Estimation non-paramétrique de mesures de risque pour des lois conditionnelles à queues lourdes avec application à des extrêmes pluviométriques

par

Jonathan EL METHNI

en collaboration avec

Laurent GARDES & Stéphane GIRARD

Congrès SMAI 2015 Les Karellis
Lundi 8 Juin 2015

Motivations

- But de la théorie des valeurs extrêmes $\Longrightarrow$ étudier et caractériser le comportement des valeurs extrêmes d'un échantillon de variables aléatoires.
- L'exemple d'application historique est l'hydrologie $\Longrightarrow$ Gumbel années 50.

La quantité de pluie est modélisée par une variable aléatoire Y ayant pour fonction de survie $\overline{F}(y) = 1 - F(y) = \mathbb{P}(Y \geq y)$.

On dispose de $Y_{1,n} \leq \cdots \leq Y_{n,n}$ un échantillon ordonné de quantités de pluies annuelles.

Les hydrologues souhaitent estimer le niveau de pluie H qui est atteint ou dépassé en moyenne une fois sur T années *i.e.* on veut estimer H tel que

$$1/T = \mathbb{P}(Y \geq H) = \overline{F}(H)$$

autrement dit on veut estimer

$$H = \overline{F}^{-1}(1/T)$$

- H sera appelé niveau de retour correspondant à une période de retour T.
- C'est la quantité standard d'intérêt dans les études environnementales.

Niveau de retour

Le quantile d'ordre $\alpha \in]0,1[$ de la fonction de survie est défini par

$$q(\alpha) = \overline{F}^{-1}(\alpha).$$

$\Longrightarrow$ Le niveau de retour $H = \overline{F}^{-1}(1/T)$ est donc un quantile d'ordre $\alpha = 1/T$.

- Que se passe-t-il si la période de retour T est supérieure à la durée d'observation ? Autrement dit que se passe-t-il si

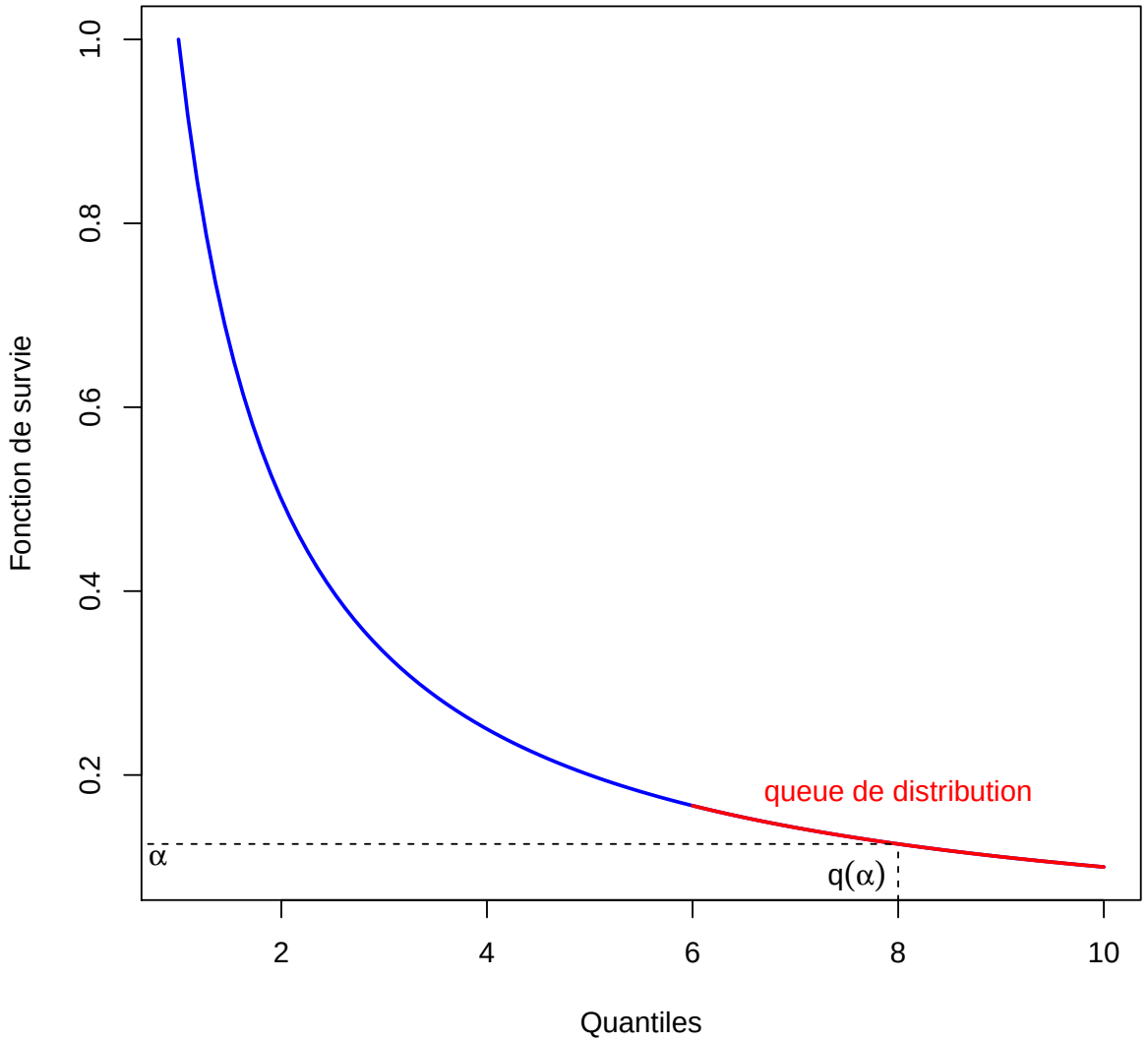
$$T > n \Longleftrightarrow \alpha = 1/T < 1/n \underset{n\to\infty}{\longrightarrow} 0 \quad ?$$

On souhaite estimer des quantiles dit extrêmes $q(\alpha_n)$ d'ordre α_n définis par

$$q(\alpha_n) = \overline{F}^{-1}(\alpha_n) \quad \text{avec} \quad \alpha_n \to 0 \quad \text{quand} \quad n \to \infty$$

- Un niveau de retour dont la période de retour est supérieure à la durée d'observation est un quantile extrême.

Quantiles extrêmes



Que se passe-t-il si $q(\alpha_n) > Y_{n,n}$?

Quantiles extrêmes

On peut montrer que l'on a

$$\mathbb{P}(q(\alpha_n) > Y_{n,n}) = \exp(-n\alpha_n(1 + o(1))$$

- 1er cas :

  $$\text{Si } n\alpha_n \to \infty \text{ alors } \mathbb{P}(q(\alpha_n) > Y_{n,n}) \to 0.$$

  Un estimateur naturel est la statistique d'ordre $Y_{n-\lfloor n\alpha_n \rfloor+1,n}$ (où $\lfloor . \rfloor$ est la fonction partie entière).

- 2ème cas :

  $$\text{Si } n\alpha_n \to 0 \text{ alors } \mathbb{P}(q(\alpha_n) > Y_{n,n}) \to 1.$$

  Dans ce cas, on ne peut pas estimer $q(\alpha_n)$ en inversant simplement la fonction de répartition empirique :

$$\widehat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \leq y\} \quad \text{car} \quad \widehat{F}_n(y) = 1 \quad \text{pour} \quad y \geq Y_{n,n}.$$

Le comportement de $Y_{n,n}$ est caractérisé par sa fonction de répartition $F_{Y_{n,n}}(y) = F^n(y)$ qui est une loi dégénérée.

Le théorème de Fisher-Tippett-Gnedenko [1943]

Théorème

Soit $(Y_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition F. S'il existe deux suites normalisantes réelles $(a_n)_{n\geq 1} > 0$ et $(b_n)_{n\geq 1} \in \mathbb{R}$ et une loi non-dégénérée $\mathcal{H}_\gamma$ telles que

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{Y_{n,n} - b_n}{a_n} \leq y\right) = \lim_{n\to\infty} F^n(a_n y + b_n) = \mathcal{H}_\gamma(y),$$

alors à une translation et un changement d'échelle près on a

$$\mathcal{H}_\gamma(y) = \exp\left(-(1+\gamma y)_+^{-1/\gamma}\right),$$

où $\gamma \in \mathbb{R}$ et $z_+ = \max(0, z)$.

- $\mathcal{H}_\gamma$ est appelée fonction de répartition de la loi des valeurs extrêmes.
- Si F vérifie le théorème de Fisher-Tippett-Gnedenko on dit que F appartient au domaine d'attraction de $\mathcal{H}_\gamma$.
- Cette loi dépend du seul paramètre de forme γ appelé indice des valeurs extrêmes ou indice de queue.

Trois domaines d'attraction

Selon le signe de γ, on distingue trois domaines d'attraction :

- si $\gamma < 0$, on dit que F appartient au domaine d'attraction de Weibull. Il contient des lois dont la fonction de survie n'a pas de queue de distribution.
- si $\gamma = 0$, on dit que F appartient au domaine d'attraction de Gumbel. Il contient les lois dont la fonction de survie est à décroissance exponentielle, *i.e.* les lois à queues légères.
- si $\gamma > 0$, on dit que F appartient au domaine d'attraction de Fréchet. Il contient les lois dont la fonction de survie est à décroissance polynomiale, *i.e.* les lois à queues lourdes.

Fréchet ($\gamma > 0$)	Gumbel ($\gamma = 0$)	Weibull ($\gamma < 0$)
Pareto	Normale	Uniforme
Student	Exponentielle	Beta
Burr	Log-normale	ReverseBurr
Chi-deux	Gamma	
Fréchet	Weibull	
Log-gamma	Logistique	
Log-logistique	Gumbell	
Cauchy		

Lois à queues lourdes $\gamma > 0$

Toutes les lois appartenant au domaine d'attraction de Fréchet peuvent se réécrire

$$\overline{F}(y) = y^{-1/\gamma}\ell(y) \quad \text{avec} \quad \gamma > 0,$$

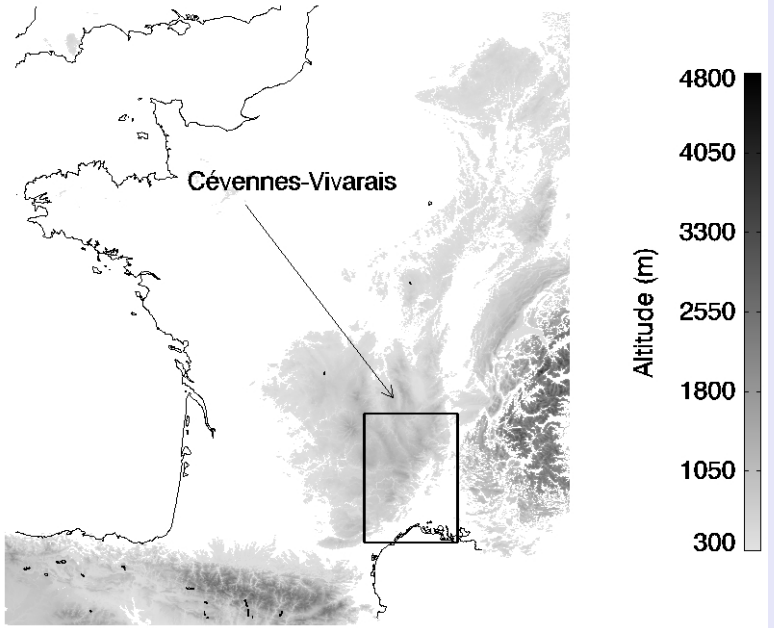
et ℓ est une fonction à variations lentes (f.v.l) à l'infini *i.e.* $\forall \lambda \geq 1$,

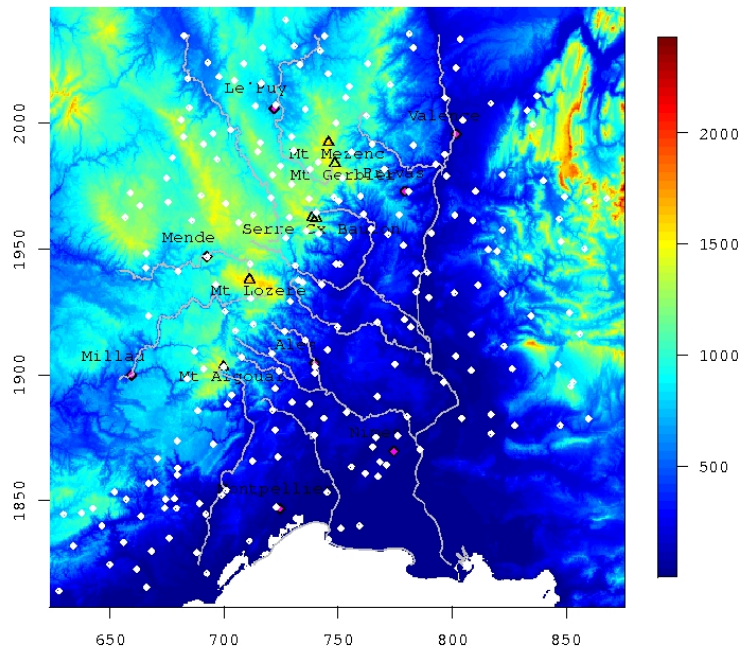
$$\lim_{y\to\infty} \frac{\ell(\lambda y)}{\ell(y)} \to 1.$$

- Le paramètre γ contrôle le comportement de la queue de la fonction de survie et donc celui des valeurs extrêmes.

Dans de nombreuses applications, la variable d'intérêt est mesurée simultanément avec une covariable.

