estimation des paramètres

- Soit $\widehat{T}_{\max}$ l'arbre maximal obtenu dans la première phase et $K_{\max}$ le nombre de ses feuilles
- $\mathcal{T}_\ell$ la ℓ^{e} feuille pour $\ell = 1, \ldots, K_{\max}$
- **Estimateur $\widehat{\theta}(\mathbf{X})$ de la fonction de régression** $\theta^*(\mathbf{X})$ donné par

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

→ Fonction constante par morceaux

Élagage : sélection de modèle

Extraire de $\widehat{T}_{\max}$ un sous-arbre qui réalise un compromis entre simplicité et bonne adéquation.

- Nombre de feuilles sélectionné

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

  $\lambda > 0$ est choisi par cross-validation

- Arbre sélectionné : $\widehat{T} = \widehat{T}_{\widehat{K}}$

Arbres de régression Pareto généralisés

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

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

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

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

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

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

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

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

  → $\boldsymbol{\theta}^*(\mathbf{x}) = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}[\phi(Y - u, \boldsymbol{\theta}) \mathbf{1}_{Y > u} \mid \mathbf{X} = \mathbf{x}]$

Théorie des valeurs extrêmes et régression

- Cadre de la régression
  - Considère une observation Y de caractéristiques $\mathbf{X}$
  - Choix d'un seuil $u(\mathbf{X})$
  - la loi des excès $Z = Y - u(\mathbf{X}) \mid (\mathbf{X}, Y \geq u(\mathbf{X}))$ converge vers une GPD de paramètres $\sigma_0(\mathbf{X})$ et $\gamma_0(\mathbf{X}) > 0$

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

- But : estimer $\theta_0(\mathbf{x}) = (\sigma_0(\mathbf{x}), \gamma_0(\mathbf{X}))$

- Rappel (méthode PoT) : sélectionner les observations $Y_i \geq u(\mathbf{X}_i)$
- Ici suppose $u(\mathbf{x}) = u \in [u_{\min}, u_{\max}]$
  - $u_{\min}$ tel que $\mathbb{P}(Y \geq u_{\min}) = k_n/n$
  - $u_{\max}$ tel que $\mathbb{P}(Y \geq u_{\max}) = u_0 k_n/n$, pour $u_0 \leq 1$.

- Application de l'algorithme CART avec log-vraisemblance négative GPD aux Z_i

Garanties théoriques

- k_n : nombre moyen de $Y_i \geq u$
- $\widehat{T}_K$: arbre avec K feuilles notées $\mathcal{T}_\ell, \ell = 1, \ldots, K$ avec les paramètres $\widehat{\theta}_\ell^K$
- T_K^* : arbre avec les mêmes feuilles que $\widehat{T}_K$ mais avec les paramètres θ_ℓ^{*K}.

Théorème 1 - Bornes de déviation [Farkas, Heranval, Lopez, and Thomas, 2021a]

$$\mathbb{P}\left(\sup_{u_{\min} \leq u \leq u_{\max}} \| \widehat{T}_K - T_K^* \|_2^2 \geq t \right) \leq 2 \left(\mathrm{e}^{-\frac{C_1 k_n t}{K\beta^2 (\log k_n)^2}} + \mathrm{e}^{-\frac{C_2 k_n t^{1/2}}{K^{1/2}\beta \log k_n}} \right) + \frac{C_3 K}{k_n^{5/2} t^{3/2}},$$

De plus,

$$\mathbb{E}\left[\sup_{u_{\min} \leq u \leq u_{\max}} \| \widehat{T}_K - T_K^* \|_2^2 \right] \leq C_4 \frac{K\beta^2 (\log k_n)^2}{k_n}.$$

Garanties théoriques

- Nombre optimal de feuilles $K^* = \arg\max_{K=1,\dots,K_{\max}} \mathbb{E}\left[\phi(Y-u, \theta^{*K}(\mathbf{X}))\mathbf{1}_{Y>u}\right]$.
- $T^* = T^*_{K^*}$
- $\widehat{T} = \widehat{T}_{\widehat{K}}$ l'arbre sélectionné correspondant

Théorème 2 - Consistance de l'étape d'élagage [Farkas, Heranval, Lopez, and Thomas, 2021a]

