Etude des données atmosphériques extrêmes

Mathis Cordier, Amandine Peillon

21 mai 2019

Table des matières

Introduction

Notre TER porte sur les quantiles des valeurs extrêmes. Le but était d'utiliser ces quantiles sur une base de données météorologiques, avec des températures et des taux d'ozone maximums pris à des endroits spécifiques aux Etats-Unis sur une année.

"*Lorsque l'on étudie un phénomène aléatoire, on s'intéresse principalement à la partie dite centrale de la loi modélisant au mieux le phénomène considéré (calcul de l'espérance, la médiane, la variance, utilisation du théorème central limite, etc...). Cependant, l'étude des "grandes" valeurs (ou de manière équivalente des "petites" valeurs) du phénomène est parfois essentielle lorsqu'il s'agit par exemple de quantifier le risque pour une compagnie d'assurance (par exemple connaître la fréquence des crues d'une rivière, etc...).*
La théorie des valeurs extrêmes propose un cadre théorique solide pour l'étude de ces grandes (ou petites) valeurs dites extrêmes. La difficulté principale réside dans l'application concrête de la théorie (estimation, inférence, etc...) puisque par nature, un évènement extrême est très peu observé. On parle d'évènement rare. L'objectif principal est l'estimation de quantiles extrêmes. "

Laurent GARDES

Nous verrons dans un premier temps la partie théorique de l'étude en introduisant la théorie des valeurs extrêmes. Cela nous permettra de comprendre les quantiles que nous allons utiliser sur nos données atmosphériques extrêmes.
Puis nous appliquerons ces quantiles dans un second temps afin de mettre en pratique les connaissances que nous aurons acquises dans la première partie.

1 Théorie des valeurs extrêmes

1.1 Quelques notions utiles

1.1.1 L'inverse généralisée d'une fonction croissante et continue à droite

Définition 1. Soit ϕ une fonction croissante et continue à droite sur $\mathbb{R}$. L'inverse généralisée de ϕ est définie par :

$$\phi^{\leftarrow}(y) := \inf\{x \mid \phi(x) \geq y\}$$

avec la convention $\inf\{\emptyset\} = +\infty$.

Dans le cas où ϕ est une fonction continue, l'inverse généralisée coïncide avec l'inverse classique. On peut alors définir l'inverse généralisée d'une fonction décroissante et continue à droite ψ par

$$\psi^{\leftarrow}(y) := \inf\{x \mid \psi(x) \leq y\}$$

L'inverse généralisée vérifie les propriétés utiles suivantes :

Proposition 2. *Soit ϕ une fonction croissante et continue à droite. L'inverse généralisée $\phi^{\leftarrow}$ est une fonction croissante et continue à gauche. On a de plus les propriétés ci-dessous :*

1. $\phi(\phi^{\leftarrow}(y)) \geq y$
2. $\phi^{\leftarrow}(y) \leq x \Leftrightarrow y \leq \phi(x)$
3. $x < \phi^{\leftarrow}(y) \Leftrightarrow y > \phi(x)$

Démonstration. Montrons tout d'abord la continuité à gauche. Pour ce faire, raisonnons par l'absurde. Supposons qu'il existe une suite (x_n) telle que $x_n \uparrow x$ avec

$$\phi^{\leftarrow}(x_n) \uparrow \phi^{\leftarrow}(x^-) < \phi^{\leftarrow}(x)$$

Il existe donc $\delta > 0$ et y tels que pour tout $n \geq 1$,

$$\phi^{\leftarrow}(x_n) < y < \phi^{\leftarrow}(x) - \delta$$

Or, $\phi^{\leftarrow}(x_n) < y$ signifie que $\inf\{z \mid \phi(z) \geq x_n\} < y$ et donc que $\phi(y) \geq x_n$ pour tout $n \geq 1$. En faisant tendre $n \longrightarrow \infty$, on a donc que $\phi(y) \geq x$, ce qui implique également que $y \geq \phi^{\leftarrow}(x)$. Ceci est bien évidemment en contradiction avec le fait que $y < \phi^{\leftarrow}(x) - \delta$.
Les propriétés 1., 2. et 3. sont des conséquences directes de la définition de l'inverse généralisée. □

1.1.2 Le quantile

Définition 3. Soient $X_1, X_2, X_3, ..., X_n,\ n \in \mathbb{N}$ des variables aléatoires indépendantes et identiquement distribuées, de fonction de répartition F. Le quantile d'ordre α de la fonction de répartition F est défini pour $\alpha \in [0,1]$ par :

$$F^{\leftarrow}(\alpha) = \inf\{x \mid F(x) \geq \alpha\}$$

La fonction quantile correspond donc à l'inverse généralisée de F au point α (les quantiles sont les valeurs réciproques de la fonction de répartition). On peut aussi expliquer les quantiles d'une variable aléatoire comme étant les valeurs que prend la variable pour des valeurs de probabilités sous le quantile considéré. Noter que le quantile d'ordre 1 est égal au point terminal de F défini par $x_F := \inf\{x \mid F(x) \geq 1\}$. En théorie des valeurs extrêmes, on s'intéressera donc à des quantiles dont l'ordre α est proche de 1. On appellera le quantile d'ordre α la quantité :

$$q(\alpha) := F^{\leftarrow}(1-\alpha) = inf\left\{x \mid \bar{F}(x) \leq \alpha\right\}$$

où $\bar{F} := 1 - F$ est appelée fonction de survie. Il est à noter que d'après le point 1. de la proposition 2,

$$\mathbb{P}(X > q(\alpha)) = 1 - F(F^{\leftarrow}(1-\alpha)) \leq \alpha$$

On a égalité si $q(\alpha)$ est un point de continuité de F.

Autrement dit, le quantile d'une distribution est un nombre $q(\alpha)$ tel qu'une proportion α des valeurs de la population soit inférieure ou égale à $q(\alpha)$.
La zone $[X_{n,n}, \infty)$ est appelée la queue de distribution de F et un quantile appartement à la queue de distribution est appelé un quantile extrême.

Définition 4. On classe les quantiles en trois catégories selon leur ordre α_n :

1. On dirait d'un quantile qu'il est classique si $n\alpha_n \longrightarrow \infty$
2. On dirait d'un quantile qu'il est intermédiaire si $n\alpha_n \longrightarrow c \in [1, \infty[$
3. On dirait d'un quantile qu'il est extrême si $n\alpha_n \longrightarrow c \in [0, 1[$

1.2 La théorie

La théorie des valeurs extrêmes est en lien avec le comportement asymptotique des valeurs extrêmes $\max(X_1, X_2, ..., X_n)$ ou $\min(X_1, X_2, ..., X_n)$ quand $n \longrightarrow \infty$.
Soit F la fonction de distribution sous-jacente et x_F son point terminal.
Ce point peut être l'infini. Alors :

$$\max(X_1, X_2, ..., X_n) \xrightarrow{P} x_F, n \longrightarrow \infty$$

En effet :

$$P(\max(X_1, ..., X_n) \leq x) = P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) = F^n(x)$$

Pour $x < x_F$, $F^n(x) \longrightarrow 0$.
Pour $x \geq x_F$, $F^n(x) \longrightarrow 1$
Par conséquent, pour obtenir une distribution asymptotique non-dégénérée, une normalisation est nécessaire.

Rappel Une fonction de répartition non-dégénérée est une fonction de répartition qui n'est pas associée à une variable constante presque sûrement.

