Modélisation des Extrêmes et des Records Application aux divers domaines

Zaher KHRAIBANI

Lebanese University-Faculty of Sciences

12 Mars 2020

Université Libanaise

laboratoire eau environnement systemes urbains

Outline

Université Libanaise

laboratoire eau environnement systemes urbains

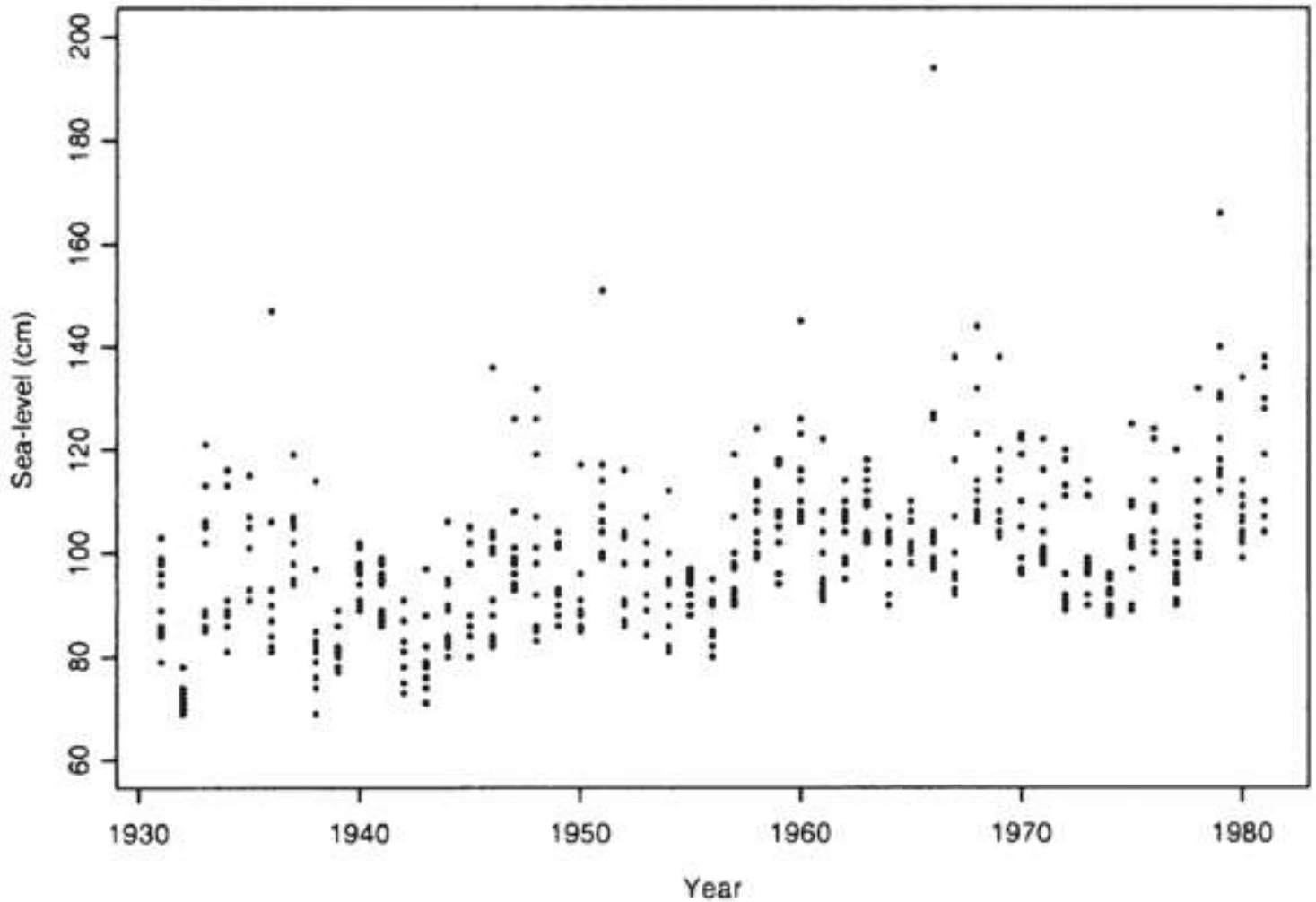
1. **Oil rig protection against wind and wave in Lebanon.** *Communications in Statistics: Case Studies, Data Analysis and Applications*, 2020, 6:2, 191-214.

2. **Application of records theory on the COVID-19 pandemic in Lebanon: Prediction and prevention.** *Epidemiology and Infection*, 2020, 148, E192.

3. Collaboration avec LEESU

Qu'est-ce qu'une valeur extrême?

On se concentre sur les valeurs extrêmes univariées, celles trouvées quand on regarde la distribution des valeurs dans une seule dimension


Sea-level (cm)
60
80
100
120
140
160
180
200
1930
1940
1950
1960
1970
1980
Year

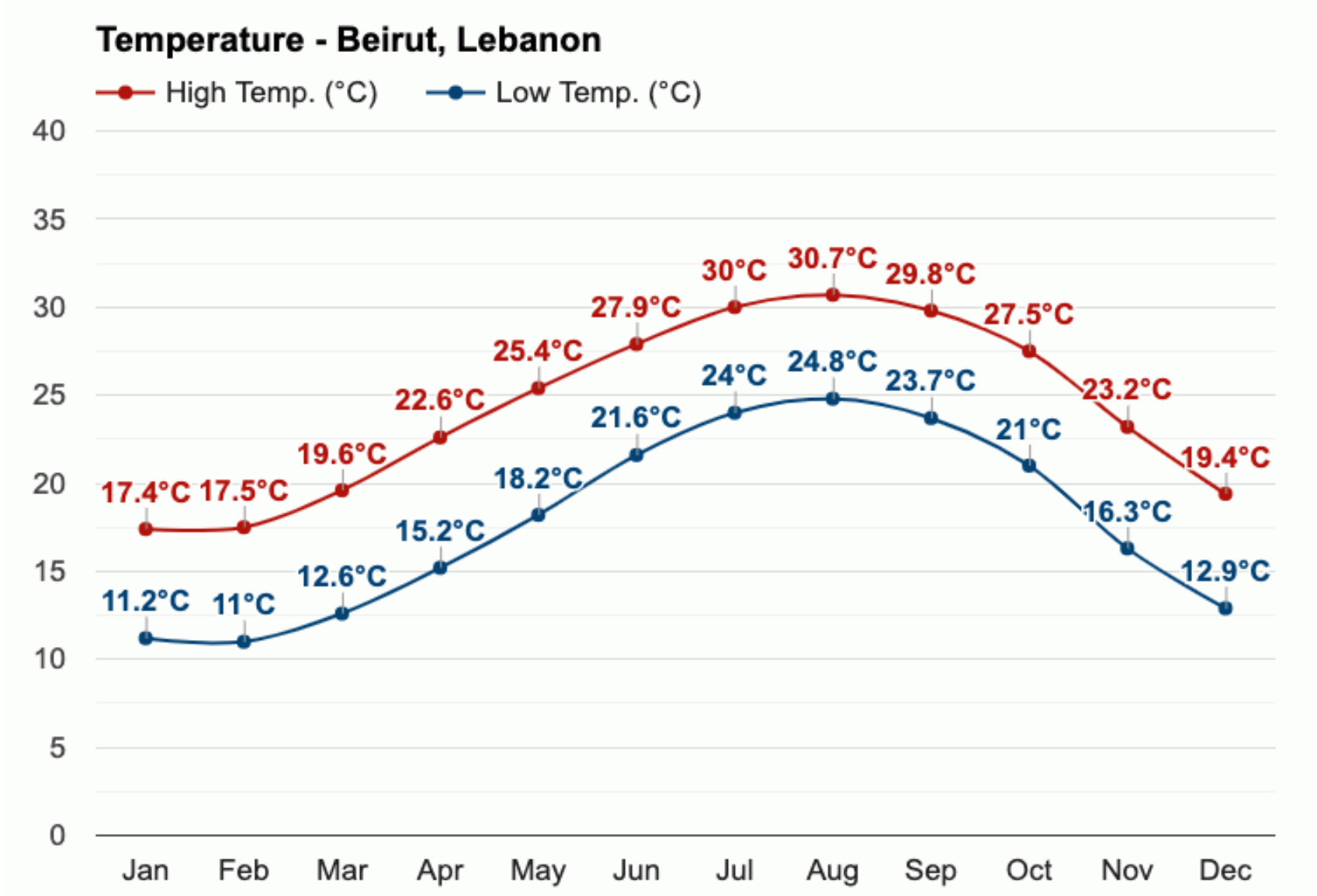
Des Applications

Les risques climatiques

- Vagues de chaleur,
- Tempêtes,
- Inondations,
- Sécheresse,
- Tremblement de terre
- ...

doivent être évalués pour être anticipés pour pouvoir mettre en place une politique de prévention.

En Climat (1)


Temperature - Beirut, Lebanon
High Temp. (°C)
Low Temp. (°C)
40
35
30
25
20
15
10
5
0
17.4°C
17.5°C
19.6°C
22.6°C
25.4°C
27.9°C
30°C
30.7°C
29.8°C
27.5°C
23.2°C
19.4°C
11.2°C
11°C
12.6°C
15.2°C
18.2°C
21.6°C
24°C
24.8°C
23.7°C
21°C
16.3°C
12.9°C
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

En Épidémiologie (2)

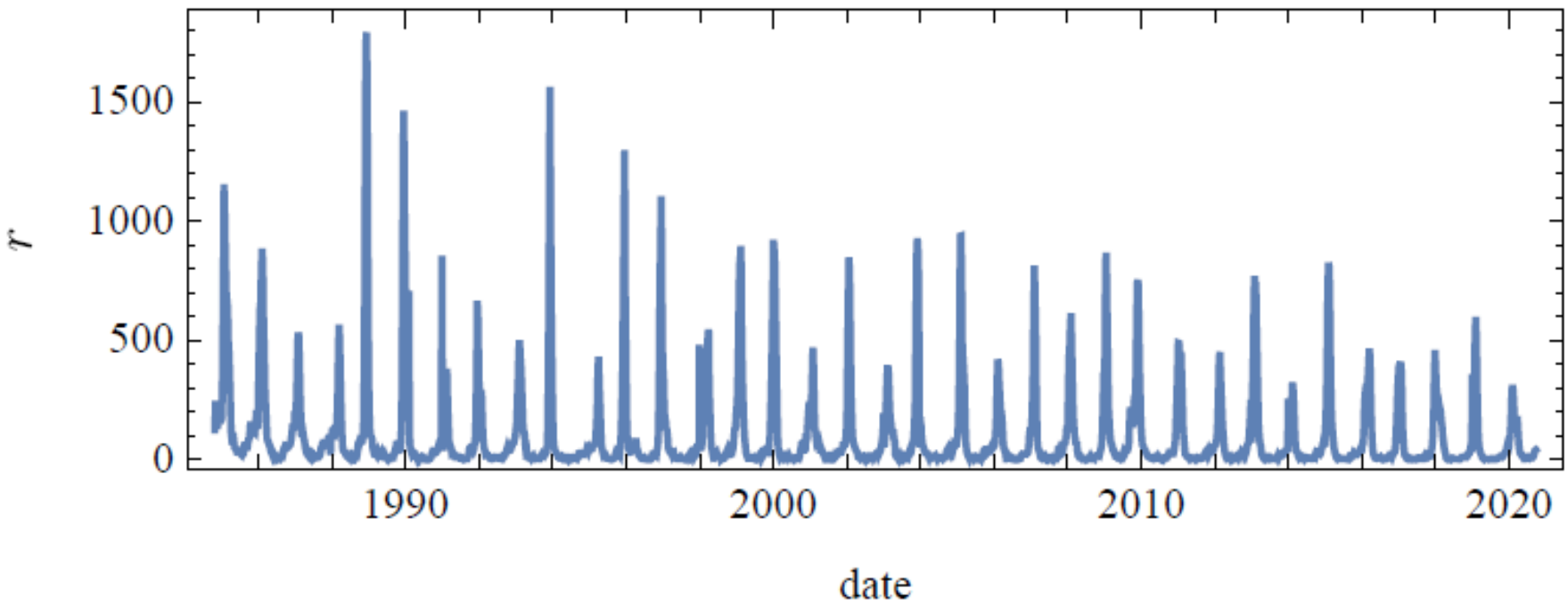


Figure 2: Taux d'incidence de la grippe en France.