La région Cévennes-Vivarais

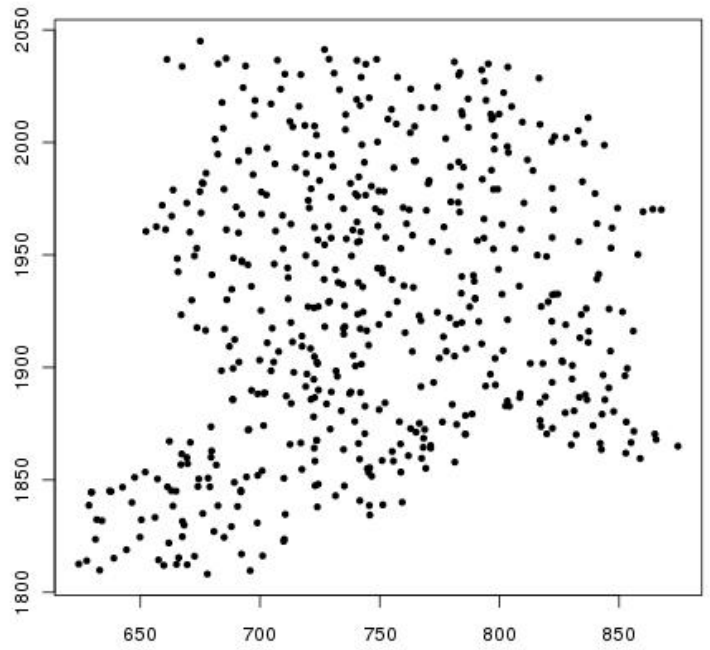




Horizontalement on a la longitude (en km), verticalement la latitude (en km) et en échelle de couleurs l'altitude (en m). Quelques stations d'observations (losanges blancs).

Description du jeu de données réelles

Les 523 stations d'intérêt



- Données fournies par Météo-France.
- Y : hauteur de pluie journalière en mm.
- $X = \{\text{longitude, lattitude, altitude}\}$.

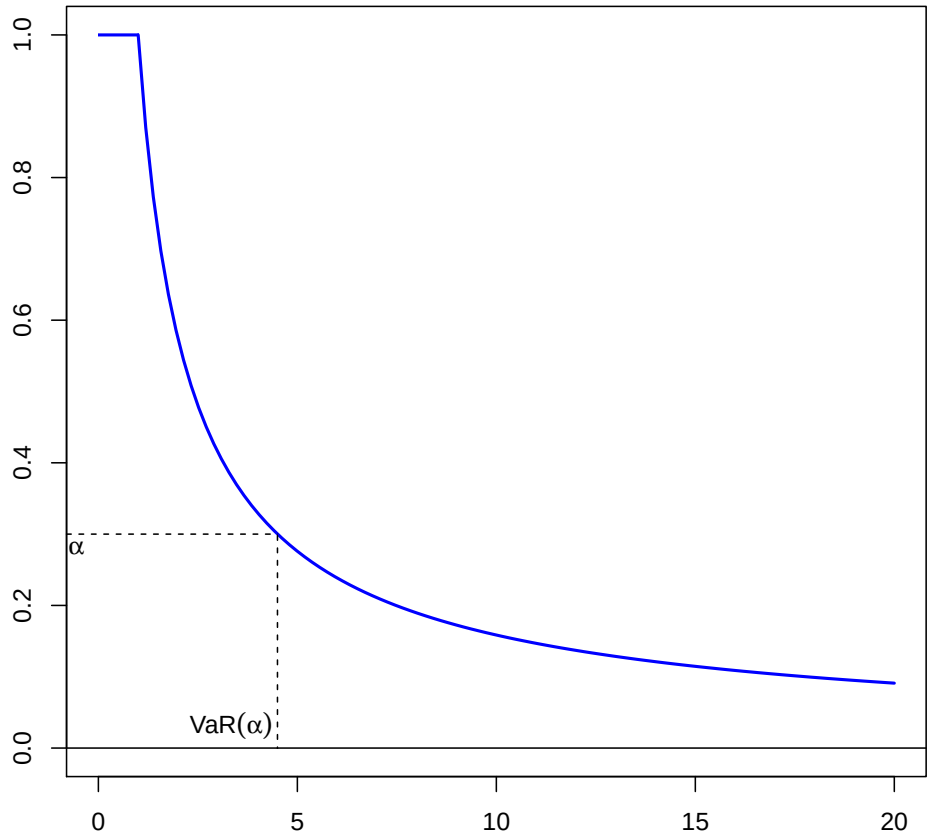
- 1958 $\Longrightarrow$ 2000 soit 43 ans de données.
- 523 stations notées $\{x_t; t = 1, \ldots, 523\}$.
- Nombre total d'observation $= 5\ 513\ 734$.

But $\Longrightarrow$ obtenir des cartes d'estimation de mesures de risque des pluies journalières correspondant à une période de retour de 100 ans en tout point de la région.

La Value-at-Risk

- Soit Y une variable aléatoire désignant un montant de pertes. La Value-at-Risk au niveau $\alpha \in]0,1[$ notée $\mathrm{VaR}(\alpha)$ introduite en 1993 est définie par

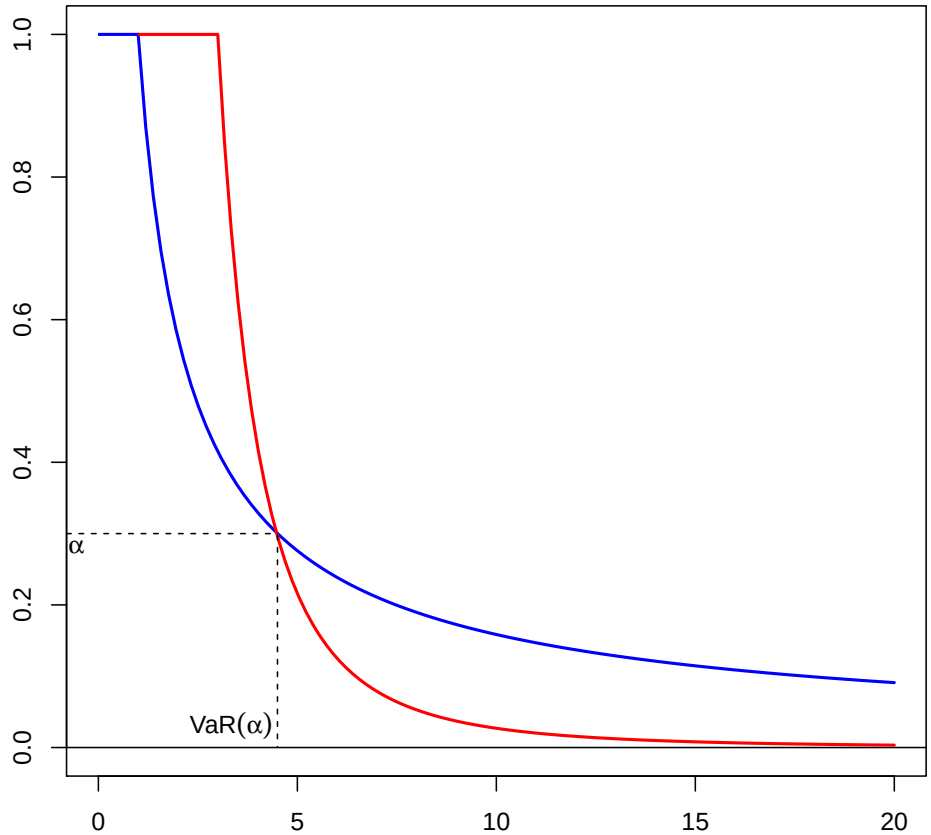
$$\mathrm{VaR}(\alpha) := \overline{F}^{-1}(\alpha).$$



- La $\mathrm{VaR}(\alpha)$ est le quantile d'ordre α de la fonction de survie de la *v.a.* Y.

Défaults de la Value-at-Risk

- Soit Y_1 et Y_2 deux *v.a.* de pertes ayant pour fonctions de survies associées $\overline{F}_1$ et $\overline{F}_2$.

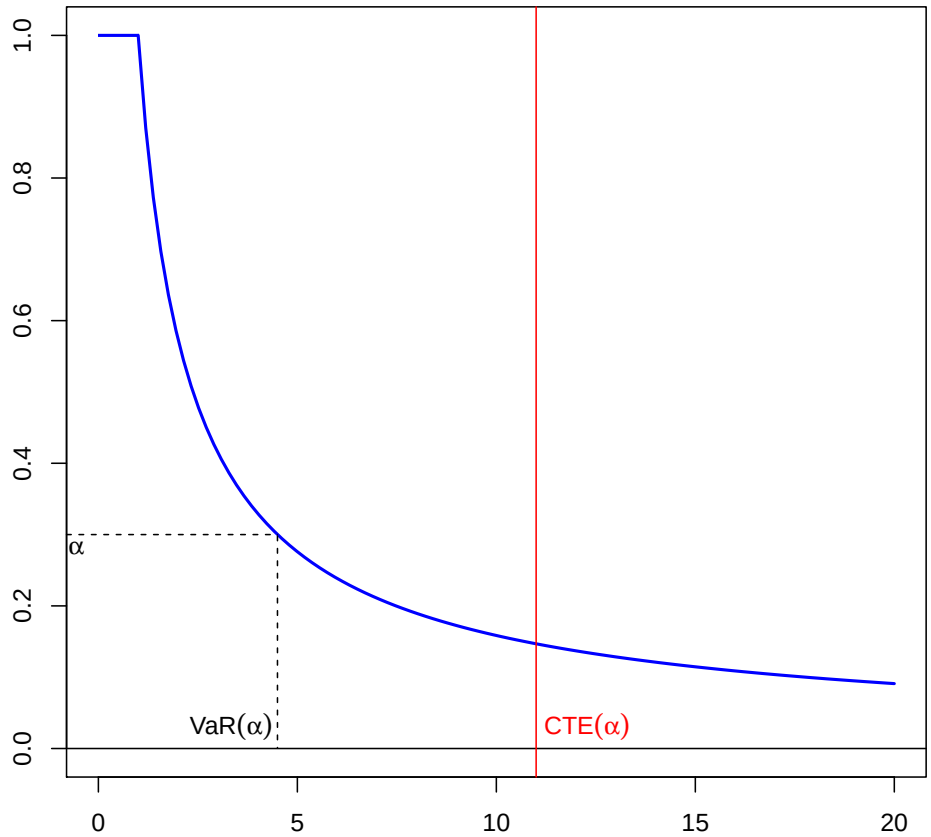


- Un des principaux reproches fait à la VaR est que des v.a à queues légères et à queues lourdes (Embrechts *et al.* [1997]) peuvent avoir la même VaR(α).

La Conditional Tail Expectation

- La Conditional Tail Expectation au niveau $\alpha \in]0,1[$ notée CTE(α) est définie par

$$\mathrm{CTE}(\alpha) := \mathbb{E}(Y|Y > \mathrm{VaR}(\alpha)).$$

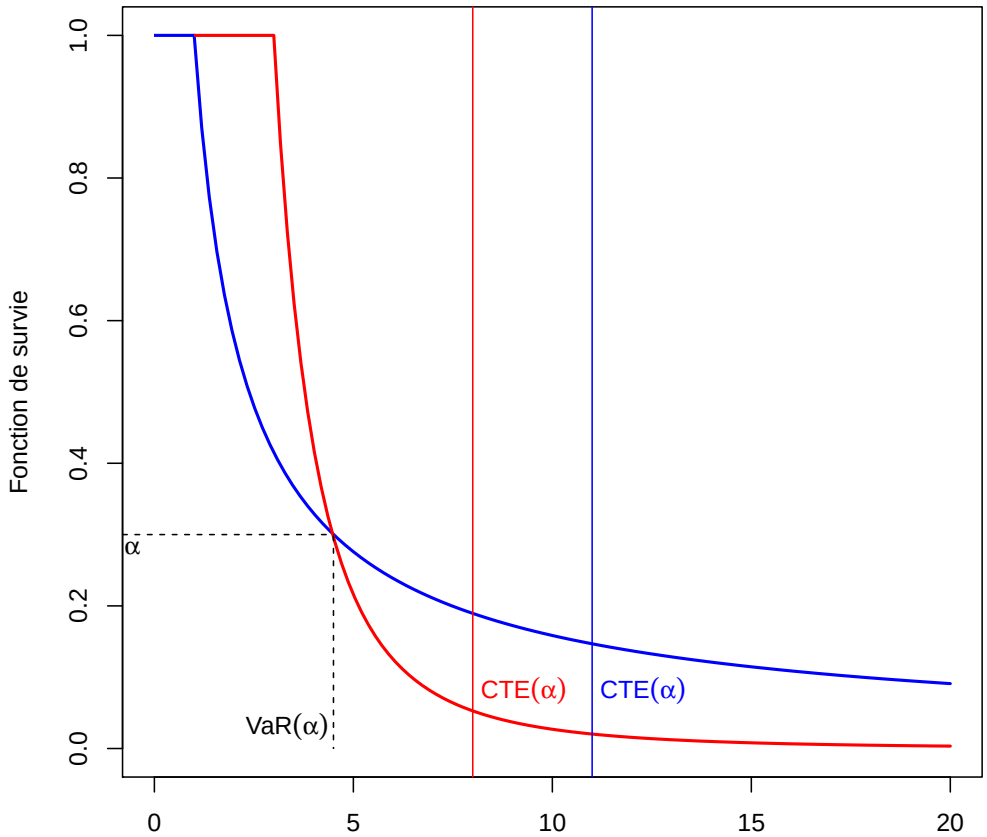


- La CTE(α), donne des informations sur la distribution de Y au delà de la VaR(α) et donc contrairement à la VaR(α), sur l'épaisseur de la queue de distribution.