Maintenant on suppose qu'il existe un $a_n > 0$ et un b_n réel $(n = 1, 2, ...)$ tels que :

$$\frac{\max(X_1, X_2, ..., X_n) - b_n}{a_n}$$

converge en loi vers une distribution non-dégénérée quand $n \longrightarrow \infty$. Ce qui équivaut à :

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \tag{1}$$

avec $x \in \mathbb{R}$ et G une fonction de distribution non-dégénérée.
Le but est de trouver toutes les fonctions de répartition G qui respectent (1). On appelle ces distributions les distributions de valeurs extrêmes. Ensuite, pour chaque distribution asymptotique, il faut trouver des conditions nécessaires et suffisantes sur la distribution intiale F telles que (1) soit vérifiée.
La classe des distributions F satisfaisant (1) est appelée le domaine d'attraction maximum :

$$(1) \Leftrightarrow \frac{\max(X_1, X_2, ..., X_n) - b_n}{a_n} \xrightarrow{\mathcal{L}} Y$$

où la variable aléatoire Y admet G comme fonction de répartition. F appartient donc au domaine d'attraction de G et on le note $F \in \mathcal{DA}(H)$.

On va maintenant utiliser (1). En prenant les logarithmes à gauche et à droite, on obtient la relation équivalente qui, pour chaque point de continuité x tel que $0 < G(x) < 1$,

$$\lim_{n\to\infty} n\, logF(a_n x + b_n) = logG(x) \tag{2}$$

Il en suit que $F(a_n x + b_n) \longrightarrow 1$, pour chaque x. Ainsi :

$$\lim_{n\to\infty} \frac{-logF(a_n x + b_n)}{1 - F(a_n x + b_n)} = 1$$

(2) est finalement équivalente à :

$$\lim_{n\to\infty} n(1 - F(a_n x + b_n)) = -logG(x)$$

Avec l'égalité :

$$\lim_{n\to\infty} \frac{1}{n(1 - F(a_n x + b_n))} = \frac{1}{-logG(x)} \tag{3}$$

Proposition 5. *Soient F et H deux fonctions de répartitions ayant des queues proportionnelles i.e. ayant même point terminal x_F et avec*

$$\lim_{x\to x_F} \frac{1 - F(x)}{1 - H(x)} = c \in]0, \infty[$$

Si F appartient au domaine d'attraction d'une fonction de répartition non dégénérée G avec (a_n) et (b_n) comme suite de normalisation, alors H appartient au domaine d'attraction de $G^{1/c}$ avec les mêmes suites de normalisation (a_n) et (b_n).

Nous allons utiliser par la suite la fonction inverse sur cette condition.

Lemme 6. *Soient f_n une suite de fonction non décroissantes et g une fonction non décroissante. On suppose que pour tout x dans un intervalle ouvert (a,b), il existe un point de continuité de g.*

$$\lim_{n\to\infty} f_n(x) = g(x) \tag{4}$$

Soient $f_n^{\leftarrow}$, $g^{\leftarrow}$ les fonctions inverses de f_n et g. Pour chaque x dans l'intervalle (g(a), g(b)) qui est un point de continuité de $g^{\leftarrow}$, on a :

$$\lim_{n\to\infty} f_n^{\leftarrow}(x) = g^{\leftarrow}(x) \tag{5}$$

On reprend (3) à laquelle on applique le Lemme 6. Soit U la fonction inverse de $1/(1 - F)$. $U(t)$ est défini pour tout $t > 1$. Il en suit que (3) est équivalente à :

$$\lim_{n\to\infty} \frac{U(nx) - b_n}{a_n} = G^{\leftarrow}(e^{-1/x}) =: D(x) \tag{6}$$

pour chaque x positif.

Le théorème suivant permet de rendre (6) plus flexible.

Théorème 7. *Soient $a_n > 0$ et b_n des suites de réels et G une fonction de distribution non dégénérée. Les affirmations suivantes sont équivalentes :*

1. *Pour tout x point de continuité de G :*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x)$$

2.

$$\lim_{t\to\infty} t(1 - F(a(t)x + b(t))) = -logG(x) \tag{7}$$

pour tout x point de continuité de G pour tout $0 < G(x) < 1$, $a(t) := a_{[t]}$ *et* $b(t) := b_{[t]}$, *avec* $[t]$ *la partie entière de t.*

3.

$$\lim_{t\to\infty} \frac{U(tx) - b(t)}{a(t)} = D(x) \tag{8}$$

pour tout $x > 0$ *point de continuité de* $D(x) = G^{\leftarrow}(e^{-1/x})$, $a(t) := a_{[t]}$ *et* $b(t) := b_{[t]}$.

Avec toutes les informations que nous venons de voir, nous sommes en mesure d'identifier la classe des distributions non dégénérées qui peuvent se comporter comme une limite dans la relation (1). Cette classe est appelée la classe des distributions des valeurs extrêmes.

Théorème 8. *La classe des distributions de valeurs extrêmes est* $G_\gamma(ax + b)$ *avec* $a > 0$, *b réel, où :*

$$G_\gamma(x) = exp(-(1 + \gamma x)^{-1/\gamma}), \quad 1 + \gamma x > 0 \tag{9}$$

avec γ *réel et où pour* $\gamma = 0$, *la partie asymptotique à droite se comporte comme* $exp(-e^{-x})$.
Ce qui est équivalent à dire que :

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}$$

pour tout $x > 0$ *et avec* $U(t) = F^{\leftarrow}(1 - 1/t) = G(1 - 1/t)$.

Ici, nous avons introduit γ qui est l'indice de valeur extrême.
En fonction du signe de γ, G non dégénérée et de même type que l'une des fonctions de répartitions suivantes :

1. Fonction de répartition de Fréchet définie pour $\gamma > 0$ par

$$\Psi_\gamma^{(F)}(x) = \begin{cases} 0 & si\ x < 0 \\ \exp(-(x)^{-1/\gamma}) & si\ x \geq 0 \end{cases}$$

Le domaine de Fréchet est l'ensemble des lois à "queues lourdes", $\bar{F}(x) \longrightarrow 0$ comme une puissance de x lorsque $x \longrightarrow \infty$.

2. Fonction de répartition de Weibull définie pour $\gamma < 0$ par

$$\Psi_{\gamma}^{(W)}(x) = \begin{cases} \exp(-(-x)^{-1/\gamma}) & si\ x \leq 0 \\ 1 & si\ x > 0 \end{cases}$$

Le domaine de Weibull est l'ensemble des lois à "queues finies", $\bar{F}(x) = 0$ pour $x > x_F$.

3. Fonction de répartition de Gumbel définie pour $\gamma = 0$ par

$$\Psi^{(G)}(x) = \exp(-e^{-x})\ si\ x \in \mathbb{R}$$

Le domaine de Gumbel est l'ensemble des lois à "queues légères", $\bar{F}(x) \longrightarrow 0$ exponentiellement vite lorsque $x \longrightarrow \infty$.

Pour une queue lourde de F, une hypothèse commune est que pour $\gamma > 0$,

$$1 - F(y) = y^{-1/\gamma} l(y) \tag{10}$$

avec $y \longrightarrow \infty$ et où l est une fonction à variations lentes satisfaisant la condition $l(ty)/l(y) \longrightarrow 1$ pour $y \longrightarrow \infty$ et pour tout $t > 0$. Cette condition est équivalente sur la fonction quantile :

$$Q(1 - 1/y) = y^{\gamma} L(y) \tag{11}$$

où L est une fonction à variations lentes liée à l.
Par conséquent, quand $x \longrightarrow 1$ et $x_n \longrightarrow 1$, $Q(x_n)/Q(x) \sim \{(1-x)/(1-x_n)\}^{\gamma}$.
C'est par ailleurs la base de l'estimateur des valeurs extrêmes de Weissman(1978) :

$$\widehat{Q}(x_n) = y_{n-k,n} \left[k/\left\{n(1-x_n)\right\}\right]^{\widehat{\gamma}}$$

où $\widehat{\gamma}$ est un estimateur de γ.