En Finance (3)

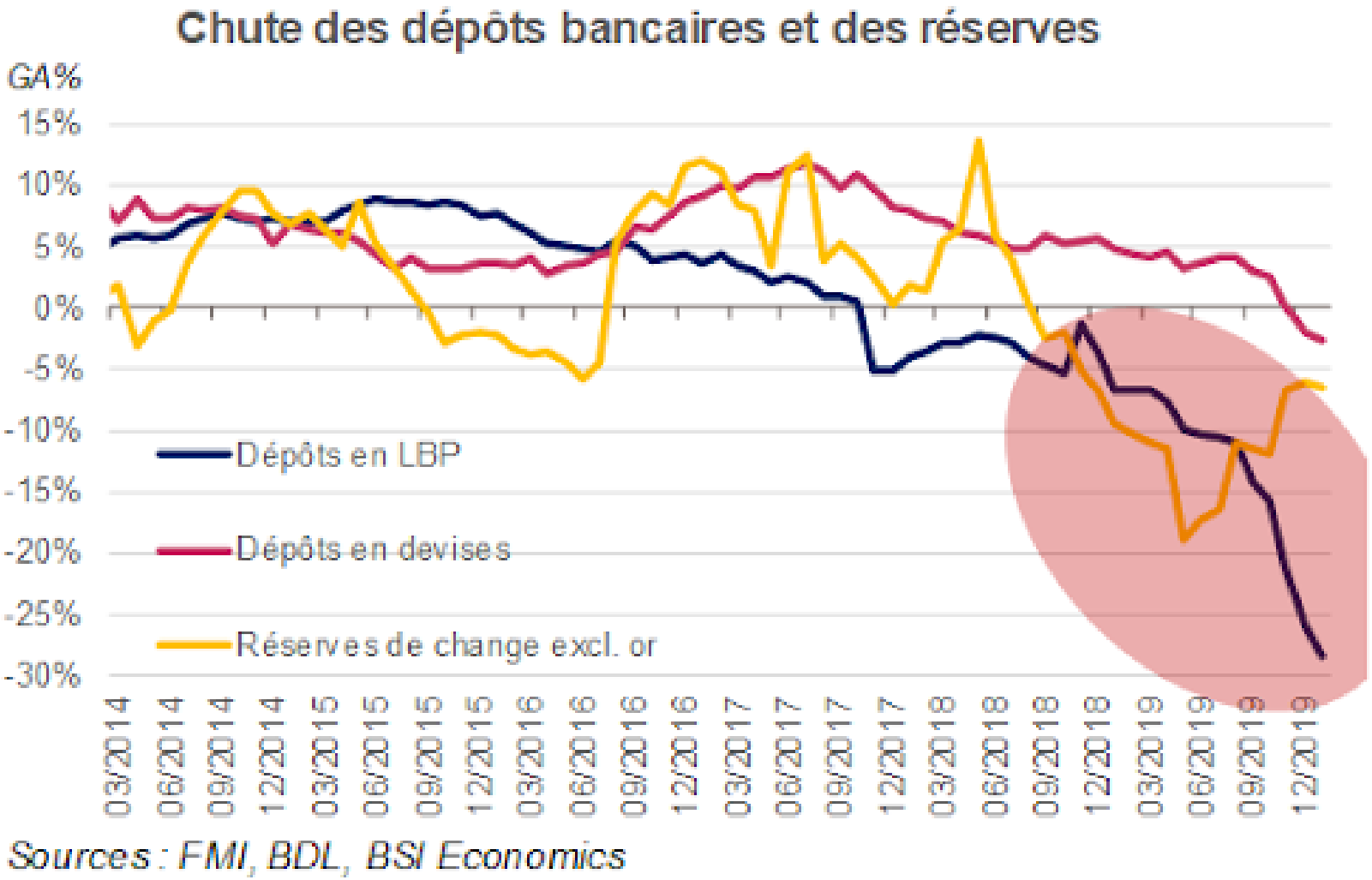


Figure 3: Crise Financière au Liban

Quelles sont les causes des valeurs extrêmes?

- Erreurs humaines, Erreurs de saisie
- Erreurs d'instrument, Erreurs de mesure
- Erreurs de traitement des données, Manipulation des données
- Erreurs d'échantillonnage, Extraction des données de mauvaises sources
- Une valeur extrême n'est pas une erreur, juste une innovation dans les données

Approche 1: Dépassements de seuils élevés

Grandeur d'intérêt X (niveau de l'eau, montant des dommages à payer par une assurance, température, émergence d'une maladie ...)

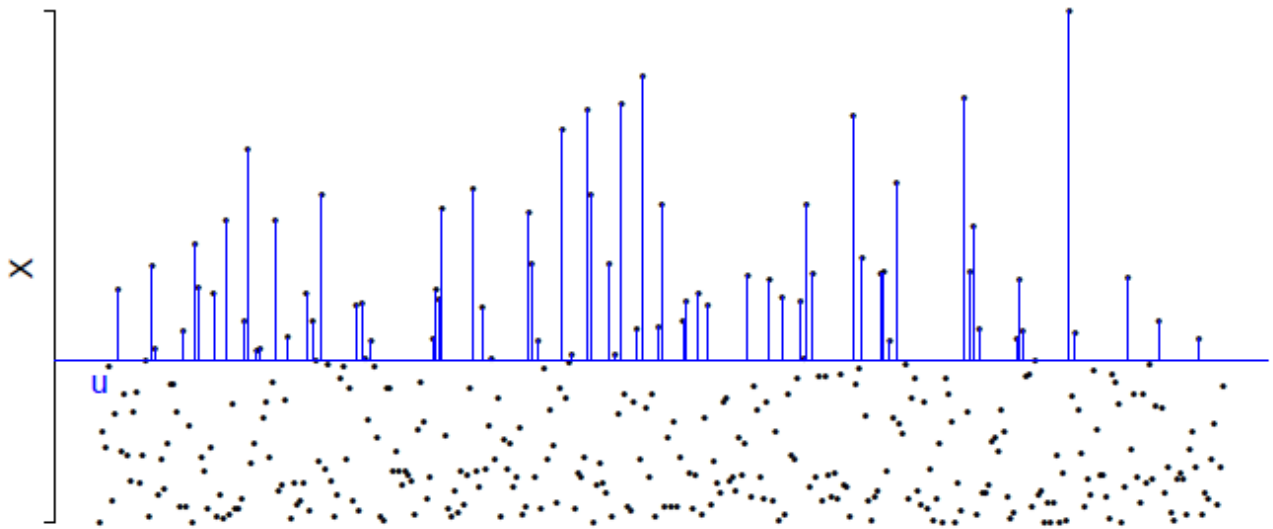


Figure 4: Méthode de dépassement seuil

Étant donné un seuil h élevé, trouver $p = P(X > h)$.