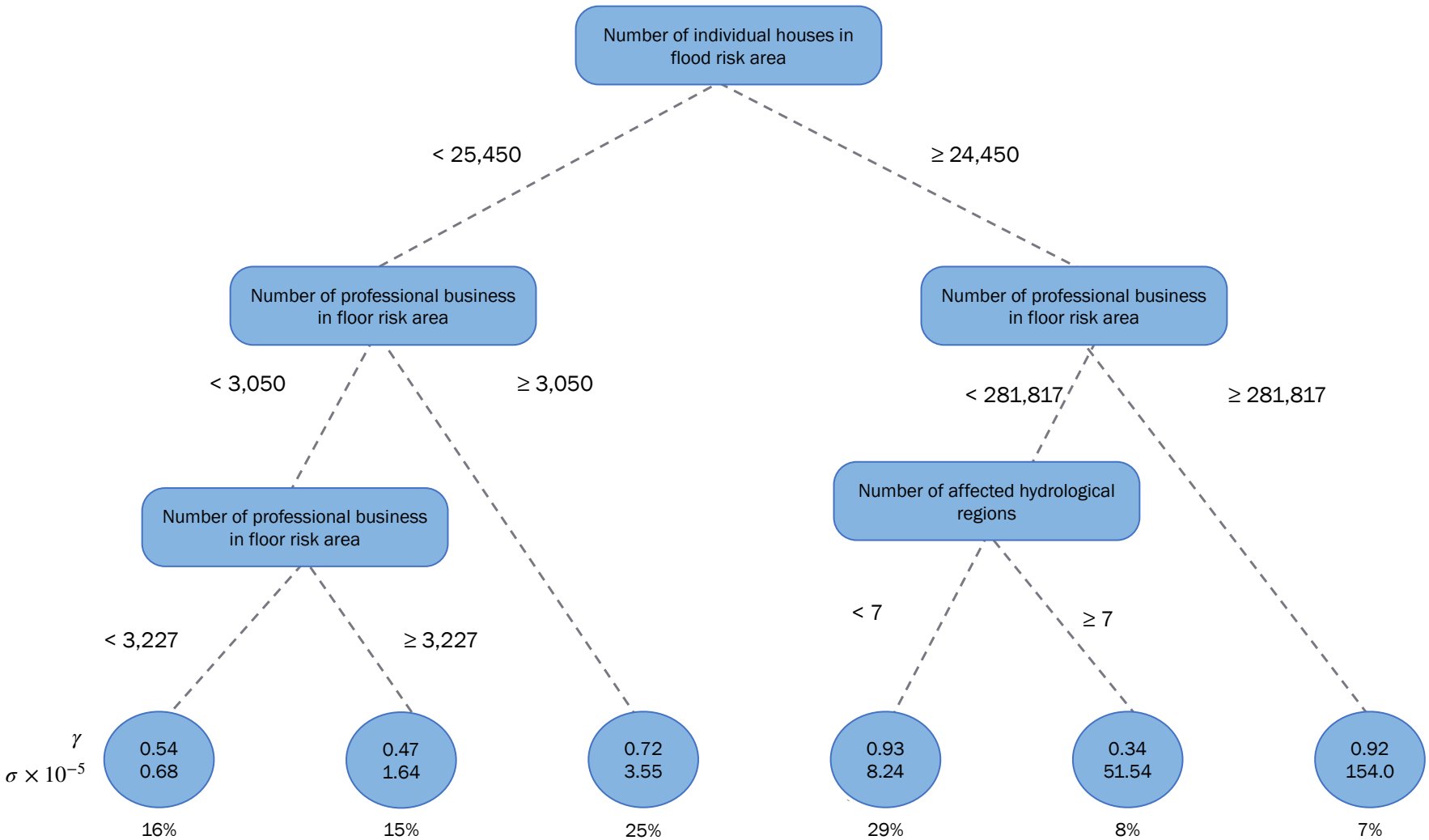
Sous certaines conditions, pour tout $u \in [u_{\min}, u_{\max}]$,

$$\mathbb{E}\left[\|\widehat{T} - T^*\|_2^2\right] \leq \frac{C_5 K^* (\log k_n)^2}{k_n},$$

Prédiction du coût d'un événement inondation

- Base de données SILECC
  - Partenariat avec la MRN (Contrat CIFRE)
  - 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 1 100 événements

Prédiction du coût d'un événement inondation



Perspectives de recherche

- **Choix du seuil u**
  - S'inspirer des techniques de [Boucheron and Thomas, 2015] pour intégrer une procédure automatique du choix du seuil u dans l'algorithme GP CART
  - Remplacer dans l'algorithme GP CART la vraisemblance GPD par la vraisemblance de la loi de Pareto généralisée étendue [Naveau et al., 2016]
- **Structures en réseaux**
  - Comprendre la structure spatio-temporelle de la surmortalité européenne à l'aide de modèles graphiques spatiaux extrêmes de [Engelke and Hitz, 2020]
  - Adapter un modèle log-normal de Poisson proposé par Chiquet et al. [2018, 2021] pour obtenir une structure de réseaux cachés de connexion entre des sociétés pour étudier la contamination d'une attaque cyber
- **Méthodes en gestion du risque**
  - Généralisation de l'approche de simulation proposée dans [Legrand et al., 2023] à une dimension $K > 2$ pour l'estimation des mesures de risque en finance
  - Arbres de régression pour les copules et utilisation en assurance paramétrique

Merci de votre attention

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