

Catastrophes naturelles et autres événements extrêmes : quelques modèles mathématiques à l'œuvre

Anne-Laure Fougères

Institut Camille Jordan, Université Lyon 1

Colloque Inter'actions

19 mai 2014

I. Problématiques associées aux modèles de valeurs extrêmes

II. Théorie des valeurs extrêmes

III. Et en pratique ?

IV. Application : collaboration avec EDF R&D

V. Conclusions

I. Problématiques de type “valeurs extrêmes”

1. Contexte d'une variable unidimensionnelle :

► **Sûreté des installations électriques en bord de mer :**

exposition aux événements climatiques extrêmes.

Objectif : évaluer la probabilité qu'une variable X (la surcote par exemple, ou la vitesse du vent, ou la houle, ou la température...) dépasse un seuil critique

$$\mathbb{P}(X > \text{seuil critique}) ,$$

avec des données tq aucune n'a atteint ce seuil (ou presque).

$$\Rightarrow \text{Estimation empirique} = \frac{\text{nombre d'observations} > \text{seuil critique}}{\text{nombre total d'observations}} = 0$$

► **Autres contextes avec la même problématique :**

- Estimer $\mathbb{P}$ (sortie de piste d'un avion) ;
- Industrie : fatigue des matériaux, corrosion...
ex : Si X = amplitude de vibration de turbines, estimer $\mathbb{P}(X > \text{seuil critique})$...
- Hauteur des digues aux Pays-Bas : la hauteur des digues doit atteindre la **“hauteur atteinte par l'eau en moyenne seulement une fois tous les 10^4 ans”** (niveau-retour à 10^4 -ans).

Dans tous ces cas de figure, on veut :

- **évaluer la probabilité** $\mathbb{P}(X > \ell)$, avec ℓ tel que très peu d'observations soient telles que $X_i > \ell$,

ou de manière équivalente

- **évaluer un quantile extrême** : x_p tq $\mathbb{P}(X > x_p) = 10^{-a}$, alors que la taille d'échantillon est de l'ordre de 10^b , avec $10^b \ll 10^a$.

2. Contexte d'une variable multidimensionnelle :

- **Risque alimentaire** : Exposition globale au risque de **contamination** par un ensemble d'aliments de taille d

Exemple : ochratoxine A (mycotoxine spécialement dangereuse pour la santé, présente dans de nombreux aliments, céréales, abats et charcuteries, vins...).

Notons $C = (C_1, \dots, C_d)$ la consommation d'un individu, et $Q = (Q_1, \dots, Q_d)$ le taux de contamination associé.

On cherche alors $\mathbb{P} \left(\sum_{k=1}^d Q_k C_k > \text{seuil critique} \right)$,
avec des **dépendances** possibles entre variables.

[Bertail, Cléménçon & Tressou (2010), Statistical analysis of a dynamic model for dietary contaminant exposure, *Journal of Biological Dynamics*].

- Une **compagnie d'assurances** possède plusieurs types de contrats : habitation, véhicules, commerces...
En cas de **catastrophe naturelle**, ces différentes branches deviennent **dépendantes**.

Problème : Comment évaluer la probabilité que cette compagnie d'assurances ait à faire face à un sinistre provoquant sa ruine ?

Dans un modèle de risque à temps discret (détails suivent), la **probabilité de ruine** s'écrit

$$\mathbb{P} \left(\max_{1 \leq k \leq d} \sum_{i=1}^k X_i Y_i > x \right), \text{ pour } x \text{ "grand"}.$$

[Fougères & Mercadier (2012), Risk measures and multivariate extensions of Breiman's Theorem, *J. Appl. Probab.*].

► **Modèles de risque à temps discret :**

$$R_0 = x, \quad R_i = R_{i-1}(1 + \rho_i) - X_i, \quad i \in \mathbb{N}^*, \quad (1)$$

où

- $x \geq 0$ représente le **capital initial** d'une compagnie d'assurance,
- $\rho_i \in (-1, \infty)$ est le **taux d'intérêt aléatoire** pour la i ème période,
- $X_i =$ **montant total de sinistres – revenus des primes (aléatoire)**,

de sorte que R_i représente le **surplus** calculé pour la période de 0 à i .