Approche 2: Maxima sur de longues périodes

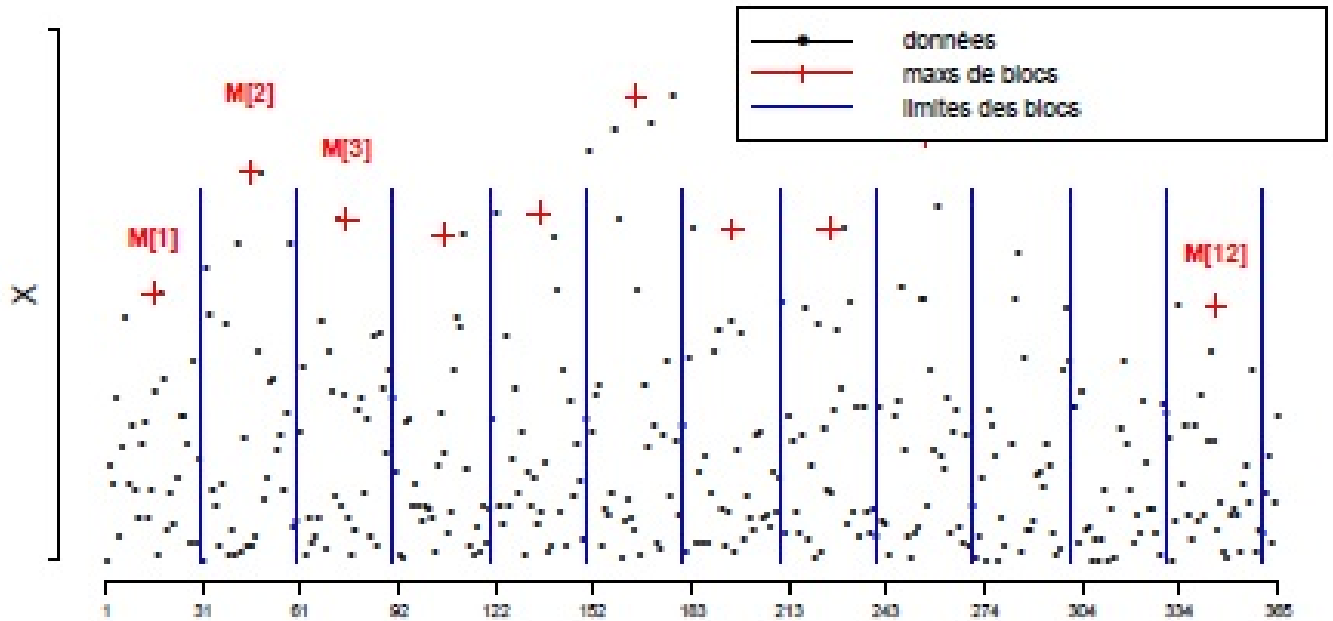


Figure 5: Méthode de Block Maxima

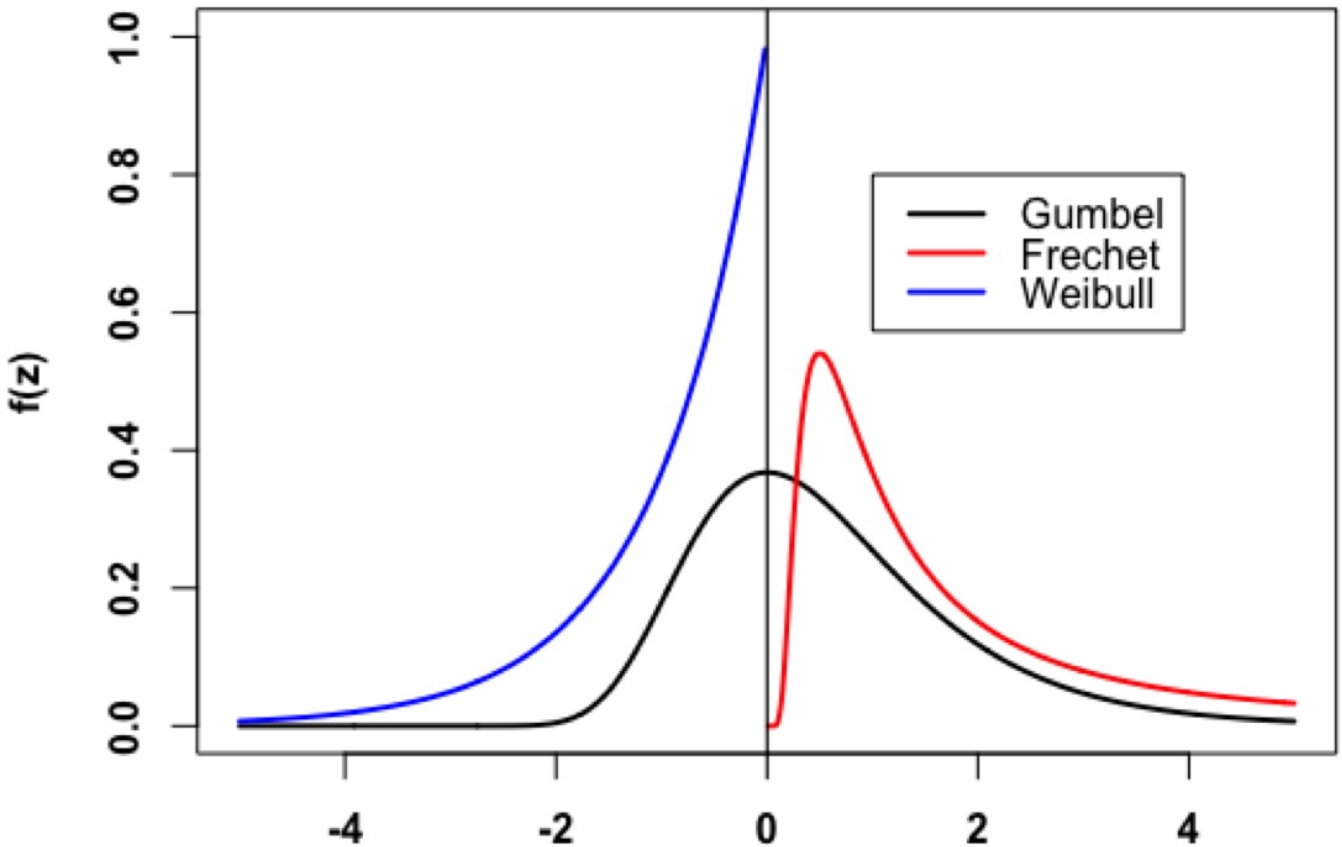
Maximum d'un bloc de taille n:

$$M_n = max_{\{t=1,\ldots,t\}} X_t$$

Loi des valeurs extrêmes

Les maximas sont distribuées selon la distribution généralisée des valeurs extrêmes (GEV) :

$$P(x, \mu, \sigma, \psi) = exp[-(1 + \psi \frac{x - \mu}{\sigma})^{-1/\sigma}]$$

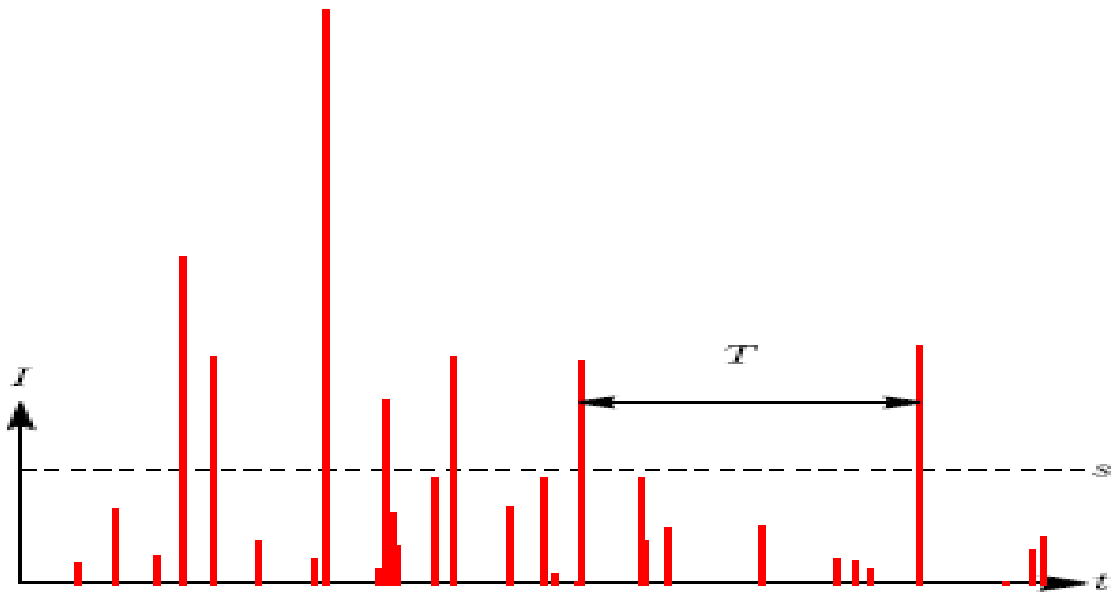


Caratéristiques des Lois

- **Loi de Gumbel**: Support non borné et à queues fines. Les maximas saisonniers ou annuels, en météorologiques, décrire la fréquence des pluies extrêmes,...
- **Loi de Weibull**: Support borné. Etude de la fiabilité des infrastructures hydrauliques.
- **Loi de Fréchet**: Support non borné et à queues épaisses. Étude des évènements extrêmes tels que le maximum annuel des précipitations journalières ou le débit des rivières,...

Période de retour

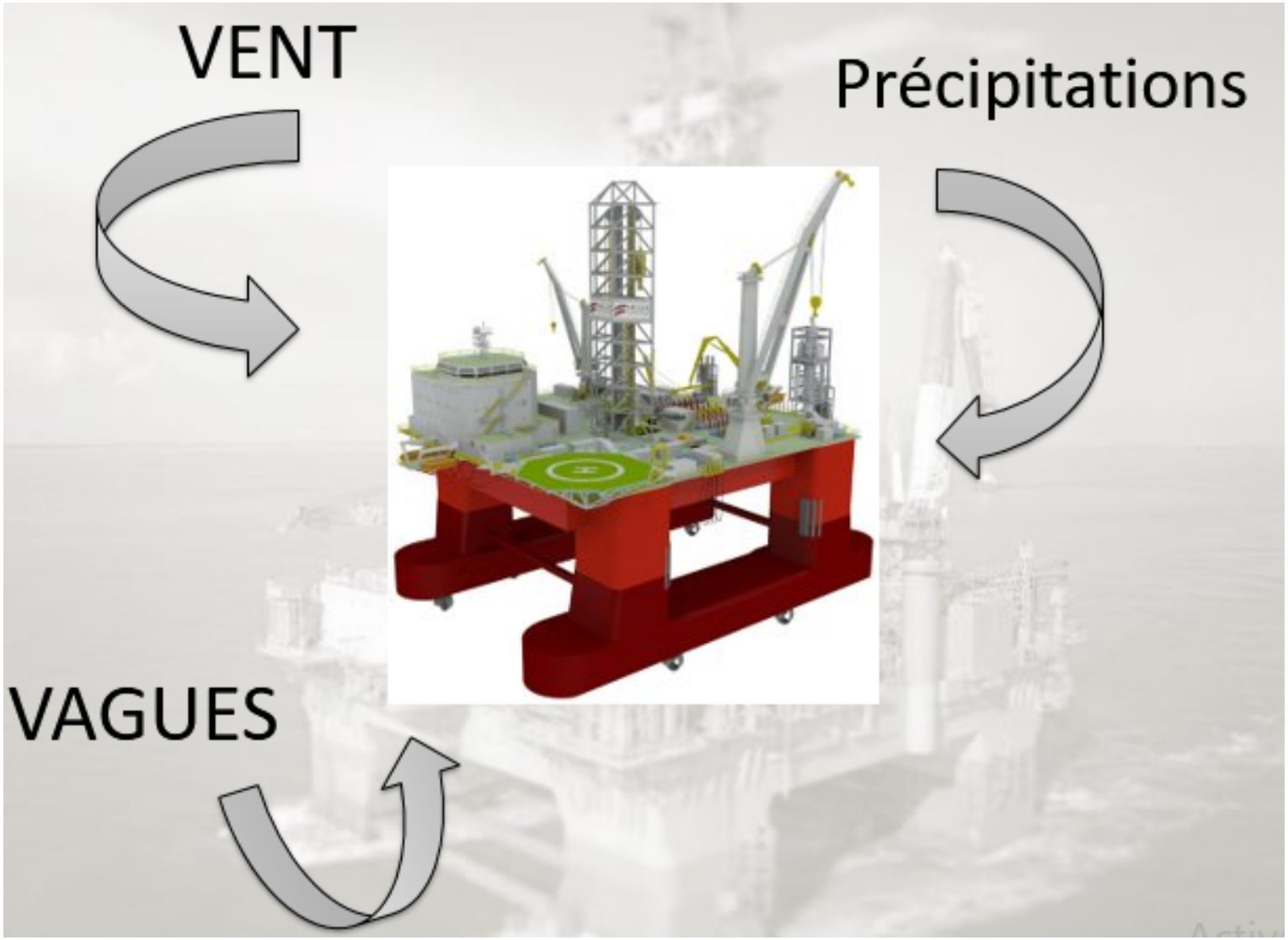
Soit T un intervalle de temps moyen entre 2 évènements, dont l'intensité atteint ou dépasse un certain seuil u.

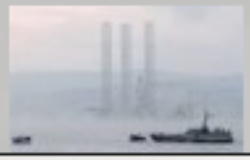


Un évènement de période de retour T a en moyenne une probabilité $1/T$ de se produire chaque année.