1.3 Estimateurs à noyau des courbes de niveau extrême

Soit (X_i, Y_i), $i = 1, ..., n$, une copie indépendante d'une paire aléatoire (X, Y) dans $\mathbb{R}^p \times \mathbb{R}$. Nous abordons le problème de l'estimation des courbes de niveau extrême, définies comme les graphes des fonctions $x \in \mathbb{R}^p \mapsto q(\alpha_n \mid x) \in \mathbb{R}$ vérifiant $\mathbb{P}(Y > q(\alpha_n \mid x) \mid X = x) = \alpha_n$, où $\alpha_n \to 0$ quand $n \to \infty$.

La fonction conditionnelle de survie de Y sachant $X = x$ est notée $\bar{F}(y \mid x) = \mathbb{P}(Y > y \mid X = x)$ et la fonction de densité de probabilité de X est notée g. L'estimateur à noyau de $\bar{F}(y \mid x)$ est défini pour tout $(x, y) \in \mathbb{R}^p \times \mathbb{R}$ par :

$$\hat{\bar{F}}_n(y \mid x) := \sum_{i=1}^{n} K_h(x - X_i) \mathbb{1}\left\{Y_i > y\right\} / \sum_{i=1}^{n} K_h(x - X_i) \tag{12}$$

où $\mathbb{1}$ est la fonction indicatrice et $h = h_n$ est une suite réelle telle que $h \longrightarrow 0$ quand $n \longrightarrow \infty$. On introduit aussi ici $K_h(t) = K(t/h)/h^p$ où K est une densité de probabilité dans $\mathbb{R}^p$. Dans ce contexte, h est la constante qui définie la fenêtre de lissage.
La distribution asymptotique de $\hat{\bar{F}}_n$ est établie lors de l'estimation des probabilités du domaine de Weibull i.e. quand $y = y_n$ tend à l'infini avec un échantillon de taille n. De la même manière, les estimateurs à noyau de quantiles conditionnels $q(\alpha \mid x)$ sont définis via la fonction inverse de $\hat{\bar{F}}_n(. \mid x)$:

$$\hat{q}_n(\alpha \mid x) := \hat{\bar{F}}_n^{\leftarrow}(\alpha \mid x) = inf\left\{t, \hat{\bar{F}}_n(t \mid x) \leq \alpha\right\} \tag{13}$$

pour tout $\alpha \in (0, 1)$.
La distribution asymptotique de $\widehat{q}_n$ est étudiée lors de l'estimation des quantiles extrêmes, i.e., quand $\alpha = \alpha_n$ tend vers 0 quand l'échantillon de taille n tend vers l'infini.

Pour la suite, on va affirmer que la fonction conditionnelle de survie satisfait :

(F.1) $\quad \bar{F}(y \mid x) = y^{-1/\gamma(x)} l(y \mid x)$

où γ est l'indice des valeurs extrêmes conditionnel, une fonction positive. Pour x fixé, $l(. \mid x)$ est une fonction à variations lentes à l'infini, i.e., pour tout $\lambda > 0$,

$$\lim_{y \to \infty} \frac{l(\lambda y \mid x)}{l(y \mid x)} = 1 \tag{14}$$

(F.1) revient à supposer que la distribution conditionnelle de Y sachant $X = x$ est dans le domaine d'attraction maximal de Fréchet.
(F.1) permet aussi de dire que $\bar{F}(\cdot \mid x)$ tend à l'infini avec $-1/\gamma(x)$.
On peut considérer que

(F.2) $l(\cdot \mid x)$ est normalisée

sans perdre d'informations car les fonctions à variations lentes ne sont intéressantes que d'un point de vue asymptotique. Cette fonction à variations lentes peut être écrite de la façon suivante :

$$l(y \mid x) = c(x) \exp\left(\int_1^y \frac{\varepsilon(u \mid x)}{u} \, \mathrm{d}u\right) \tag{15}$$

où $c(\cdot)$ est une fonction positive, et $\varepsilon(y \mid x) \longrightarrow 0$ quand $y \longrightarrow \infty$. Ainsi, $l(\cdot \mid x)$ est différentiable et la fonction auxiliaire est donnée par $\varepsilon(y \mid x) = y l'(y \mid x)/l(y \mid x)$.

Cette fonction joue un rôle important dans la théorie des valeurs extrêmes puisque qu'elle donne la vitesse de convergence de $\lim_{y\to\infty} \frac{l(\lambda y|x)}{l(y|x)} = 1$ et plus généralement le biais des estimateurs des valeurs extrêmes.
Ici, nous nous limitons à supposer que

(F.3) $|\varepsilon(\cdot \mid x)|$ est continue et finalement non croissante.

Certaines conditions Lipschitziennes sont ici requises. Pour tout $(x, x') \in \mathbb{R}^p \times \mathbb{R}^p$, la distance Euclidienne entre x et x' est notée $d(x, x')$ et les hypothèses suivantes sont introduites :

(H.1) Il existe $c_\gamma > 0$ tel que $\left|\frac{1}{\gamma(x)} - \frac{1}{\gamma(x')}\right| \leq c_\gamma d(x, x')$

(H.2) Il existe $c_l > 0$ et $y_0 > 1$ tels que $\sup_{y\geq y_0} \left|\frac{log\, l(y|x)}{logy} - \frac{log\, l(y|x')}{logy}\right| \leq c_l d(x, x')$

(H.3) Il existe c_g tel que $|g(x) - g(x')| \leq c_g d(x, x')$

La dernière hypothèse est standard dans le cadre de l'estimation du noyau.

(K) K est une densité de probabilité bornée sur $\mathbb{R}^p$, le support S étant inclus dans la boule unité de $\mathbb{R}^p$.

1.3.1 Quelques résultats importants

Voyons d'abord l'estimation des probabilités à queue finie $\bar{F}(y_n \mid x)$ quand $y_n \longrightarrow \infty$. Le résultat suivant fournit des conditions suffisantes pour la normalité asymptotique de $\hat{\bar{F}}(y_n \mid x)$.

Théorème 9. *On suppose (F.1), (L.1), (L.2), (L.3), et (K) retenues.*
Introduisons :
- $0 < a_1 < a_2 < ... < a_J \in \mathbb{R}$ *où* J *est un entier positif.*
- $y_n \longrightarrow \infty$ *tel que* $nh^p\bar{F}(y_n \mid x) \longrightarrow \infty$ *et* $nh^{p+2}log^2(y_n)\bar{F}(y_n \mid x) \longrightarrow 0$ *quand* $n \longrightarrow \infty$.
- $y_{n,j} = a_j y_n$ *pour* $j = 1, ..., J$.

Puis, pour tout $x \in \mathbb{R}^p$ *tel que* $g(x) > 0$, *le vecteur aléatoire*

$$\left\{\sqrt{nh^p\bar{F}(y_n \mid x)}\left(\frac{\hat{\bar{F}}_n(y_{n,j} \mid x)}{\bar{F}(y_{n,j} \mid x)} - 1\right)\right\}_{j=1,...,J}$$

est asymptotiquement gaussien, centré avec comme matrice de covariance $\frac{\|K\|_2^2}{g(x)}C(x)$ *où* $C_{j,j'}(x) = a_{j\wedge j'}^{1/\gamma(x)}$ *pour* $(j,j') \in \{1,...,J\}^2$.