La Conditional Tail Expectation

- La Conditional Tail Expectation au niveau $\alpha \in]0,1[$ notée CTE(α) est définie par

$$\mathrm{CTE}(\alpha) := \mathbb{E}(Y|Y > \mathrm{VaR}(\alpha)).$$

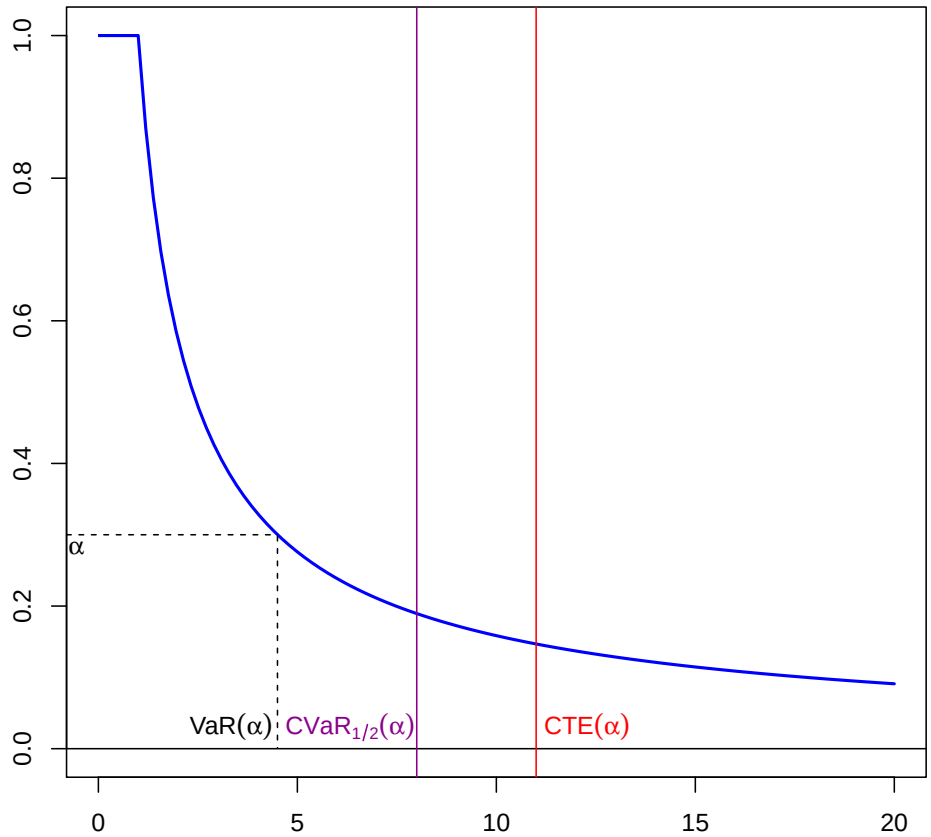


- La CTE(α), donne des informations sur la distribution de Y au delà de la VaR(α) et donc contrairement à la VaR(α), sur l'épaisseur de la queue de distribution.

La Conditional-Value-at-Risk

- La Conditional-Value-at-Risk au niveau $\alpha \in]0,1[$ notée $\mathrm{CVaR}_\lambda(\alpha)$ introduite par Rockafellar et Uryasev [2000] est définie par

$$\mathrm{CVaR}_\lambda(\alpha) := \lambda \mathrm{VaR}(\alpha) + (1-\lambda)\mathrm{CTE}(\alpha) \quad \text{avec} \quad 0 \leq \lambda \leq 1.$$

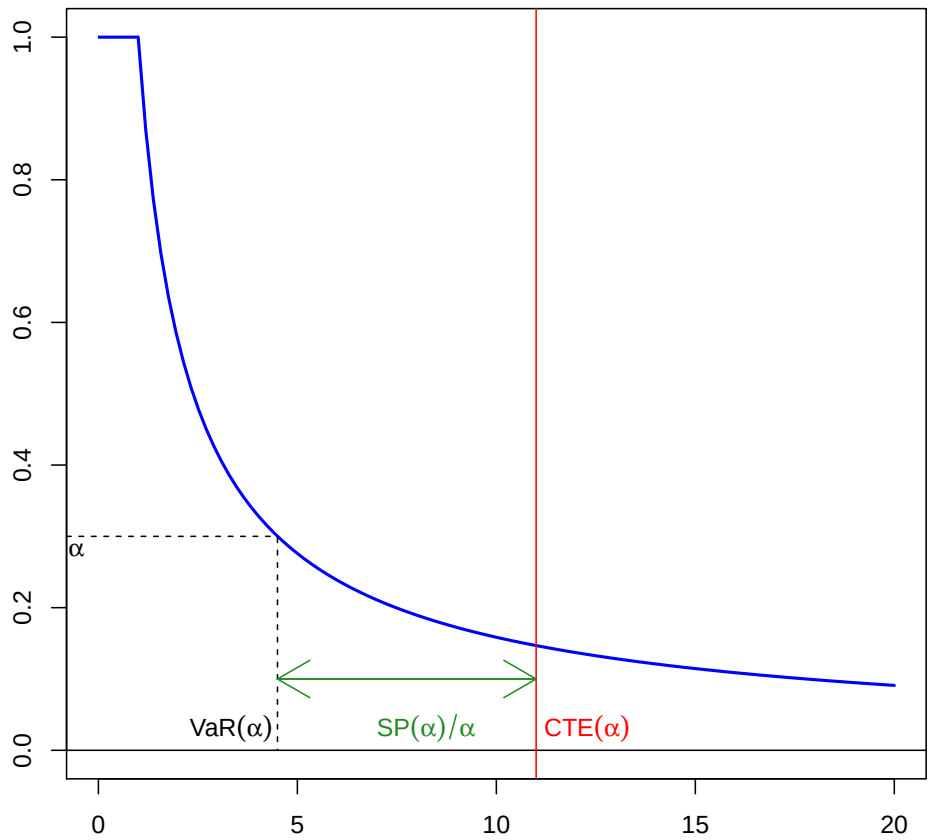


- On peut remarquer que $\mathrm{CVaR}_1(\alpha)$=VaR(α) et que $\mathrm{CVaR}_0(\alpha)$=CTE(α).

La Stop-loss Premium reinsurance

- La mesure de rique Stop-loss Premium reinsurance avec un niveau de rétention égal à VaR(α) (voir Cai et Tan [2007]) est définie par :

$$\mathrm{SP}(\alpha) := \alpha\,(\mathrm{CTE}(\alpha) - \mathrm{VaR}(\alpha))\,.$$



- Cette mesure de risque permet ainsi de mettre en évidence les cas dangereux.

Cadre de travail : pertes extrêmes et régression

L'apport nouveau de nos travaux consiste en l'ajout de deux difficultés supplémentaires dans le cadre de l'estimation de mesures de risque.

1. On ajoute la présence d'une covariable aléatoire $X \in \mathbb{R}^p$.

2. On s'intéresse à l'estimation de mesures de risque dans le cas de pluies extrêmes.

$\Longrightarrow$ Pour cela on remplace l'ordre fixé $\alpha \in]0,1[$ par une suite $\alpha_n \to 0$ quand la taille de l'échantillon $n \to \infty$.

En notant par $\overline{F}(.|x)$ la fonction de survie conditionnelle de Y sachant $X = x$, on définit la Regression Value-at Risk par :

$$\mathrm{RVaR}(\alpha_n|x) := \overline{F}^{-1}(\alpha_n|x),$$

et la Regression Conditional Tail Expectation par :

$$\mathrm{RCTE}(\alpha_n|x) := \mathbb{E}(Y|Y > \mathrm{RVaR}(\alpha_n|x), X = x).$$

$\Longrightarrow$ Les mesures de risque dépendent alors seulement de RVaR et de RCTE.

$$\begin{aligned} \mathrm{RCVaR}_\lambda(\alpha_n|x) &= \lambda \mathrm{RVaR}(\alpha_n|x) + (1-\lambda)\mathrm{RCTE}(\alpha_n|x), \\ \mathrm{RSP}(\alpha_n|x) &= \alpha_n(\mathrm{RCTE}(\alpha_n|x) - \mathrm{RVaR}(\alpha_n|x)). \end{aligned}$$

$\Longrightarrow$ On veut estimer toutes les mesures de risque mentionnées.

Il nous faut obtenir la loi jointe asymptotique de $\left\{\widehat{\mathrm{RCTE}}_n(\alpha_n|x),\ \widehat{\mathrm{RVaR}}_n(\alpha_n|x)\right\}$.

On définit le moment conditionnel d'ordre $a \geq 0$ de Y sachant $X = x$ par

$$\varphi_a(y|x) = \mathbb{E}\left(Y^a \mathbb{I}\{Y > y\}|X = x\right),$$

où $\mathbb{I}\{.\}$ est la fonction indicatrice.

En remarquant que $\varphi_0(y|x) = \overline{F}(y|x)$ on obtient

$$\begin{aligned} \mathrm{RVaR}(\alpha_n|x) &= \varphi_0^{-1}(\alpha_n|x), \\ \mathrm{RCTE}(\alpha_n|x) &= \frac{1}{\alpha_n}\varphi_1(\varphi_0^{-1}(\alpha_n|x)|x). \end{aligned}$$

Estimateur de $\varphi_a(.|x)$:

Pour estimer le moment d'ordre $a \geq 0$ de Y sachant $X = x$, on utilise un estimateur à noyau défini pour $(x, y) \in \mathbb{R}^p \times \mathbb{R}$ par

$$\widehat{\varphi}_{a,n}(y|x) = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) Y_i^a \mathbb{I}\{Y_i > y\} \Big/ \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right).$$

- La fonction K est appelée noyau. C'est une densité bornée sur $\mathbb{R}^p$, de support S inclu dans la boule unité de $\mathbb{R}^p$.
- (h_n) est une suite non aléatoire telle que $h_n \to 0$ quand $n \to \infty$ appelée paramètre de lissage.

Estimateur de $\varphi_a^{-1}(.|x)$:

Comme $\hat{\varphi}_{a,n}(.|x)$ est une fonction décroissante on donc peut définir un estimateur de $\varphi_a^{-1}(.|x)$ par

$$\hat{\varphi}_{a,n}^{-1}(\alpha|x).$$

Lois à queues lourdes en présence de covariable

(F) On suppose que la fonction de survie conditionnelle de Y sachant $X = x$ est à queue lourde et admet une fonction de densité.

Ce qui revient à supposer que

$$\forall y > 0, \quad \text{on a} \quad \overline{F}(y|x) = y^{-1/\gamma(x)} \ell(y|x)$$

où dans ce contexte,

- $\gamma(.)$ est une fonction inconnue et positive de la covariable x et sera appelée indice de queue conditionnel puisqu'elle contrôle la lourdeur de la queue de la loi conditionnelle de Y sachant $X = x$.
- $\ell(.|x)$ est une fonction à variations lentes à l'infini. On a (à x fixé), pour tout $\lambda > 0$,

$$\lim_{y \to \infty} \frac{\ell(\lambda y|x)}{\ell(y|x)} = 1.$$

Cette hypothèse revient à supposer que la loi conditionnelle de Y sachant $X = x$ est à queue lourde.

Loi jointe de nos estimateurs

Théorème 1

Supposons que **(F)**, soit vérifiée et que pour tout $x \in \mathbb{R}^p$ tel que $g(x) > 0$ et $0 < \gamma(x) < 1/2$ on ait $(\alpha_n)_{n\geq 1}$ telle que $\alpha_n \to 0$ et $nh_n^p\alpha_n \to \infty$ quand $n \to \infty$. alors le vecteur aléatoire

$$\sqrt{nh_n^p\alpha_n}\left\{\left(\frac{\widehat{\mathrm{RCTE}}_n(\alpha_n|x)}{\mathrm{RCTE}(\alpha_n|x)} - 1\right), \left(\frac{\widehat{\mathrm{RVaR}}_n(\alpha_n|x)}{\mathrm{RVaR}(\alpha_n|x)} - 1\right)\right\}$$

est asymptotiquement Gaussien, centré, avec une matrice de variance-covariance

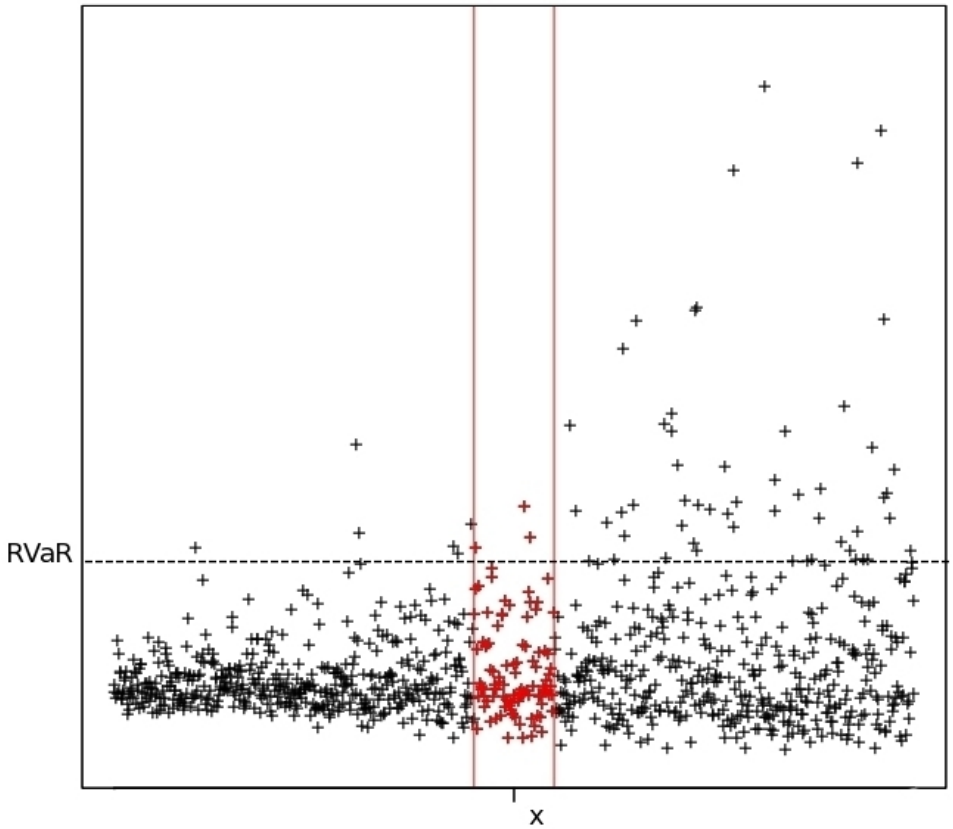
$$\Sigma(x) = \gamma^2(x)\frac{\|K\|_2^2}{g(x)}\begin{pmatrix} \frac{2(1-\gamma(x))}{1-2\gamma(x)} & 1 \\ 1 & 1 \end{pmatrix}.$$

- $\Sigma(x)$ est proportionnelle à $\gamma^2(x) \Longrightarrow$ si $\gamma(x)$ augmente (*i.e.* plus la queue est lourde) plus la variance de nos estimateurs augmente.

