UNIVERSITÉ DE LILLE, FRANCE.

Faculté des sciences : UFR de Mathématiques

M1 MAS

TER

présenté par

Alexandre LANGLOIS et Antony PUDLICKI

Statistique spatiale des extrêmes: Estimation non-paramétrique de l'indice de queue

---

dirigée par Sophie DABO

Soutenu le 29 mai 2019

Résumé

Ce TER concerne la modélisation d'événements spatiaux extrêmes, un domaine de la statistique porteur de nombreuses applications en environnement, météorologie, océanographie, épidémiologie, économie et bien d'autres. Prenons par exemple les inondations en France en janvier 2018.

Les pluies et les grandes crues qui ont eu lieu sont rarement observées, pourtant lorsqu'elles ont lieu, elles peuvent créer de véritables catastrophes. C'est pour cela qu'il serait très utile d'avoir un moyen de les prévoir. C'est là tout le rôle de la théorie des valeurs extrêmes (TVE). En effet, cette théorie se base sur les phénomènes "rares" (cracks boursiers, catastrophes naturelles...) et fournit des outils qui permettent de prédire, ou tout du moins mieux comprendre leur apparition et leur occurence.
De nos jours, avec le développement des nouvelles technologies, de nombreuses données extrêmes sont désormais collectées en plusieurs positions géographiques. Etudier ce type de données ne peut donc se faire sans tenir compte d'éventuelles dépendances spatiales. La statistique spatiale a ainsi connu ces dernières années un déploiement dynamique de méthodes statistiques (paramétriques et non paramétriques) pour la modélisation d'événements extrêmes spatiaux.
Dans la suite, on s'attardera un peu plus sur la construction d'outils de la TVE spatiale, leur pertinence, puis on résumera ce qui existe déjà pour en toute fin, finir sur notre contribution et quelques résultats numériques.

Mots-Clefs

Théorie des valeurs extrêmes; statistique spatiale; estimateur à noyau; estimateur de Hill; indice de queue d'une distribution extrême; stationnarité; mélange (mixing)

Remerciements

Nous tenons à remercier profondément Sophie DABO-NIANG pour nous avoir proposé ce sujet de TER et pour nous avoir accompagné dans sa réalisation tout au long de ces 5 mois. Nous souhaitons aussi la remercier pour son aide dans la compréhension du sujet, la rédaction du TER et la réalisation des simulations.

Contents

Chapter 1

Introduction à la théorie des valeurs extrêmes et à la statistique spatiale

Résumé

Dans cette partie, on va s'attarder sur les éléments clés de la TVE et la statistique spatiale qui donnent les techniques de calculs des quantiles ou des valeurs extrêmes et la prise en compte de la dépendance spatiale.

1.1 Concepts clés de la théorie des valeurs extrêmes

Considèrons n variables aléatoires réelles $X_1, X_2, \cdots, X_n$ indépendantes et identiquement distribuées (i.i.d) de fonction de répartition $F(x) = \mathbb{P}(X_1 \leq x)$. La TVE s'intéresse à des valeurs qui sont supérieures à un certain quantile x_p, à savoir :

$$\mathbb{P}(X_i > x_p) < \epsilon \quad \epsilon \, très \, petit$$

Si x_p est le maximum de l'échantillon, estimer la probabilité précédente est celle d'observer un phénomène extrême ayant une valeur plus grande que le maximum de l'échantillon. Cette probabilité peut être exprimée sous forme d'un quantile dont l'estimation dépend du paramètre de la queue de distribution. Le théorème fondamental de la TVE (connu sous le nom de Théorème de Fisher et al. (1928)) donne les lois limites possibles du maximum de l'échantillon et permet d'avoir une certaine connaissance sur la queue de distribution.
Pour réaliser cette partie, nous nous somme inspirès des travaux de Gardes and Girard (2004), Drees (1995), Gardes and Girard (2005);

Soit $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ la statistique d'ordre associée à ces n variables aléatoires. Pour tout $i = 1, \cdots, n$, la variable aléatoire $X_{(i)}$ s'appelle la $(n-i+1)ème$ statistique d'ordre de l'échantillon. Il existe deux statistiques d'ordre qui sont particulièrement intéressantes pour l'étude des événements extrêmes. Ce sont les statistiques d'ordre extrême qui correspondent à la plus petite statistique d'ordre $X_{(1)}$ (ou statistique du minimum).

$$X_{(1)} = \min(X_1; \cdots; X_n)$$

et à la plus grande statistique d'ordre $X_{(n)}$ (ou statistique du maximum).

$$X_{(n)} = \max(X_1; \cdots; X_n).$$

Le comportement asymptotique du maximum $X_{(n)}$ (resp. minimum $X_{(1)}$) permet de rendre compte sur la fin de la distribution. Ces deux statistiques d'ordre sont liées l'une à l'autre à l'aide de la relation

$$X_{(1)} = -\max(-X_1, \cdots, -X_n).$$

Étant donnée la fonction F et que les variables aléatoires sont i.i.d alors la fonction de répartition F^n du maximum est donnée par

$$\begin{aligned} F^n(x) &= \mathbb{P}\left(X_{(n)} \leq x\right) = \mathbb{P}\left(X_1 \leq x, \cdots, X_n \leq x\right) \\ &= \prod_{i=1}^{n} \mathbb{P}\left(X_i \leq x\right) = \left[F(x)\right]^n. \end{aligned} \tag{1.1}$$

Puisque F étant souvent inconnue, on ne peut généralement déterminer la distribution du maximum à partir du précédent résultat car les extrêmes se trouvent, à droite et à la fin du support de la distribution.

Il faut donc connaitre la distribution des valeurs extrêmes, pour l'estimer et en déduire les quantiles extrêmes.

1.1.1 Approche par la fonction de survie et domaine d'attraction

On définit la fonction de survie d'une variable T par

$$\mathbb{P}(T \geq t) = 1 - \mathbb{P}(T \leq t) = 1 - F(t)$$

avec $F(t)$ la fonction de répartition de T.
Soit $X_1, ..., X_n$ des variables aléatoires i.i.d de fonction de répartition F. On définit $F^n(t)$ comme $F^n(t) = \mathbb{P}(max(X_1, ..., X_n) \leq t) = \mathbb{P}(\cap_i\{X_i \leq t\}) = \mathbb{P}(X_1 \leq t)^n$ car les X_i sont i.i.d.

Une des approches de la TVE est d'estimer la fonction de survie $S^n(t) = 1 - F^n(t)$ du maximum.

Dans la littérature, on peut lire souvent que la TVE donne une estimation de l'indice de queue, c'est-à-dire une estimation de la forme de la queue de la distribution, qui permet de savoir si les valeurs extrêmes seront plus des minimums ou des maximums. On notera par la suite cet indice de queue γ.

Definition 1.1. *Soient 2 suites $a_n > 0$ et b_n telles que pour tout $x \in \mathbb{R}$, $F^n(a_n x + b_n) \longrightarrow H(x)$, H une fonction de répartition non dégénérée. On dit alors que F appartient au domaine d'attraction de H, noté DAM(H).*
On a de plus : Si $F \in DAM(H)$, alors $nF(a_n x + b_n) \longrightarrow -ln\, H(x)$.
H appartient à l'un des trois types suivants de domaine d'attraction :

Type I (Gumbel) : $\exp\left(-e^{-x}\right), \quad \forall x \in \mathbb{R},$

Type II (Fréchet): $\begin{cases} 0 & si\ x \leq 0 \\ \exp\left(-x^{-\alpha}\right) & si\ x > 0 \end{cases}$

Type III (Weibull): $\begin{cases} \exp\left(-(-x)^{\alpha}\right) & si\ x \leq 0 \\ 1 & si\ x > 0 \end{cases}$

Ces trois types de distribution sont différenciés par un unique paramètre noté γ et appelé indice des valeurs extrêmes, il contrôle la "lourdeur" de la queue de distribution. Ils donnent l'ensemble des lois limites H en considérant les variables du type $dX + m$, où X suit une loi de Weibull, de Gumbel ou de Fréchet., d, m des réels. Plus généralement, voir par exemple Zhou (1986), De Haan and Ferreira (2007), on nomme la distribution des valeurs extrêmes (EVD)

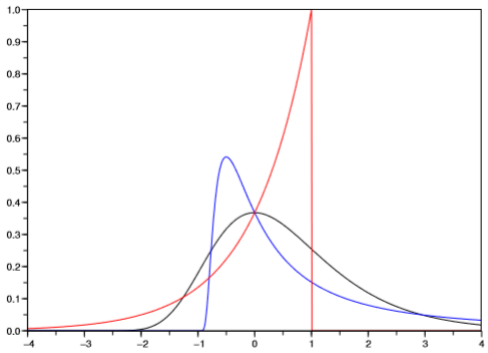
$$H_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right), & \gamma \neq 0 \quad 1+\gamma x > 0 \\ \exp(-\exp(-x)), & \gamma = 0 \quad -\infty \leq x \leq +\infty. \end{cases} \tag{1.2}$$

Les trois domaines d'attractions associés à γ sont donc :

Weibulll	Fréchet	Gumbel
$\gamma < 0$	$\gamma > 0$	$\gamma = 0$
Uniforme,Beta...	Cauchy,Pareto,Student...	Normale,Exponentielle,Lognormal...

Ces domaines d'attractions correspondent donc à 3 types d'indices de queue de distribution différents, et donc à 3 estimations différentes de la fonction de survie.

L'indice de queue détermine la forme de la queue de la distribution, la figure suivante en donne une illustration :

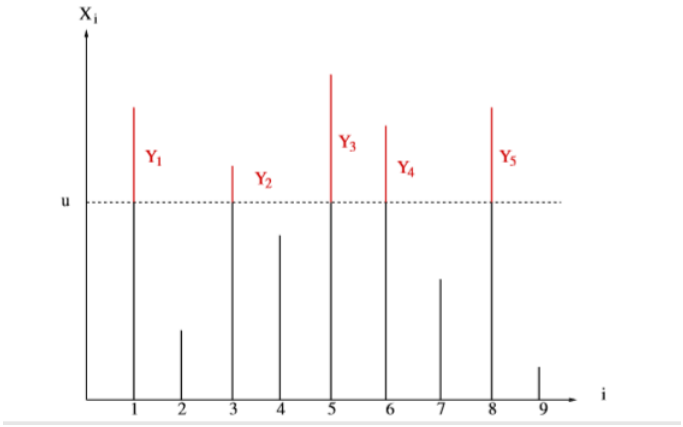


En rouge $\gamma = -1$, en bleu $\gamma = 1$, en noir $\gamma = 0$

1.1.2 Approche par l'échantillon d'excès

En pratique, il y a très peu de valeurs extrêmes (minimales ou maximales) et donc il est difficile d'avoir un échantillon assez grand pour pouvoir estimer la distribution des extrêmes.

Une des solutions est de fixer un seuil et de prendre toutes les valeurs au-dessus (ou en-dessous dans le cas du minimum), comme le graphique suivant :



Les valeurs $(Y_i)_{i=1,...,p}$ forment un échantillon d'excès. A partir de cet échantillon, il est possible alors d'estimer une fonction de distribution d'excès.