Occupons nous maintenant de l'estimation du quantile extrême $q(\alpha_n \mid x)$ quand $\alpha_n \longrightarrow 0$ pour $n \longrightarrow \infty$. La normalité asymptotique de $\hat{q_n}(\alpha_n \mid x)$ peut être établie sous des conditions similaires :

Théorème 10. *On suppose (F.1), (F.2), (L.1), (L.2), (L.3), et (K) retenues. Introduisons :*

- $0 < \tau_J < \tau_{J-1} < ... < \tau_1 \in \mathbb{R}$ *où* J *est un entier positif.*
- $\alpha_n \longrightarrow 0$ *tel que* $nh^p\alpha_n \longrightarrow \infty$ *et* $nh^{p+2}\alpha_n log^2(\alpha_n) \longrightarrow 0$ *quand* $n \longrightarrow \infty$.
- $\alpha_{n,j} = \tau_j\alpha_n$ *pour* $j = 1, ..., J$.

Puis, pour tout $x \in \mathbb{R}^p$ *tel que* $g(x) > 0$, *le vecteur aléatoire*

$$\left\{ \sqrt{nh^p\alpha_n} \left(\frac{\hat{q_n}(\alpha_{n,j} \mid x)}{q(\alpha_{n,j} \mid x)} - 1 \right) \right\}_{j=1,...,J}$$

est asymptotiquement gaussien, centré avec comme matrice de covariance $\|K\|_2^2 \frac{\gamma^2(x)}{g(x)}\Sigma$ *où* $\Sigma_{j,j'} = 1/\tau_{j\wedge j'}$ *pour* $(j,j') \in \{1,...,J\}^2$.

Introduisons une version de l'estimateur à noyau de Hill pour $\gamma(x)$:

$$\hat{\gamma}_n^H(x) = \sum_{j=1}^{J} \left[\log \hat{q}_n(\tau_j\alpha_n \mid x) - \log \hat{q}_n(\alpha_n \mid x) \right] / \sum_{j=1}^{J} \log(1/\tau_j)$$

où (τ_j) est une séquence décroissante de poids.

Corollaire 11. *On suppose (F.1), (F.2), (F.3), (L.1), (L.2), (L.3) et (K) retenues. Soit* $1 = \tau_1 > \tau_2 > ... > \tau_J > 0$ *où* J *est un entier positif. Si* $\sigma_n \longrightarrow 0$, $\sigma_n^{-1}h \log \alpha_n \longrightarrow 0$ *et* $\sigma_n^{-1}\varepsilon(q(\alpha_n \mid x) \mid x) \longrightarrow 0$ *quand* $n \longrightarrow \infty$, *puis, pour tout* $x \in \mathbb{R}^p$ *tel que* $g(x) > 0$, $\sigma_n^{-1}(\hat{\gamma}_n^H(x) - \gamma(x))$ *converge vers un vecteur aléatoire centré Gaussien avec variance* $\|K\|_2^2\gamma^2(x)V_J/g(x)$ *où :*

$$V_J = \left(\sum_{j=1}^{J} \frac{2(J-j)+1}{\tau_j} - J^2 \right) / \left(\sum_{j=1}^{J} \log(1/\tau_j) \right)^2$$

L'estimateur à noyau de quantiles extrêmes $\hat{q}_n(\alpha_n \mid x)$ requiert des conditions strictes sur l'ordre α_n du quantile, car il ne peut pas extrapoler par construction au-delà de l'observation maximale dans la boule $B(x,h)$. Pour surmonter cette limitation, un estimateur de type Weissman peut être dérivé :

$$\hat{q}_n^W(\beta_n \mid x) = \hat{q}_n(\alpha_n \mid x)(\alpha_n/\beta_n)^{\hat{\gamma}_n(x)}$$

Ici $\hat{q}_n(\alpha_n \mid x)$ est l'estimateur à noyau de quantile extrême jusqu'ici considéré et $\hat{\gamma}_n(x)$ est un estimateur de l'indice conditionnel de queue $\gamma(x)$.
Le facteur d'extrapolation $(\alpha_n/\beta_n)^{\hat{\gamma}_n(x)}$ permet d'estimer les quantiles extrêmes d'ordre β_n arbitraires petits.

2 Programmation des quantiles

Afin de mettre en pratique les quantiles vus précédemment, nous avions à notre disposition une base de données dans laquelle se trouvent des températures maximales et des taux d'ozone maximums, pris à des endroits différents durant toute une année. Le but est de pouvoir représenter les variations de température et de taux d'ozone dans une région sur l'année. Pour cela, nous avons utilisé le logiciel R.
Toutes les explications seront annoncées dans le cas général. Nous les accompagnerons avec le code R correspondant à notre base de données.
Pour le code R, nous utiliserons les packages dplyr, ggplot2 et RColorBrewer.

2.1 Optimisation du calcul

Soit un tableau de données :

$$\begin{array}{c} \begin{array}{cc|ccc} Y & & X_1 & \dots & X_p \end{array} \\ \left[\begin{array}{cc|ccc} \vdots & & \vdots & & \vdots \end{array}\right] \end{array}$$

Nous chercherons à optimiser le temps de calcul. Nous avons une variable quantitative à expliquer Y et une variable explicative X de dimension strictement supérieure à 1. Nous souhaitons regarder l'évolution de Y lorsqu'un unique élément du vecteur X varie et que tous les autres éléments de X sont fixés. En effet, si l'on cherche à représenter l'évolution sur un élément, il n'est pas raisonnable de calculer pour chaque valeur observée, les quantiles classiques et quantiles de Weissman. On répéterait alors plusieurs fois les mêmes calculs et le temps de calcul serait proportionnel au nombre de valeurs observées.
L'objectif des deux premières sous-parties est de calculer $\hat{\tilde{F}}_n$ pour toutes les valeurs de l'élément variant de X.

Nous diviserons le processus en deux parties : la première peut se faire sans connaître la valeur des éléments fixés, la seconde nécessite que ces éléments soient connus. On peut alors modifier les éléments fixés sans avoir à répéter la procédure faite en première partie.

2.1.1 Calculs préliminaires

Introduction des notations : Soient ci-dessous les différentes notations que nous utiliserons dans cette partie :

1. N le nombre d'observations (lignes) de la base de données,

2. M_X la matrice de taille $N \times \dim(X)$ avec sur chaque ligne le vecteur X correspondant à l'observation, on considèrera maintenant $\dim(X) = p$,
3. M'_X la matrice M_X privée de la colonne correspondant à l'élément variant,
4. J le vecteur normalisé et ordonné des points que l'on représentera en abscisses,
5. *champ* toutes les valeurs distinctes de Y, la variable à expliquer,
6. $\mathbb{U}$ l'intervalle $[-1, 1]$, on notera $\mathbb{U}^+$ l'intervalle $[0, 1]$,
7. β le risque du quantile des valeurs extrêmes, il est fixé par l'utilisateur.

Normalisations : Nous créons une nouvelle base de données identique à l'originale, puis nous normalisons les colonnes des variables explicatives. Soit L_X la liste des colonnes de la base de données correspondant aux variables explicatives. Alors nous obtenons :

$$(M_X)_{ij} = \frac{\max(L_X)_j - (M_X)_{ij}}{\max(L_X)_j - \min(L_X)_j}$$

$\forall (i, j) \in [\![1, N]\!] \times [\![1, p]\!]$

```
normalize = function(df){
  df2 = df
  for (i in 3:5){
    df2[,i] = (df2[,i]-min(df2[,i])) / (max(df2[,i])-min(df2[,i]))
  }
  return(df2)
}

normalize_val = function(x,i){
  return((x-min(df_base[,i])) / (max(df_base[,i])-min(df_base[,i])) )
}
```

Fonctions sur R : normalize pour normaliser le tableau de données et normalize_val pour normaliser une valeur.