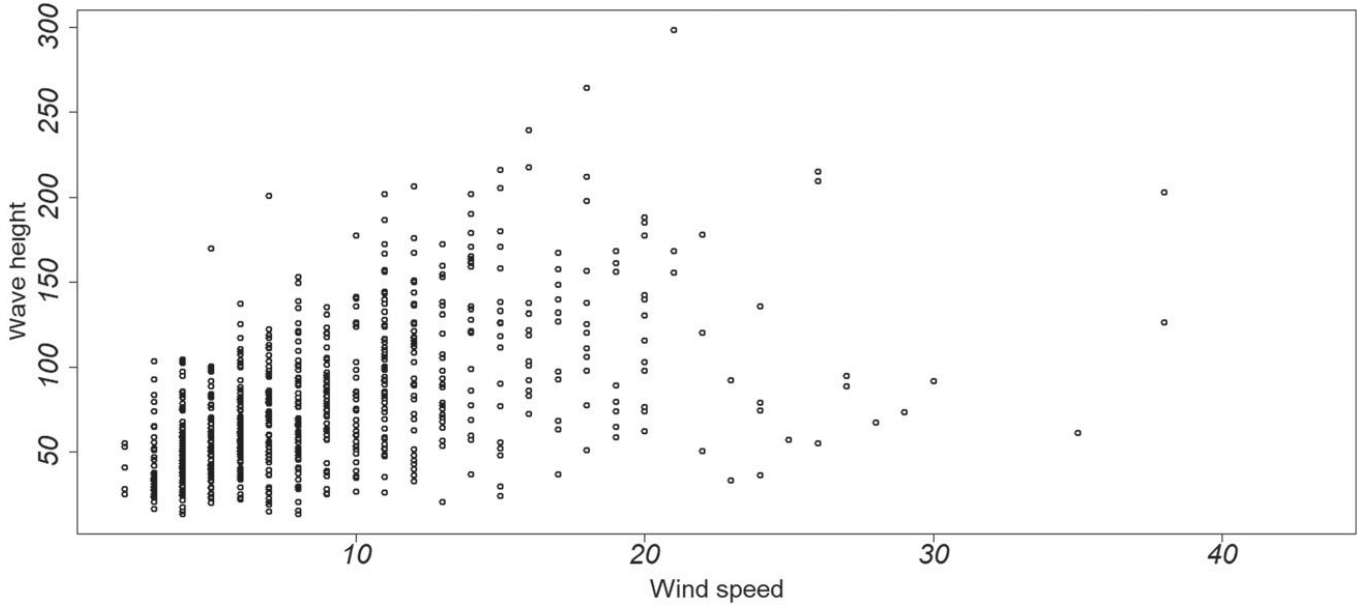
Exemple d'application 1

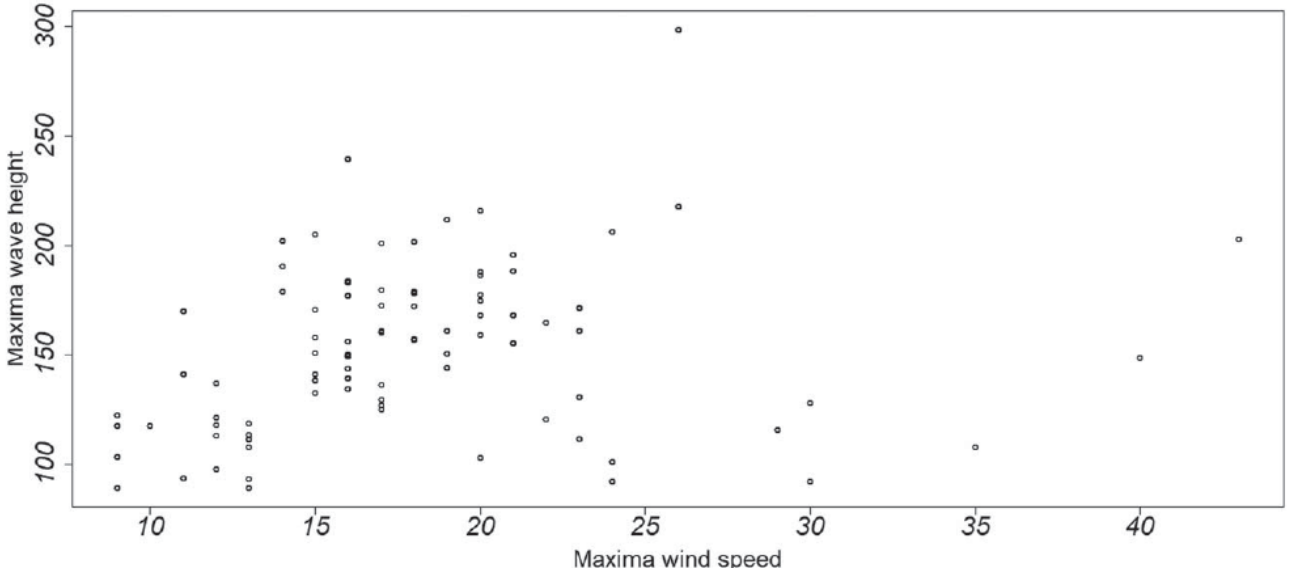
Oil rig protection against wind and wave in Lebanon.

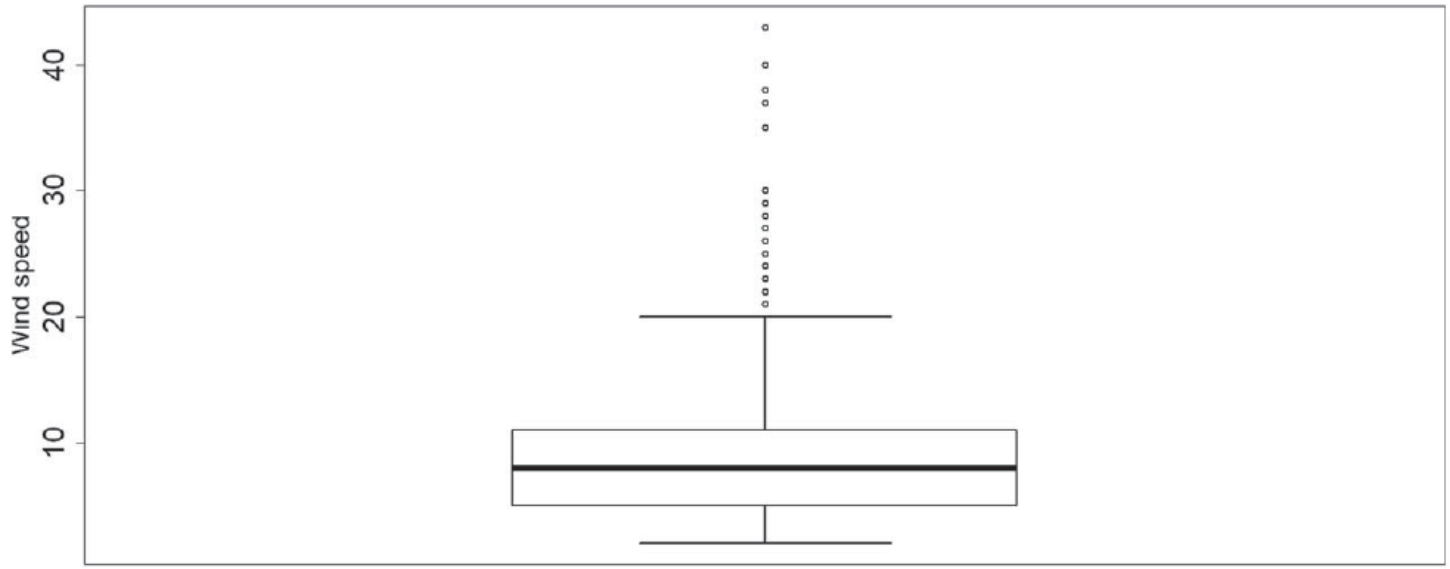


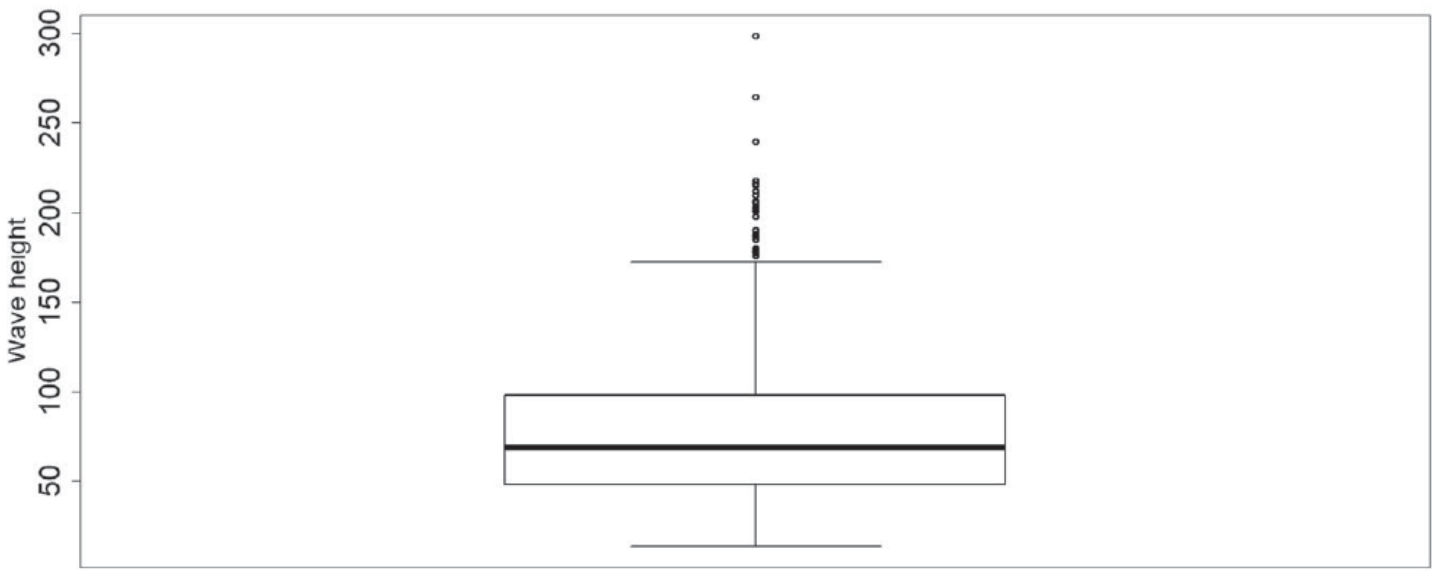
Nom de la plateforme	Image	Profondeur	Lieu	Cause	Bilan
1-kolskaya (autoélévatrice, 2011)		De 20 jusqu'à 100/120m	Russie	Tempête	67 morts
2-Usumacinta (autoélévatrice, 2007)		De 20 jusqu'à 100/120m	Golfe de Mexique	Tempête→ Déplacement → Collision → Explosion	22 morts
3-Platefomes pétrolières (2005)			Golfe de Mexique	Ouragan Katrina	58 plateformes endommagées 30 perdues
4-Seacrest (Navire de forage,1989)		>500m Mer profonde et ultraprofonde	Golfe de Thailande	Typhon, Vents violents, Vagues de 12m	91 morts
5-Ocean Ranger (1982)		>500 m Mer profonde et Ultraprofonde	Eaux Canadiennes	Vagues	84 morts

Data:2000-2015 (n=1353)