- La densité $g(x)$ de la covariable est au dénominateur de la matrice de variance-covariance asymptotique $\Longrightarrow$ moins il y a de point (*i.e.* plus la densité est faible) plus la variance des estimateurs sera grande.

Conditions sur les suites α_n et h_n

- $nh_n^p\alpha_n \to \infty$: condition nécessaire et suffisante pour qu'il y ait presque sûrement au moins une observation dans la région $B(x, h_n) \times [\mathrm{RVaR}(\alpha_n|x), +\infty[$ de $\mathbb{R}^p \times \mathbb{R}$.



Si $\alpha_n = \alpha$ fixé on retrouve la condition de normalité asymptotique classique : $nh_n^p \to \infty$.
Si $h_n = h$ fixé on retrouve la condition de normalité asymptotique classique : $n\alpha_n \to \infty$.

Extrapolation

- Dans le Théorème 1, la condition $nh_n^p\alpha_n \to \infty$ nous donne une restriction sur l'ordre des mesures de risque que l'on peut estimer.
- Cette restriction est due à l'utilisation de nos estimateurs à noyau qui ne permettent pas d'extrapoler au-delà de la plus grande observation dans la boule $B(x, h_n)$.
- Par conséquent, α_n doit correspondre à un ordre de quantile extrême se trouvant dans l'échantillon.

Definition

Soient $(\alpha_n)_{n\geq 1}$ et $(\beta_n)_{n\geq 1}$ deux suites positives telles que $\alpha_n \to 0$, $\beta_n \to 0$ et $0 < \beta_n < \alpha_n$. On peut ainsi définir un estimateur permettant d'extrapoler par

$$\widehat{\mathrm{RCTE}}_n^{W}(\beta_n|x) = \widehat{\mathrm{RCTE}}_n(\alpha_n|x)\left(\frac{\alpha_n}{\beta_n}\right)^{\hat{\gamma}_n(x)}$$

Extrapolation

- Dans le Théorème 1, la condition $nh_n^p\alpha_n \to \infty$ nous donne une restriction sur l'ordre des mesures de risque que l'on peut estimer.
- Cette restriction est due à l'utilisation de nos estimateurs à noyau qui ne permettent pas d'extrapoler au-delà de la plus grande observation dans la boule $B(x, h_n)$.
- Par conséquent, α_n doit correspondre à un ordre de quantile extrême se trouvant dans l'échantillon.

Definition

Soient $(\alpha_n)_{n\geq 1}$ et $(\beta_n)_{n\geq 1}$ deux suites positives telles que $\alpha_n \to 0$, $\beta_n \to 0$ et $0 < \beta_n < \alpha_n$. On peut ainsi définir un estimateur permettant d'extrapoler par

$$\widehat{\mathrm{RCTE}}_n^{W}(\beta_n|x) = \widehat{\mathrm{RCTE}}_n(\alpha_n|x)\left(\frac{\alpha_n}{\beta_n}\right)^{\hat{\gamma}_n(x)}$$

où

- $\widehat{\mathrm{RCTE}}_n(\alpha_n|x)$ est l'estimateur à noyau précédent.

Extrapolation

- Dans le Théorème 1, la condition $nh_n^p\alpha_n \to \infty$ nous donne une restriction sur l'ordre des mesures de risque que l'on peut estimer.
- Cette restriction est due à l'utilisation de nos estimateurs à noyau qui ne permettent pas d'extrapoler au-delà de la plus grande observation dans la boule $B(x, h_n)$.
- Par conséquent, α_n doit correspondre à un ordre de quantile extrême se trouvant dans l'échantillon.

Definition

Soient $(\alpha_n)_{n\geq 1}$ et $(\beta_n)_{n\geq 1}$ deux suites positives telles que $\alpha_n \to 0$, $\beta_n \to 0$ et $0 < \beta_n < \alpha_n$. On peut ainsi définir un estimateur permettant d'extrapoler par

$$\widehat{\mathrm{RCTE}}_n^{W}(\beta_n|x) = \widehat{\mathrm{RCTE}}_n(\alpha_n|x)\left(\frac{\alpha_n}{\beta_n}\right)^{\hat{\gamma}_n(x)}$$

où

- $\widehat{\mathrm{RCTE}}_n(\alpha_n|x)$ est l'estimateur à noyau précédent.
- $(\alpha_n/\beta_n)^{\hat{\gamma}_n(x)}$ est le terme permettant l'extrapolation.

Extrapolation

- Dans le Théorème 1, la condition $nh_n^p\alpha_n \to \infty$ nous donne une restriction sur l'ordre des mesures de risque que l'on peut estimer.
- Cette restriction est due à l'utilisation de nos estimateurs à noyau qui ne permettent pas d'extrapoler au-delà de la plus grande observation dans la boule $B(x, h_n)$.
- Par conséquent, α_n doit correspondre à un ordre de quantile extrême se trouvant dans l'échantillon.

Definition

Soient $(\alpha_n)_{n\geq 1}$ et $(\beta_n)_{n\geq 1}$ deux suites positives telles que $\alpha_n \to 0$, $\beta_n \to 0$ et $0 < \beta_n < \alpha_n$. On peut ainsi définir un estimateur permettant d'extrapoler par

$$\widehat{\mathrm{RCTE}}_n^{W}(\beta_n|x) = \widehat{\mathrm{RCTE}}_n(\alpha_n|x)\left(\frac{\alpha_n}{\beta_n}\right)^{\hat{\gamma}_n(x)}$$

où

- $\widehat{\mathrm{RCTE}}_n(\alpha_n|x)$ est l'estimateur à noyau précédent.
- $(\alpha_n/\beta_n)^{\hat{\gamma}_n(x)}$ est le terme permettant l'extrapolation.
- $\hat{\gamma}_n(x)$ est un estimateur de l'indice de queue conditionnel.

Normalité asymptotique de $\widehat{\mathrm{RCTE}}_n^{W}(\beta_n|x)$

Théorème 2

Supposons que les hypothèses du Théorème 1 soient vérifiées. Considérons $\hat{\gamma}_n(x)$ un estimateur de l'indice de queue conditionnel tel que

$$\sqrt{nh_n^p\alpha_n}(\hat{\gamma}_n(x) - \gamma(x)) \xrightarrow{d} \mathcal{N}\left(0, v^2(x)\right),$$

avec $v(x) > 0$. Si $(\beta_n)_{n\geq 1}$ est une suite positive telle que $\beta_n \to 0$, $\beta_n/\alpha_n \to 0$ quand $n \to \infty$, alors pour tout $x \in \mathbb{R}^p$, on a

$$\frac{\sqrt{nh_n^p\alpha_n}}{\log(\alpha_n/\beta_n)}\left(\frac{\widehat{\mathrm{RCTE}}_n^{W}(\beta_n|x)}{\mathrm{RCTE}(\beta_n|x)} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, v^2(x)\right).$$

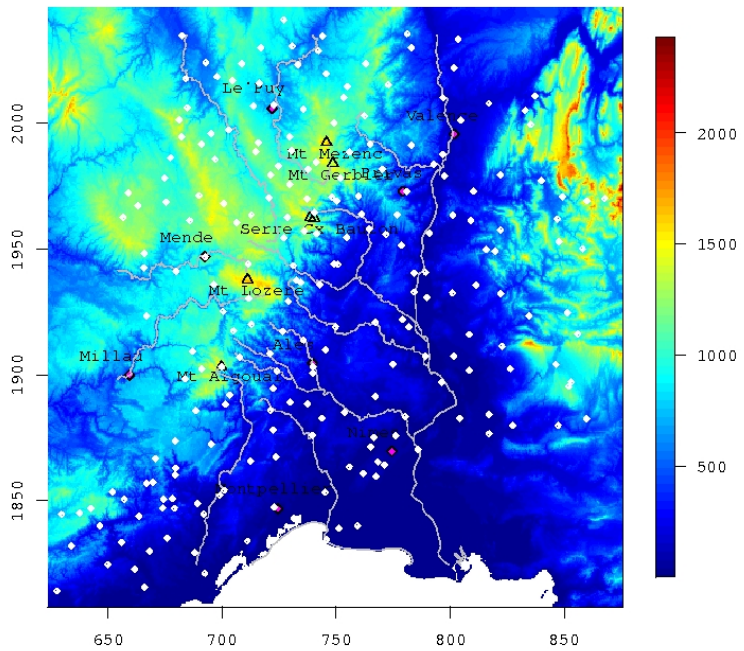
- La condition $\beta_n/\alpha_n \to 0$ permet d'extrapoler et donc de choisir un ordre β_n arbitrairement petit.
- Daouia et al. [2011] ont établi la normalité asymptotique de

$$\widehat{\mathrm{RVaR}}_n^{W}(\beta_n|x) = \widehat{\mathrm{RVaR}}_n(\alpha_n|x)\left(\frac{\alpha_n}{\beta_n}\right)^{\hat{\gamma}_n(x)}$$

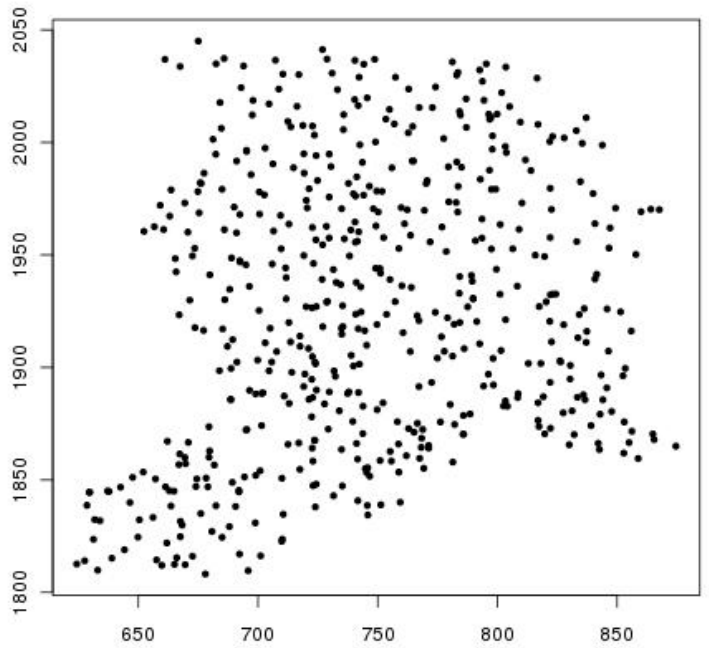
Applications à un jeu de données hydrologique

Y : hauteur de pluie journalière en mm. $X = \{\text{longitude, latitude, altitude}\}$. 1958 $\Longrightarrow$ 2000.

La région Cévennes-Vivarais



Les 523 stations d'intérêt

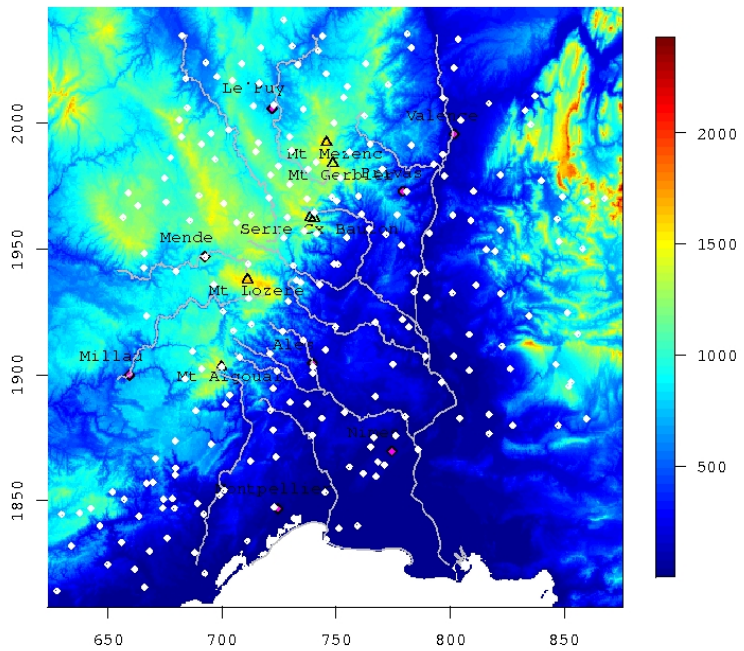


But $\Longrightarrow$ estimer $\mathrm{RVaR}(\beta_n|x)$ et $\mathrm{RCTE}(\beta_n|x)$ pour $\beta_n = 1/(100 * 365.25)$.