En résumé, la TVE se base, soit sur une estimation de la fonction de survie, soit sur une estimation de la distribution d'un échantillon d'excès. On verra par la suite que ces deux méthodes convergent et peuvent se substituer.

1.1.3 L'approche paramétrique de la TVE dans le cadre i.i.d

Cette approche consiste à estimer la fonction de survie d'une distribution via une suite de variables indépendantes et identiquement distribuées (i.i.d) et l'indice de queue γ, qui permet de connaître la forme de la queue de la distribution.

1.1.3.1 Estimation de la fonction de survie

On peut déjà définir le théorème central limite (TCL) pour la loi du maximum.

Soit un échantillon $(X_1, ..., X_n)$ i.i.d de fonction de répartition F. Le TCL nous donne :

$$\frac{\sqrt{n}\left(\frac{1}{n}\sum X_i, - m\right)}{\sigma} \rightarrow N(0,1)$$

avec $(X_1, ..., X_n)$ de moyenne m et de variance σ^2.
Par analogie, le TCL peut s'étendre à la loi du maximum. Soient deux suites $a_n > 0$ et b_n, on a alors

$$\frac{1}{a_n}\left(max(X_1, ..., X_n) - b_n\right) \longrightarrow Z,$$

de fonction de répartition H.

On peut aussi définir la fonction de répartition empirique qui nous servira par la suite pour estimer la fonction de survie et l'indice de queue. Soit $(X_i)_{i=1,...,n}$ une suite de variables i.i.d, on définit sa fonction de survie empirique comme :

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \geq x}$$

Théorème 1.1. ***(Théorème des valeurs extrêmes) :*** *Si $F \in DAM(G)$, alors G est du même type que :*

$$H_\gamma(x) \;=\; \begin{cases} exp(-(1+\gamma x)_+^{-1/\gamma}) \; si \; \gamma \;\neq 0 \\ exp(-e^x) \; si \; \gamma \;= 0 \end{cases}$$

avec $x_+ \;=\; max(0,x)$

Remarque 1.1. *La preuve de ce théorème est laissée en annexe, voir Gardes and Girard (2004)*

Ce théorème permet d'estimer la distribution de la valeur extrême (EVD) si on connaît la valeur de γ. C'est pourquoi il est important d'estimer au mieux (sans biais) l'indice de valeur extrême γ. Une fois cet indice estimé, il est facile d'estimer les quantiles extrêmes de la loi associée. Il existe essentiellement deux méthodes pour approcher un quantile extrême :
- utiliser la loi de la théorie des valeurs extrêmes
- utiliser la loi de Pareto Généralisée, donnée dans la Définition 1.3

1.1.3.2 Résultat pour l'estimation de la fonction de survie

Soit x_i les réalisations des variables X_i ordonnées. On suppose qu'elles admettent une fonction de répartition $F \in DAM(H)$, H étant une fonction de répartition non dégénérée. D'après le théorème des valeurs extrêmes que l'on peut appliquer au maximum, on a :

$$S(y_n) = \frac{1}{n}(1+\gamma\, y_n^{-\frac{1}{\gamma}})\, 1_{\gamma\neq 0} \;+\; \frac{1}{n} exp(-y_n)\, 1_{\gamma=0}$$

On peut poser : $y_n = a_n x_n + b_n$. On obtient alors une approximation de la fonction de survie pour le maximum :

$$S(X_n) = \frac{1}{n}\left(1+\gamma\,(\frac{(x_n-b_n)}{a_n})^{-\frac{1}{\gamma}}\right)\, 1_{\gamma\neq 0} \;+\; \frac{1}{n} exp(\frac{-x_n-b_n}{a_n})\, 1_{\gamma=0}$$

où (X_n) est un échantillon de maximas.

En général, les paramètres a_n, b_n et γ ne sont pas connus. On peut cependant les estimer selon trois méthodes :
- maximum de vraisemblance
- percentiles
- méthode des moments pondérés

Parmi ces trois méthodes, les moments pondérés sont les plus utiles pour de petits échantillons.

Definition 1.2. *Soit Z une variable aléatoire de fonction de répartition F. Si Z est intégrable, le moment pondéré d'ordre $r \in \mathbb{N}$ et $s \in \mathbb{N}$ de Z est :*

$$WM_Z(r,s) = \mathbb{E}[ZF^r(Z)(1-F(Z)^s]$$

Soit $(Y_1,...,Y_k)$ un échantillon de k maximas de fonction de survie $S_{a,b,\gamma}(y)$. D'après la définition d'au-dessus, on peut calculer les moments pondérés d'ordre r suivants :

$$\mu_r = \mathbb{E}(\, Y\, S^r_{a,b,\gamma}(Y)\,) \;= \frac{1}{1+r}\,(\,a - \frac{b}{\gamma}\,(1-(r+1)^\gamma\,\Gamma(1-\gamma))\,)\; 1_{\gamma<1}.$$

L'avantage d'estimer $S_{a,b,\gamma}$ par la méthode des moments est qu'il suffit de calculer les moments d'ordre 1, 2 et 3 pour avoir un estimateur des 3 paramètres a_n , b_n et γ.
Dans la suite, on posera **r=0**.

1.1.4 Fonction de répartition d'excès

Dans cette section, nous abandonnons la fonction de survie empirique d'un échantillon de maximas pour passer à la fonction de répartition d'un échantillon d'excès au dessus d'un seuil u. Il nous faut d'abord deux définitions.

Definition 1.3. *La loi de Pareto Généralisée (GPD) de paramètre $\gamma \in \mathbb{R}$ et $\sigma > 0$ est :*

$$G_{\gamma,\sigma}(x) = \begin{cases} 1 - (1 + \gamma\frac{x}{\sigma})^{-\frac{1}{\gamma}} & \gamma \neq 0 \\ 1 - exp(-\frac{x}{\gamma}) & \gamma = 0 \end{cases}$$

Definition 1.4. *On définit la fonction de répartition d'excès de X au-dessus d'un seuil u par :*

$$\bar{F}_u(x) \, = \, \mathbb{P}(X - u \leq x \mid X > u) \text{ pour } u \leq x_F \text{ et } x \geq 0$$

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

Théorème 1.2. ***Théorème de Pickands***

$$\forall\, y \geq 0\,, \bar{F}_u(y) \, = \frac{\bar{F}(u+y)}{\bar{F}(u)} \, \simeq \bar{G}_{\gamma,\sigma}(y)$$

la fonction de survie de la GPD. En posant le changement de variable $x = u + y$ et $\alpha = \mathbb{P}(X \geq u) \, = \bar{F}(u)$, on obtient $\forall\, x \geq 0$:

$$\bar{F}(x) \, \simeq \, \bar{F}(u)\bar{G}_{\gamma,\sigma}(x-u) \, \simeq \, \alpha\, \bar{G}_{\gamma,\sigma}(x - F^{-1}(\alpha)\,)$$

On a alors :

$$\lim_{x \to x_F} \sup_{0<y<x_F-u} \mid \overline{F}(x) - \overline{G}_{\gamma,\sigma}(x) \mid = 0$$

Preuve 1.1. *Afin de démontrer le théorème de Pickand nous avons besoin du lemme et de la proposition ci-dessous*

Proposition 1.1. *La fonction de répartition F de point terminal x_F appartient au domaine d'attraction de H_γ si et seulement si il existe une fonction positive $a(.)$ telle que :*

$$\lim_{u \to x_F} \frac{\overline{F}(u + xa(u))}{\overline{F}(u)} = 1 - G_{\gamma,1}(x)$$

Lemme 1.1. *(Lemme de Dini): Soit $(\phi_n)_{n\geq 1}$une suite de fonctions croissantes de $\mathbb{R} \to \mathbb{R}$ et soit ϕ une fonction continue sur $\mathbb{R}$.*

Si $\phi_n(x) \to \phi(x)$pour tout $x \in \mathbb{R}$ alors:

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \mid \phi_n(x) - \phi(x) \mid = 0$$

□

En remarquant que pour tout $0 < x < (x_F - u)/a(u)$,

$$\frac{\overline{F}(u + xa(u))}{\overline{F}(u)} = 1 - F(xa(u))$$

La proposition assure que F appartient au domaine d'attraction de H_γ si et seulement si il existe une fonction positive $a(.)$ telle que:

$$\lim_{u \to x_F} \mid F_u(xa(u)) - G_{\gamma,1}(x) \mid = 0$$

Puis d'après le lemme de Dini, cette convergence est uniforme en x puisque $G_{\gamma,1}$ est une fonction continue. En posant $y = xa(u)$, on a donc:

$$\lim_{x \to x_F} \sup_{0<y<x_F-u} \mid F_u(y) - G_{\gamma,a(u)}(y) \mid = 0,$$

ce qui est attendu.

Ce théorème est le résultat important qui permet d'approcher la loi des excès au-dessus d'un seuil par une loi de Pareto Généralisée (GPD). On a alors le résultat suivant :

$$S_{\gamma,\sigma}(x) = \left(1 + \gamma \frac{x}{\sigma}\right)^{-\frac{1}{\gamma}} 1_{\gamma \neq 0} \; + \; exp\left(-\frac{x}{\sigma}\right) 1_{\gamma=0}$$

Avec γ, σ des paramètres à estimer via des méthodes de vraisemblance ou des moments pondérés.

Le théorème donne aussi la fonction de survie d'un échantillon d'excès, que l'on note $\bar{S}$:

$$\bar{S}(X) \; = \; \alpha \left(1 + \gamma \left(\frac{x - F^{-1}(\alpha)}{\sigma}\right)\right)^{-\frac{1}{\gamma}} 1\gamma \neq 0 \; + \; \alpha \, exp\left(-\frac{x - F^{-1}(\alpha)}{\sigma}\right) 1_{\gamma=0}$$

1.1.4.1 Estimateurs des moments pondérés pour la loi GPD

Proposition 1.2. *Soit Z une loi de Pareto Généralisée de paramètre γ et σ. Si $\gamma < 1$ $\forall \, s \in \mathbb{N}$, on a le moment pondéré suivant*

$$WM_Z(0, s) = \frac{\sigma}{(s+1)(s+1-\gamma)}.$$

Soit $(Y_1, ..., Y_k)$ un échantillon de k excès au dessus du seuil u et $S_{\gamma,\sigma}(Y)$ la loi GPD. Par la Proposition 1.1, on obtient le moment pondéré d'ordre s suivant pour $S_{\gamma,\sigma}(Y)$:

$$\nu_s = \frac{\sigma}{(s+1)(s+1-\gamma)}.$$

En combinant les moments $\nu_0 \, et \, \nu_1$ on a :

$$\gamma \; = \; \frac{4\nu_1 - \nu_0}{2\nu_1 - \nu_0}, \sigma \; = \; \frac{2\nu_1\nu_0}{\nu_0 - 2\nu_1}$$

1.1.4.2 Convergence entre EVD et GPD

Proposition 1.3. *La fonction de répartition F de point terminal x_F appartient au domaine d'attraction de H_γ si et seulement s'il existe une fonction positive $a(.)$ telle que :*

$$\lim_{u \to x_F} \frac{\bar{F}(u + xa(u))}{\bar{F}(u)} = 1 - G_{\gamma,1}(x)$$

Grâce au théorème de Pickands et à la Proposition 1.2, on a une convergence de la fonction estimée EVD vers la fonction de survie estimée de GPD. Cela est utile en pratique car la loi de Pareto permet de caractériser l'ensemble des trois domaines d'attraction.

$$b_n \longrightarrow \sigma \text{ et } \bar{F}^{-1}(\alpha) \longrightarrow a_n.$$

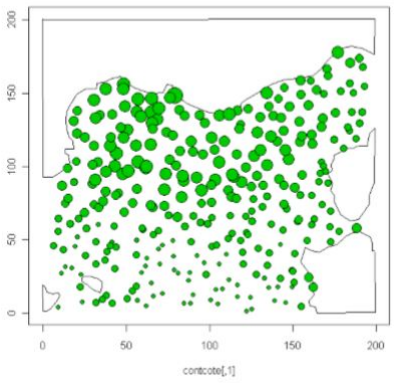
Les avantages d'utiliser la GPD sont :
- il est plus facile d'avoir un échantillon d'excès que d'extrêmas
-$\bar{F}^{-1}(\alpha)$ est un quantile facile à estimer par inversion de la fonction de survie empirique
- Souvent $\alpha = \frac{k}{n}$ la proportion d'excès

1.2 Introduction à la statistique spatiale

Le but de la statistique spatiale est d'étudier des phénomènes (températures, étude d'une population, étude des villes,...) sur un ensemble spatiale $S \subset \mathbb{R}^d$, $d \geq 2$.
Il y a une dépendance en espace de ces données. On va donc poser Z, un champ aléatoire (une famille de variables aléatoires) sur S avec $Z = \{Z_s, s \in S\}$ composé de variables indexé par S.