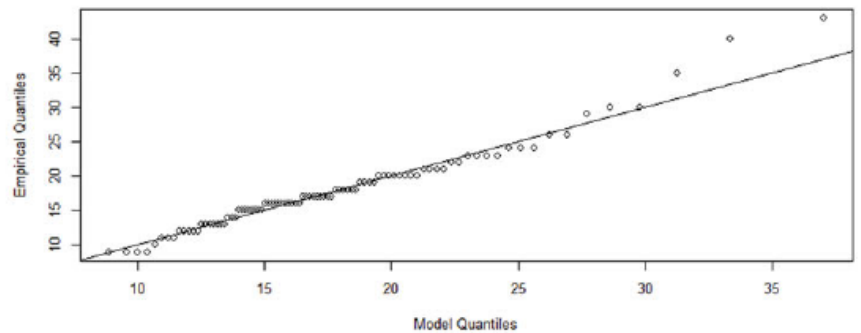


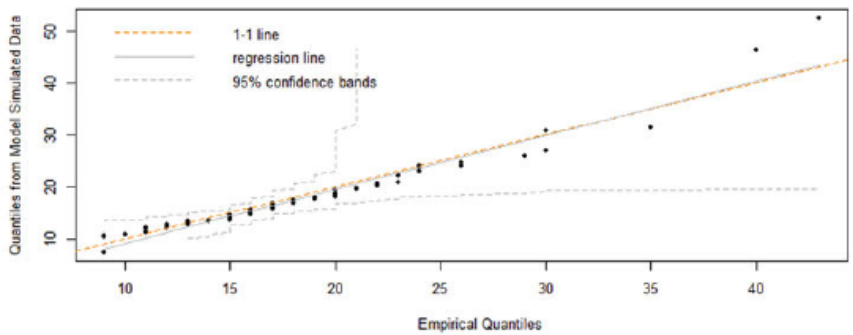


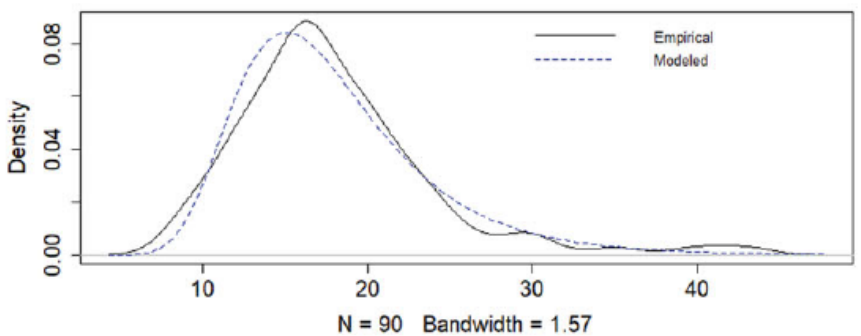

Wind speed
10
20
30
40

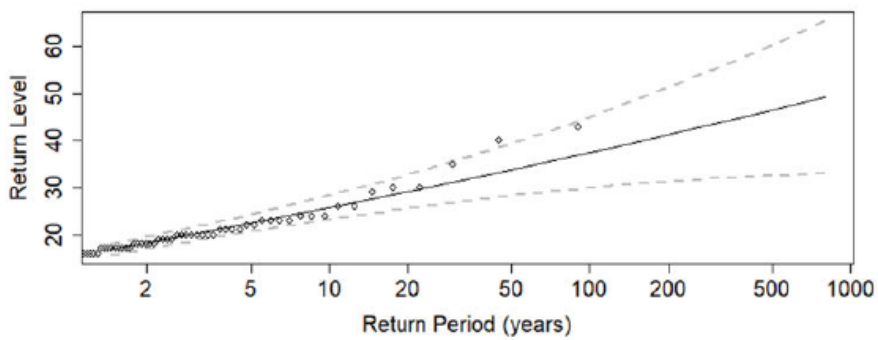

Wave height
50
100
150
200
250
300

Distribution des Maxima: Vitesse de vent









Niveau de retour:(Vent;Vague)

Table 5. Return levels for $X^\star$ and $Y^\star$.

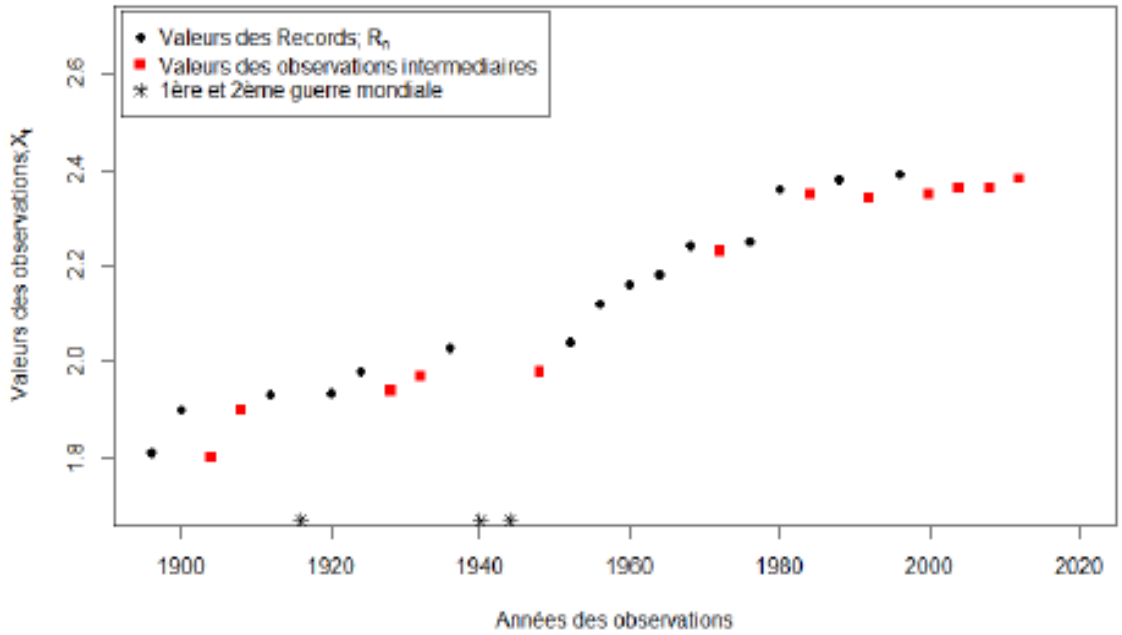
T (years)	$\hat{q}_{X^\star,T}$	$\hat{q}_{Y^\star,T}$
50	43.570	284.839
100	47.530	298.510
500	57.197	327.238

Table 6. Confidence interval of the return levels for $X^\star$ and $Y^\star$.

T (years)	Confidence interval of $q_{X^\star,T}$	Confidence interval of $q_{Y^\star,T}$
50	[32.014,55.126]	[237.160,332.517]
100	[32.831,62.228]	[240.854,356.166]
500	[33.271,81.123]	[244.034,410.442]

Pourquoi les records?

- Les records font partie de la culture populaire.
- Sont souvent plus facilement accessibles que les données sur lesquelles ils sont construits.



- Les records sont des valeurs extrêmes de valeurs extrêmes.

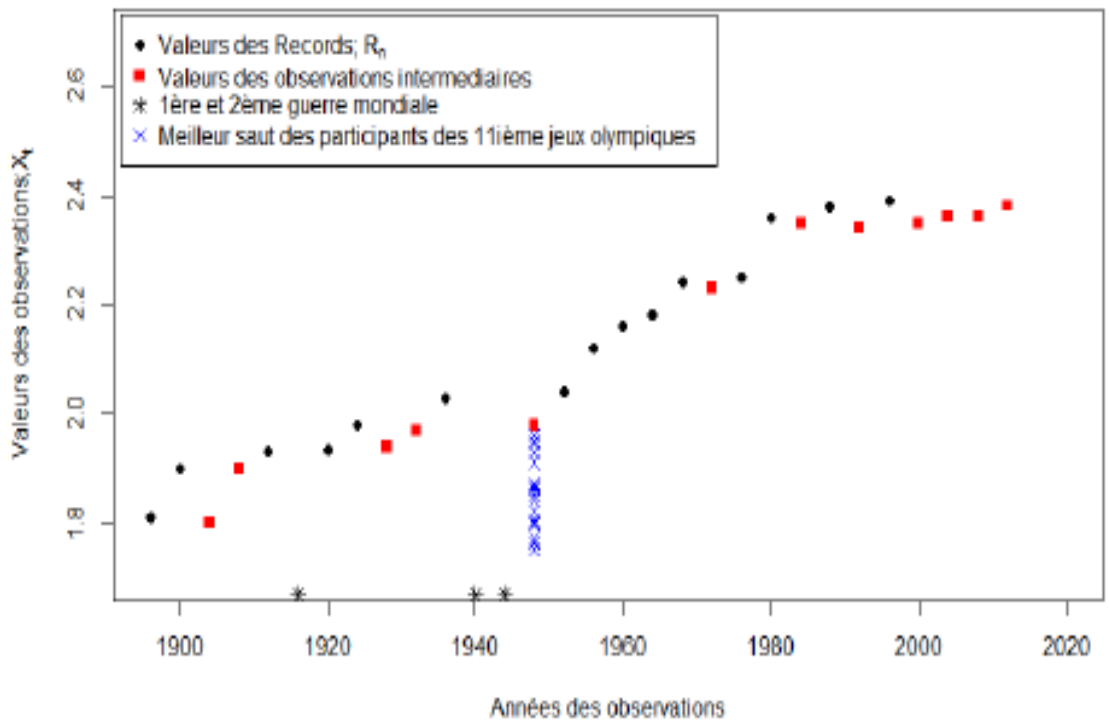


Figure 3 – Saut (en hauteur homme (Mètres)) du gagnant de chacun des jeux olympique (Les X_t)

Application des records

Intérêt des records successifs dans différentes domaines tels que :

- Sports (records olympiques, Mellon, B. 1988).
- Economie et finance (gestions de produits financiers, d'assurances Paul E. 2008).
- Phénomènes naturels (crues, pressions atmosphériques, séismes,...Richard M.2003, Khraibani et al.2018).
- Contrôle de systèmes industriels (évaluer les facteurs de panne et limiter leur intensité).
- Épidémiologie (Khraibani et al. 2014).

Caractéristiques des records

1. Robustesse dans le cas de v.a.i.i.d. Lois exactes a n fini par rapport a la théorie classique des valeurs extrêmes (TVE).
2. Pas déraisonnable de modéliser les X_n par les distributions (GEV).Gumbel,Weibull, Frechet
3. Nevzorov 2014, les utilisent pour créer un test détectant les valeurs aberrantes dans un jeu de données "normal".
4. Le processus de records représente la tendance maximale observée d'un tel phénomène.
5. Khraibani 2014, test non-paramétrique basé sur le nombre de records observés.

Travaux sur les records

1. Demand forecasting using the Records theory: Evidence form French spatial data infrastructures (en cours).
2. Application of records theory on the COVID-19 pandemic in Lebanon: Prediction and Prevention. *Epidemiology and Infection*, 2020.
3. Analysis of the extreme and records values for temperature and precipitation in Lebanon. *Communications in Statistics*, 2020.
4. Inférence fondée sur la vraisemblance pour des modèles de records. *Comptes Rendus Mathématique*, 2018·
5. A non parametric exact test based on the number of records for an early detection of emerging events: Illustration in epidemiology. *Communications in Statistics*, 2015.
6. Application of the Records Method to Identify the Sporadicity of Percnon gibbesi Distribution in Greece. *Journal of Scientific Research and Reports*, 2014.