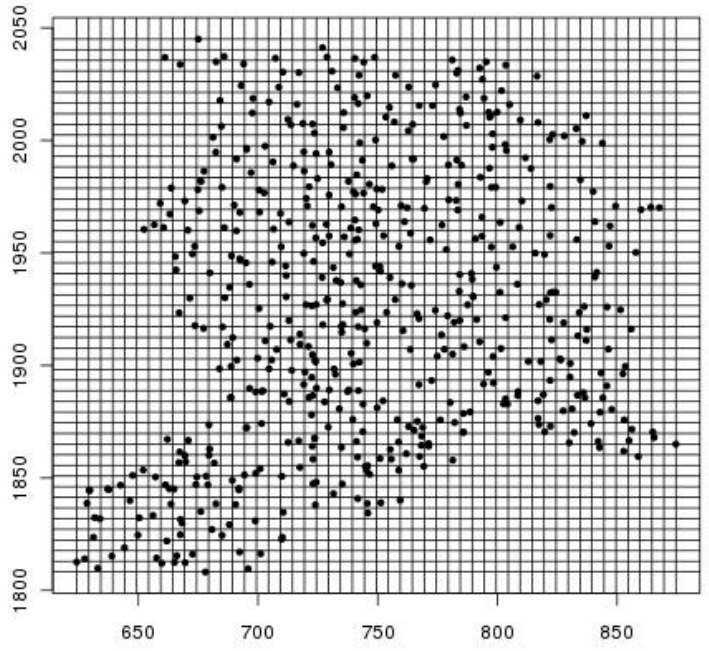
Applications à un jeu de données hydrologique

Y : hauteur de pluie journalière en mm. $X = \{\text{longitude, latitude, altitude}\}$. 1958 $\Longrightarrow$ 2000.

La région Cévennes-Vivarais



Grille 200 $\times$ 200 points

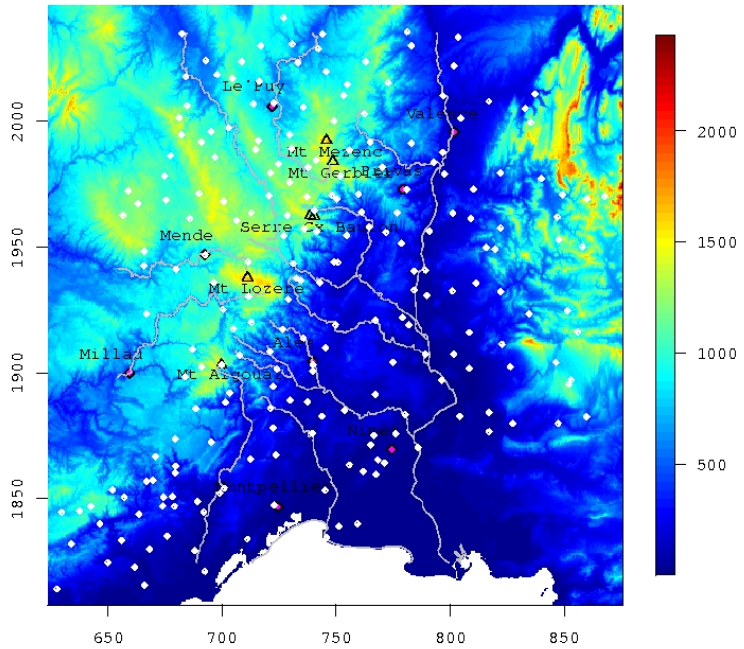


But $\Longrightarrow$ estimer $\mathrm{RVaR}(\beta_n|x)$ et $\mathrm{RCTE}(\beta_n|x)$ pour $\beta_n = 1/(100 * 365.25)$.

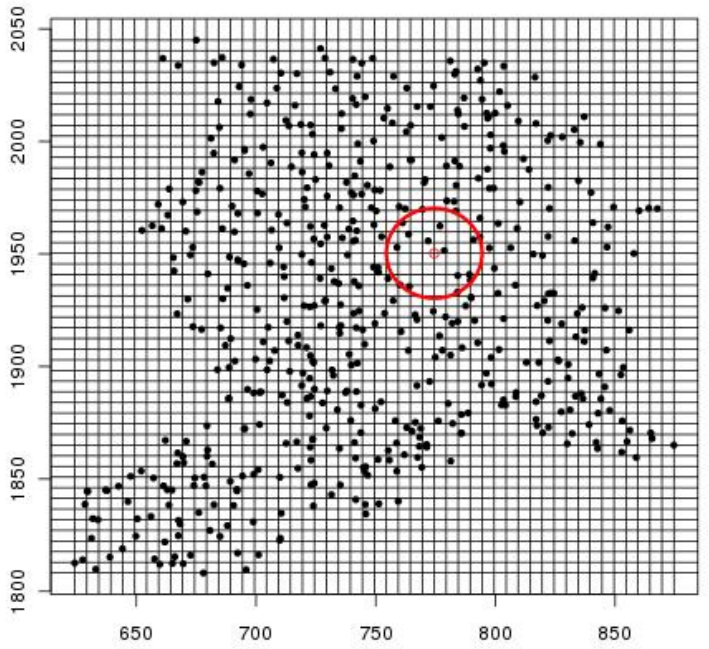
Applications à un jeu de données hydrologique

Y : hauteur de pluie journalière en mm. $X = \{\text{longitude, latitude, altitude}\}$. 1958 $\Longrightarrow$ 2000.

La région Cévennes-Vivarais



Travail dans $B(x, h_n)$



But $\Longrightarrow$ estimer $\mathrm{RVaR}(\beta_n|x)$ et $\mathrm{RCTE}(\beta_n|x)$ pour $\beta_n = 1/(100 * 365.25)$.

Procédure afin de choisir (h_n, α_n)

Nos estimateurs dépendent des deux paramètres de contrôle :

- h_n paramètre de lissage : problème récurrent en statistique non-paramétrique.
- α_n nombre de statistiques d'ordre utilisées : problème classique du compromis biais/variance en statistique des valeurs extrêmes.

Si α_n grand $\Longrightarrow$ biais important car on sort de la queue de distribution.
Si α_n petit $\Longrightarrow$ variance importante car on utilise peu d'observations.

$\Longrightarrow$ Mise en place d'une procédure afin de choisir (h_n, α_n).

Notre procédure est basée sur l'estimation de la fonction $\gamma(x)$ puisqu'elle contrôle :

- la lourdeur de la queue de la loi,
- ainsi que l'extrapolation.

Le principe de notre procédure est de choisir le couple empirique $(h_{emp}, \alpha_{emp}) \in \mathcal{H} \times \mathcal{A}$ pour lequel deux estimations différentes de l'indice $\gamma(x_t)$ en chaque station t coïncident.

- Hydrologues $\Longrightarrow$ important de combiner des informations locales et régionales.

Estimateur de Hill

- Dans un cadre sans covariable

Soit $(\alpha_n)_{n\geq 1}$ une suite positive telle que $\alpha_n \to 0$, l'estimateur de Hill [1975] est défini par :

$$\hat{\gamma}_{n,\alpha_n} = \frac{1}{\lfloor n\alpha_n \rfloor - 1} \sum_{i=1}^{\lfloor n\alpha_n \rfloor - 1} \log\left(Y_{n-i+1,n}\right) - \log\left(Y_{n-\lfloor n\alpha_n \rfloor + 1,n}\right)$$

où $\lfloor . \rfloor$ est la fonction partie entière.

- Adaptation de l'estimateur de Hill à la présence d'une covariable

Soit $(\alpha_n)_{n\geq 1}$ une suite positive telle que $\alpha_n \to 0$. Un estimateur à noyau de type Hill (Gardes et Girard [2008]) est donné par

$$\hat{\gamma}_{n,\alpha_n}(x) = \sum_{j=1}^{J} (\log(\widehat{\mathrm{RVaR}}_n(\tau_j\alpha_n|x)) - \log(\widehat{\mathrm{RVaR}}_n(\tau_1\alpha_n|x))) \Big/ \sum_{j=1}^{J} \log(\tau_1/\tau_j),$$

où $J \geq 1$ et $(\tau_j)_{j\geq 1}$ est une suite de poids strictement positive décroissante.

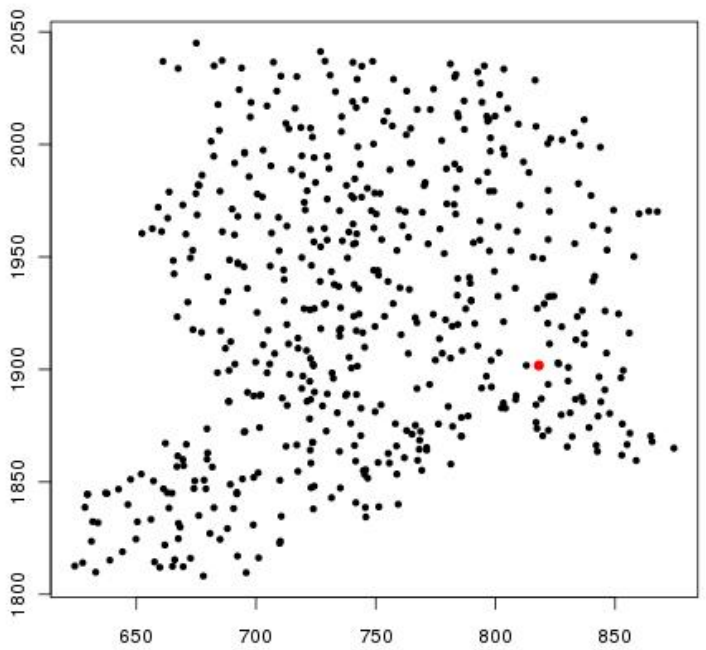
Procédure de type validation croisée pour choisir h_n et α_n : Etape 1

- Double boucle sur $\mathcal{H} = \{h_i; i = 1, \ldots, M\}$ et sur $\mathcal{A} = \{\alpha_j; j = 1, \ldots, R\}$.
- Boucle sur toutes les stations $\{x_t; t = 1, \ldots, 523\}$.

Procédure de type validation croisée pour choisir h_n et α_n : Etape 1

- Double boucle sur $\mathcal{H} = \{h_i; i = 1, \ldots, M\}$ et sur $\mathcal{A} = \{\alpha_j; j = 1, \ldots, R\}$.
- Boucle sur toutes les stations $\{x_t; t = 1, \ldots, 523\}$.

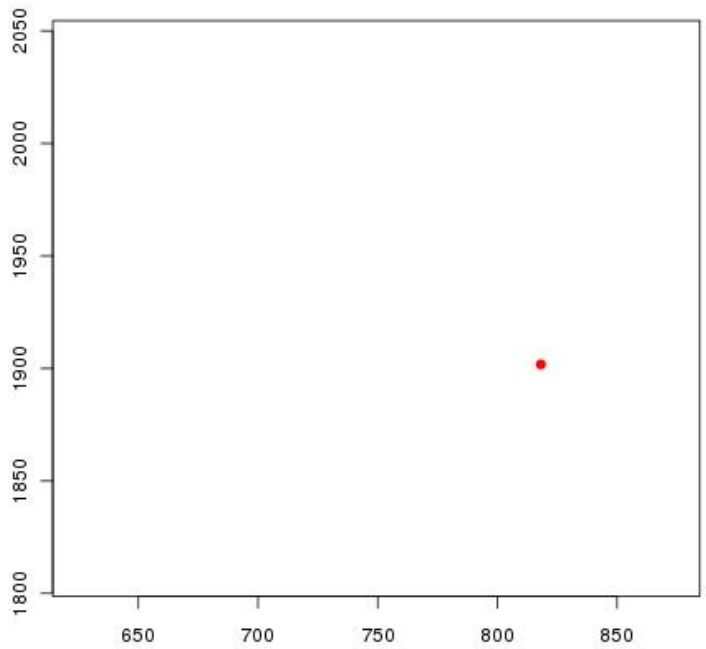
On se place en une station x_t



Procédure de type validation croisée pour choisir h_n et α_n : Etape 1

- Double boucle sur $\mathcal{H} = \{h_i; i = 1, \ldots, M\}$ et sur $\mathcal{A} = \{\alpha_j; j = 1, \ldots, R\}$.
- Boucle sur toutes les stations $\{x_t; t = 1, \ldots, 523\}$.

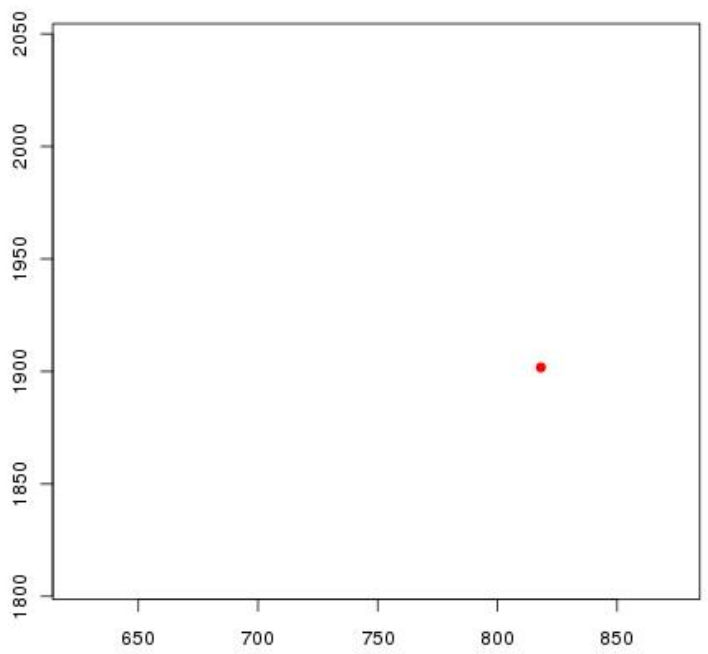
On retire toutes les autres stations



Procédure de type validation croisée pour choisir h_n et α_n : Etape 1

- Double boucle sur $\mathcal{H} = \{h_i; i = 1, \ldots, M\}$ et sur $\mathcal{A} = \{\alpha_j; j = 1, \ldots, R\}$.
- Boucle sur toutes les stations $\{x_t; t = 1, \ldots, 523\}$.

