Olivier CÔME
Gueladio NIASSE
N° étudiant Olivier CÔME: 22110708
N° étudiant Gueladio NIASSE: 21714307
olivier.come@etu.umontpellier.fr
gueladio.niasse@etu.umontpellier.fr
Master 2
Statistique et Science des Données
Université de Montpellier

Projet: Statistique des évènements extêmes et applications

HAX005X : Valeurs extrêmes
Rédigé le 26 Février 2023 en LaTeX

Sommaire

1 Introduction

Dans le cadre de ce projet, nous allons modéliser le comportement extrême des vagues dans le golfe du lion à l'aide des méthodes vue à travers l'unité d'enseignement HAX005X "valeurs extrêmes". D'après la source [1], le golf s'étale sur 220 kilomètres de la Camargue à la frontière espagnole. La côte, essentiellement sableuse, a été façonnée par la houle (la mer gagnant souvent les terres par élévation du niveau marin) et l'érosion côtière. L'apport de sédiments en provenance des fleuves a également permis de faire avancer le rivage pendant de longues périodes. Des formations de lagunes "comme les graus" (parfois temporaires) ont pu apparaitre et ont permis de faire communiquer les étangs littoraux avec la mer. Le golf du Lion est donc un milieu naturellement dynamique. C'est dans ce contexte que nous étudions le comportement extrême des vagues à cet endroit, de façon univariée dans un premier temps puis de façon bivariée dans un second temps.

2 Présentation des données

Les données que nous avons à notre disposition pour ce projet sont les suivantes :

- **DonneesStations**
- **DonneesVagues**

Pour réaliser les analyses, nous utiliserons essentiellement les données en provenance du dataframe *DonneesVagues* qui correspondent à des enregistrements de hauteurs de vagues significatives horaires de 20 stations situées dans le golf du Lion. Ces mesures horaires ont été enregistrée de 1961 à 2012. Ce dataframe est constitué de 464280 observations (en lignes) et de 21 variables (en colonnes). Decrivons un peu plus en détaille les colonnes de ce dataframe *DonneesVagues* :

- **date** nous renseignent sur la date (format année mois jours) et l'heure précise (format heure minute seconde) à laquelle a été enregistrée la mesure.
- **station 1 à 20** nous renseigne sur les hauteurs de la vagues mesurées

Le deuxième dataframe *DonneesStations*, quant à lui nous renseigne sur les coordonnées géographique des 20 stations. Il est composé de 20 observation (lignes) et de 5 variables (colonnes). Décrivons un peu plus en détaille les colonnes de ce dataframe.

- **lon :** floatant correspondant à la longitude de la station
- **lat :** floatant correspondant à la latitude de la station
- **depth :** floatant correspondant à la profondeur de la station
- **stationName :** Chaine de caractère indiquant le nom de la station

Les figures 1 et 2 sont des captures d'écran d'une portion des dataframe *DonneesVagues* et *DonneesStations*. La figure 3 correpond au nuage de points des données du dataframe *DonneesVagues*.

	date	station1	station2	station3	station4	station5	station6	station7	station8
1	1961-01-01 02:00:00	0.5120391	0.13847668	0.1775285	0.1967454	0.3011920	0.5258080	0.2734331	0.3345910
2	1961-01-01 03:00:00	0.5465476	0.08934399	0.1447777	0.1692832	0.3607951	0.5600442	0.2940980	0.3694929
3	1961-01-01 04:00:00	0.5894736	0.06477053	0.1429445	0.1745569	0.4378489	0.6025565	0.3268866	0.4230243
4	1961-01-01 05:00:00	0.6452147	0.05305877	0.1727725	0.2093817	0.5270881	0.6563302	0.3789678	0.4797138
5	1961-01-01 06:00:00	0.7143217	0.04773302	0.2210754	0.2612950	0.6135621	0.7224682	0.4465642	0.5400973
6	1961-01-01 07:00:00	0.7871489	0.04815382	0.2775594	0.3227399	0.6846716	0.7935590	0.5217415	0.6003126
7	1961-01-01 08:00:00	0.8617782	0.05834556	0.3420269	0.3958148	0.7518139	0.8599985	0.5987842	0.6319415
8	1961-01-01 09:00:00	0.9462464	0.08159746	0.4178586	0.4796206	0.8359310	0.9354858	0.6765540	0.6421875
9	1961-01-01 10:00:00	1.0216480	0.11668154	0.4989500	0.5657550	0.9561845	1.0071346	0.7399866	0.6414675
10	1961-01-01 11:00:00	1.0811893	0.16052766	0.5806396	0.6535683	1.1047519	1.0663446	0.7936877	0.6346608
11	1961-01-01 12:00:00	1.1294807	0.21032791	0.6651449	0.7451116	1.2539893	1.1148334	0.8430105	0.6261106
12	1961-01-01 13:00:00	1.1729137	0.26388204	0.7553571	0.8425921	1.4017688	1.1577642	0.8891504	0.6206052
13	1961-01-01 14:00:00	1.2247154	0.32077914	0.8476549	0.9395052	1.4871451	1.2081485	0.9299158	0.6271747
14	1961-01-01 15:00:00	1.2902025	0.38350457	0.9268072	1.0191746	1.5202062	1.2706363	0.9638701	0.6566994