Exemple d'application 2

Application of records theory on the COVID-19 pandemic in Lebanon: Prediction and prevention

- Détecter l'émergence d'une nouvelle maladie pour laquelle on a peu d'observations (cas d'une première émergence).
- Déterminer s'il s'agit d'une maladie sporadique ou d'une maladie émergente en se basant sur l'approche par processus de records.

Qu'est ce qu'une maladie émergente ?

- Maladie dont l'incidence réelle augmente de manière significative dans une population, période donnée (début d'épidémie).
- Infections nouvelles, causées par l'évolution ou la modification d'un agent pathogène ou d'un parasite existant.
- L'événement de démarrage de l'émergence est la transition de la stabilité de l'état 0 pathogène à l'instabilité de cet état.

Définition mathématique d'une maladie émergente

- Soit Y_n le pourcentage de cas cliniques dans une population au temps n, tel que $Y_n = f(Y_{n-1})$ et $f(0) = 0$. En utilisant le développement de Taylor a l'ordre 1 au voisinage de 0 et en supposant Y_0 petit:

Définition mathématique d'une maladie émergente

- Soit Y_n le pourcentage de cas cliniques dans une population au temps n, tel que $Y_n = f(Y_{n-1})$ et $f(0) = 0$. En utilisant le développement de Taylor a l'ordre 1 au voisinage de 0 et en supposant Y_0 petit:
- $f(Y_0) = f(0) + Y_0 f'(0) + O(Y_0^2) \Rightarrow f(Y_0) \approx Y_0 f'(0)$.

Définition mathématique d'une maladie émergente

- Soit Y_n le pourcentage de cas cliniques dans une population au temps n, tel que $Y_n = f(Y_{n-1})$ et $f(0) = 0$. En utilisant le développement de Taylor a l'ordre 1 au voisinage de 0 et en supposant Y_0 petit:
- $f(Y_0) = f(0) + Y_0 f'(0) + O(Y_0^2) \Rightarrow f(Y_0) \approx Y_0 f'(0)$.
- $Y_1 = f(Y_0) = Y_0 f'(0) + O(Y_0^2)$

Définition mathématique d'une maladie émergente

- Soit Y_n le pourcentage de cas cliniques dans une population au temps n, tel que $Y_n = f(Y_{n-1})$ et $f(0) = 0$. En utilisant le développement de Taylor a l'ordre 1 au voisinage de 0 et en supposant Y_0 petit:
- $f(Y_0) = f(0) + Y_0 f'(0) + O(Y_0^2) \Rightarrow f(Y_0) \approx Y_0 f'(0)$.
- $Y_1 = f(Y_0) = Y_0 f'(0) + O(Y_0^2)$
- Donc $Y_2 = f(f(Y_0)) = f(Y_0) f'(0) + O(f(Y_0)^2) = Y_0 (f'(0))^2 + f(Y_0) O(Y_0^2) + O(Y_0^2)$ Pour $f'(0) < 1$,

Définition mathématique d'une maladie émergente

- Soit Y_n le pourcentage de cas cliniques dans une population au temps n, tel que $Y_n = f(Y_{n-1})$ et $f(0) = 0$. En utilisant le développement de Taylor a l'ordre 1 au voisinage de 0 et en supposant Y_0 petit:
- $f(Y_0) = f(0) + Y_0 f'(0) + O(Y_0^2) \Rightarrow f(Y_0) \approx Y_0 f'(0)$.
- $Y_1 = f(Y_0) = Y_0 f'(0) + O(Y_0^2)$
- Donc $Y_2 = f(f(Y_0)) = f(Y_0) f'(0) + O(f(Y_0)^2) = Y_0 (f'(0))^2 + f(Y_0) O(Y_0^2) + O(Y_0^2)$ Pour $f'(0) < 1$,
- Dans le cas de non émergence:

$$f'(0) < 1 \Rightarrow f(Y_0) < Y_0 \Rightarrow Y_1 < Y_0$$

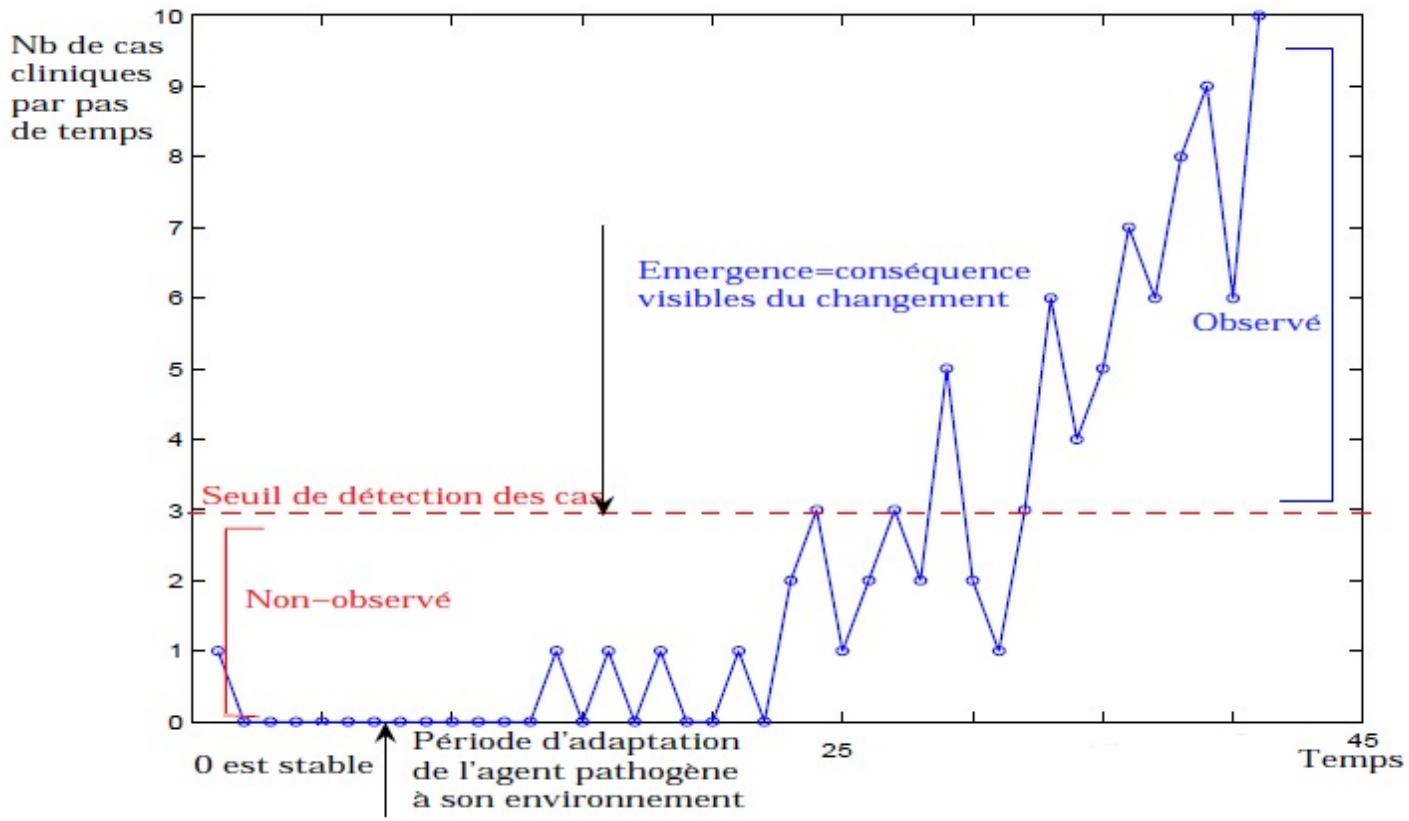
et $f(Y_1) = f(f(Y_0)) \approx f(Y_0) f'(0) < Y_0$. On obtient par récurrence que $Y_n := f(n)(Y_0)$ reste du même ordre de grandeur que Y_0.

- Dans le cas d'une émergence: $f'(0) > 1$ et on obtient :

$$f(Y_0) > Y_0 \Rightarrow Y_1 > Y_0$$

donc Y_n ne reste pas négligeable.

Figure 6: Émergence d'une maladie



Orientations méthodologiques

- Méthodes classiques pour détecter l'émergence ou la reémergence :

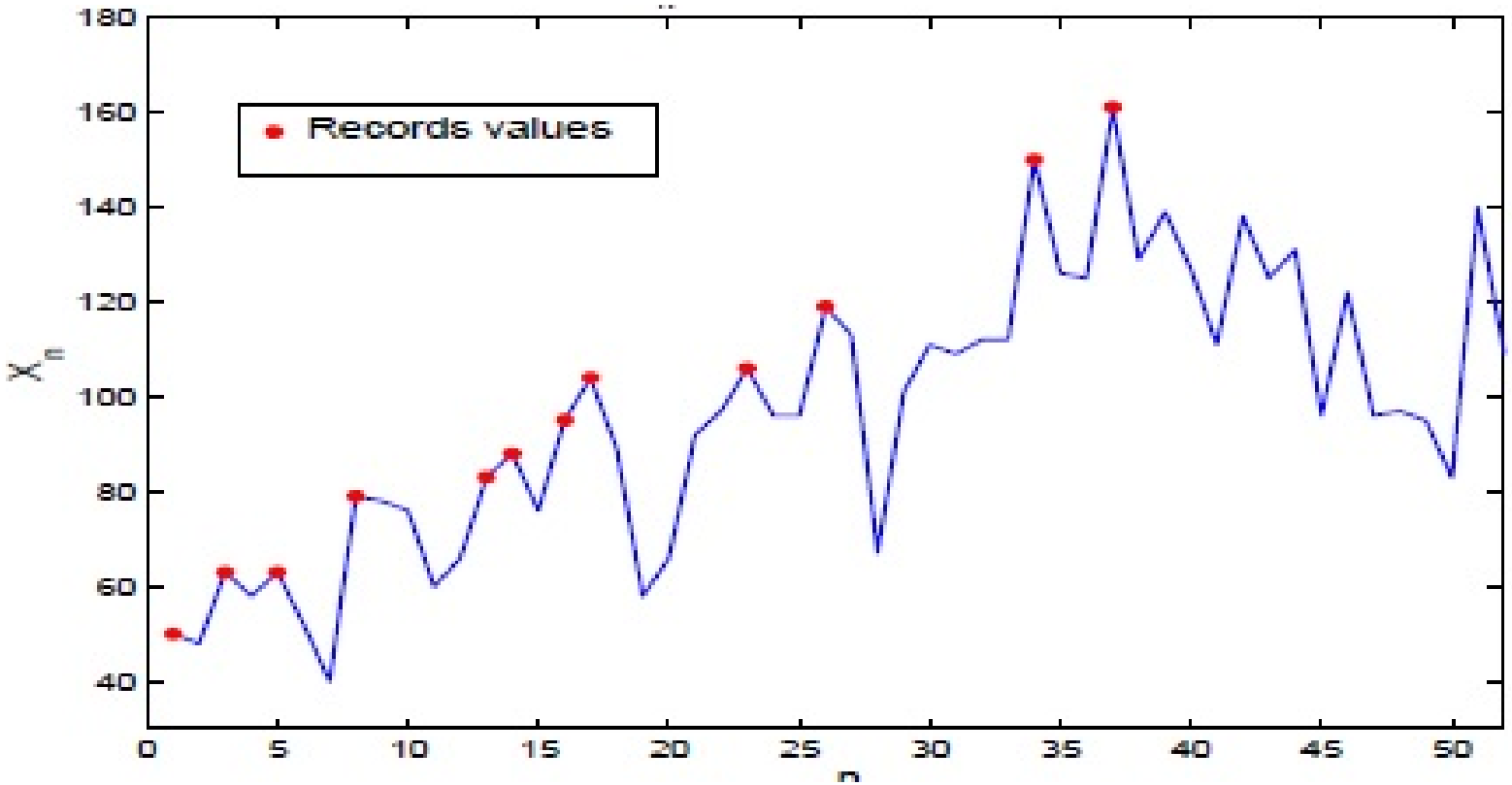
  - Méthodes asymptotiques (Colses,S.2001).