► **Probabilité de ruine** à l'horizon de temps fini $[0, d]$:

$$\psi(x, d) = \mathbb{P} \left(\min_{1 \leq k \leq d} R_k < 0 \mid R_0 = x \right) .$$

Introduisons le **“facteur d’actualisation”** pour la période i à 0

$$Y_i = \prod_{j=1}^i (1 + \rho_j)^{-1}, \quad i \in \mathbb{N}^*,$$

on peut réécrire

- ▶ le **modèle de risque à temps discret** (1) comme

$$R_0 = x, \quad R_i = Y_i^{-1} \left(x - \sum_{j=1}^i X_j Y_j \right), \quad i \in \mathbb{N}^* ;$$

- ▶ la **probabilité de ruine** comme

$$\psi(x, d) = \mathbb{P} \left(\max_{1 \leq k \leq d} \sum_{i=1}^k X_i Y_i > x \right) .$$

► Prédiction des inondations côtières :

Conditions extrêmes en mer	conjuguent	grandes hauteurs de vagues HmO	et	grands niveaux de surcote SWL
----------------------------------	-------------------	---	-----------	--

Données :

828 tempêtes observées sur 13 ans au large des côtes hollandaises
(localisation proche de Petten)

Problème : Protection d'une digue

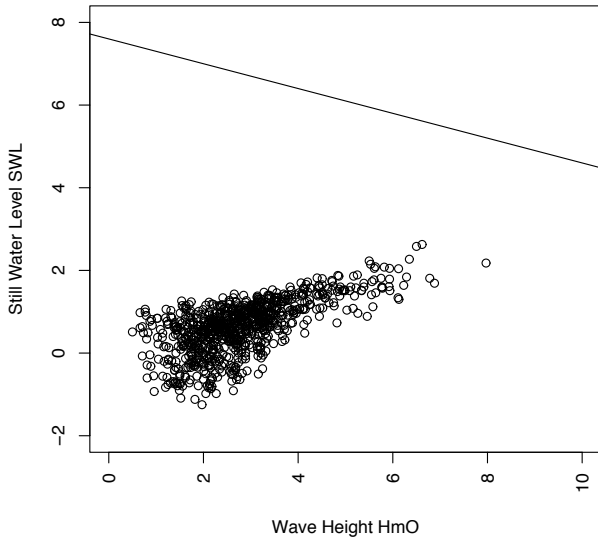
Concentration sur une variable de structure

$$\Delta(\mathbf{X}, \nu) = 0.3 \text{ HmO} + \text{SWL} - \nu, \quad \text{où } \mathbf{X} = (\text{HmO}, \text{SWL})$$

Objectif : Evaluation de $P(\mathbf{X} \in A_\nu)$,

$$\text{où } A_\nu = \{\mathbf{x} \in \mathbb{R}^2 : \Delta(\mathbf{x}, \nu) > 0\}.$$

**Wave Height (H_{m0}) and Still Water Level (SWL)
recorded during 828 storm events for the Dutch coast**



Références sur ce problème :

Différentes modélisations effectuées dans le cadre d'un projet de l'Union Européenne (*Neptune* 1995–1997) :

- ▶ De Haan & de Ronde (1998)
- ▶ Bruun & Tawn (1998)
- ▶ Draisma *et al.* (2004)
- ▶ De Haan & Ferreira (2006, chap. 8)

Important :

Même si la variable d'intérêt est une variable de structure *unidimensionnelle* (par exemple $\Delta(\mathbf{X}, \nu)$), prendre en compte la **dépendance** entre les composantes $X_1, \dots, X_d$ du vecteur est **crucial**.

Conclusions de cette introduction:

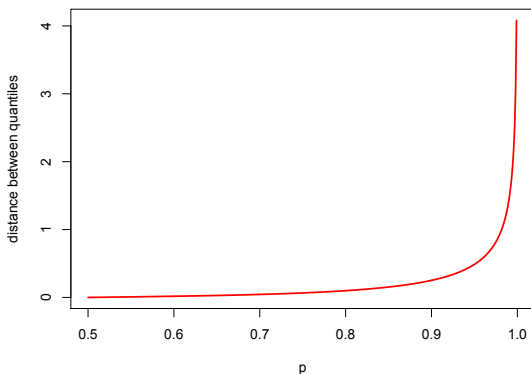
- De nombreux exemples existent dans des contextes uni- ou multi-dimensionnels, où l'on veut
 - ▶ connaître la **probabilité** qu'un vecteur aléatoire $\mathbf{X}$ soit dans un **ensemble extrême** A , tel que typiquement aucune observation $\mathbf{X}^i (i = 1, \dots, n)$ ne tombe dans A .
 - ▶ connaître un **quantile** de petit ordre, ou une région associée à une petite probabilité.
- L'estimation empirique ne peut être utile **sans modélisation supplémentaire**.
- Prendre en compte la **dépendance** entre les composantes de $\mathbf{X}$ est **crucial**.