FIGURE 1 – Extrait du dataframe *DonneesVagues*

	indexStation	lon	lat	depth	stationName
1	1	4.6400	42.0600	2326.11743	Lion
2	3	4.1400	43.4600	8.91505	EspigobsLa
3	4	6.7070	43.2960	146.92639	StMaximeLa
4	6	6.7400	43.3000	221.32100	StMaximeGr
5	7	7.8000	43.4000	2324.22217	61001
6	8	4.7000	42.1000	2337.85327	61002
7	9	6.2070	42.9333	622.45752	61004
8	10	9.2810	43.0680	368.80164	61005
9	11	9.0833	43.8500	1119.25525	61010
10	12	7.2290	43.6349	201.81612	61187
11	13	3.1683	42.4883	18.77695	61188
12	14	3.7785	43.3710	29.17235	61190
13	15	3.1250	42.9167	39.00000	61191
14	16	3.6500	41.9170	1222.56653	61196

FIGURE 2 – Extrait du dataframe *DonneesStations*

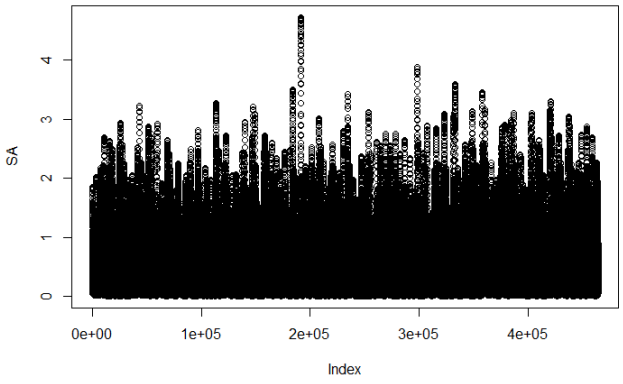


FIGURE 3 – Nuage de points des données *DonneesVagues*

Dans ce projet nous réaliserons dans un premier les analyses dans le cas univarié, c'est à dire en ne considérant qu'une seule station puis nous ferons une analyse bivarié en considérant cette fois-ci plusieurs stations.

3 Partie Univariée

Dans cette partie nous serons dans le cas univarié en nous considérerons la station 2. Pour mener à bien les analyses nous mettrons en œuvre deux approches. Dans un premier temps nous utiliserons la méthode *GEV* (Generalized Extreme Value) puis dans un second temps nous utiliserons la méthode *GPD*(Generalized Pareto distribution).