  - Méthodes descriptives (lissages, Tager IB 1996).

  - Méthodes exactes (tests non paramétriques (Run) William H. 2008).

Représentation des valeurs de records

- Le processus de records représente la tendance maximale observée de l'épidémie.



Définition du processus de records

- Par Définition, $\{R_n : n \geq 1\}$ et $\{L_n : n \geq 1\}$ sont respectivement la suite des valeurs des records et la suite des indices des records. Plus précisément :

$$\begin{aligned} L_1 &= 1 \\ L_n &= \inf\{j > L_{n-1} : X_j > X_{L_{n-1}}\} \\ R_n &= X_{L_n} \end{aligned}$$

- Soit N_n le nombre total des records parmi $\{X_1; ...; X_n\}$ avec $N_1 = 1$:

$$N_n = \sum_{j=1}^{n} \delta_j;$$

où δ_j est l'indicatrice de record tel que :

$$\delta_j = \begin{cases} 1, & \text{si } X_j > max(X_1, ..., X_{j-1}) \\ 0, & \text{sinon} \end{cases}$$

Nombre de records

Les principaux résultats de la Théorie de Record dans le cas i.i.d. ont été obtenus au cours de la période 1952-1983 (Voir Chandler (1953), Arnold (1998) et Nevzorov (2001)):

- Les v.a. $\{\delta_n\}_{n\geq 1}$ sont indépendantes et $\delta_n \sim$ *Bernoulli*$(1/n)$*avec* :

$$\begin{aligned} P_n &= \text{taux de record (la probabilité que } X_n \text{ soit un record)} \\ &= \mathbb{P}[\delta_n = 1] \\ &= 1/n \end{aligned}$$

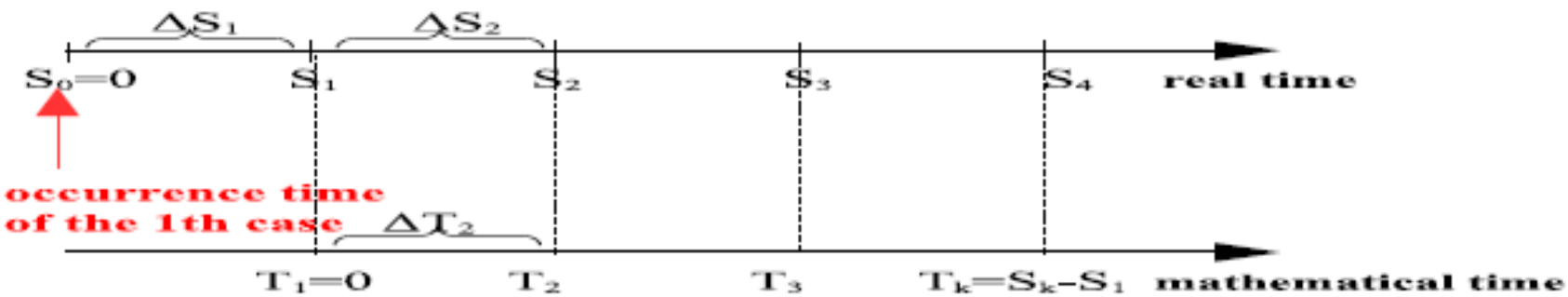
- La distribution exacte de N_n est donnée par:

$$\mathbb{P}[N_n = m] = \frac{s(n,m)}{n!},\ 0 \leq m \leq n$$

- En plus Arnold et Nevzorov donnent les distributions marginales et jointes des suites $\{R_n\}_{n\geq 1}$ et $\{L_n\}_{n\geq 1}$.

Formalisation mathématique du problème

Figure 7: Processus Ponctuel



- $\{S_n\}_{n\geq 0}$: instants d'occurrence successifs des cas cliniques (processus de renouvellement).
- $X_n = (\Delta S_n)^{-1}$, $\Delta S_n = S_n - S_{n-1}$,temps d'attente entre l'arrivée de deux cas successifs. Dans le cas d'une émergence $\{\Delta S_n\} \searrow \Leftrightarrow \{(\Delta S_n)^{-1}\} \nearrow \Rightarrow$ records de $X_n \Rightarrow \{T_n, X_n\}$: Processus Ponctuel.

Formalisation des hypothèses à tester

1. H_0 (cas sporadiques): $\{\Delta S_n\}$ i.i.d,

   $P(\Delta S_n \leq s) := E(s) = 1 - \exp(-\lambda s)$ (loi sans mémoire).

Formalisation des hypothèses à tester

1. H_0 (cas sporadiques): $\{\Delta S_n\}$ i.i.d,

   $P(\Delta S_n \leq s) := E(s) = 1 - \exp(-\lambda s)$ (loi sans mémoire).

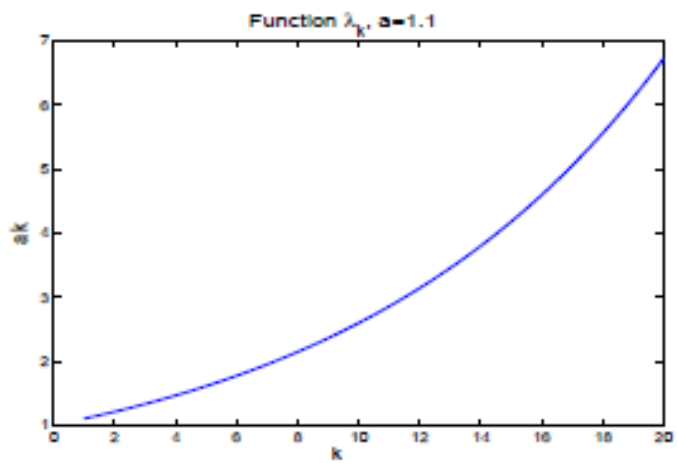
2. H_1 (émergence): $\{\Delta S_n\}$ indépendantes, $\bar{E}_n = \bar{E}_n^{\rho_n}$,

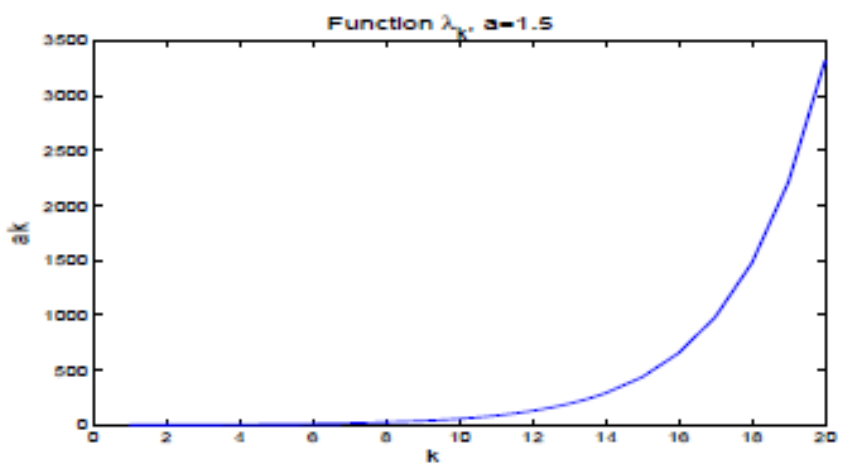
   $(\bar{E}_n = 1 - E)$, $\{\rho_n\}_n$ suite positive croissante.

Fréquence de l'émergence

- $E(\Delta T_k) = (\lambda_k)^{-1}$ où $\lambda_k = \lambda.\rho_k = \lambda.a^k$ est la fréquence du nombre de cas par unité de temps au temps T_k ,$a > 1$, croissance exponentielle d'une maladie infectieuse.

Figure 8: Fonction $\{\lambda_k\}$ où $\lambda_k = a^k$, pour $a = (1.1, 1.5)$ et $\lambda = 1$.





Application du test de records sur le COVID-19

- Test de H_0 contre H_1: Définitions et Rappels

  Les hypothèses à tester:

  -H_0 (maladie sporadique): $\{X_k\}$ i.i.d., $X_1 \sim F$, F continue.

  -H_1 (maladie émergente): $\{X_k\}$ indépendantes $X_k \sim F_k$

- Statistique de test utilisée : N_n.

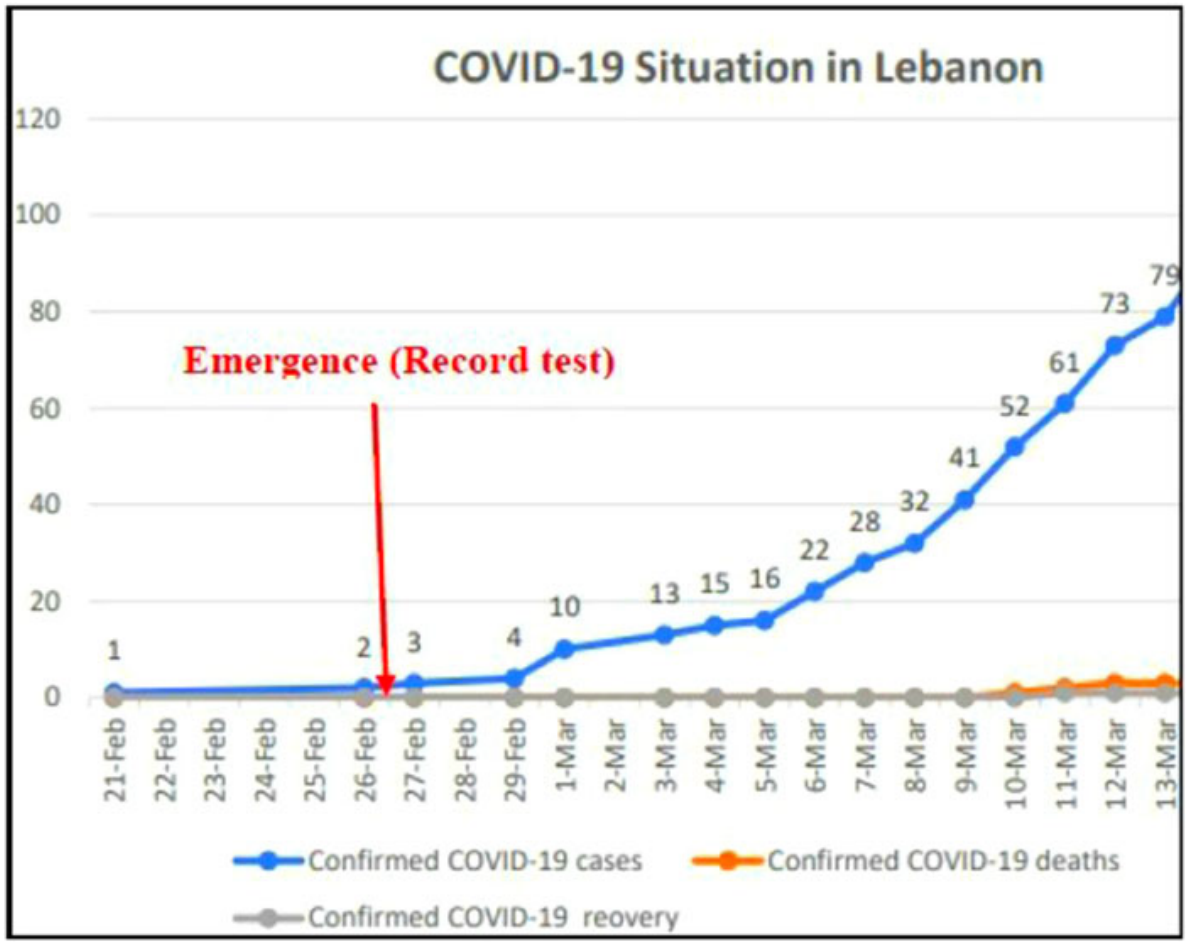
- Risque d'erreur: $\alpha = P_{H_0}(\textit{rejeter } H_0) = P_{H_0}(N_n \geq N_\alpha)$, (la