Definition 1.5. *Un champ aléatoire est une famille de variables aléatoires sur un domaine fini $S \subset \mathbb{R}^d$, $d \geq 2$.*

On considère alors les données spatiales comme des réalisations de champs aléatoires. Prenons par exemple un échantillon de données de chlorophile récolté par des bâteaux, on observe graphiquement :



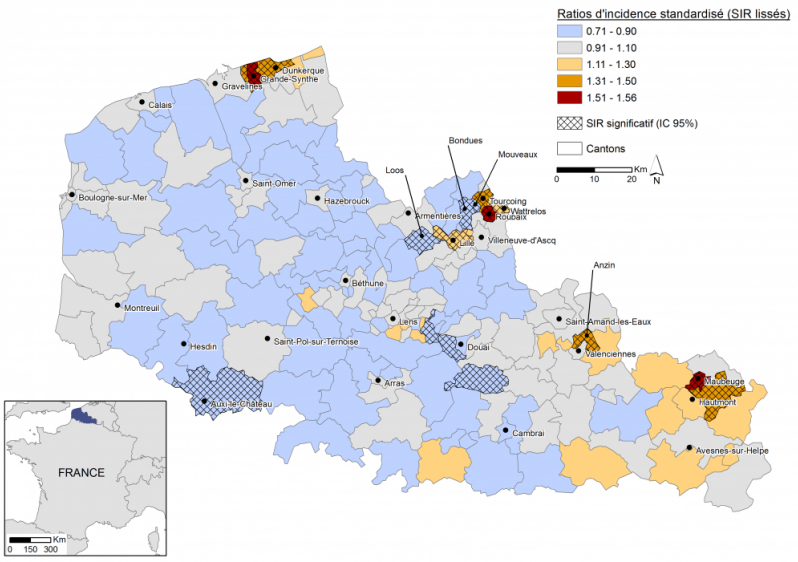
On peut étudier la répartition de la chlorophile dans l'espace grâce à un variogramme, qui sera défini et présenté plus loin. Dans notre cas, la chlorophile fait partie des données géostatistiques, mais il existe d'autres types de données spatiales que l'on verra par la suite.

1.2.1 Les types de données spatiales

Il existe trois types de données spatiales: les données géostatistiques, les données latticielles et les processus ponctuels spatiaux.

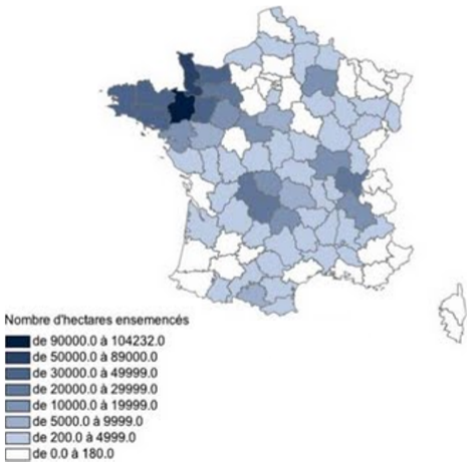
1.2.1.1 Les processus ponctuels spatiaux

Pour ce type de donnée, on a une répartition aléatoire spatiale des points de S où ont lieu les observations.
Le but de ce processus est de savoir si la répartition des sites est aléatoire, régulière ou si elle présente des agrégats. Il sera utilisé pour expliquer la distribution des caractéristiques des objets en fonction de leur position relative comme par exemple la taille d'aiguilles de pin dans un sous-bois.



1.2.1.2 Les données latticielles

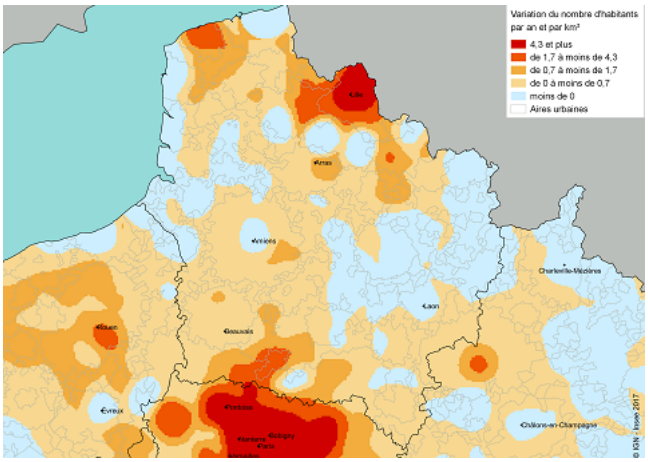
Pour ce type de donnée, le champs Z est observé sur un ensemble S qui est fixé et discret. On représente les données en leur centre (par exemple des régions, des villes,...). Il n'y a pas forcément de dépendance entre tous les sites. On va utiliser ces données pour des études sur les régions ou les villes comme le pourcentage d'individus du groupe sanguin A dans les regions. Ce type de données s'illustrera par exemple de cette façon:



Nous voyons ici qu'il n'y a pas forcement de dépendance entre les cantons.

1.2.1.3 Les données géostatistiques

Pour ce type, les données sont mesurées sur un nombre fini de sites. Le champ $Z = (Z_s, s \in S)$e st observé en n sites fixés $\{s_1, s_2, ..., s_n\} \subset S$ et $Z_s \in \mathbb{R}$ est supposé ici à valeurs réelles. On va utiliser ces données pour des études de températures, de précipitations, sur la propriété des sols. Ce type de données s'illustrera par exemple de cette façon:



La TVE est souvent appliquée à des données indépendantes ou temporelles, or il existe ne nombreuses situations où la dimension spatiale est très importante. Notre but est d'étendre l'estimation d'indices de queue qui existe dans le domaine temporel au cadre spatial.
Dans la suite, nous allons résumer les éléments essentiels qui concernent les données spatiales ainsi que leur dépendance. D'une part avec une approche paramétrique puis par une approche non paramétrique.

1.2.2 Approches paramétriques de la dépendance spatiale en géostatistique

A partir de maintenant nous allons développer quelques propriétés sur le domaine auquel nous nous intéresserons plus particulièrement : les données géostatistiques et présenter les outils qui permettent d'étudier la variabilité et d'interpoler au mieux possible les données.

1.2.2.1 Processus stationnaire fort

Un processus spatial Z est dit stationnaire si la loi jointe des Z_{s_i} est invariante par translation c'est à dire :

$$\mathbb{P}\left(\mathrm{Z_{s_1}} \leq \mathrm{z_1}, ..., \mathrm{Z_{s_k}} \leq \mathrm{z_k}\right) = \mathbb{P}(\mathrm{Z_{s_1+h}} \leq \mathrm{z_1}, ..., \mathrm{Z_{s_k+h}} \leq \mathrm{z_k})$$

1.2.2.2 Champ stationnaire faible ou second ordre

Un champ Z est dit stationnaire du second ordre sur S, si les deux premiers moments existent et si :
$E\left(Z_s\right) = \mu$, $\forall s \in S$ et $cov\left(Z_s, Z_{s+h}\right) = C(h)$ où C fonction de covariance stationnaire.

Definition 1.6. *Un champ Z est dit intrinsèque si ses h-accroissements sont stationnaires c'est a dire le processus X^h, tel que $X_s^h = Z_{s+h} - Z_s$ est stationnaire.*

Un champ stationnaire est intrinsèque mais la réciproque n'est pas vraie.

1.2.2.3 Variogramme pour des processus intrinsèques

On s'intéresse aux processus intrinsèques car il est parfois difficile de vérifier la stationnarité d'ordre 2 (L_2). Pour cela, on définit la notion de variogramme pour des processus intrinsèques comme :

$$\gamma(h) = \frac{1}{2} Var(Z_{s+h} - Z_s)$$

Avec ce variogramme, on obtient les propriétés suivantes :
- $\gamma(h) = 0$
- $\gamma(h) = \gamma(-h)$
- $\gamma \geq 0$
- $\lim_{\|h\| \to \infty} \gamma(h) = l < +\infty \Longrightarrow$ processus stationnaire.

Il est à noter que si le processus est stationnaire alors on a : $\gamma(h) = C(h) - C(0)$.