3.1 Approche GEV

Pour réaliser l'étude avec l'approche GEV nous allons nous aider de la méthode par blocs.
Au départ, on suppose avoir des réalisations indépendantes et de même loi F d'un certain phénomène d'intérêt : $X_1, X_2, \ldots, X_k$.
Dans notre cas nous avons la station 2 comme station de référence (SA). Pour obtenir un échantillon de max, on découpe les données en m blocs de même taille n :

$$\underbrace{X_1, X_2, \ldots, X_n}_{Z1} \,|\, \underbrace{X_{n+1}, X_{n+2}, \ldots, X_{2n}}_{Z2} \,|\, \ldots \,|\, \underbrace{X_{(m-1)n+1}, \ldots, X_{mn}}_{Zm} \,|$$

On obtient alors un échantillon de m réalisations de loi GEV : $Z_1, Z_2, \ldots, Z_m$.
Afin de faire des analyses de bonnes qualités il est nécessaire de sélectionner une taille de bloc n adéquat. comme il a été dit précédemment, le dataframe *DonneesVagues* est composé de 464280 lignes. Afin de choisir une bonne valeur de n, nous avons fait un script (disponible en annexe) qui affiche tous les diviseurs de 464280, ces derniers sont affichés en figure 4. À partir de là nous avons pu tester plusieurs valeurs de m et nous avons finalement pris $m = 2980$. Nous aurons donc 2980 blocs tous de même taille $n = 159$. Nous allons maintenant justifier pourquoi nous avons décidé de retenir cette valeur.
Les points du graphique de la figure 5, correspondent aux maximums de chaque blocs. Comme on peut le voir, le nuage de point obtenu est assez dispersé ce qui nous conforte dans l'idée que le paramètre n a été bien choisi.
D'autre part, nous allons également nous appuyer sur le graphique du *Quantile plot* afin de voir si le paramètre n a bien été choisi. On rappel que le Quantile plot est le nuage de points :

$$\left\{\left(\widehat{G}^{-1}\left(\frac{i}{m+1}\right), z_{i,m}\right), i = 1, \ldots, m\right\}$$

Où

$$\widehat{G}(\frac{i}{m+1}) = \widehat{\mu} - \frac{\widehat{\sigma}}{\widehat{\gamma}}\left[1 - \left(-log\left(\frac{i}{i+1}\right)\right)^{-\widehat{\gamma}}\right]$$

Si l'on observe le graphique de la figure 6, on s'aperçoit que les points du nuages sont proches de la diagonale (ligne bleu) donc l'ajustement est de bonne qualité, la taille des blocs retenu pour l'approche GEV a donc bien été choisi.

```
> print(li_divisor)
 [1]     1     2     3     4     5     6     8    10    12    15    20
[12]    24    30    40    53    60    73   106   120   146   159   212
[23]   219   265   292   318   365   424   438   530   584   636   730
[34]   795   876  1060  1095  1272  1460  1590  1752  2120  2190  2920
[45]  3180  3869  4380  6360  7738  8760 11607 15476 19345 23214 30952
[56] 38690 46428 58035 77380 92856 116070 154760 232140 464280
```