Restriction du champ de Y : Pour les calculs de quantile, nous cherchons des valeurs maximales dans la queue de distribution de F_n. Alors nous n'avons pas besoin d'une grande précision pour les valeurs observées les plus faibles. De plus, il se peut qu'un grand nombre de valeurs soient observées très peu fréquemment, on peut alors ne pas conserver ces valeurs dans le champ s'il existe une valeur proche apparaissant beaucoup plus fréquemment.
En restreignant ces champs, nous perdons un peu d'information notamment de précision mais nous gagnons considérablement en temps de calcul.

Calcul des K_h pour toutes les valeurs de l'élément variant de M_X : On note V la colonne $(M_X)_j$ correspondant à l'élément variant de X.
On crée une grille en répétant N fois le vecteur J et N_j fois le vecteur d :

$$M_K = \begin{bmatrix} J \\ J \\ \vdots \\ J \end{bmatrix} - \begin{bmatrix} V & V & \dots & V \end{bmatrix} \in \mathcal{M}_{N,N_J}(\mathbb{U})$$

On applique ensuite K_h sur tous les éléments de la matrice M_K :

$$(M_K)_{ij} = K_h((M_K)_{ij}) \in \mathcal{M}_{N,N_J}(\mathbb{U}^+)$$

On obtient alors sur un quadrillage des valeurs de K_h en fonction du temps.

```
K = function(x){
  return(15/16*(1-x^2)^2)
}

Kh = function(x){
  tmp = x/h
  tmp[abs(tmp)>1]=1
  tmp = K(tmp)
  return(tmp/h)
}
```

Fonctions sur R : K est la fonction de répartition et Kh la fonction de calcul de K_h.

Calcul des indicatrices : Soit $ind \in \mathcal{M}_{N,N_{champ}}(\mathbb{R})$
$\forall i,j$, ind_{ij} est un booléen, chaque colonne de la matrice ind est l'indicatrice correspondant à la valeur du champ.
Cette matrice représente $\mathbb{I}_{\{Y_i>y\}}$ dans le calcul de $\hat{\bar{F}}_n$.
Pour calculer ind nous créons deux matrices $I_1, I_2 \in \mathcal{M}_{N,N_{champ}}(\mathbb{U}^+)$:

$$I_1 = \begin{bmatrix} V & V & \dots & V \end{bmatrix}, I_2 = \begin{bmatrix} champ \\ champ \\ \vdots \\ champ \end{bmatrix}$$

Puis

$$\forall i,j, \quad ind_{ij} = \begin{cases} 1 & si \quad (I_1)_{ij} > (I_2)_{ij} \\ 0 & sinon \end{cases}$$

Alors :

$$ind = \begin{bmatrix} 1 & 1 & 0 & \dots \\ 1 & 0 & 1 & \\ \vdots & & & \ddots \end{bmatrix}$$

```
indic_oz = function(N,champ){
  Y = matrix(rep(df$oz,length(champ)),ncol=length(champ))
  tmp = t(matrix(rep(champ,N),ncol=N))
  return(Y>tmp)
}

indic_tp = function(N,champ){
  Y = matrix(rep(df$tp,length(champ)),ncol=length(champ))
  tmp = t(matrix(rep(champ,N),ncol=N))
  return(Y>tmp)
}
```

Fonctions sur R : indic_ oz permet d'obtenir l'indicatrice pour la variable réponse ozone et indic_ tp permet d'obtenir celle pour la variable réponse tmp.

```
ind_tp = indic_tp(N,champ_tp)
ind_oz = indic_oz(N,champ_oz)
```

Fonctions sur R : applications des fonctions ci-dessus pour la température et le taux d'ozone.

2.1.2 Calculs dépendants des éléments fixés de X

Normalisations : Nous recevons des valeurs réelles correspondant aux mesures étudiées. Il faut alors normaliser ces valeurs pour les adapter aux calculs.
Notons x le vecteur des éléments fixés et L_X la liste des colonnes de la matrice M'_X. Alors nous obtenons :

$$\forall i \in [\![1, \dim(x)]\!], \quad x_i = \frac{\max(L_X)_i - x_i}{\max(L_X)_i - \min(L_X)_i}$$

Soit U la liste des colonnes de M_X qui correspondent à la variable fixée. Nous calculons maintenant pour tous les x_i les vecteurs :

$$v_i = \begin{bmatrix} x_i \\ x_i \\ \vdots \\ \vdots \\ x_i \end{bmatrix} - \begin{bmatrix} \boxed{U_i} \end{bmatrix}$$

Nous obtenons alors le vecteur v des valeurs de K_h pour les éléments fixés, tel que :

$$\forall j \in [\![1, N]\!],\ \ v_j = \prod_{i=1}^{\dim(x)} K_h((v_i)_j)$$

v contient naturellement beaucoup de valeurs nulles. En effet, v_j est non nul seulement si tous les x_i sont assez proches des $(U_i)_j$. Cette proximité est définie par la valeur de h qui définit la fenêtre de lissage.
Si v est le vecteur nul, aucune donnée n'est dans la fenêtre de lissage de x. On ne peut alors pas calculer de quantile.
Alors on obtient la matrice des éléments résultants de K_h par l'opération :

$$\forall (i, j) \in [\![1, N]\!] \times [\![1, N_J]\!], \quad (M_K)_{ij} = v_i \times (M_K)_{ij}$$

Puis $M_K = M_K^T \in \mathcal{M}_{N_J,N}(\mathbb{U}^+)$
Alors, on note S_K le vecteur résultant de la somme sur les lignes de M_K, soit

$$(S_K)_i = \sum_{j=1}^{N} (M_K)_{ij}$$

On passe ensuite M_K dans l'indicatrice :

$$M_K = M_K \cdot ind$$

Où $\cdot$ représente le produit matriciel. On a alors $M_K \in \mathcal{M}_{N_J,N_{champ}}(\mathbb{U}^+)$
Nous pouvons maintenant calculer la matrice F correspondant à $\hat{\bar{F}}_n$ par l'opération :

$$\forall (i, j) \in [\![1, N_J]\!] \times [\![1, N_{champ}]\!], \quad F_{ij} = (M_K)_{ij}/(S_K)_i$$

F représente l'évolution de $\hat{\bar{F}}_n$ en fonction des variables explicatives et de la variable à expliquer. Ainsi :

$$\forall (i, j) \in [\![1, N_J]\!] \times [\![1, N_{champ}]\!], \quad F_{ij} = \hat{\bar{F}}_n(champ_j|X_i)$$

```
# Températures
Ki_tp = M%*%ind_tp
Fn_tp = Ki_tp/SK
Qtp = apply(Fn_tp,1,qnW_tp)


# Ozone
Ki_oz = M%*%ind_oz
Fn_oz = Ki_oz/SK
Qoz = apply(Fn_oz,1,qnW_oz)
```

Fonctions sur R : fonctions qui retournent $\hat{\bar{F}}_n$ pour la température et le taux d'ozone.

2.2 Calcul vectoriel des quantiles

2.2.1 Calcul de $\hat{q}_n$

Soit F, la matrice issue de l'optimisation. On cherche alors

$$\forall i,\ \inf\{j, F_{ij} \leqslant \alpha\}$$

On obtient un vecteur, noté qc, de taille N_J représentant l'évolution du quantile en fonction de la valeur de J.

```
qnc = function(alpha,Fn,champ){
  return(champ[min(which(Fn<=alpha))])
}
```

Fonctions sur R : fonctions qui retournent $\hat{q}_n$ pour la température et le taux d'ozone.