Avec ce genre de variogramme, on a trois outils essentiels :

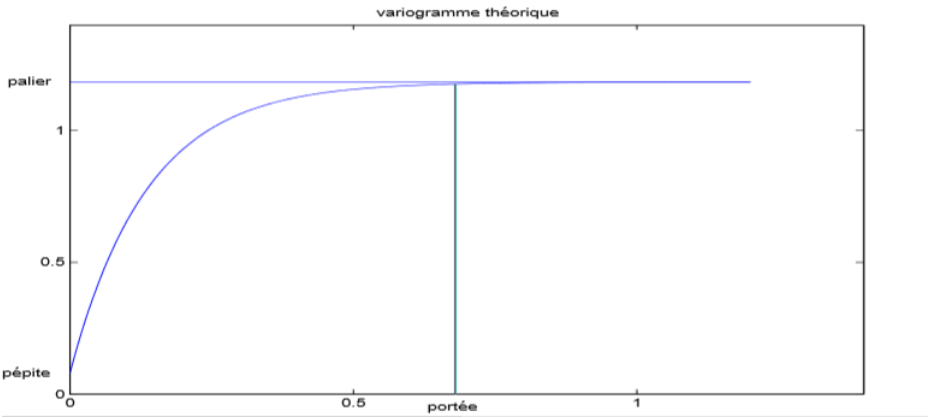
- **Portée** : on la définit comme la distance à partir de laquelle deux observations ne se ressemblent plus du tout en moyenne, elles ne sont plus liées linéairement (covariance nulle). À cette distance, la valeur du variogramme correspond à la variance du processus.
- **Palier** : on le définit comme l'écart de variance le plus grand en moyenne. On a

$$C = \lim_{h \to \infty} \gamma(h)$$

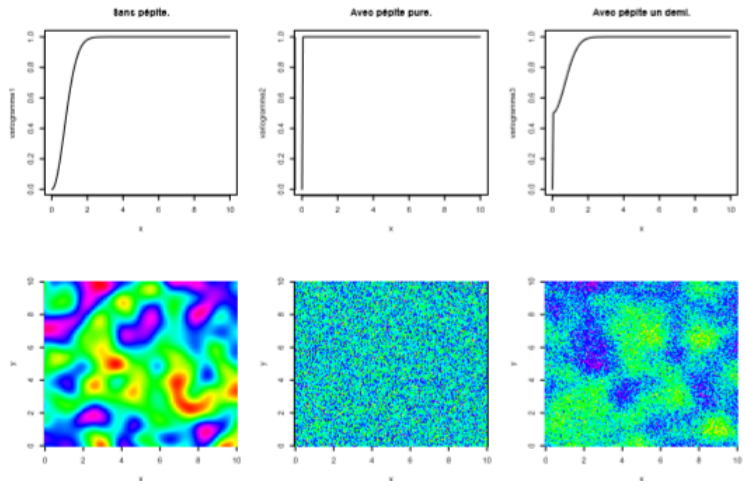
- **Pépite** :on la définit comme la variance entre deux points infiniment proches. On a la propriété suivante :

$$c = \lim_{h \to 0} \gamma(h)$$

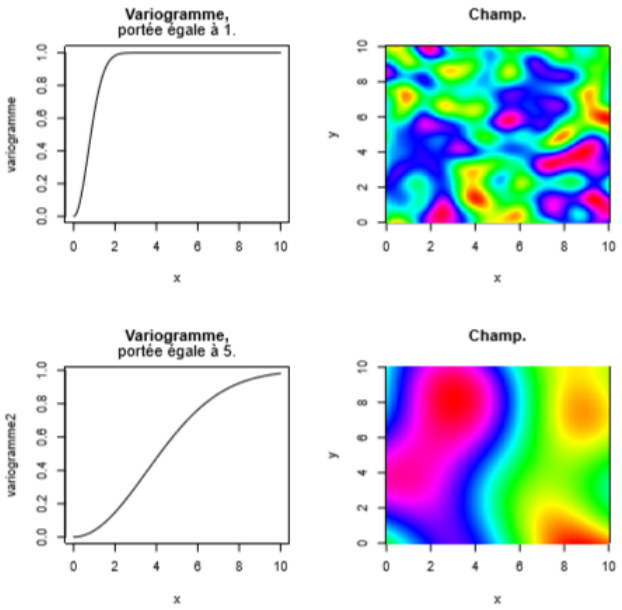
Remarque: Si la pépite $c = 0$ alors le processus est dit régulier.



Pour mieux comprendre l'utilité de ces outils, on peut les illustrer comme suit.



A gauche, on observe un processus régulier, c'est-à-dire avec une pépite nulle. Au centre, avec une pépite pure (la plus grande erreur possible localement), le processus n'est pas lisible à cause de l'erreur commise. A droite, avec une pépite plus petite, on commence à mieux apercevoir le processus.



Avec une portée égale à 1, la variance est réduite et on a un champ plus précis. Au contraire, avec une portée égale à 5, on voit sur le champ que les zones sont plus grandes, donc la dépendance est plus grande.

1.2.2.4 Quelques exemples de variogramme

Il existe différents variogrammes isotropiques, c'est-à-dire des variogrammes qui ne dépendent pas de la direction dans laquelle on se dirige. En voici quelques exemples ;

- exponentielle : $\gamma(h) = C\Big(1 - exp(-\frac{||h||}{a})\Big) \qquad C > 0,\ a > 0$
- gaussien : $\gamma(h) = C\left(1 - exp(-\frac{||h||^2}{a})\right) \qquad C > 0,\ a > 0$
- puissance : $\gamma(h) = C\,||h||^2$

Tous ces variogrammes sont des cas particuliers des variogrammes de la classe de Matern, définis comme :

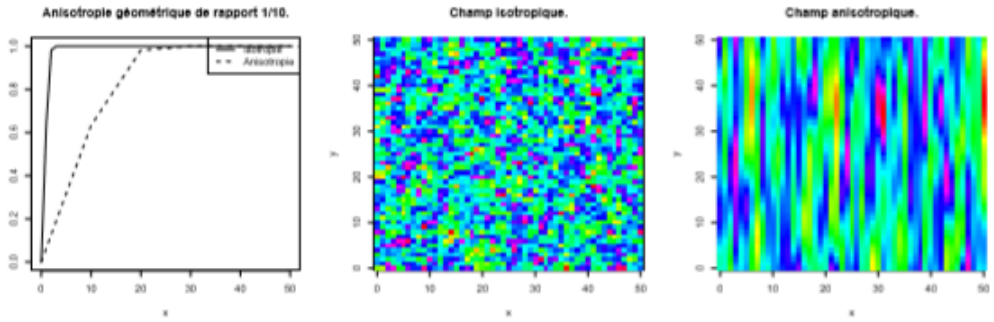
$$\gamma(h) = C\Big(1 - \frac{1}{2^{\upsilon-1}\Gamma(\nu)}\Big(\frac{2\nu^{1/2}||h||}{\rho}\Big)\kappa_\nu\Big(\frac{2\nu^{1/2}||h||}{\rho}\Big)\Big)$$

avec ν qui règle la stationnarité et κ une fonction de Bessel.

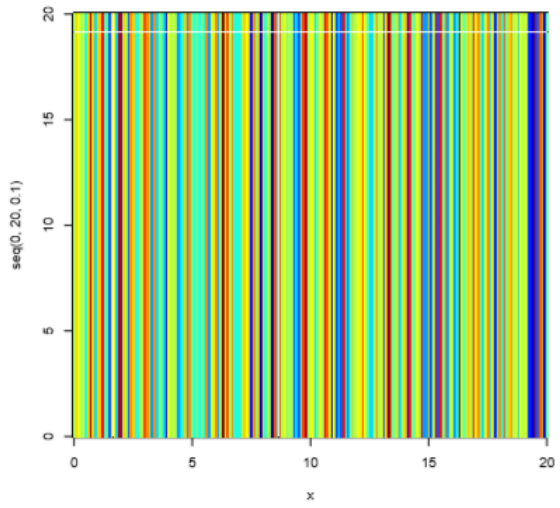
P.S : si $\nu = 1/2$ on retrouve le variogramme gaussien.

Definition 1.7. *Au contraire d'isotropique, on définit un champ Z anisotropique si au moins deux variogrammes directionnels sont différents.*
Le variogramme directionnel d'un champ intrinsèque dans la direction $\overrightarrow{e}$ est defini par: $2\gamma(h) = Var(Z_{s+h\overrightarrow{e}} - Z_s)$ *pour* $h \in \mathbb{R}$

Remarque 1.2. *L'anisotropie permet d'étudier la continuité spatiale. Au contraire d'isotroprie, un champ est anisotropique si les données dépendent de la direction spatiale dans laquelle on se dirige. Il existe deux types d'anisotropie : l'anisotropie géometrique et l'anisotropie zonal. Ci-dessous la différence visuelle de ces deux anisotropies :*



Comme un champ isotropique ne dépend pas de la direction, on a une pépite quasiment pure. En effet, on peut observer deux points très proches mais de directions différentes. Au contraire, une anisotropie géométrique dépend de la direction donc on observe un peu plus de zones étirées dans une direction.



L'anisotropie zonale fait apparaître des bandes.

1.2.2.5 Estimation d'un variogramme

En pratique, on ne connaît pas le variogramme des données et on cherche à l'estimer, on utilise alors le variogramme expérimental défini par :

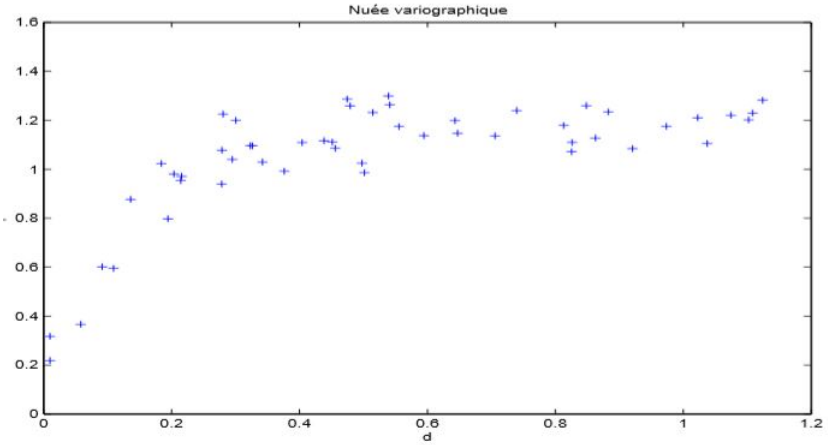
$$\widehat{\gamma_n}(h) = \frac{1}{2\#N(h)} \sum_{s_i,s_j \in N(h)} (Z_{s_i} - Z_{s_j})^2 h \in \mathbb{R}$$

avec : $N(h) = \{(s_i, s_j) : h - \Delta \leq s_i - s_j \leq h + \Delta; i, j = 1, ..., n\}$.

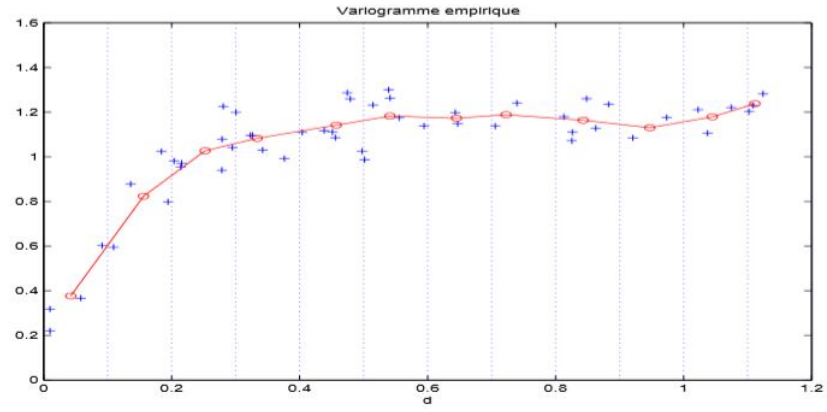
Propriété 1.1. *Avec ce variogramme expérimental, on possède différentes propriétés :*
- $\hat{\gamma_n}(h)$ est sans biais
- $\hat{\gamma_n}(h)$ ne nécessite pas l'estimation de la moyenne
- Si Z est Gaussien, $\hat{\gamma_n}(h)$ est une somme de $\chi^2(1)$

1.2.2.6 Exemple : construction d'un variogramme

Considérons une nuée variographique : $\gamma^*_{i,j} = \frac{(Z_{si} - Z_{sj})^2}{2}$ muni de la distance $d_{ij} = ||s_i - s_j||$. On observe :



Le variogramme expérimental calculé avec la formule au-dessus donne un variogramme estimé :