On retire toutes les autres stations

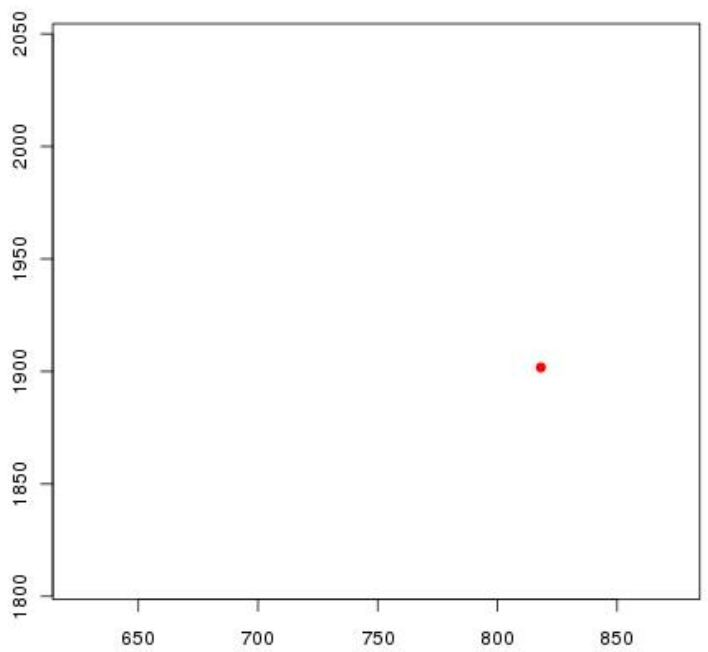


- On estime $\gamma > 0$ en utilisant l'estimateur de Hill classique.
- Il dépend seulement de α_j.
- Les α_j sont choisis de sorte que l'on soit dans la queue de distribution $\max_{j \in \{1,\ldots,R\}} (\alpha_j) < 0.1$

Procédure de type validation croisée pour choisir h_n et α_n : Etape 1

- Double boucle sur $\mathcal{H} = \{h_i; i = 1, \ldots, M\}$ et sur $\mathcal{A} = \{\alpha_j; j = 1, \ldots, R\}$.
- Boucle sur toutes les stations $\{x_t; t = 1, \ldots, 523\}$.

On retire toutes les autres stations

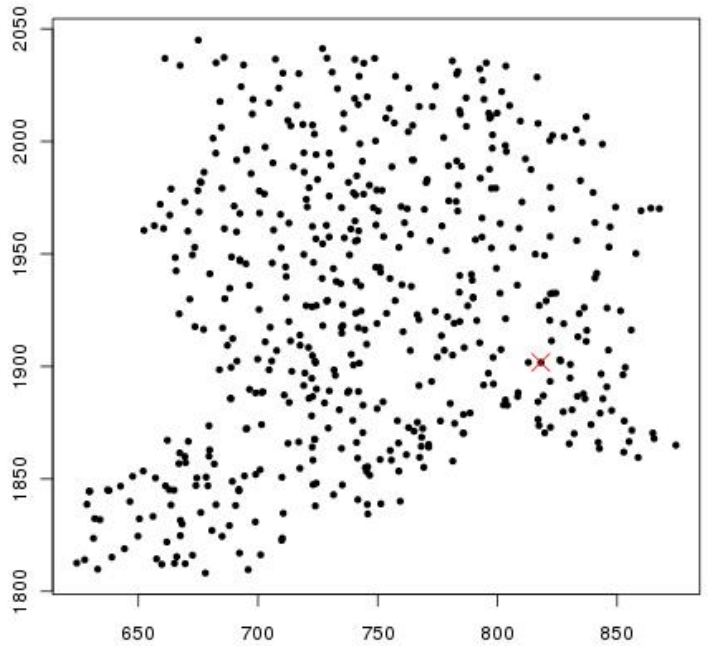


- On estime $\gamma > 0$ en utilisant l'estimateur de Hill classique.
- Il dépend seulement de α_j.
- Les α_j sont choisis de sorte que l'on soit dans la queue de distribution $\max_{j \in \{1,\ldots,R\}} (\alpha_j) < 0.1$

$\Longrightarrow$ On obtient $\hat{\gamma}_{n,t,\alpha_j}$

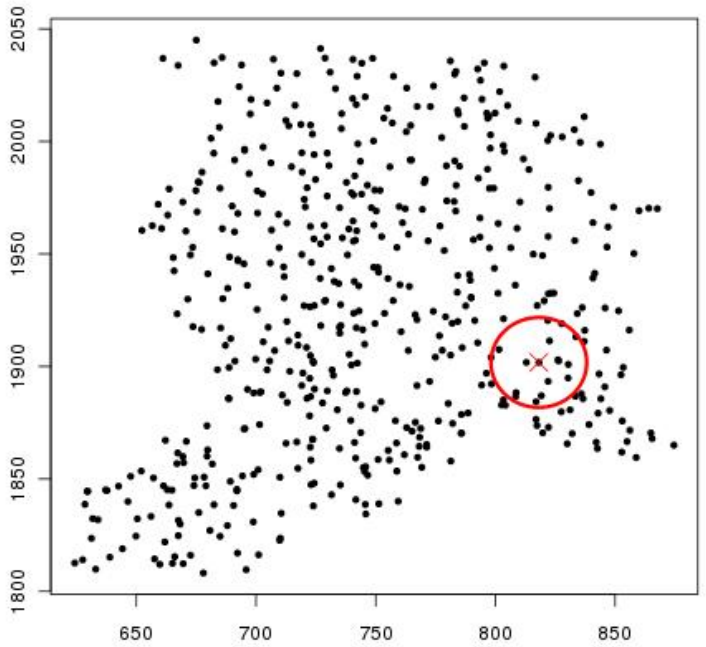
Procédure de type validation croisée pour choisir h_n et α_n : Etape 2

On retire cette station x_t



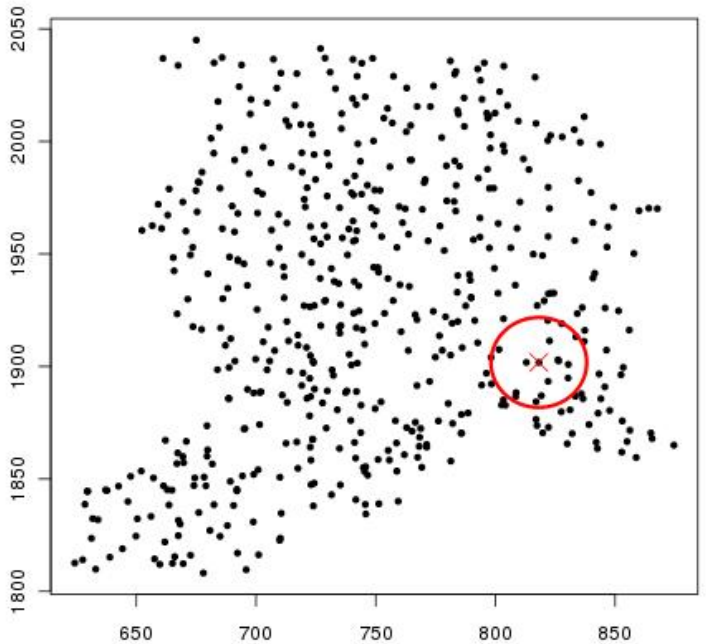
Procédure de type validation croisée pour choisir h_n et α_n : Etape 2

On travaille dans $B(x_t, h_i) \setminus \{x_t\}$



Procédure de type validation croisée pour choisir h_n et α_n : Etape 2

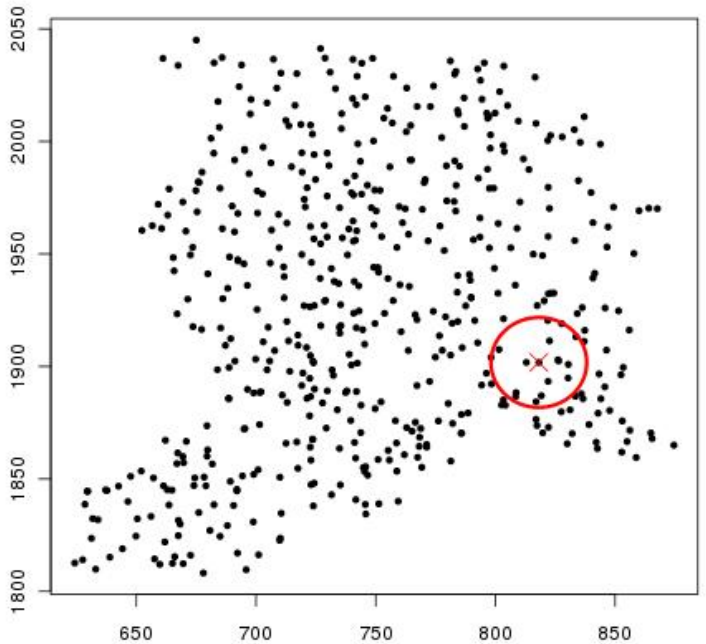
On travaille dans $B(x_t, h_i) \setminus \{x_t\}$



- On estime $\gamma(x) > 0$ en utilisant l'estimateur de Hill conditionnel.
- Il dépend de α_j et de h_i.
- Les h_i sont choisis de sorte qu'il y ait au moins une station dans $B(x_t, h_i) \setminus \{x_t\}$.

Procédure de type validation croisée pour choisir h_n et α_n : Etape 2

On travaille dans $B(x_t, h_i) \setminus \{x_t\}$

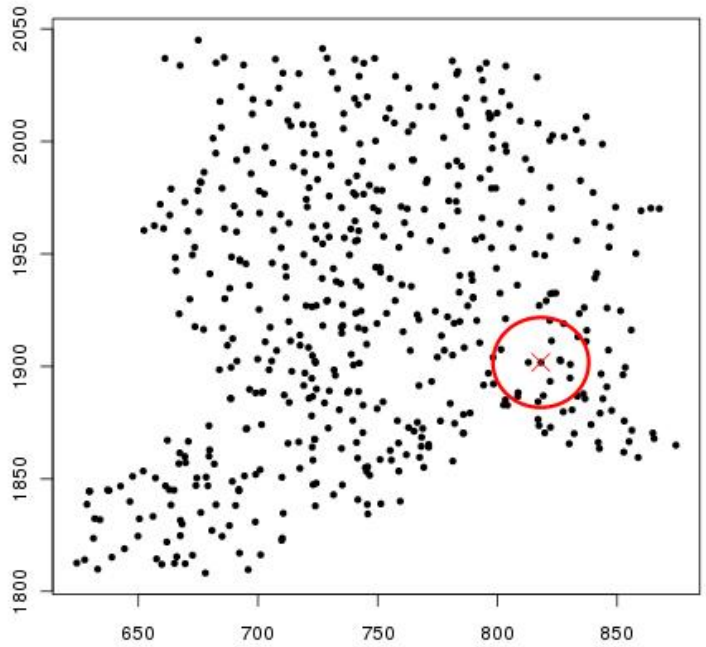


- On estime $\gamma(x) > 0$ en utilisant l'estimateur de Hill conditionnel.
- Il dépend de α_j et de h_i.
- Les h_i sont choisis de sorte qu'il y ait au moins une station dans $B(x_t, h_i) \setminus \{x_t\}$.

$\Longrightarrow$ On obtient $\hat{\gamma}_{n,h_i,\alpha_j}(x_t)$

Procédure de type validation croisée pour choisir h_n et α_n : Etape 2

On travaille dans $B(x_t, h_i) \setminus \{x_t\}$



- On estime $\gamma(x) > 0$ en utilisant l'estimateur de Hill conditionnel.
- Il dépend de α_j et de h_i.
- Les h_i sont choisis de sorte qu'il y ait au moins une station dans $B(x_t, h_i) \setminus \{x_t\}$.