Objectif : obtenir des **modèles** donnant des réponses pertinentes à ces questions.

Remarque

Attention ! un modèle (paramétrique ou non paramétrique) s'ajustant bien à *l'ensemble des données* peut être un mauvais ajustement *pour la queue* de la distribution.

Exemple : distance entre les quantiles d'ordre p d'une fonction de répartition gaussienne F et d'une fonction de répartition de Student G à 4 degrés de liberté. Graphe de $p \mapsto F^{-1}(p) - G^{-1}(p)$:



II. Théorie des valeurs extrêmes

On **observe** :

des réalisations indépendantes et de même loi (i.i.d) $\mathbf{X}^1, \dots, \mathbf{X}^n$ issues d'un vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_d)$ à valeurs dans $\mathbb{R}^d$.

On **souhaite** :

- connaître la **probabilité** que $\mathbf{X}$ appartienne à un **ensemble "extrême"** A , extrême au sens où typiquement aucune observation $\mathbf{X}^i (i = 1, \dots, n)$ ne tombe dans A :

$$\mathbb{P}(\mathbf{X} \in A) = o(1/n).$$

- connaître un **quantile** d'ordre plus petit que $1/n$ (ou plus grand que $1 - 1/n$), ou une région associée à une petite proba.
- connaître la **probabilité conditionnelle** que $\mathbf{X}$ appartienne à un ensemble A , sachant que l'une des composantes du vecteur est extrême : $\mathbb{P}(\mathbf{X} \in A \mid X_k > x)$ pour x "grand".

Absence d'observations $\Rightarrow$ nécessité de se donner un **modèle**.

Théorie des valeurs extrêmes :

1. Hypothèse d'existence d'un domaine d'attraction univarié

Soit $X_1, \dots, X_n$ i.i.d. de fonction de répartition $F(\cdot) = \mathbb{P}(X_i \leq \cdot)$.

On **suppose** que

$$\mathbb{P}\left(\frac{\max_i X_i - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \rightarrow G(x),$$

quand $n \rightarrow \infty$, avec G non dégénérée (**attracteur** de F).

Alors **les limites possibles G sont de la forme**, pour $\mu, \xi \in \mathbb{R}$, et $\sigma > 0$:

$$G(x) = \exp \left\{ - \left(1 + \xi \frac{x - \mu}{\sigma} \right)_+^{-1/\xi} \right\},$$

Lois de valeurs extrêmes généralisées (GEV).

Quelles sont ces lois ?

Elles diffèrent principalement selon le signe de ξ .

Notons $G_\xi(x) = \exp\{-(1 + \xi x)_+^{-1/\xi}\}$.

1. Si $\xi > 0$: alors $G_\xi(x) < 1 \quad \forall x$. Et l'on a au voisinage de ∞ :

$$1 - G_\xi(x) \sim \frac{1}{(\xi x)^{1/\xi}}.$$

cq : les moments $\mathbb{E}|X|^p$ d'ordre $p \geq 1/\xi$ n'existent pas.

2. Si $\xi = 0$: alors on a au voisinage de ∞ :

$$1 - G_0(x) \sim e^{-x}.$$

cq : tous les moments existent.