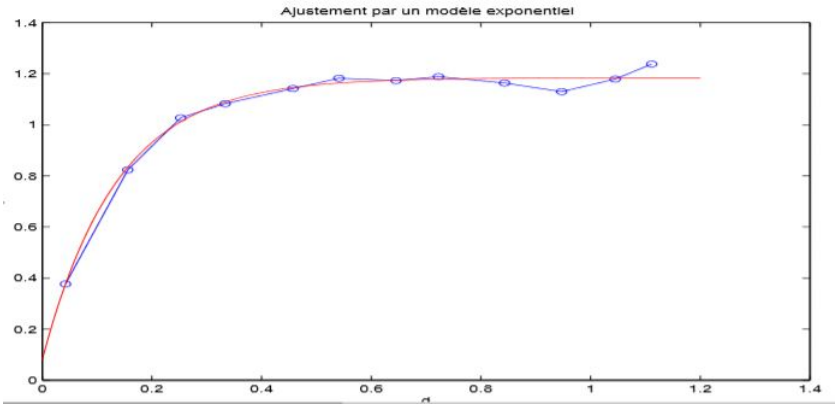


Un variogramme est lié à la variance et une variance est positive, alors un variogramme admissible est telle que la variance de toutes combinaisons linéaires des variables Z_s soit positive. Or ici, on observe une pente négative pour $d \in [0.8, 1]$, ce variogramme empirique n'est alors pas admissible et on peut le modifier pour le rendre admissible. En pratique,

on cherche une fonction de variogramme admissible paramétrique de type gaussien, exponentiel, de Matern,..., qui ajuste le mieux le variogramme empirique.
On pose alors γ_θ un variogramme admisible, solution du problème :

$$\min_\theta \sum_{k=1}^{K} (\gamma_\theta(d_k) - \hat{\gamma_n}(d_k))^2$$

On obtient alors ce variogramme :



1.2.3 Approches non-paramétriques de la dépendance en géostatistique

Dans cette approche non paramétrique, on définit les mélanges "mixing". Cette notion vient de l'étude des processus stochastiques temporels. Elle est utile car elle permet de déterminer la dépendance d'un échantillon de données.
Il existe plusieurs sortes de "mixing" mais seulement un nous intéresse, car il sera utilisé par la suite pour définir un échantillon de données spatiales.

Definition 1.8. *Soit (Z_k) une suite de variables aléatoires réelles. On dit que cette suite est α - mixing si :*

$$\alpha(m) = \sup_k \sup_{A \in \mathscr{F}_{m+k}^{\infty}} \{\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\} \underset{m\to\infty}{\longrightarrow} 0$$

$A \in \mathscr{F}_{m+k}^{\infty}$ la tribu borélienne engendrée par les X_i, $m + k \leq i \leq \infty$.
$B \in \mathscr{F}_1^k$ la tribu borélienne engendrée par les X_i, $1 \leq i \leq k$.

Definition 1.9. *Soit X un N-échantillon de données dont la covariance $Cov(X_i, X_j) \neq 0 \quad \forall\, i, j \in \{1, ..., N\}$. On dit que X est $\beta -$ mixing si :*

$$\beta(m) = \sup_{p>0} \mathbb{E}\Big(\sup_{C \in B_{p+m+1}^{m}} (\mathbb{P}(C|B_1^p) - \mathbb{P}(C)\,) \Big) \underset{m\to\infty}{\longrightarrow} 0$$

avec B_1^p une tribu borélienne engendrée par les X_i, $1 \leq i \leq p$

On étudie ici la dépendance entre un évènement $C|B_1^p$ (évènement qui se produit sachant ce qui s'est produit avant) et C (évènement seul). Si $\beta(m) \to 0$, cela signifie qu'à partir de la distance m, les données de B_1^p et celles de B_{1+m}^p deviennent indépendantes. Dans le cadre temporel, on peut considérer m comme un écart de temps, dans le cadre spatial, on peut le considérer comme une distance entre des sites. Pour illuster la notion de β-mixing ou α-mixing, on peut choisir l'exemple des séries temporelles en finance ou de la vitesse du vent dans le domaine spatial. La notion de α -mixing est plus faible que le β - mixing, on privilégie donc le β - mixing pour plus de généralités.

Chapter 2

Etat de l'art sur l'estimation des indices de queue

Résumé

Dans cette partie, on va faire un état de l'art non exhaustif des estimateurs d'indice de queue et de la fonction de densité qui existent dans les cadres i.i.d et temporel.

2.1 Estimateurs d'indices de queue et de densité dans le cas i.i.d

Nous allons voir ici des estimateurs non-paramétriques d'indices de queue dans le cas i.i.d et des estimateurs de régression non-paramétriques. Ces deux classes d'estimateurs permettent de mieux comprendre la distribution d'un échantillon de maximas ou de valeurs extrêmes.

2.1.1 Estimateur de Hill

Soit un échantillon $(X_1, ..., X_n)$ de variables aléatoires i.i.d .Cet estimateur est le point de départ de tous les autres. En se restreignant au domaine de Fréchet ($\gamma > 0$) et en se basant sur un modèle semi-paramétrique, c'est-à-dire qu'on utilise $\bar{F}^{-1}(p) \simeq \bar{F}^{-1}(\alpha)\,(\frac{p}{\alpha})^{-\gamma}$, on définit l'estimateur de Hill pour l'indice de queue γ :

$$\hat{\gamma}(k) \;=\; \tfrac{1}{k}\sum_{i=1}^{k-1}\,(\,ln\,X_{n-i+1,n}\;-ln\,X_{n-k+1,n}\,)$$

où les $X_{i,n}$ sont les observations ordonnées et indépendantes identiquement distribuées, voir Hill (1975), Gomes et al. (2008).
Par la suite, nous utiliserons une version revisitée de cet estimateur dans le cadre spatial avec une prise en compte de la dépendance.
Drees et al. (2000) ont établit la normalité asymptotique de l'estimateur de Hill :

$$\sqrt{k}(\hat{\gamma}_k^H - \gamma) \xrightarrow{loi} \mathcal{N}(\tfrac{\lambda}{1-\rho}, \sigma^2)$$

Où σ^2 possède une expression simple dans le cas i.i.d et $\frac{\lambda}{1-\rho} \to 0$.

2.1.2 Estimateur par noyau de Parzen-Rosenblatt

Rosenblatt et Parzen ont introduit en 1956 et 1962 respectivement l'estimateur par noyau de la densité d'une variable aléatoire. C'est une généralisation de la méthode d'estimation de la densité par histogramme. Elle permet d'estimer la densité en tout point du support.

Definition 2.1. *Soit $(X_1, ..., X_n)$ un échantillon i.i.d d'une variable aléatoire X admettant une densité f non-dégénérée. On définit l'estimateur non-paramétrique pour la densité f :*

$$\hat{f}_h(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right),$$

voir Parzen (1962), Bean and Tsokos (1980), Rosenblatt (1956) pour plus de détails. Cet estimateur a été motivé par l'estimateur de la fonction de répartition empirique. En effet, la fonction de répartition F de l'échantillon, est de fonction de répartition empirique

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\mathrm{X_i \leq x}\}}$$

Comme $f(x)$ est la dérivée de $F(x)$:

$$f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h} = \lim_{h\to 0} \frac{P[x-h < X \leq x+h]}{2h},$$

l'estimateur a été le premier proposé par Parzen (1962) et Rosenblatt (1956).

$$\begin{aligned} \widehat{f}(x) = & \frac{\sharp\{i : x-h < X_i \leq x+h\}/n}{2h} = \frac{\frac{1}{n}\sum_{i=1}^{n} 1_{]x-h,x+h]}(X_i)}{2h} \\ = & \frac{\widehat{F}(x+h) - \widehat{F}(x-h)}{2h} \end{aligned}$$

On peut écrire cet estimateur comme :

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K_R(\tfrac{X_i - x}{h})$$

où K_R : $\mathrm{t} \mapsto \frac{1}{2} 1_{\{-1<\mathrm{t}\leq 1\}}$ est le noyau rectangulaire. L'estimateur avec le noyau ci-dessus est une fonction discontinue, pour avoir un estimateur de la densité plus lisse, on peut choisir un noyau plus régulier K, comme la densité d'une gaussienne standard.

2.1.3 Estimateur de régression par lissage par noyau de Nadaraya (1964) Watson

Nadaraya (1964) et Watson ont étendu l'estimateur à noyau de la densité au cadre de la régression, voir Kempeners (2010), Tsybakov (2003)

Ainsi , on obtient l'estimateur de $g(x) = \mathbb{E}(Y|X = x)$ de Nadaraya-Watson :

$$\widehat{g^{NW}}(x) = \frac{\hat{v}(x)}{\hat{f}(x)}$$

avec : $\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K_i(x)$; $\hat{v}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} Y_i K_i(x)$; $K_i(x) = K(\frac{(x-X_i)}{h})$

Cet estimateur fait partie des estimateurs linéaires de régression non paramétrique, plus précisément c'est un estimateur obtenu par lissage par noyau.

Remarque 2.1. *Cet estimateur peut être vu comme une méthode des k plus proches voisins.*

Remarque 2.2. *Pour trouver le pas h qui minimise l'erreur* $\mathbb{L}^2$*, on peut utiliser la méthode dite de validation croisée.*

On peut souhaiter assigner des poids différents aux X_i, on quitte alors le domaine i.i.d.

Les estimateurs à noyau de la densité et de la régression ont inspirés beaucoup de travaux utilisant les fonctions de noyau comme c'est le cas de l'estimateur des indices de queue qui nous intéresse.

2.2 Estimateurs non-paramétrique d'indices de queue et de la densité dans le cas non i.i.d

2.2.1 Estimateur par noyau de la densité de Robinson

L'inférence non-paramétrique est très utilisée en pratique du fait qu'elle ne suppose pas d'hypothèses paramétriques comme c'est le cas des modèles linéaires. La littérature non-paramétrique est très abondante, en particulier pour des données indépendantes. Dans le cas dépendant, on peut citer le travail de Robinson (2011) qui a étendu l'estimateur de Nadaraya-Watson. Dans le domaine spatial, nous remarquons que l'estimateur de Nadaraya-Watson utilisé pour des données i.i.d garde certaines propriétés dans le cas spatial telle que la distribution limite. La cohérence et la théorie asymptotique de la distribution de cet estimateur ont été montrées dans un cadre conçu pour être appliqué à divers types de données spatiales, voir Robinson (2011).
Dans Robinson (2011), les variables spatiales considérées X_i en n sites notés i sont rangés dans un ordre géographique (lexicographique) dans une grille (de gauche a droite, de bas en haut).
Soit $g(X_i) = \mathbb{E}(Y_i|X_i)$ et l'erreur $U_i = Y_i - g(X_i)$ avec comme propriété :

$$Var(Y_i|X_i) = Var(U_i|X_i) = \sigma^2(X_i)$$

où $\sigma^2(X_i) \neq \sigma^2(X_j) \qquad \forall\, i \neq j$

qui introduit une hétéroscédasticité conditionnelle. Avec ces nouveaux éléments, Robinson (2011) a prouvé la consistance et la normalité asymptotique de l'estimateur de Nadaraya-Watson. En se basant sur les travaux de ce dernier, on peut étendre les travaux non-paramétriques des cadres i.i.d et temporel pour estimer l'indice de queue ainsi que les quantiles extrêmes dans un cadre spatial.