  région de rejet de H_0 est déterminée par les grandes valeurs de

  N_n).

Illustration graphique

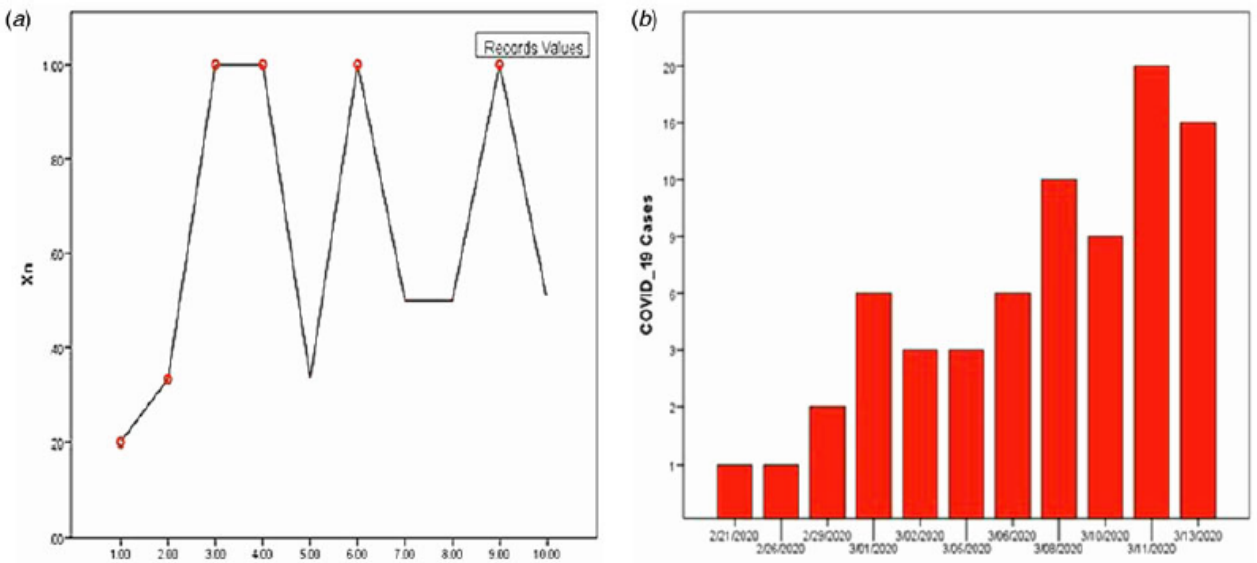
Figure 9: Daily cumulative/emerged number of confirmed, fatal and recovered cases of Coronavirus Disease 2019 (COVID-19) in Lebanon.



Description

- Nouvelle maladie, observée dans différents pays, pas d'information épidémiologique.
- 10 cas détectés au Liban de 21 Fev.2020 au 1 Mars 2020.
- ΔT_n, le temps d'interarrivée entre le *nième* et le $(n+1)$*ième* cas.

Figure 10: (a) Records values of $X_n = (\Delta T_n)^{-1}$. (b) Number of observed COVID-19 cases per day.



Valeurs et nombre de records

Table 1. Waiting times between two successive cases and number of COVID-19 cases in Lebanon per day

T_n	21/02	26/02	29/02	01/03	02/03	05/03	06/03	08/03	10/03	11/03	13/03
COVID	1	1	2	6	3	3	6	10	9	20	16
(ΔT_n)	5	3	1	1	3	1	2	2	1	2	–
$(\Delta T_n)^{-1}$	0.2	0.33	1	1	0.33	1	0.5	0.5	1	0.5	–

- Le nombre de records observés $N_n = 6$

Table 2. $P(N_n \geq m)$, for $n = 10$, 20 and for different values of m

	m						
n	1	2	3	4	5	6	7
10	1	0.9000	0.6171	0.2939	0.0945	0.0203	0.0029
20	1	0.9500	0.7726	0.4978	0.2470	0.0944	0.0280

- $P_{H_0}(N_{10} \geq 6) = 0.0203 \Rightarrow$ Rejet H_0, (À partir de quelques cas observes au début de l'épidémie, nous concluons que le COVID-19 est une maladie émergente au Liban).

Test H_0, H_1

Table 3. $P(N_n \geq m)$ under H_1, for $n = 10, 20, 30$ and for different values of m and a

N	10	20	30
m	6	7	8
$P(N_n \geq m)$, $a = 1.1$	0.0455	0.1108	0.14261
$P(N_n \geq m)$, $a = 1.5$	0.2683	0.7554	0.9352

- On pourrait remarquer que le COVID-19 au Liban émerge très rapidement avec une forte probabilité de dépasser certaines valeurs records pour $n = 10, 20, 30$.

Prévision de l'émergence

- En prenant en compte les futurs COVID-19 records, nous calculons la probabilité de temps d'attente ($\Delta T_n{}^*$) pour observer un nouveau records:

$$P(\Delta T_n{}^* > n^*) = \frac{n^{'}}{n^{'} + n^*}$$

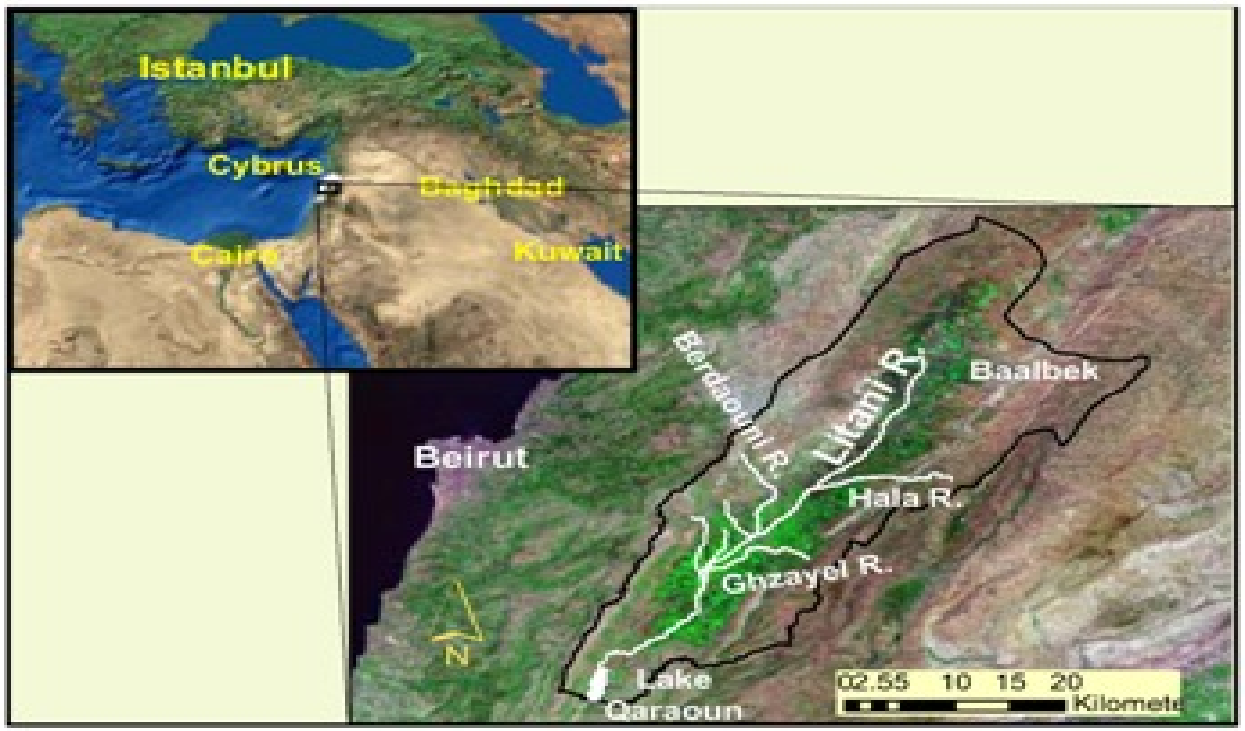
- Pour ($n^* = 5$) et $n^{'} = 22$ jours (21 février 2020-13 mars 2020):
- $P(\Delta_n{}^* > 5) = 22/(22+5) \approx 0,82 \Rightarrow$ Croissance rapide de la fréquence du COVID-19 au Liban sur une courte période.

Conclusion

1. Nouvelle méthodologie pour détecter l'émergence d'une nouvelle maladie.
2. Tester H_0 (maladie sporadique) contre H_1 (émergence de la maladie).
3. N_n: statistique de test (robuste) indépendante de F sous H_0 et H_1, exactement calculée pour chaque valeur de n.
4. Sous H_1 la distribution du N_n dépend seulement du paramètre $\rho_k = a^k$ (croissance exponentielle du nombre de cas d'une maladie infectieuse).
5. L'approche par les records est particulièrement adéquate pour n petit.

Collaboration avec LEESU

Thèse en cours (Alya ATOUI): Evaluation des impacts environnementaux au Liban: Eau-sol-air.
Directeur: Régis Moilleron, Co-Directeur: Samir Abbad Andaloussi



Quelques annonces...

1. Lancement du premier AaP de l'Institut des Mathématiques pour la Planète Terre.
2. Appel à projets 2021 : Programme IntenSciF (AUF)
3. Bourses Doctorales CNRS-L/Ambassade de France au Liban
4. Projet CEDRE: Partenariat Hubert Curien (PHC) franco-libanais