3. Si $\xi < 0$: alors $G_\xi(-1/\xi) = 1$ borne sup du support finie.
Et l'on a au voisinage de 0 :

$$1 - G_\xi\left(-\frac{1}{\xi} - x\right) \sim (-\xi x)^{-1/\xi}.$$

Un ingrédient clé : **Théorème de convergence vers les types** :

Soit $A, B, Z_n (n \in \mathbb{N})$ des variables aléatoires, et $a_n > 0, \alpha_n > 0$ et $b_n, \beta_n \in \mathbb{R}^n$ des constantes. Supposons que

$$a_n^{-1}(Z_n - b_n) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} A.$$

Alors
$$\alpha_n^{-1}(Z_n - \beta_n) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} B \quad (\star)$$

si et seulement si

$$\lim_{n \rightarrow \infty} \frac{a_n}{\alpha_n} = a \in [0, \infty), \quad \lim_{n \rightarrow \infty} \frac{b_n - \beta_n}{\alpha_n} = b \in \mathbb{R}.$$

On a de plus : $(\star) \Rightarrow B = aA + b$ en loi, et a et b sont les uniques constantes pour lesquelles ceci est vrai ;

$(\star) \Rightarrow A$ non dégénérée si et seulement si $a > 0$, auquel cas A et B sont de même type.

Theorème : lois limites pour maxima

Fisher & Tippett (1928), Gnedenko (1943)

Soit $X_1, \dots, X_n$ indépendantes et de même loi.

S'il existe $a_n > 0$, $b_n \in \mathbb{R}$ et G non dégénérée tq

$$a_n^{-1} \left(\max_{i=1, \dots, n} X_i - b_n \right) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} G, \quad (DA)$$

alors G est du même type que l'une de ces 3 lois :

Fréchet : $\Phi_\alpha(x) = \exp \{-x^{-\alpha}\}$, $x > 0$ ($\alpha > 0$)

Weibull : $\Psi_\alpha(x) = \exp \{-(-x)^\alpha\}$, $x \leq 0$ ($\alpha < 0$)

Gumbel: $\Lambda(x) = \exp \{-e^{-x}\}$, $x \in \mathbb{R}$.

Grandes lignes de la démonstration :

(DA) implique que pour tout $t > 0$,

$$F^{[nt]}(a_{[nt]}x + b_{[nt]}) \rightarrow G(x), \quad x \in \mathbb{R}.$$

Mais on a aussi

$$F^{[nt]}(a_n x + b_n) = (F^n(a_n x + b_n))^{[nt]/n} \rightarrow G^t(x), \quad x \in \mathbb{R}.$$

Par le théorème de convergence vers les types, on en déduit qu'il existe des fonctions $\gamma(t) > 0$ et $\delta(t) \in \mathbb{R}$ telles que

$$\lim_{n \rightarrow \infty} \frac{a_n}{a_{[nt]}} = \gamma(t) \in [0, \infty), \quad \lim_{n \rightarrow \infty} \frac{b_n - b_{[nt]}}{a_{[nt]}} = \delta(t),$$

et de plus

$$G^t(x) = G(\gamma(t)x + \delta(t)).$$

$$G^t(x) = G(\gamma(t)x + \delta(t)) .$$

On en déduit que, pour tout $s, t > 0$,

$$\gamma(st) = \gamma(s)\gamma(t), \quad \delta(st) = \gamma(t)\delta(s) + \delta(t) .$$

La solution de ces équations fonctionnelles conduit aux trois types $\Phi_\alpha, \Lambda, \Psi_\alpha$.



L'hypothèse d'existence d'un domaine d'attraction est liée à la notion de **variation régulière**.

Pour la suite de l'exposé,
nous supposerons pour simplifier que $\xi > 0$.

Définition : $U : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ est dite **à variation régulière d'indice $\alpha \in \mathbb{R}$** (notation : $U \in RV_\alpha$) si pour tout $x > 0$,

$$\lim_{t \rightarrow \infty} \frac{U(tx)}{U(t)} = x^\alpha .$$

Théorème :

F est **dans le domaine d'attraction d'une loi GEV avec $\xi > 0$**
si et seulement si **$1 - F$ est à variation régulière d'indice $-1/\xi$**
et $x^* = \sup\{x : F(x) < 1\}$ est infini.

Exemples : lois de Cauchy, Fréchet, Pareto, Student, ...

Conséquence :

On cherche à estimer la probabilité d'un événement extrême, i.e.

$$\mathbb{P}(X > tx), \text{ pour } x > 0 \text{ et pour } t \text{ très grand,}$$

de sorte que l'on n'a pas d'observation dans cette zone d'intérêt.

Sous l'hypothèse d'existence d'un domaine d'attraction, on a

$1 - F \in \text{RV}_{-1/\xi}$, d'où :

$$\mathbb{P}(X > tx) \approx x^{-1/\xi} \mathbb{P}(X > t)$$

et l'on choisit un t de manière à avoir un nombre suffisant d'observations plus grandes que t .