FIGURE 4 – liste des diviseurs de 464280

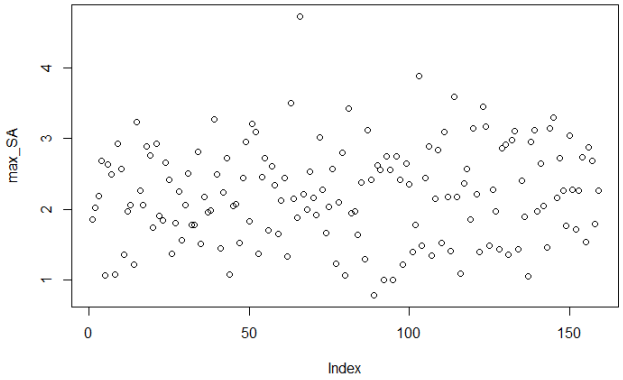


FIGURE 5 – nuage de points obtenu avec la méthode par bloc

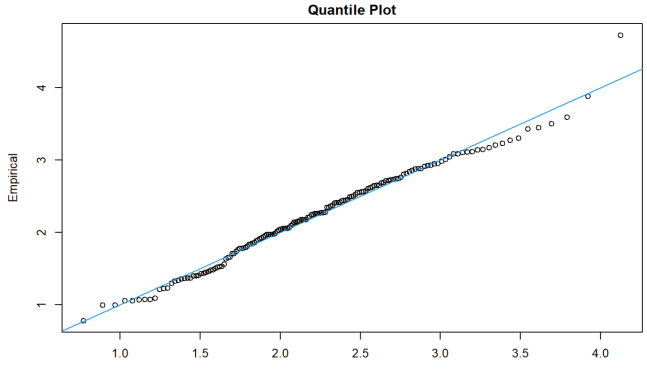


FIGURE 6 – Quantile plot

3.1.1 Niveaux de retour associés aux périodes de retour 100, 500 et 1000

Dans cette sous-section, nous allons nous intéresser aux niveaux de retour $x_{\frac{1}{T}}$ associés aux périodes de retour $T = \frac{1}{q}$ pour $T \in \{100, 500, 1000\}$. L'on souhaite connaître le niveau qui sera dépassé tout les 100, 500 et 1000 ans.

Périodes ou Années T / Niveau de retour	$T = 100$	$T = 500$	$T = 1000$
$x_{\frac{1}{T}}$	3.99003	4.413046	4.563588

TABLE 1 – Niveaux de retour associés aux périodes de retour T

Périodes ou Années T / Intervalles de confiance	$T = 100$	$T = 500$	$T = 1000$
IC de niveau $1 - \alpha = 95\%$	$[3.691942, 4.288119]$	$[3.974492, 4.851601]$	$[4.059768, 5.067409]$

TABLE 2 – Intervalles de confiance des niveaux de retour $x_{\frac{1}{T}}$

En regardant les valeurs des niveaux de retour affichés dans le tableau 1, on constate qu'ils sont bien tous compris dans leurs intervalles de confiance respectifs (voir tableau 2). Cela nous confirme une fois de plus que la taille n des blocs choisis pour l'approche GEV est bonne.
D'autre part, on remarque que plus la période de retour T est grande et plus le niveau de retour augmente. Cette augmentation des hauteurs de vagues peut s'expliquer par la monté du niveau de la mer au file des années. On pourrait donc se demander si des plages comme celles de Carnon, Palavas ou encore La Grande-Motte existeront toujours dans 1000 ans ?

3.2 Approche GPD

Pour mener cette fois-ci l'étude avec l'approche GPD, il faut déterminer la valeur du seuil à utiliser. Une fois de plus, si l'on souhaite réaliser de bonnes analyses, il faut être vigilant au seuil que l'on choisit. Afin de déterminer ce dernier, nous allons nous aider de la figure 7, sur laquelle est tracée le diagramme de durée de vie résiduelle moyenne (Mean residual life plot).

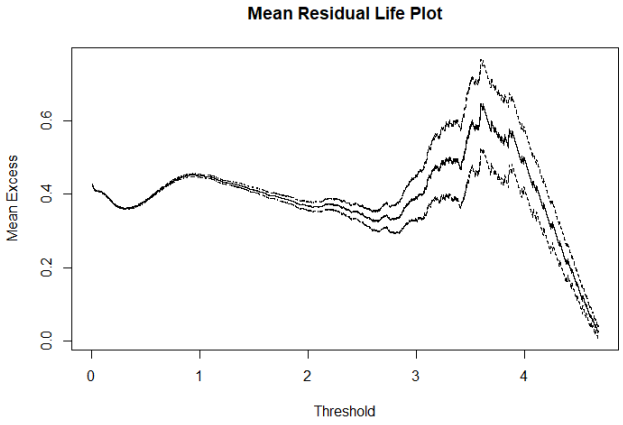
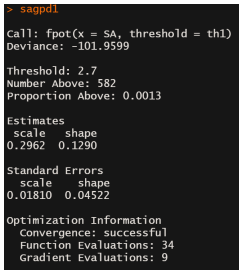
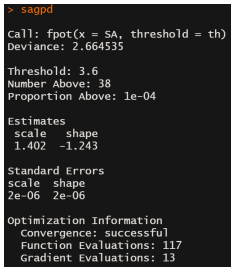


FIGURE 7 – diagramme de durée de vie résiduelle moyenne

Sur la figure 7, on voit que les pics significatifs sont atteints pour des valeurs de seuil (Threshold) valant approximativement 2.7 et 3.65. Nous avons donc testé ces deux valeurs de seuil et nous avons finalement retenu la valeur 3.65. Nous allons maintenant justifier pourquoi nous avons décidé de retenir cette dernière valeur.