2.2.2 Calcul de $\hat{\gamma}_n^H$

Soient $\hat{\gamma}_n^H$ l'estimateur de Hill de γ et $\tau = \left[1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{9}\right]^T$. Nous calculons $\hat{q}_n$ pour le risque $\tau_i\alpha$ pour chaque valeur τ_i de τ. Nous appliquons ensuite la formule théorique du calcul de $\hat{\gamma}_n^H$.

```
gam_ncH = function(champ,Fn){
  tau = 1:9
  Sdiv = sum(log(tau))
  Snum = c()
  for (t in alpha/tau){
    Snum = c(Snum,qnc(t,Fn,champ))
  }
  return(sum(log(Snum)-log(Snum[1]))/Sdiv)
}
```

Fonctions sur R : fonctions qui retournent $\hat{\gamma}_n^H$ pour la température et les taux d'ozone.

2.2.3 Calcul de $\hat{q}_n^W$

On cherche à calculer $\hat{q}_n^W$ l'estimateur de Weissman pour le quantile des valeurs extrêmes. Nous avons le vecteur qc obtenu précédemment et nous notons g_i la valeur de $\hat{\gamma}_n^H$ correspond à qc_i. Pour toute valeur qc_i de qc, on calcule la valeur qw_i de qw.

$$qw_i = qc_i \left(\frac{\alpha}{\beta}\right)^{g_i}$$

Ainsi on obtient deux vecteurs qc et qw de taille N_J représentant l'évolution de ces deux quantiles en fonction de la valeur de J.

```
qnW = function(Fn,beta,champ){
  res = qnc(alpha,Fn,champ)*(alpha/beta)^gam_ncH(champ,Fn)
  return(res)
}
```

Fonctions sur R : fonctions qui retournent $\hat{q}_n^W$ pour la température et le taux d'ozone.

3 Représentations graphiques

3.1 Introduction à notre base de données

	oz	tp	lat	lng	day
9108	0.000	15.0000000	29.76800	-95.22058	331
66133	0.000	28.3333333	37.28329	-83.20932	276
72350	0.000	28.3333333	38.23887	-82.98810	210
78687	0.000	10.5555556	38.92185	-77.01318	344
78699	0.000	10.0000000	38.92185	-77.01318	345
79894	0.000	3.3333333	39.05528	-76.87833	360
82195	0.000	8.3333333	39.31083	-76.47444	344
82202	0.000	9.4444444	39.31083	-76.47444	345
86189	0.000	5.0000000	39.75118	-104.98762	98
86230	0.000	16.6666667	39.75176	-105.03068	253
88213	0.000	7.7777778	39.92304	-75.09762	344
90836	0.000	6.1111111	40.59664	-74.12525	344
91311	0.000	13.3333333	40.73614	-73.82153	1
91570	0.000	12.2222222	40.73614	-73.82153	3

Représentation du tableau de données

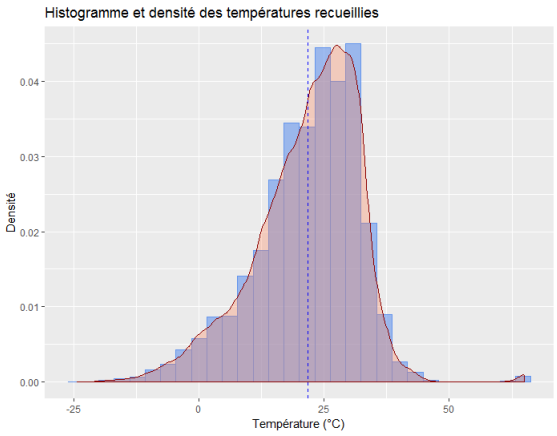
La base de données sur laquelle nous travaillons est composée de deux variables à expliquer (température et taux d'ozone) et trois variables explicatives (latitude, longitude et journée). Il y a 126 969 observations la composant.

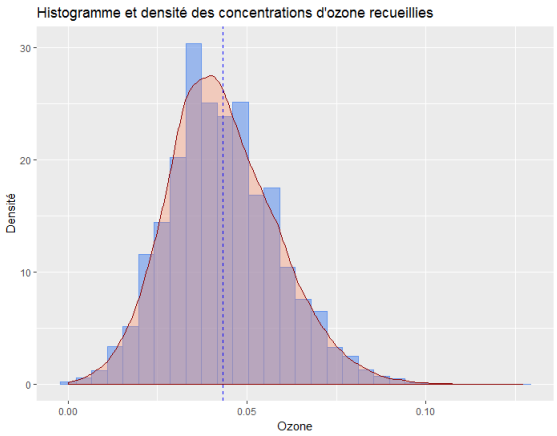
```
df = df[df$lat<62,] # Suppression de l'Alaska
df = df[order(df$oz),] # Tri par température
df$tp = (df$tp-32)*5/9 # Passage en Celsius
df_base = df
df = normalize(df)
```

Fonctions sur R : df_ base est la base de donnée initiale, df est celle que l'on a normalisé.

Dans les lignes de code ci-dessus, il est marqué que nous avons supprimé l'Alaska. En effet, il n'y avait qu'une borne en Alaska et donc ce n'était pas représentatif de la région. Nous avons aussi passer les températures en Celsius par convention.

Nous pouvons représenter un aperçu de la densité des températures et des concentrations d'ozone par les graphiques suivants :





Une fois les variables explicatives normalisées, nous obtenons *df* :

	oz	tp	lat	lng	day
9108	0.000	15.0000000	0.1703551	0.52137247	0.906593407
66133	0.000	28.3333333	0.5007062	0.73434592	0.755494505
72350	0.000	28.3333333	0.5427108	0.73826841	0.574175824
78687	0.000	10.5555556	0.5727326	0.84421063	0.942307692
78699	0.000	10.0000000	0.5727326	0.84421063	0.945054945
79894	0.000	3.3333333	0.5785978	0.84660159	0.986263736
82195	0.000	8.3333333	0.5898313	0.85376300	0.942307692
82202	0.000	9.4444444	0.5898313	0.85376300	0.945054945
86189	0.000	5.0000000	0.6091879	0.34819159	0.266483516
86230	0.000	16.6666667	0.6092132	0.34742816	0.692307692
88213	0.000	7.7777778	0.6167423	0.87817572	0.942307692
90836	0.000	6.1111111	0.6463517	0.89541691	0.942307692
91311	0.000	13.3333333	0.6524838	0.90080221	0.000000000
91570	0.000	12.2222222	0.6524838	0.90080221	0.005494505

Les individus ici présentés ont été recueillis par 426 stations dont nous avons récupéré les coordonnées géographiques grâce à la fonction ci-dessous :

```
DF = dplyr::summarise(dplyr::group_by(df_base,lat), lng=mean(lng))
DF = cbind(rownames(DF),DF)
colnames(DF)[1]="station"
DF$station = paste("Station",DF$station)
```

	station	lat	lng
1	Station 1	25.89252	-97.49383
2	Station 2	26.22621	-98.29107
3	Station 3	26.30986	-98.18310
4	Station 4	27.42670	-97.29830
5	Station 5	27.51745	-99.51522
6	Station 6	27.76534	-97.43426
7	Station 7	27.83241	-97.55538
8	Station 8	28.83617	-97.00553
9	Station 9	29.04376	-95.47295
10	Station 10	29.25447	-94.86129