On écrit plus classiquement cette **approximation** un peu différemment, pour t grand :

$$\mathbb{P}(X > t + z\psi(t) | X > t) \approx (1 + \xi z)_+^{-1/\xi}$$

Loi de Pareto généralisée (GPD)
approximant les dépassements d'un seuil.

En pratique :

On choisit alors le seuil t empiriquement (pas toujours facile !) et on **estime** ξ et $\psi(t)$ (par maximum de vraisemblance, par exemple).

2. Hypothèse d'existence d'un domaine d'attraction multivarié

Soit $\mathbf{X}^i = (X_1^i, \dots, X_d^i)$ ($i = 1, \dots, n$) n vecteurs aléatoires i.i.d. de $\mathbb{R}^d$, de f.d.r. F . On suppose que

$$\mathbb{P} \left\{ \mathbf{a}_n^{-1} (\max_i \mathbf{X}^i - \mathbf{b}_n) \leq \mathbf{x} \right\} = F^n(\mathbf{a}_n \mathbf{x} + \mathbf{b}_n) \rightarrow G(\mathbf{x}),$$

quand $n \rightarrow \infty$, avec G f.d.r à marges non dégénérées $G_1, \dots, G_d$.

Sous cette hypothèse (notée **MEVT**), on a pour G :

- ▶ des **lois marginales paramétriques** GEV ;
- ▶ une structure de dépendance s'exprimant en terme d'une **fonction de dépendance caudale** ℓ , tq pour tout $\mathbf{v} > \mathbf{0}$,

$$\begin{aligned} \ell(\mathbf{v}) &= \lim_{t \rightarrow \infty} t \mathbb{P} \left[F_1(X_1) > 1 - \frac{v_1}{t} \text{ ou } \dots \text{ ou } F_d(X_d) > 1 - \frac{v_d}{t} \right] . \\ &= -\log G \left[G_1^{\leftarrow}(e^{-v_1}), \dots, G_d^{\leftarrow}(e^{-v_d}) \right] . \end{aligned}$$

Comment utiliser ce résultat ?

Illustration sur deux exemples simples

Supposons que l'on souhaite estimer $\mathbb{P}(\mathbf{X} \in A)$, où

$\mathbf{X}$ a des marges Pareto standard ($1 - F_j(t) = \frac{1}{t}$),
de f.d.r. jointe F dans le domaine d'attraction de G , et
soit A un ensemble “extrême”.

- Supposons $A = (0, n\mathbf{u}]^c$, avec n “très grand”.

MEVT assure l'approximation suivante :

$$\mathbb{P}(\mathbf{X} \in A) = 1 - F(n\mathbf{u}) \approx -\frac{1}{n} \log G(\mathbf{u}) = \frac{1}{n} \ell \left(\frac{1}{\mathbf{u}} \right) .$$

► Supposons $d = 2$ et $A = (nu_1, \infty) \times (nu_2, \infty)$. Alors :

$$\begin{aligned}\mathbb{P}(\mathbf{X} \in A) &= 1 - F_1(nu_1) - F_2(nu_2) + F(nu_1, nu_2) \\ &= \frac{1}{nu_1} + \frac{1}{nu_2} - \{1 - F(nu_1, nu_2)\} ,\end{aligned}$$

de sorte que, pour n “très grand”,

$$\mathbb{P}(\mathbf{X} \in A) \approx \frac{1}{nu_1} + \frac{1}{nu_2} - \frac{1}{n} \ell\left(\frac{1}{u_1}, \frac{1}{u_2}\right) .$$

Remarque :

En cas d'indépendance asymptotique, on a $\ell(\mathbf{v}) = v_1 + v_2$,
d'où...

$$\mathbb{P}(\mathbf{X} \in A) \approx 0 !$$

Ici, **MEVT** n'aide pas à estimer $\mathbb{P}(\mathbf{X} \in A)$, et il faut faire des hypothèses supplémentaires.

III. Et en pratique ?

Le cadre **MEVT** est bien défini, mais...

s'ajuste-t-il correctement en pratique ?