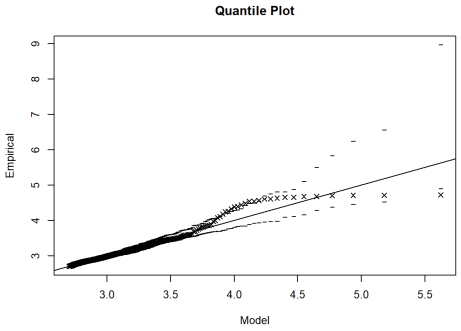


(a) seuil ≈ 2.7

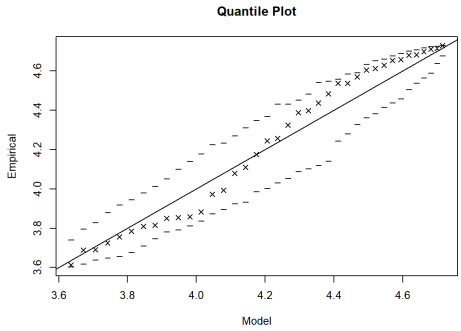


(b) seuil ≈ 3.65

FIGURE 8 – Résultats donnés en sortie de la fonction *fpot* pour les 2 seuils



(a) seuil ≈ 2.7



(b) seuil ≈ 3.65

FIGURE 9 – Quantile plot des 2 seuils

En observant les résultats affichés sur les captures d'écran de la figure 8, on voit que la proportion de valeurs dépassant le seuil (proportion above) est plus faible dans le cas où ce dernier vaut ≈ 3.65 que dans le cas où il vaut ≈ 2.7.
De plus, en observant les graphiques (*Quantile plot*) de la figure 9, on remarque que le nuage de point du seuil ≈ 3.65 est plus proche de la droite en diagonale que le nuage de points du seuil ≈ 2.7.
Ces résultats nous ont donc poussé à choisir le seuil= 3.65 car il nous permettra de faire une analyse de meilleur qualité qu'avec l'autre valeur de seuil.

3.2.1 Niveaux de retour associés aux périodes de retour 100, 500 et 1000

Maintenant que nous avons déterminer le seuil (threshold), nous allons nous intéresser comme dans le cas de l'approche GEV, aux niveaux de retour $x_{\frac{1}{T}}$ associés aux périodes de retour $T = \frac{1}{q}$ pour $T \in \{100, 500, 1000\}$. L'on souhaite connaître le niveau qui sera dépassé tout les 100, 500 et 1000 ans avec l'approche GPD.

Périodes ou Années T / Niveau de retour	$T = 100$	$T = 500$	$T = 1000$
$x_{\frac{1}{T}}$	4.686475	4.72502	4.723653

TABLE 3 – Niveaux de retour associés aux périodes de retour T

Quand on regarde le tableau 3, on constate que le niveau de retour pour la période $T = 500$ est supérieur à celui pour la période $T = 100$. Cependant, pour la période $T = 1000$, la valeur de $x_{\frac{1}{T}}$ est inférieur à celle trouvée pour $T = 500$. Ce qui est différent des résultats obtenus avec l'approche GEV. En effet, pour rappel, avec l'approche GEV on avait trouvé que plus la période de retour T était grande et plus le niveau de retour augmentait, ce qui était cohérent avec la monté des eaux. Dans le cas de ces données, l'approche GEV est donc peut être préférable car elle fournit des résultats plus cohérents que ceux obtenus avec

l'approche GPD.

4 Partie bivariée

Dans cette nouvelle partie, on se place maintenant dans le cas bivarié. Pour mener à bien les analyses nous mettrons en œuvre l'approche *GEV* (Generalized Extreme Value). L'objectif sera de prédire la valeur du quantile extrême z_p vérifiant :

$$\mathbb{P}(X > z_p | Y > y)$$

avec p petit et où Y correspond à la hauteur de vagues associées à la station 2 (S_A) et X à la hauteur de vagues associées à la station 9 (S_B).

5 Bibliographie

[1] `https://reporterre.net/Le-golfe-du-Lion-est-tres-vulnerable-a-la-montee-des-eaux`
page web rédigée par Alexandre Brun et Benoît Devillers