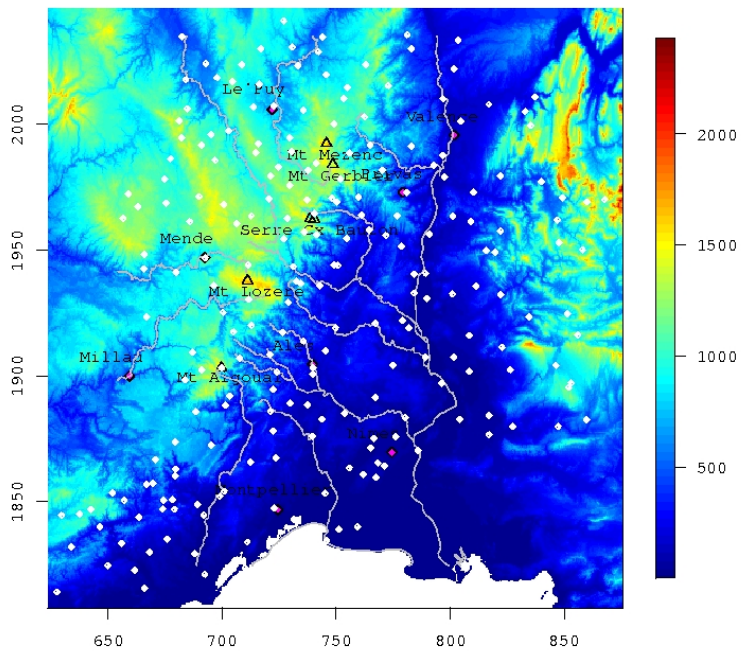
$\Longrightarrow$ On obtient $\hat{\gamma}_{n,h_i,\alpha_j}(x_t)$

$$(h_{emp}, \alpha_{emp}) = \underset{(h_i,\alpha_j)\in\mathcal{H}\times\mathcal{A}}{\arg\min} \ \text{mediane} \ \left\{\left(\hat{\gamma}_{n,t,\alpha_j} - \hat{\gamma}_{n,h_i,\alpha_j}(x_t)\right)^2, t \in \{1,\ldots,523\}\right\}.$$

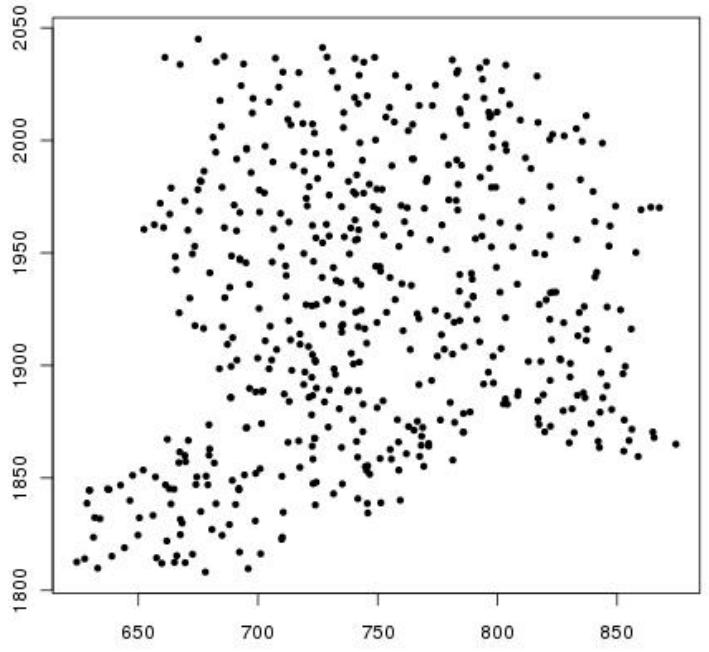
Applications à un jeu de données hydrologique

Y : hauteur de pluie journalière en mm. $X = \{\text{longitude, latitude, altitude}\}$. 1958 $\Longrightarrow$ 2000.

La région Cévennes-Vivarais



Les 523 stations d'intérêt

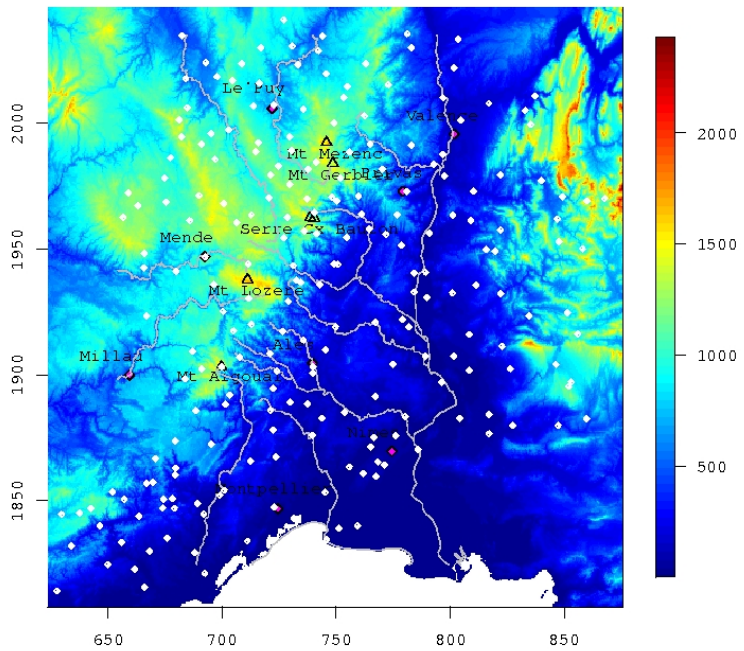


$\Longrightarrow$ Résultats de la procédure : $(h_{emp}, \alpha_{emp}) = (24, 1/(3 * 365.25))$.

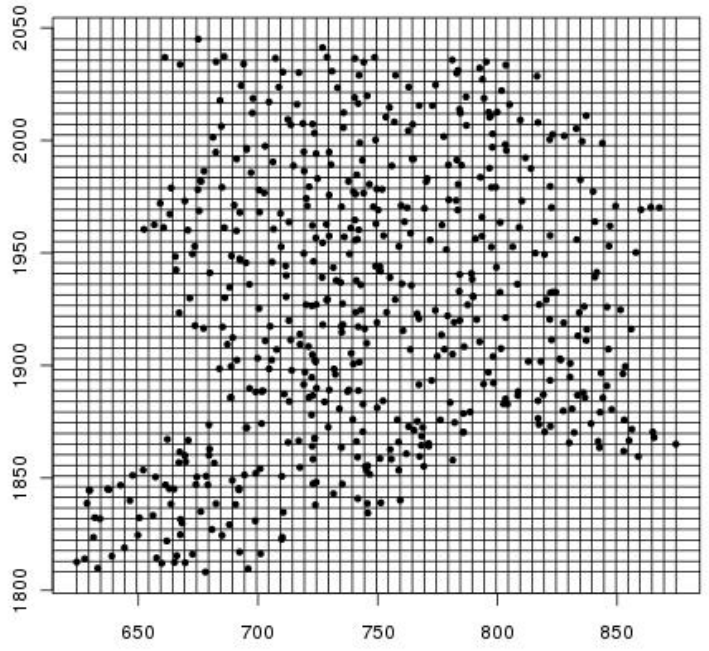
Applications à un jeu de données hydrologique

Y : hauteur de pluie journalière en mm. $X = \{\text{longitude, latitude, altitude}\}$. 1958 $\Longrightarrow$ 2000.

La région Cévennes-Vivarais



Grille 200 $\times$ 200 points

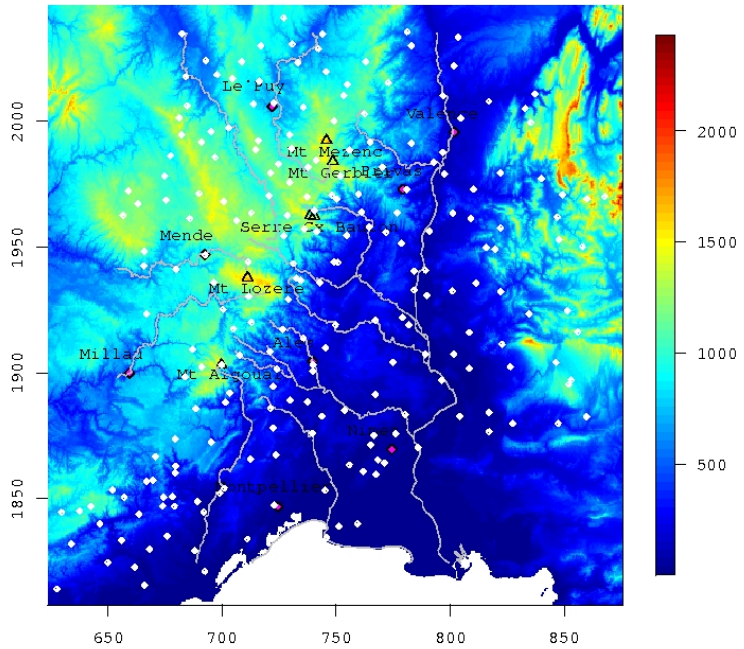


$\Longrightarrow$ Résultats de la procédure : $(h_{emp}, \alpha_{emp}) = (24, 1/(3 * 365.25))$.

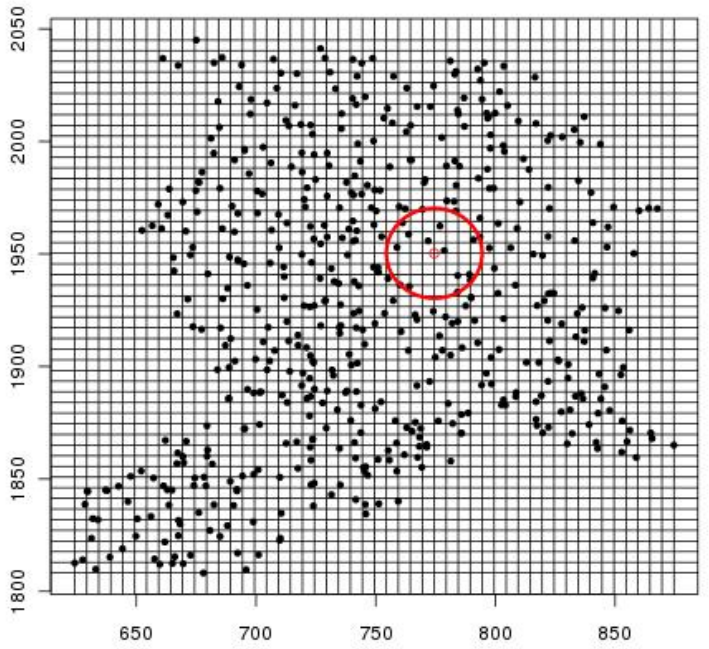
Applications à un jeu de données hydrologique

Y : hauteur de pluie journalière en mm. $X = \{\text{longitude, latitude, altitude}\}$. 1958 $\Longrightarrow$ 2000.

La région Cévennes-Vivarais



Travail dans $B(x, h_n)$

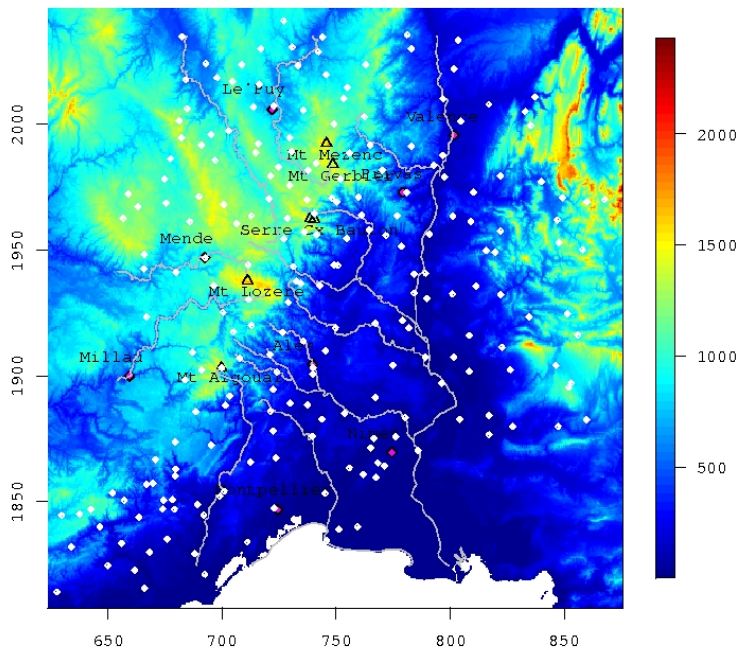


$\Longrightarrow$ Résultats de la procédure : $(h_{emp}, \alpha_{emp}) = (24, 1/(3 * 365.25))$.

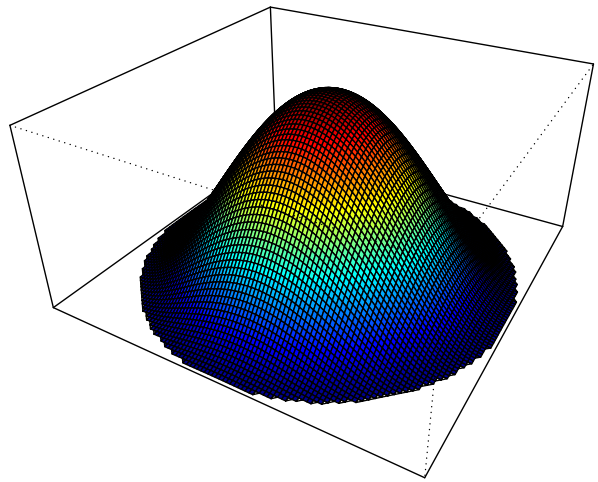
Applications à un jeu de données hydrologique

Y : hauteur de pluie journalière en mm. $X = \{\text{longitude, latitude, altitude}\}$. 1958 $\Longrightarrow$ 2000.