2.2.2 Estimateur d'indice de queue de Chavez-Demoulin and Guillou (2018)

Chavez-Demoulin and Guillou (2018) ont appliqué la Théorie des Valeurs Extrêmes pour des processus temporels stationnaires β-mélangeant. Ils ont construit un estimateur d'indice de queue débiaisé. Soit $(X_1, ..., X_n)$ un échantillon β-mélangeant, on rappelle qu'il vérifie :

$$\beta(m) = \sup_{p>0} \mathbb{E}\Big(\sup_{C \in B_{p+m+1}^m} (\mathbb{P}(C|B_1^p) - \mathbb{P}(C)\,) \Big) \underset{m\to\infty}{\longrightarrow} 0$$

Dans cette partie, la construction de l'estimateur asymptotiquement sans biais de Chavez-Demoulin and Guillou (2018) sera développée et étendu au domaine spatial dans le chapitre

suivant.
Pour construire leur estimateur, ces auteurs transforment l'estimateur de Hill pour corriger son biais et proposent une classe d'estimateurs d'indice de queue parmi laquelle ils construisent un estimateur asymptotiquement sans biais.
Chavez-Demoulin and Guillou (2018) se positionnent d'abord dans un cadre temporel avec une série temporelle $\beta - mixing$.
On suppose que cette série appartient au domaine de Fréchet, et donc qu'elle possède une fonction de répartition F et la fonction de quantile extrême U qui vérifie :

$U = (\frac{1}{1-F})^{\leftarrow}$ est la fonction inverse généralisée de $\frac{1}{1-F}$ telle que $\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{\gamma} \quad \forall\, x > 0$

Pour la suite, on a les conditions suivantes :
- Condition de Second Ordre (CSO) pour spécifier l'ordre de convergence de la fonction U : Soit A telle que $\lim_{t\to\infty} A(t) = 0 \quad et \quad \rho < 0$ tel que

$$\lim_{t\to\infty} \frac{1}{A(t)} \left(\ln U(tx) - \ln U(t) - \gamma \ln(x) \right) = \frac{x^{\rho}-1}{\rho},\ \forall\, x > 0$$

- Condition de régularité (CR) pour étudier la normalité asymptotique de l'estimateur : $\exists\, \epsilon > 0$ une fonction $r(.,.)$ et une suite l_n telles que
$a)$ $\frac{\beta(l_n)}{l_n} n + l_n \frac{ln(k)^2}{\sqrt{k}} \underset{n\to\infty}{\longrightarrow} 0$

$b)$ $\frac{n}{l_n k}\, cov\left(\sum_{i=1}^{l_n} (\mathbf{1}_{X_i > F(1-\frac{kx}{n})}), \sum_{i=1}^{l_n} (\mathbf{1}_{X_i > F(1-\frac{ky}{n})}) \right) \longrightarrow r(x,y)\ la\ fonction\ covariance$

$c)$ $\frac{n}{l_n k} \mathbb{E}\left(\left[\sum_{i=1}^{l_n} \mathbf{1}_{F(1-\frac{ky}{n}) < X_i < F(1-\frac{kx}{n})} \right]^4 \right) \leq C(y-x)$

A l'aide de ces hypothèses, Chavez-Demoulin and Guillou (2018) introduisent une famille d'estimateur de l'indice de queue γ de la forme :

$$\hat{\gamma}_k(K) = T_k(Q_n) = \int_0^1 ln(\tfrac{Q_n(t)}{Q_n(1)}) d(tK(t))$$

avec $Q_n(t) = X_{n-kt,n}$ (les $X_{i,n}$ sont les observations ordonnées),

$$T_k(z) = \int_0^1 ln\left(\frac{z(t)}{z(1)}\right) d(tK(t)),$$

K est un noyau qui vérifie $\int_0^1 K(t)dt = 1$ (CK).

Remarque 2.3. *La forme de l'estimateur construit ressemble à l'estimateur de Hill, en effet :*

$$\hat{\gamma}_k(K) = T_k(Q_n) = \int_0^1 ln(\tfrac{Q_n(t)}{Q_n(1)}) d(tK(t))$$

avec $Q_n(t) = X_{n-kt,n}$.
L'estimateur de Hill vu dans le chapitre précédent est :

$$\hat{\gamma}(k) \ = \ \tfrac{1}{k} \sum_{i=1}^{k-1} \left(\ln X_{n-i+1,n} - \ln X_{n-k+1,n} \right)$$

Par une formule de quadrature, on peut retrouver à partir de $\hat{\gamma}_k(K)$ *une formule similaire à celui de Hill.*

Les auteurs ont prouvé les divers résultats asymptotiques suivants sur $\hat{\gamma}_k(K)$.

Théorème 2.1. ***Normalité asymptotique de*** $\hat{\gamma}_k(K)$. *On suppose* $(X_1, X_2, ...)$ *un échantillon* $\beta -$ *mixing de fonction de répartition* F, *qui vérifie CSO,CR et CK. On a le résultat suivant :*

$$\sqrt{k}(\hat{\gamma}_k(K) - \gamma - A(\tfrac{n}{k}) \int_0^1 t^{-\rho} K(t)dt\,) \xrightarrow{loi} \gamma \int_0^1 [\,t^{-1}W(t) - W(1)\,]d(tK(t))$$

avec W(t) un processus gaussien centré.

De ce théoréme découle un corollaire :

Corollaire 2.1. *Sous les hypothèses du théorème prédédent, on a*

$$\sqrt{k}(\hat{\gamma}_k(K) - \gamma) \longrightarrow \mathcal{N}(\lambda AB(k), A\nu(k)\,)$$

Avec $AB(k) = \int_0^1 t^{-\rho} K(t)dt\,,$
$A\nu(K) = \gamma^2 \int_0^1 \int_0^1 (\,\frac{r(t,s)}{ts} - \frac{r(t,1)}{t} - \frac{r(1,s)}{s} - r(1,1)\,)\, dtK(t)\, dsK(s)$

Détermination du noyau optimal

Soit $K_\Delta(t) = \Delta K_1(t) + (1-\Delta)K_2(t)$ une combinaison de deux noyaux vérifiant (CK). Par le corollaire ci-dessus, avoir un estimateur asymptotique sans biais équivaut à $\hat{\gamma}_k(K_\Delta) = 0$:

$$\frac{\lambda}{\sqrt{k}} AB(k) = \frac{\lambda}{\sqrt{k}}(\,\Delta AB(K_1) + (1-\Delta)AB(K_2)\,) = 0 \Longleftrightarrow \Delta^* = \frac{AB(K_2)}{AB(K_2) - AB(K_1)}.$$

De cette manière, l'estimateur $\hat{\gamma}_k(K_{\Delta^*})$ est asymptotiquement sans biais.

Détermination de la variance minimale

Chavez-Demoulin and Guillou (2018) ont repris le travail de Goegebeur and Guillou (2013) pour avoir le noyau optimal :

$$K_{\Delta^*_{opt}}(t) = (\tfrac{1-\rho}{\rho})^2 - \tfrac{(1-\rho)\,(1-2\rho)}{\rho^2}\, t^{-\rho},\ t\,\epsilon\,[0,1]$$

qui implique la variance minimale $A\nu(K_{\Delta^*_{opt}}(t)\,) = \gamma^2(\frac{1-\rho}{\rho})^2$.
Comme ρ est inconnu, il peut étre approché par un estimateur $\hat{\rho}_k$.
Ainsi parmi la famille d'estimateurs proposée par Chavez-Demoulin and Guillou (2018), celui sans biais et de variance minimale est :

$$\hat{\gamma}_k(K_{\Delta^*_{opt}}) \text{ tel que } \sqrt{k}(\hat{\gamma}_k(K_{\Delta^*_{opt}}) - \gamma) \xrightarrow{loi} \mathcal{N}(0, A\nu(K_{\Delta^*_{opt}})\,)$$

Chapter 3

Estimateur de Chavez et Guillou étendu au domaine spatiale

Résumé

Dans ce chapitre, nous allons proposer un estimateur d'indice de queue pour des données spatiales à partir de l'estimateur de Chavez-Demoulin and Guillou (2018). Rappelons que cet estimateur proposé dans le domaine temporel est :

$$\hat{\gamma}_k(K) = T_k(Q_n) = \int_0^1 ln(\frac{Q_n(t)}{Q_n(1)})d(tK(t))$$

avec $Q_n(t) = X_{n-kt,n}$
Nous étendrons cet estimateur au domaine spatial en tenant compte de la dépendance spatiale, en complétant avec des simulations pour connaître le comportement de ce nouvel estimateur. Pour la suite, nous nous situons dans le domaine de Fréchet, c'est-à-dire que nous travaillons avec des lois du même type que les loi de Cauchy, Pareto... d'indice de queue $\gamma > 0$. Nous travaillerons donc avec des valeurs extrêmes qui sont des maximums.

3.1 Application de l'estimateur au domaine spatiale

Les variables spatiales considérées X_i en n sites notés i sont rangées dans un ordre géographique (lexicographique) dans une grille (de gauche a droite, de bas en haut), comme par exemple un relevé de températures en France, ou un relevé des crues des rivières. Puis on applique la propriété du β mixing :

$$\beta(m) = \sup_{p>0} \mathbb{E}\Big(\sup_{C \,\epsilon\, B_{p+m+1}^m} (\mathbb{P}(C|B_1^p) - \mathbb{P}(C)\,) \Big) \underset{m\to\infty}{\longrightarrow} 0$$

Remarque 3.1. *La différence de cet estimateur avec la littérature est qu'il tient compte de la dépendance spatiale et que son point de départ est un estimateur d'indice de queue pour une série de données stationnaires spatiales.*

Par analogie, les propriétés et les théorèmes établis pour l'estimateur de Chavez et Guillou peuvent aussi s'appliquer dans notre cas. Avec nos notations, l'estimateur construit appartient à la famille d'estimateurs de Chavez et Guillou de la forme :

$$\hat{\gamma}_k(K) = T_k(Q_n) = \int_0^1 ln(\frac{Q_n(t)}{Q_n(1)})d(tK(t))$$

avec $Q_n(t) = X_{(n-kt,n)}, t \in [0,1]$,

$T_k(z) = \int_0^1 ln(\frac{z(t)}{z(1)}) d(tK(t))$,

un noyau K qui vérifie $\int_0^1 K(t)dt = 1$ (CK).