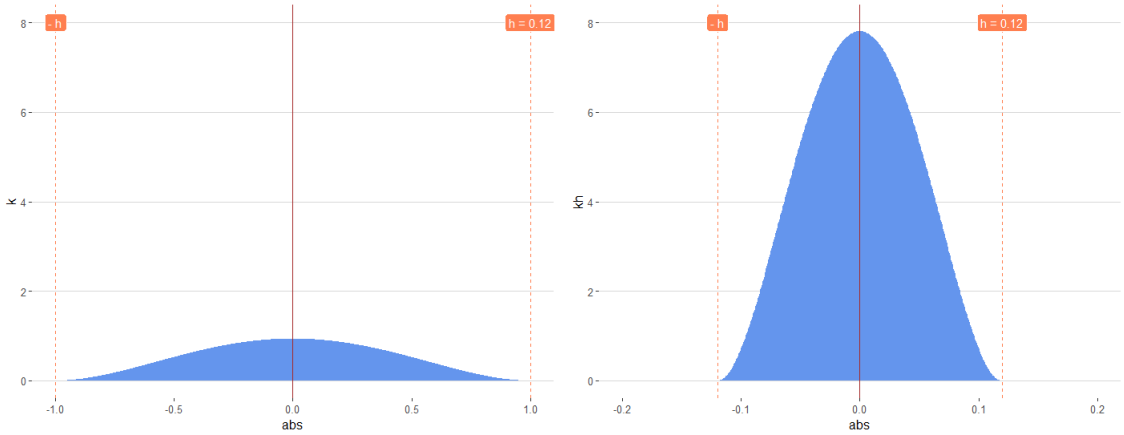
Comme énoncé dans nos sources bibliographiques, nous définissons α de la manière suivante :

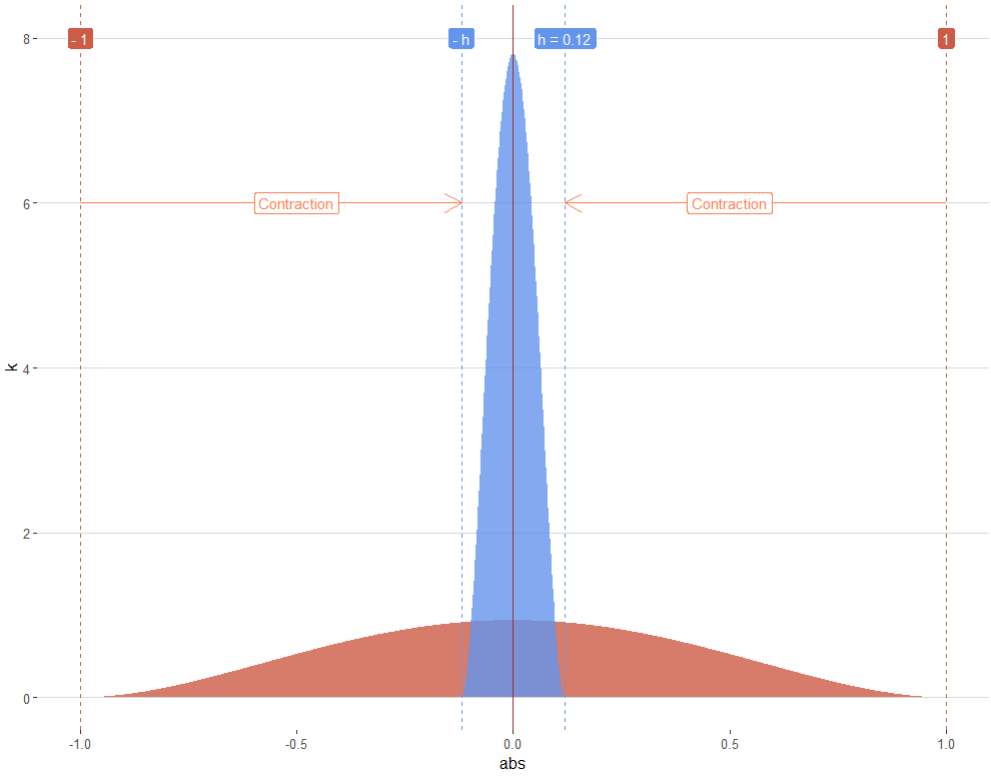
$$\alpha = \zeta \frac{logN}{N} = \zeta \frac{log(126969)}{126969}$$

Avec ζ une constante choisie arbitrairement. Nous définissons $\zeta = 11$, ainsi nous obtenons $\alpha \simeq 0.001$

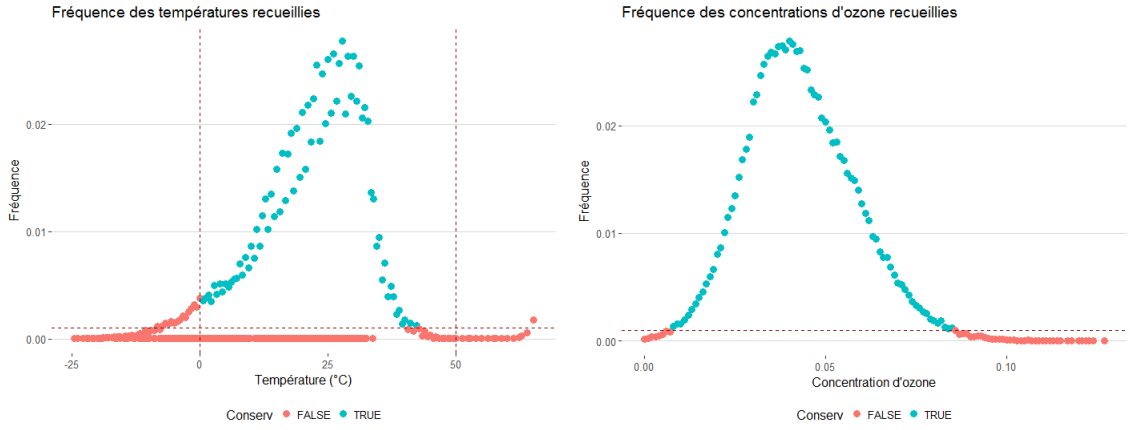
Le paramètre $\beta \in]0, \alpha]$ est choisi par l'utilisateur lors de l'exécution du programme.

h est donné, fixé à 0.12 et on a également : $K(x) = \frac{15}{16}(1 - x^2)^2$. On a alors :





Lorsque nous cherchons les valeurs de *champ*, nous obtenons plus de 400 valeurs distinctes pour chaque variable *tp* et *oz*. Cependant, un grand nombre de ces valeurs ne sont représentées que très peu fréquemment. On se propose alors, dans le but d'accélérer le temps de calcul, de ne pas prendre en compte ces valeurs dans la variable *champ*. Ainsi nous pouvons représenter les valeurs que nous conservons par ce procédé :



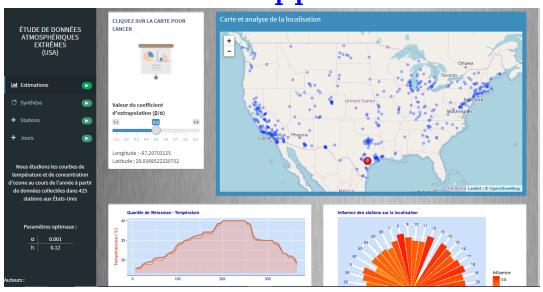
Cependant, nous avons une grosse perte d'information à droite. C'est pourtant ici que le quantile des valeurs extrêmes va avoir tendance à se situer. Nous complétons

donc nos champs par des valeurs périodiques jusqu'à la valeur maximale possible. Par cette méthode, nous avons certes perdu un peu de précision mais nous avons considérablement réduit nos temps de calcul. Nous obtenons alors des tailles de champ au minimum divisées par 4.