La région Cévennes-Vivarais



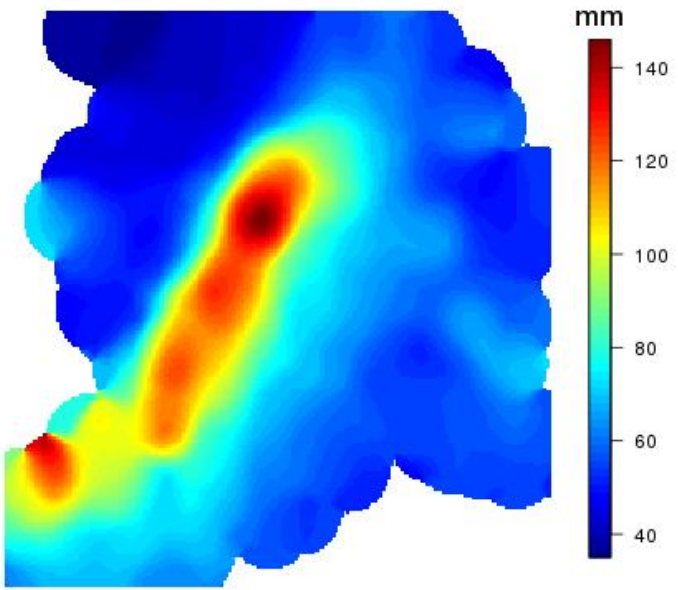
Noyau bi-quadratique



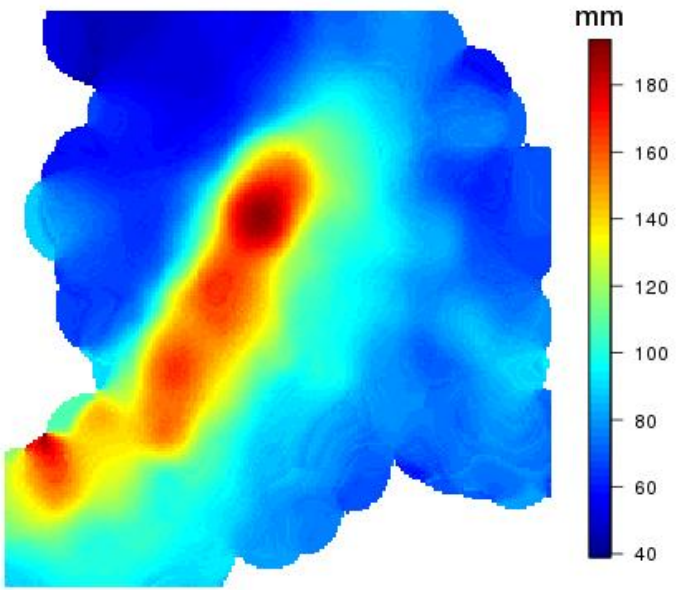
$\Longrightarrow$ Résultats de la procédure : $(h_{emp}, \alpha_{emp}) = (24, 1/(3 * 365.25))$.

Mesures de risque non extrapolées dans la région Cévennes-Vivarais

$\widehat{\mathrm{RVaR}}_n(1/(3*365.25)|x)$

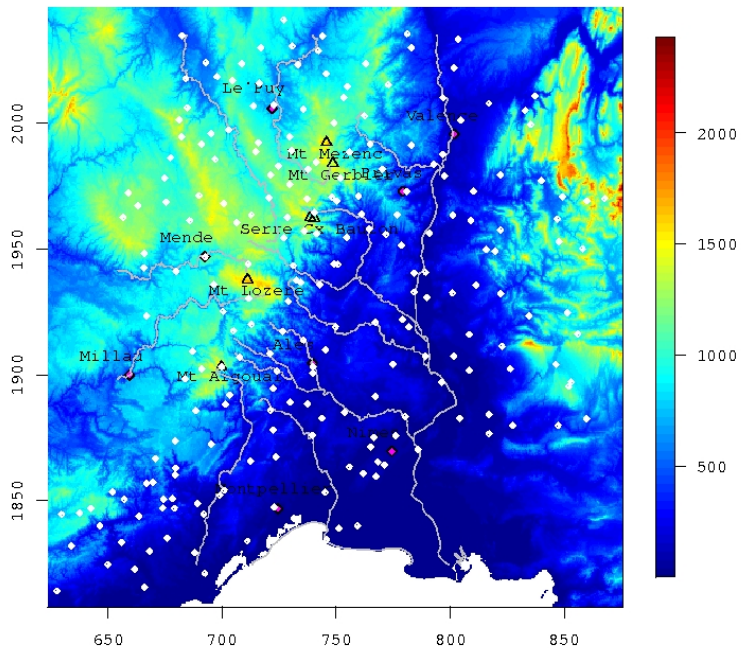


$\widehat{\mathrm{RCTE}}_n(1/(3*365.25)|x)$

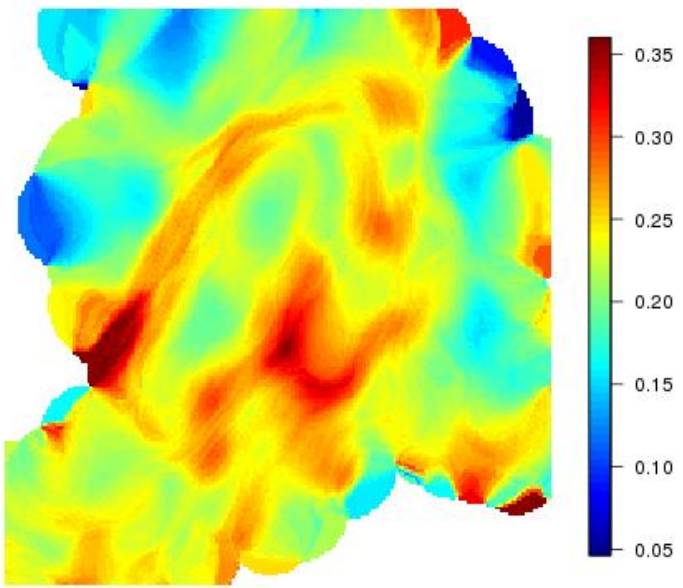


Mesures de risque extrapolées dans la région Cévennes-Vivarais

La région Cévennes-Vivarais

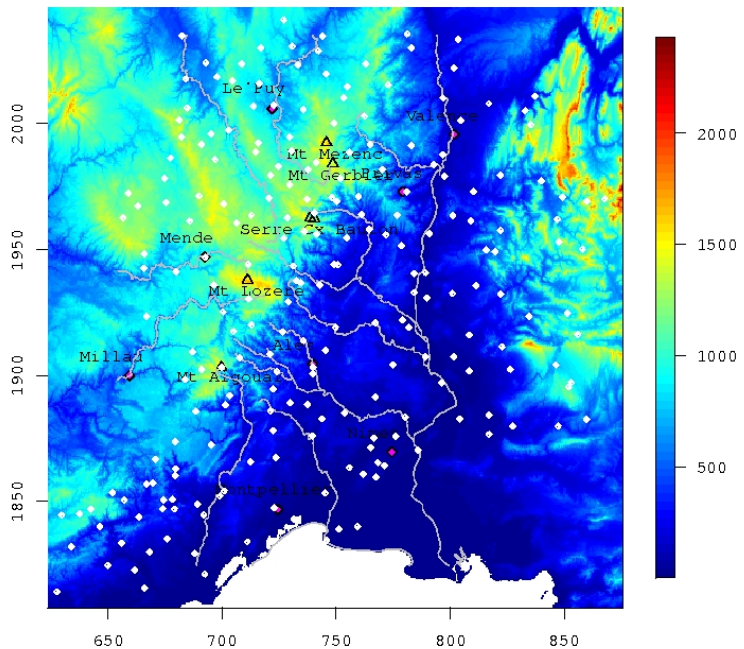


$\hat{\gamma}_{n,(1/(3*365.25))}(x)$

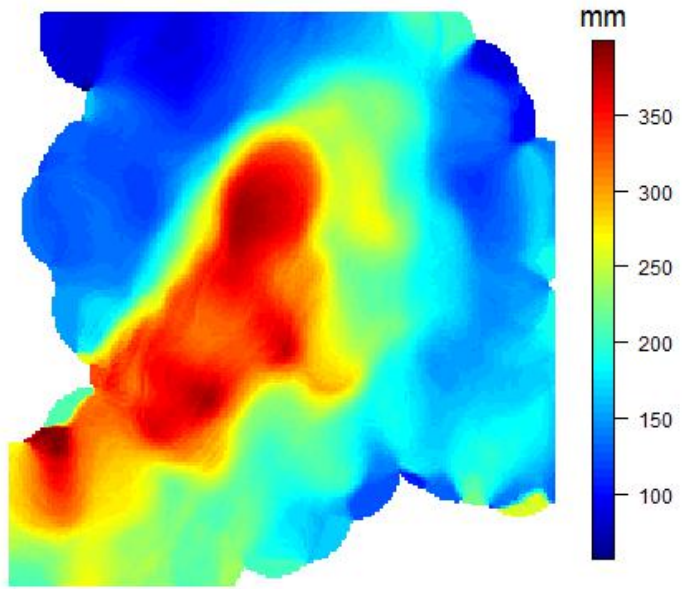


Mesures de risque extrapolées dans la région Cévennes-Vivarais

La région Cévennes-Vivarais

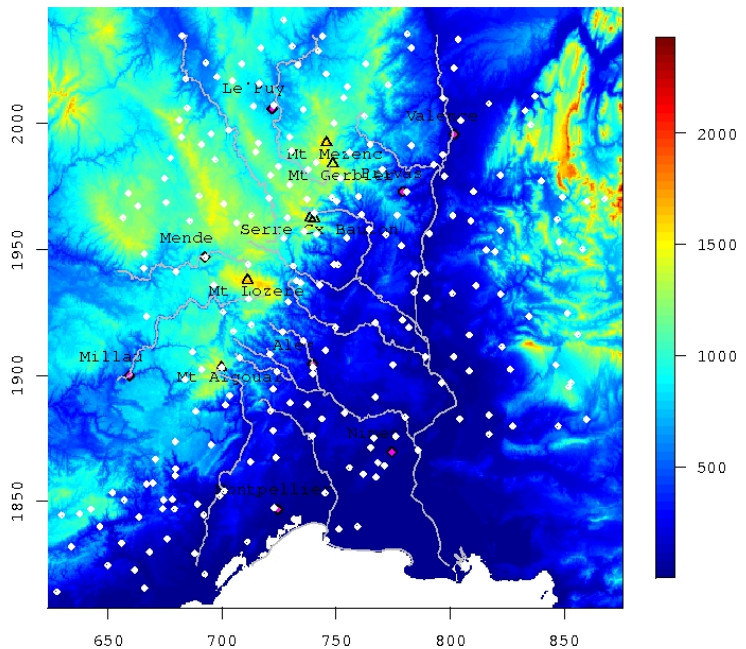


$\widehat{\mathrm{RVaR}}_n^W(1/(100*365.25)|x)$

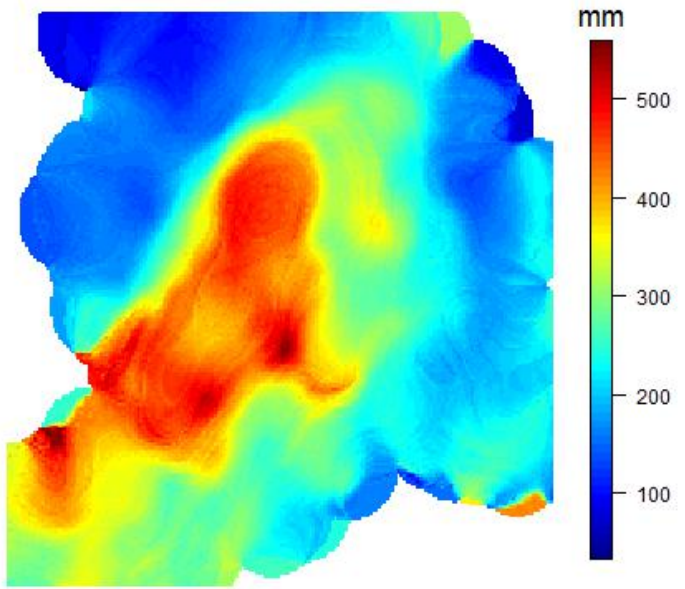


Mesures de risque extrapolées dans la région Cévennes-Vivarais

La région Cévennes-Vivarais



$\widehat{\mathrm{RCTE}}_n^W(1/(100*365.25)|x)$



Conclusions

Points clés

- Estimer VaR, CTE, CVaR et SP dans le cas de pluies extrêmes en présence d'une covariable aléatoire dans le cas d'une loi à queue lourde.
- Extrapoler ces mesures de risque à des ordres abitrairement petits.
- Application à un jeu de données réelles issu de l'hydrologie.

Commentaires

+ Cartes obtenues cohérentes du point de vu des hydrologues.
+ Nouvel outil dans la prévention des risques en hydrologie.
+ Propriétés théoriques similaires au cas univarié non extrême et/ou sans covariable.
– Fléau de la grande dimension.

Perspectives

- Estimer toutes les mesures de risque présentées dans tous les domaines d'attraction.
  $\Longrightarrow$ Cas : $\gamma(x) \in \mathbb{R}$.

Lectures

Les travaux de cette présentation ont fait l'objet d'un article de recherche qui a été publié en 2014 :

 El Methni, J., Gardes, L. and Girard, S. (2014). Nonparametric estimation of extreme risk measures from conditional heavy-tailed distributions, *Scandinavian Journal of Statistics*, **41**(4), 988–1012.

Sur ma page web personnelle vous trouverez les liens vers le Preprint sur Hal ou vers ma Thèse En Ligne TEL

 El Methni, J. (2013). Contributions à l'estimation de quantiles extrêmes. Applications à des données environnementales. *http ://tel.archives-ouvertes.fr/tel-00924293/fr/.*

Merci de votre attention

Principales références

Cai, J. and Tan, K. (2007). Optimal retention for a stop-loss reinsurrance under the VaR and CTE risk measures, *Astin Bulletin*, **37**(1), 93–112.

Daouia, A., Gardes, L., Girard, S. and Lekina, A. (2011). Kernel estimators of extreme level curves, *Test*, **20**, 311–333.

Gardes, L. and Girard, S. (2008). A moving window approach for nonparametric estimation of the conditional tail index, *Journal of Multivariate Analysis*, **99**, 2368–2388.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *The Annals of Statistics*, **3**, 1163–1174.

Embrechts, P., Kluppelberg, C. and Mikosh, T. (1997). Modelling Extremal Events, *Springer editions*.

Rockafellar, R.T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, **2**, 21–42.

Valdez, E.A. (2005). Tail conditional variance for elliptically contoured distributions, *Belgian Actuarial Bulletin*, **5**, 26–36.