- ▶ En pratique :
 - dépendance à court terme (temporelle)
 - saisonnalité
 - tendance
- ▶ En théorie : observations indépendantes et de même loi
extensible au cadre stationnaire
i.e. $\text{Loi}(X_1, X_2, \dots, X_k) = \text{Loi}(X_{t+1}, X_{t+2}, \dots, X_{t+k})$
+ hypothèse $D(u_n)$ [dépendance "pas trop longue"]

Conséquence : Trois choix possibles

- ▶ travailler sur des périodes **homogènes** (étude par saison...)
- ▶ travailler après extraction d'une tendance et d'une saisonnalité... extrapolation délicate
- ▶ faire appel à d'autres modèles, prenant en compte la **non stationnarité**.

Dacunha-Castelle, Hoang & Parey (2009, 2010).

**Regardons ce qui peut être fait dans le premier cas,
sous l'hypothèse de stationnarité**

**Regardons ce qui peut être fait dans le premier cas,
sous l'hypothèse de stationnarité**

1. Sous une hypothèse " $D(u_n)$ " (Leadbetter, 1974), on a :
 - les lois limites possibles pour les maxima sont encore les lois GEV ;

Regardons ce qui peut être fait dans le premier cas, sous l'hypothèse de stationnarité

1. Sous une hypothèse “ $D(u_n)$ ” (Leadbetter, 1974), on a :

- les lois limites possibles pour les maxima sont encore les lois GEV ;

Théorème : Soit (X_i) stationnaire, et $(\tilde{X}_i)$ i.i.d. de même loi que (X_i) . Soit $M_n = \max(X_1, \dots, X_n)$ et $\tilde{M}_n = \max(\tilde{X}_1, \dots, \tilde{X}_n)$. Sous certaines hypothèses,

$$\mathbb{P}\{a_n^{-1}(\tilde{M}_n - b_n) \leq y\} \rightarrow G_1(y)$$

pour $a_n > 0$, $b_n \in \mathbb{R}$ et G_1 fdr non dégénérée si et seulement si

$$\mathbb{P}\{a_n^{-1}(M_n - b_n) \leq y\} \rightarrow G_2(y),$$

avec $G_2(y) = G_1^\theta(y)$, pour $\theta \in (0, 1]$ **indice extrême**l.

Regardons ce qui peut être fait dans le premier cas :
sous l'hypothèse de stationnarité

1. Sous une hypothèse " $D(u_n)$ " (Leadbetter, 1974), on a :
 - les lois limites possibles pour les maxima sont encore les lois GEV ;
 - les lois limites possibles pour les dépassements de seuil sont encore les lois GPD.
2. Mais les dépassements de seuil arrivent **par paquets**, de taille moyenne limite $1/\theta$.

Quelles en sont les conséquences ?

La notion de niveau retour devient plus complexe !

Définition : Soit $(X_t)_t$ stationnaire, et T un entier fixé.

Le **niveau-retour** associé à la période T est le niveau x_T tel que, en moyenne, le maximum annuel dépasse x_T une fois en T années.

$$\mathbb{P}(\max_{t \in \text{An}} X_t > x_T) = \frac{1}{T} .$$

Calcul du niveau-retour x_T :

$$\begin{aligned} 1 - \frac{1}{T} &= \mathbb{P} \left(\max_{t \in \text{An}} X_t \leq x_T \right) \simeq \mathbb{P} \left(\max_{t \in \text{An}} \tilde{X}_t \leq x_T \right)^\theta \\ &= \mathbb{P} \left(\tilde{X}_t \leq x_T \right)^{n_0 \cdot \theta} = \mathbb{P} (X_t \leq x_T)^{n_0 \cdot \theta} , \end{aligned}$$

où n_0 représente le nombre d'observations formant un "An", soit $n_0 = 365$ en cas de données journalières. Autrement dit :

x_T est le quantile d'ordre $(1 - 1/T)^{1/(n_0 \cdot \theta)}$ associé à (X_t) .