3.2 Simulations

3.2.1 Simulation d'une loi Log-Laplace et estimateur de Hill

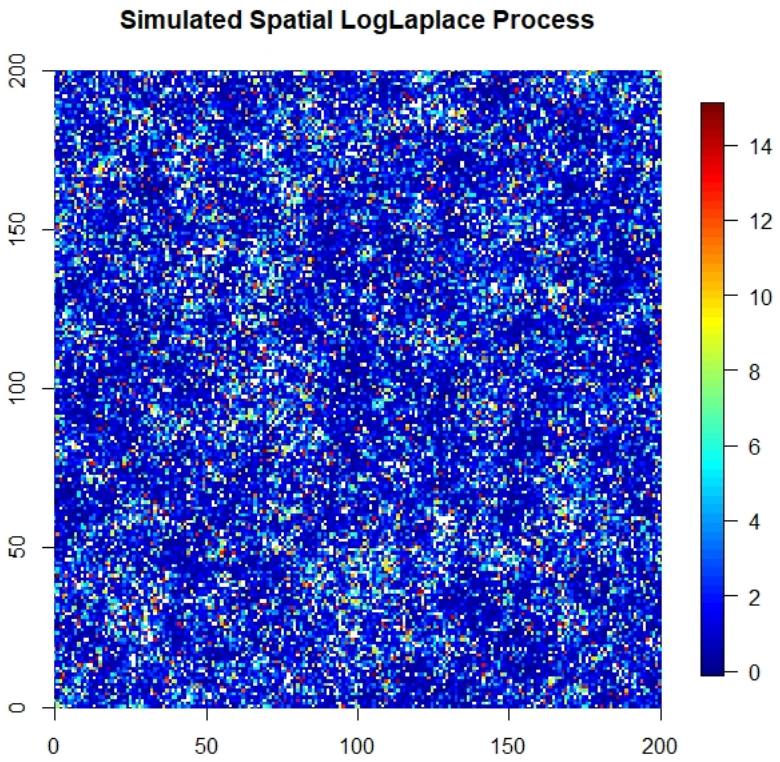
Definition 3.1. *Soit $Z = (Z(s), s \in S \subset \mathbb{R}^2)$ un processus spatial gaussien stationnaire au second ordre, centrée de fonction de covariance $\xi(h) = corr(Z(s+h), Z(s)), s+h \in S$. Soit Y une variable aléatoire telle que Y^2 suit une loi exponentielle de paramètre 2, indépendante de Z. Alors le processus $X = (X(s), s \in S)$ défini par*

$$X(s) = YZ(s)$$

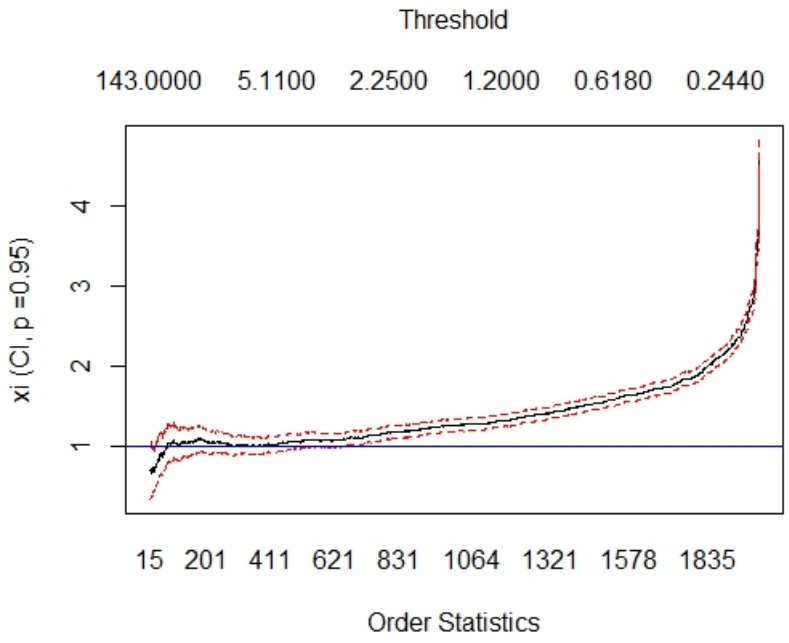
est dit processus spatial de Laplace $L(\xi)$, voir Opitz (2016). Si $Z(s)$ est une normale standard alors X est dit processus de Laplace Standard. La fonction de covariance de X est 2ξ.

Definition 3.2. *Si X est un processus de Laplace standard alors $U = \{U(s),\ s \in S \subset \mathbb{R}^2\}$, $U_s = e^{X_s}$ est un processus dont les maginales sont log Laplace d'indice de queue $\gamma = 1$ et de densité*

$$f_{U(s)} = \begin{cases} 1/(2t^2), & t \geq 1 \\ 1/2, & 0 \leq t \leq 1 \\ 0 & sinon \end{cases}$$



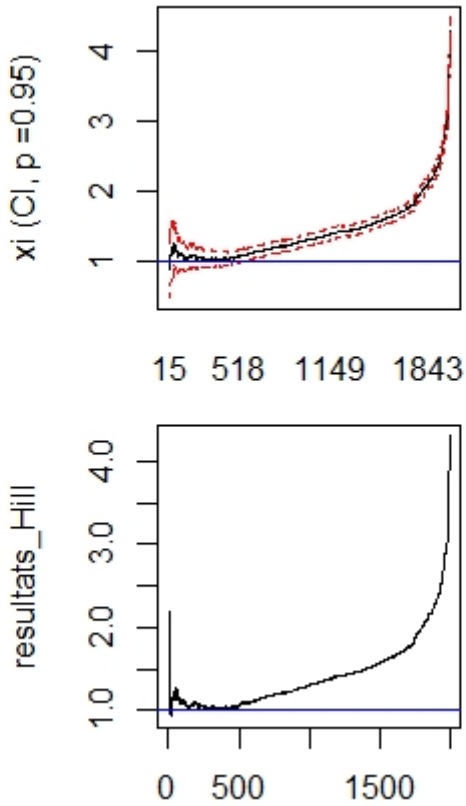
On peut apercevoir ci-dessus un processus spatial Log-Laplace, prenant ses valeurs dans $[0, 6]$ principalement. On peut constater qu'il n'y a pas beaucoup de grandes valeurs. Une loi de Log-Laplace est utilisée en pratique pour modéliser la température de l'air, faire des études sur le climat, modéliser les échanges commerciaux ou les tailles d'entreprises. La valeur de l'indice de queue γ est connue pour ce processus LogLaplace ($\gamma = 1$). Nous essayons de le retrouver à l'aide de simulations sur le logiciel R. Premièrement, nous utilisons la fonction "hill" du package "evir". Soit $X_i, i = 1, ..., n$ les données ordonnées d'un processus spatial LogLaplace :



Ce graphique donne la valeur de l'indice de queue pour $k \in 1, ..., n$, où k représente le nombre de plus grandes valeurs retenues dans l'échantillon. En bleu, se trouve la droite d'ordonnée $h = \gamma = 1$. On peut remarquer que le choix de k est important, en effet plus k augmente, plus l'estimateur diverge vers l'infini.

Pour mieux comprendre la création de cet estimateur, nous pouvons créer notre propre fonction "Hill" de l'estimateur de Hill suivant :

$$\hat{\gamma}(k) \; = \; \frac{1}{k} \sum_{i=1}^{k-1} \left(ln\, X_{n-i+1,n} \; - ln\, X_{n-k+1,n} \right)$$



On peut voir ici une comparaison entre l'estimateur du package "evir" (celui du haut) et notre fonction (celui du bas). Une vraie ressemblance peut être aperçue. On peut remarquer plus précisément en regardant le second graphique que l'estimateur commence à diverger à partir de $k = 500$. En effet, k étant le nombre de plus grandes valeurs retenues, il influence beaucoup sur la valeur de γ_k^H, qui s'écrit en termes de logarithmes. Plus k est grand, plus on garde de petites valeurs pour construire l'estimateur et l'estimateur diverge alors. Par exemple, pour un échantillon de 2000 variables i.i.d suivant une loi Log-Laplace, pour k =500, nous obtenons grâce à notre fonction Hill $\gamma = 1.080455$ alors que pour une valeur de k=1800, $\gamma = 1.861251$, et pour une valeur proche de 2000 (prenons 1990), on obtient $\gamma = 3.327252$. Sachant que théoriquement $\gamma = 1$ pour cette loi, on remarque que plus $k \to n$, plus le γ estimé augmente.

3.2.2 Simulation d'une loi de Log-Laplace et notre estimateur

Rappelons que notre estimateur s'écrit de la forme :

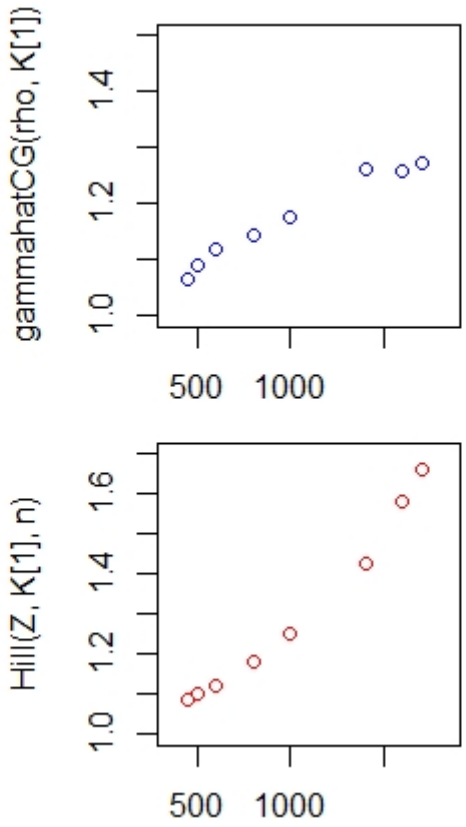
$$\hat{\gamma}_k(K) = T_k(Q_n) = \int_0^1 ln(\frac{Q_n(t)}{Q_n(1)}) d(tK(t))$$

On travaille toujours avec un échantillon de taille n=2000 d'une loi Log-Laplace d'indice de queue $\gamma = 1$. On va tester notre estimateur pour retrouver une approximation de cette valeur et le comparer à l'estimateur de Hill. Dans la construction de notre estimateur, il nous faut un noyau. Dans le chapitre état de l'art, nous avons vu qu'il existait un estimateur d'indice de queue optimal $\gamma_k(K_{\Delta^*_{opt}})$ de noyau donné par

$$K_{\Delta^*_{opt}}(t) = (\frac{1-\rho}{\rho})^2 - \frac{(1-\rho)(1-2\rho)}{\rho^2} t^{-\rho},\ t\,\epsilon\,[0,1]$$

Par la suite, on fixera la valeur de $\rho = -4$. Dans le cas de notre estimateur, pour une taille d'échantillon n=2000, k=500, nous obtenons une valeur de $\gamma = 1.089926$. On

peut comparer notre estimateur à celui de Hill pour plusieurs valeurs de k pour le même échantillon. On obtient graphiquement :



A première vue, l'estimateur de Hill tend à diverger quand $k \to n$ beaucoup plus vite que notre estimateur. On peut vérifier notre hypothèse avec le tableau des valeurs ci-dessous.

valeur de K	Estimateur Hill	Estimateur Chavez Guillou
450	1.0871	1.0664
500	1.1039	1.0899
600	1.1227	1.1210
800	1.1834	1.1451
1000	1.2542	1.1761
1400	1.4258	1.2616
1600	1.5805	1.2571
1700	1.6619	1.2741

Le tableau vérifie bien notre hypothèse, selon laquelle l'estimateur de Hill diverge plus vite, et donc est moins performant que celui de Chavez et Guillou. Cela peut s'expliquer par le fait que l'estimateur de Hill est biaisé, alors que celui de Chavez et Guillou est un estimateur débiaisé. De plus, on sait que l'estimateur de Hill est construit à partir de données indépendantes, or ici nous prenons en compte un processus spatial stationnaire, avec comme fonction de corrélation :

$$\xi(h) = corr(Z(s+h), Z(s)), s = h \in S$$

Empiriquement, on a montré que l'estimateur de Chavez et Guillou appliqué au cas spatial dans le domaine de Fréchet est plus efficace que celui de Hill.