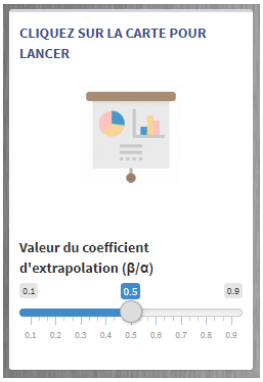
3.2 Application interactive

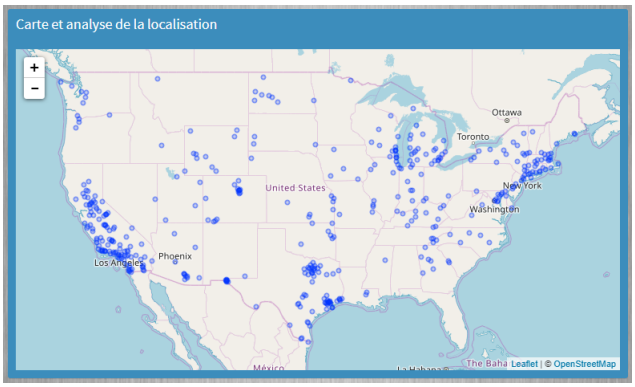
On se propose d'utiliser le package shiny de R afin de développer le code sous forme d'application interactive. Pour mieux présenter les résultats, nous utiliserons également les packages shinydashboard pour la mise en page et leaflet pour l'utilisation de cartes.
Nous n'expliquerons en détail que la partie 'Estimateurs' de l'application puisque c'est elle qui est au centre du projet.

Retrouver l'application ici



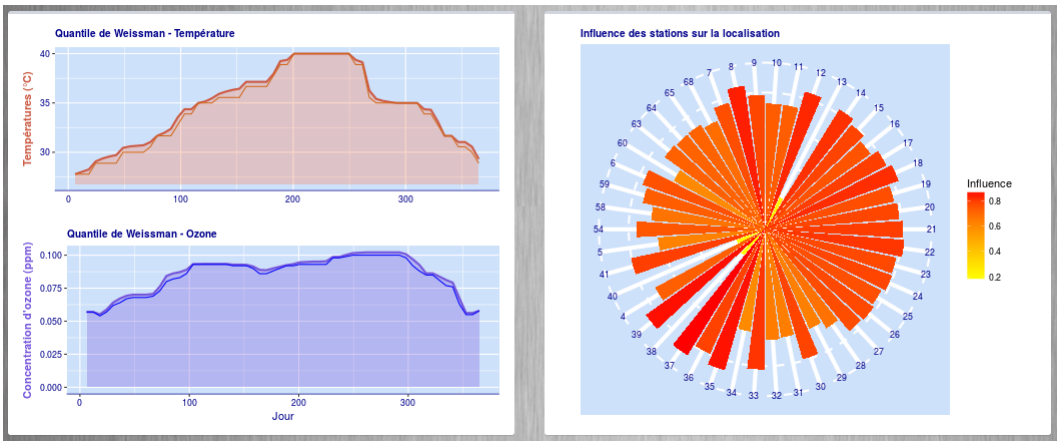
Nous demandons à l'utilisateur de saisir la valeur de β au niveau du curseur et la localisation à étudier en cliquant sur la carte.



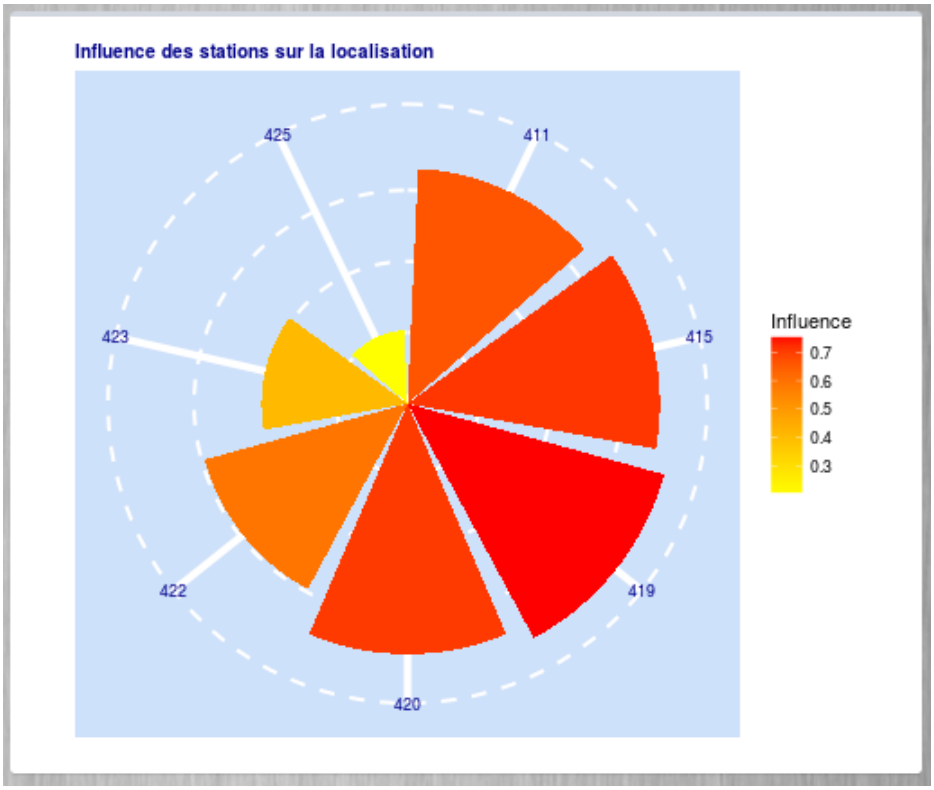


À partir de ces informations, nous calculons simultanément les évolutions des quantiles classiques et de Weissman au cours de l'année. Nous obtenons alors en

sortie :



Le graphique en sortie à gauche représente l'évolution de ces quantiles pour les données de température et d'ozone. Nous pouvons alors faire varier, pour une localisation donnée, le coefficient d'extrapolation. La différence obtenue lors de la représentation entre les deux valeurs extrêmes possibles de ce coefficient est en effet assez importante.
Le graphique en sortie à droite représente les stations qui ont influé sur le calcul des quantiles. Cette notion d'influence ne prend en compte que la distance à laquelle se trouvent les stations qui participent au calcul du quantile.
Pour la calculer, nous restreignons la base de données aux stations qui sont dans la fenêtre de lissage associée à h. Nous calculons ensuite la distance, à la fois verticalement et horizontalement, entre les localisations des lignes restantes et la localisation étudiée (Nous supposons les distances euclidiennes car nous regardons les localisations dans le plan et les distances étant assez courtes la courbure de la surface terrestre semble négligeable). Pour calculer l'influence nous passons alors ces valeurs dans la fonction K puis nous faisons le produit des distances horizontales et verticales pour chaque ligne. Nous obtenons des valeurs comprises entre 0 et 1. Plus la valeur de l'influence est proche de 1, plus l'influence est forte. Dans un cas assez simple :



Nous avons peu de stations qui influent sur le calcul des quantiles ici. La station qui apporte le plus d'informations est la station 419.

Dans le menu à gauche, vous trouverez 3 autres types d'interaction :

1. Synthèse : Représentation des moyennes sur l'année
2. Stations : Représentation de l'évolution pour une station donnée
3. Jours : Représentation des valeurs pour un jour donné. Il est possible de visualiser l'évolution en cliquant sur le bouton "Play" situé sous la barre du curseur à droite.

Bibliographie

Source accessible en cliquant sur son titre

Mathématique

Théorie des valeurs extrêmes, Laurent Gardes

Cours, Alexandre LEKINA

Kernel estimators of extreme level curves, Abdelaati DAOUIA, Laurent GARDES, Stéphane GRIMARD, Alexandre LEKINA

Extreme Value Modeling and Risk Analysis : Methods and Applications, Jun YAN, Dipak K. DEY

Extreme value theory : An intriduction, Laurens DE HAAN, Ana FERREIRA

Logistique

Support LaTeX

Site officiel RStudio

Pakage Shiny

Package ShinyDashboard

Synthèse fonctions ggplot2

Aide à la programmation ggplot2

Package Leaflet

Forum d'aide à la programmation