Une méthode voisine, le “declustering” :
fondée sur la décomposition

$$\mathbb{P} \left(\max_{t \in A_n} X_t \leq x_T \right) \simeq \mathbb{P} \left(\max_{j=1, \dots, c} \max_{t \in \mathcal{C}_j} X_t \leq x_T \right) ,$$

- ▶ c représente le **nombre moyen de clusters par A_n**
- ▶ $\mathcal{C}_j$ représente le j -ème cluster

Soit $Z_j := \max_{t \in \mathcal{C}_j} X_t$, et supposons que les clusters $\mathcal{C}_j$ sont tels que **les Z_j sont i.i.d.** de taille c . Alors

$$\mathbb{P} \left(\max_{t \in A_n} X_t \leq x_T \right) \simeq [\mathbb{P}(Z_j \leq x_T)]^c .$$

Autrement dit, **x_T est le quantile d'ordre $(1 - 1/T)^{1/c}$ des Z_j .**

IV. Application : collaboration avec EDF R&D

- ▶ réalisée en étroite collaboration avec **Cécile Mercadier** (Institut Camille Jordan), dans le cadre d'un contrat de recherche avec EdF R&D (MADONE).
- ▶ a bénéficié du travail de **Juan Juan Cai** (post-doc), **Gilles Nicolet** (stagiaire) et **Quentin Sebillé** (stagiaire, actuellement doctorant à l'Institut Camille Jordan).

Partenaires EDF : Pietro Bernadara (LNHE), Anne Dutfoy (MRI), Marie Gallois (MRI), Nicolas Malleron (LNHE), Sylvie Parey (MFEE) et Nicolas Roche (LNHE).

► **Contraintes de confidentialité :**

⇒ on ne décrira pas les variables environnementales utilisées.

⇒ on suppose que les données, après prétraitement éventuel, forment une série stationnaire (X_t, Y_t) .

► **Objectif de l'une des études :**

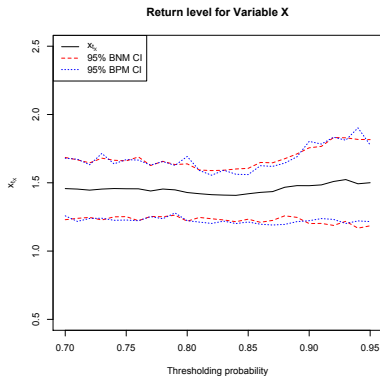
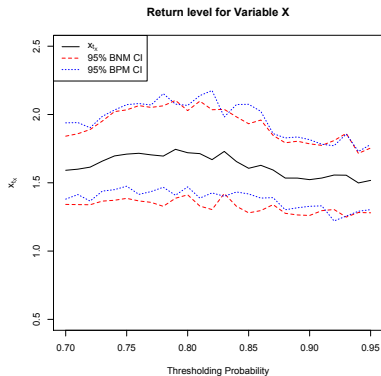
estimer une **probabilité de défaillance** de la forme

$$\mathbb{P}(X_t > x_{t_X}, Y_t > y_{t_Y}) ,$$

- où
- $x_{t_X} =$ *niveau-retour à t_X années de X (t_X fixé)*
 - $y_{t_Y} =$ *niveau-retour à t_Y années de Y (t_Y fixé)*

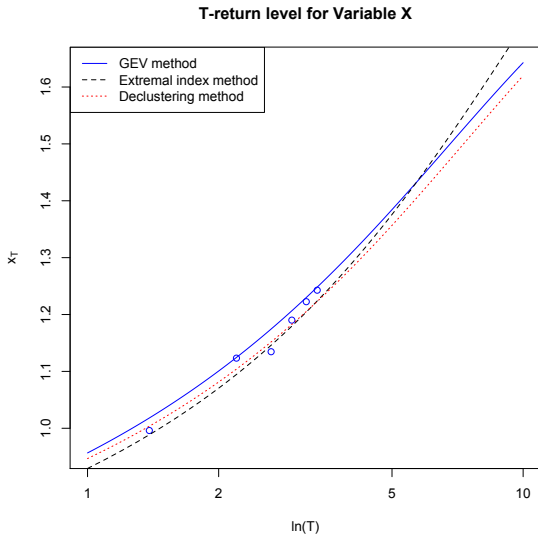
[Cai, Fougères & Mercadier (2013), Environmental data: multivariate Extreme Value Theory in practice. *Journal de la Société Française de Statistique*.]

Première partie de l'étude : estimer les niveaux-retour



Estimation du **niveau-retour** x_{t_x} , via la méthode de l'indice extrême (à gauche) et le “declustering” (à droite), en fonction de la probabilité de dépassement du seuil.

Estimation du niveau-retour x_T en fonction de $\log(T)$
via les méthodes GEV, indice extrémal et “declustering”.