Conclusion

A travers ce TER, nous avons pu comprendre l'utilité et la mise en application de la Théorie des valeurs extrêmes. Notamment dans le cadre spatial, un cas avec moins d'investigations que le cas temporel car plus délicat. En effet, dans le cas temporel, la dépendance estliée à la notion d'ordre dans le temps, or dans le cas spatial, la notion d'ordre n'est pas évidente, une valeur en un site donné ne dépend pas de la valeur précédente, mais de valeurs à des endroits proches, dans un rayon défini. Nous avons pu aussi découvrir la construction de l'estimateur de l'indice de queue et les résultats déjà existant pour son estimation. A partir de là, nous avons essayé d'étendre l'estimateur Chavez-Demoulin and Guillou (2018) au domaine spatial en supposant un échantillon d'un processus spatial stationnaire, dont on avait ordonné les données selon un ordre lexicographique. Cependant,nous n'avons pas pu allé très loin dans le développement de cet estimateur, par la suite, il sera possible d'essayer notre estimateur à un échantillon de données réelles (températures, montée des eaux...). Pour encore aller plus loin, il serait intéressant d'essayer cet estimateur dans un cadre spatio-temporel et donc de prendre en compte la double contrainte du temps et de l'espace.

Annexes

Preuve .1. ***Théorème des valeurs extrêmes***
Cette preuve est celle présente dans le cours de Gardes and Girard (2004) Pour bien démontrer le théorème, il faut introduire un lemme et une proposition.

Proposition .1. *Si $F \in DAM(G)$ et $F \in DAM(H)$ alors nécessairement G et H sont du même type :*

$$G(x) = H(ax + b)$$

Plus précisément, si $a_n > 0$, $u_n > 0$, b_n et v_n sont telles que :

$$F^n(a_n x + b_n) \to G(x)$$

et

$$F^n(u_n x + v_n) \to H(x),\ n \to \infty,\ \forall\, x \in \mathbb{R}$$

alors on a nécessairement :

$$\frac{u_n}{a_n} \to a$$

et

$$\frac{(v_n - b_n)}{a_n} \to b,\ n \to \infty$$

Lemme .1. *Soit F une fonction de répartition non dégénérée. S'il existe $a > 0$ et $b \in \mathbb{R}$ tels que :*

$$F(ax + b) = F(x)\, \forall\, x \in \mathbb{R}$$

, alors $a = 1, b = 0$

Soit $F \in DAM(G)$, il existe deux suites $a_n$et b_n telle que :

$$F^{[nt]}(a_{[nt]} x + b_{[nt]}) \to G(x) \tag{1}$$

quand $n\to \infty$,$\forall t > 0$ et $x\in \mathbb{R}$.
De plus,

$$F^{[nt]}(a_n x + b_n) = (F^n(a_n x + b_n))^{[nt]/n} \to G^t(x) pour [nt]/n \to t quand n \to \infty. \tag{2}$$

Au vue de (1),(2) et de la proposition, cela implique que :

$$\frac{a_n}{a\,[nt]} \to \alpha(t) > 0, \frac{b_n - b_{[nt]}}{a_{[nt]}} \to \beta(t)$$

et que G et G^tsont de même type :

$$G^t(x) = G(\alpha(t)x + \beta(t)) \tag{3}$$

Cette preuve consiste à résoudre l'équation (3). Pour tout t > 0 et s> 0 ,$G^{st}(x)$ peut être réecrit de trois façons différentes:

$$G(\alpha(st)x + \beta(st)) \tag{4}$$

$$G^s(\alpha(t)x + \beta(t)) = G(\alpha(s)\alpha(t)x + \alpha(s)\beta(t) + \beta(s)) \tag{5}$$

$$G^t(\alpha(s)x + \beta(s)) = G(\alpha(s)\alpha(t)x + \alpha(t)\beta(s) + \beta(t)) \tag{6}$$

De (4),(5),(6) et d'après le Lemme, on a que $\forall s > 0, t > 0$:

$$\alpha(st) = \alpha(s)\alpha(t) \tag{7}$$

$$\beta(st) = \alpha(s)\beta(t) + \beta(s) \tag{8}$$

$$\beta(st) = \alpha(t)\beta(s) + \beta(t) \tag{9}$$

Il est facile de montrer que l'unique solution de l'équation (7) est $\alpha(t) = t^A$ pour tout t > 0 et A$\in \mathbb{R}$.
Il existe alors trois cas :

Cas 1: A=0

On a alors $\alpha(t) = 1$ pour tout t>0. On peut donc réecrire (9) comme :$\beta(st) = \beta(s) + \beta(t)$. L'unique solution de cette équation est $\beta(t) = \beta(e)log(t)$ pour tout t>0.
On remplace cette solution trouvée dans (3) , on obtient alors une nouvelle équation pour G:

$$\forall x \in \mathbb{R}, t > 0\, G^t(x) = G(x + \beta(e)log(t))$$

De plus comme G est une fonction non dégénéée, il existe $x_0 \in \mathbb{R}$ tel que 0<$G(x_0)$<1.Ainsi :

$$\forall t > 1\, G(x_0) > G^t(x_0) = G(x_0 + \beta(e)log(t))$$

Cela implique que $\beta(e) < 0$. On note $\sigma = -\beta(e) > 0$. Maintenant prouvons que $\forall x \in \mathbb{R}\ 0 < G(x) < 1$
On prend $x_0 \in \mathbb{R}, G(x_0) = 1$, il suit que $\forall t > 0, G^t(x_0) = 1$ et $G(x_0 - \sigma log(t)) = 1$, on a alors G:=1 ce qui est absurde. Donc $\forall x \in \mathbb{R}$, $G(x) < 1$ et de manière similaire on obtient $\forall x \in \mathbb{R}\ \mathrm{G(x)} > 0$
Donc on a bien $\forall x \in \mathbb{R}$, $0 < G(x) < 1$.
En introduisant $\theta = -log(G(0)) > 0$, on rapelle que $\forall x \in \mathbb{R}, t > 0\, G^t(x) = G(x - \sigma log(t))$
On a en particulier pour x=0 :

$$exp(-\theta t) = G^t(0) = G(-\sigma log(t))$$

Maintenant on pose $u = -\sigma log(t)$.Ainsi on a :

$$\forall u \in \mathbb{R}\ G(u) = exp(-\theta exp(-u/\sigma))$$

Cela montre que G est du même type que H_0 c'est a dire:

$$G(x) = H_0(x/\sigma + log(\theta)$$

Cas 2: A<0

On introduit $\gamma = -A > 0$, on a alors $\forall t > 0\, \alpha(t) = t^{-\gamma}$.
De (8) et (9) on a :

$$\forall s > 0, t > 0\, \alpha(s)\beta(t) + \beta(s) = \alpha(t)\beta(s) + \beta(t)$$

$$\iff (1 - \alpha(s))\beta(t) = (1 - \alpha(t))\beta(s)$$

$$\iff c = \frac{\beta(t)}{1 - \alpha(t)} = \frac{\beta(s)}{1 - \alpha(s)}$$

On a alors $\beta(t) = c\,(1 - \alpha(t)) = c(1 - t^{-\gamma})$ avec $c \in \mathbb{R}$ et $t > 0$. En remplaçant dans l'équation (3), on a :
$G^t(x) = G(t^{-\gamma}(x - c) + c)\, \forall x \in \mathbb{R} et\, t > 0$
Si on pose y=x-c, et $G_*(y) = G(y + c)$ la fonction de répartition translatée en y+c, on obtient :

$$G_*^t(y) = G_*(t^{-\gamma}y)\, \forall\, t > 0\, et\, y \in \mathbb{R}$$

.
De plus, comme $G_* et G$ sont du même type, il est suffisant de résoudre l'équation du dessus. Soit $y_0 \in \mathbb{R} tel$ que $G_*(y_0) < 1$. En faisant tendre $t \to \infty$, l'équation précédente devient :

$$G_*(0) = 0 \implies G_*(y) = 0\, \forall\, y \leq 0.$$

Une preuve similaire au cas A=0 peut être faite en montrant que $0 < G(1) < 1$ en introduisant $\sigma = -lnG_*(1) > 0$. En considérant le cas $y = 1 dans$ l'équation, on obtient :
$G_*(t^{-\gamma}) = G_*^t(1) = exp(-\sigma t)\, \forall\, t > 0.$
Par conséquent, en procédant au changement de variable $u = t^{-\gamma}$, on obtient :

$$G_*(u) = exp(-\sigma u^{-\frac{1}{\gamma}})\, \forall\, u > 0$$

On peut voir que G est du même type que H_γ, soit :

$$G_*(t) = H_\gamma(x\gamma^{-1}\sigma^{-\gamma} - \gamma^{-1})\, \forall\, x \in \mathbb{R}$$

Cas 3 : A>0

Ce cas est similaire au cas A<0, il suffit de poser $\gamma = A > 0$.

Bibliography

Bean, S. J. and Tsokos, C. P. (1980). Developments in nonparametric density estimation. *International Statistical Review/Revue Internationale de Statistique*, pages 267–287.

Chavez-Demoulin, V. and Guillou, A. (2018). Extreme quantile estimation for β-mixing time series and applications. *Insurance: Mathematics and Economics*, 83:59–74.

De Haan, L. and Ferreira, A. (2007). *Extreme value theory: an introduction*. Springer Science & Business Media.

Drees, H. (1995). Refined pickands estimators of the extreme value index. *The Annals of Statistics*, pages 2059–2080.

Drees, H. et al. (2000). Weighted approximations of tail processes for $\backslash\beta$-mixing random variables. *The Annals of Applied Probability*, 10(4):1274–1301.

Gardes, L. and Girard, S. (2004). A pickands type estimator of the extreme value index. *arXiv preprint math/0403299*.

Gardes, L. and Girard, S. (2005). Asymptotic distribution of a pickands-type estimator of the extreme-value index. *Comptes Rendus Mathematique*, 341(1):53–58.

Goegebeur, Y. and Guillou, A. (2013). Asymptotically unbiased estimation of the coefficient of tail dependence. *Scandinavian Journal of Statistics*, 40(1):174–189.

Gomes, M. I., e Castro, L. C., Alves, M. I. F., and Pestana, D. (2008). Statistics of extremes for iid data and breakthroughs in the estimation of the extreme value index: Laurens de haan leading contributions. *Extremes*, 11(1):3–34.

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

Kempeners, T. (2010). *Quelques modÃ¨les de rÃ©gression*.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Opitz, T. (2016). Modeling asymptotically independent spatial extremes based on laplace random fields. *Spatial Statistics*, 16:1–18.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Robinson, P. M. (2011). Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics*, 165(1):5–19.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837.

Tsybakov, A. B. (2003). *Introduction à l'estimation non paramétrique*, volume 41. Springer Science & Business Media.

Zhou, S. (1986). An experimental assessment of resource queue lengths as load indices. Technical report, CALIFORNIA UNIV BERKELEY COMPUTER SCIENCE DIV.