Deuxième partie de l'étude : estimer la proba de défaillance

$$\mathbb{P}(X_t > x_{t_X}, Y_t > y_{t_Y})$$

où $x_{t_X} = \text{niveau-retour à } t_X \text{ années de } X$ et $y_{t_Y} = \text{niveau-retour à } t_Y \text{ années de } Y$, pour t_X et t_Y fixés.

Deux modélisations ont été implémentées :

- ▶ variation régulière bivariable [BRV]: Draisma *et al.* (2004);
- ▶ modèle conditionnel [CEV]: Heffernan & Tawn (2004).

Ces deux méthodes autorisent l'indépendance asymptotique.

Sans préciser les détails, évoquons succinctement le modèle **BRV**.

- Hypothèse **MEVT**: [+ comportement des marges]

$$\lim_{t \rightarrow \infty} t \mathbb{P} \left[1 - F_X(X) < \frac{x}{t} \text{ ou } 1 - F_Y(Y) < \frac{y}{t} \right] = \ell(x, y)$$

- On suppose de plus l'existence pour $x, y > 0$ de

$$\lim_{t \rightarrow \infty} \frac{\mathbb{P}(1 - F_X(X) < x/t \text{ et } 1 - F_Y(Y) < y/t)}{\mathbb{P}(1 - F_X(X) < 1/t \text{ et } 1 - F_Y(Y) < 1/t)} =: c(x, y)$$

Alors c est homogène d'ordre $1/\eta$ pour $\eta \in (0, 1]$.

Le cas $\eta < 1$ correspond à l'indépendance asymptotique.

L'**estimation de η** est un point clé, et le **choix de seuil** associé est **délicat**.

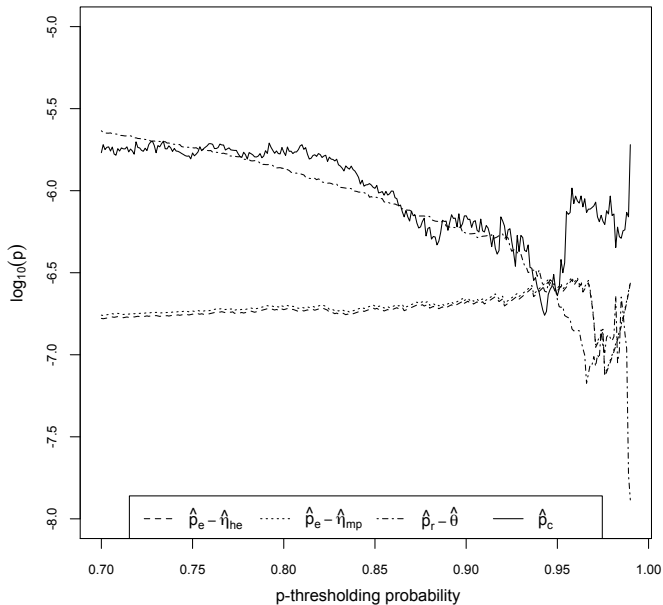
Les résultats ont été conservés pour **deux seuils différents**, correspondant à **deux probabilités de dépassement 0.9 et 0.95**.

Conclusion : Deux estimations de la probabilité de défaillance 6.32×10^{-7} et 2.46×10^{-7} .

Intervalles de confiance associés :

	Estimé ponctuel	95% BPM CI
(0.9)	6.32×10^{-7}	$[2.45 \times 10^{-7}, 1.51 \times 10^{-6}]$
(0.95)	2.46×10^{-7}	$[6.12 \times 10^{-8}, 1.13 \times 10^{-6}]$

Failure probability of (X,Y) with both η, θ -thresholding probability equal to 0.95



V. Pour conclure...

- ▶ Les **modèles de valeurs extrêmes multivariés** permettent d'**évaluer des probabilités de défaillance**.
- ▶ Ne pas oublier toutefois que **certaines hypothèses sont difficiles à tester !**
- ▶ Prendre en compte la **dépendance entre les variables d'intérêt** est **crucial**.
- ▶ **Certains types de dépendance** requièrent des **traitements spécifiques** : dépendance spatiale par exemple.
- ▶ De nombreuses questions sont encore **sans réponse** en ce qui concerne la **modélisation spatio-temporelle des extrêmes**.
Affaire à suivre !