CENTRALE NANTES

Nantes Université

DIPLÔME D'INGÉNIEUR GÉNÉRALISTE
PROJET PRATIQUE D'OPTIONS MATHÉMATIQUES APPLIQUÉES

Théorie des valeurs extrêmes avec applications environnementales et en finance

Romane **TÉZÉ**
Elyas **BENYAMINA**
Zihao **GUO**

Tuteur :
MATHIEU **Ribatet**
mathieu.ribatet@ec-nantes.fr

May 25, 2023

CONTENTS

INTRODUCTION

Ce rapport s'appuie sur les résultats théoriques présentés dans le précédent. Il présente les résultats obtenus en appliquant la Théorie des Valeurs Extrêmes à des situations concrètes. En plus de la partie de théorie détaillant de nouveaux résultats absents du précédent rapport, celui-ci se compose de trois autres parties distinctes, chacune étant une application réalisée par l'un des membres du groupe. La première application concerne le calcul de la VaR à l'aide de la théorie des Valeurs Extrêmes et du package R *evd*. La seconde présente d'abord en détails le package python *PyExtremes* puis l'utilise dans le cadre du calcul des montants extrêmes de pertes causées par des incendies au Danemark. Enfin, la dernière partie est une application de la théorie des valeurs extrêmes au machine learning, et s'intéresse à la classification binaire des covariables dont les valeurs sont supérieures à un certain seuil.

1 THÉORIE

1.1 Méthode du seuil

MOTIVATIONS L'approche par blocs, grandement étudiée dans le rapport précédent, présente le défaut de pouvoir conduire à une perte d'information. En effet, il est possible qu'au sein d'un même bloc se trouvent deux valeurs extrêmes, et on gardera alors uniquement la plus grande des deux. L'approche par seuil permet de contourner ce problème. Dans celle-ci, on ne sépare plus les données en blocs, et on considère désormais que les évènements extrêmes sont ceux dépassant une certaine valeur seuil u. En d'autres termes, on s'intéresse à la probabilité conditionnelle suivante :

$$\begin{aligned} P(X > u + y | X > u) &= \frac{P(X > u + y, X > u)}{P(X > u)} \\ &= \frac{P(X > u + y)}{P(X > u)} \\ &= \frac{1 - F(u + y)}{1 - F(u)} \end{aligned}$$

où F est la fonction de répartition de la variable X.

Si F est connue, alors la probabilité ci-dessus est connue et la distribution des évènements extrêmes l'est également. En pratique, F est inconnue. On a donc besoin d'approximations.

MODÈLE ASYMPTOTIQUE On considère $X_1, ..., X_n$ n variables aléatoires indépendantes et identiquement distribuées, et $M_n = max(X_1, ..., X_n)$. On suppose que pour n assez grand, on a

$$P(M_n \leqslant z) \approx G(z),$$

où G appartient à la famille des distributions GEV. Alors, pour u suffisamment grand, la distribution de $X - u$ conditionnellement à l'évènement $X > u$ peut être approchée par

$$H(y) = 1 - (1 + \frac{\xi y}{\tilde{\sigma}})^{\frac{-1}{\xi}},$$

définie sur $\{y : y > 0\}$ et $\{(1 + \frac{\xi y}{\tilde{\sigma}}) > 0\}$, avec $\tilde{\sigma} = \sigma + \xi(u - \mu)$.

La famille de distribution de la forme de H est appelée la famille Généralisée de Pareto (que l'on notera GP dans la suite, pour *Generalized Pareto*). Ainsi, si M_n admet pour distribution limite une distribution de la famille GEV, alors les excès de seuil ont pour distribution une distribution de la famille GP. On note également que le paramètre de forme ξ des deux distributions sont identiques.

La distribution ci-dessus est valable dans le cas où $\xi \neq 0$. Si $\xi = 0$, alors on a

$$H(y) = 1 - exp(-\frac{y}{\tilde{\sigma}}),$$

avec $y > 0$, ce qui correspond à une loi exponentielle.

SÉLECTION DU SEUIL Pour pouvoir modéliser les excès de seuil, il faut d'une part posséder n réalisations iid d'une variable aléatoire X, et une valeur de seuil u. On ne conserve que les réalisations telles que $X_i > u$, et on calcule les excès de seuil $Y_i = u - X_i$. D'après le théorème précédent, les Y_i suivent une loi de Pareto Généralisée. La sélection du seuil est un enjeu majeur de la modélisation des excès de seuil. En effet, un seuil trop bas n'est pas compatible avec les hypothèses de base du théorème précédent, valable uniquement pour un seuil suffisament élevé. D'un autre côté, plus le seuil est haut, et moins de réalisations X_i le dépasseront. On n'aura pas alors suffisamment de données pour correctement inférer les paramètres du modèle des excès de seuil. Il existe deux méthodes principales pour faire un choix. La première doit être réalisée avant d'estimer les paramètres du modèle, tandis que la seconde se base sur les paramètres estimés pour différentes valeurs de seuil.

Détaillons d'abord la première méthode. Celle-ci se base sur la moyenne d'une distribution de Pareto Généralisée, définie par $E(Y) = \frac{\tilde{\sigma}}{1-\xi}$ si $\xi < 1$. Dans le cas contraire, la moyenne est infinie. On considère que l'on dispose d'observations $X_1, ..., X_n$ et d'une valeur de seuil u_0 tels que les X_i suivent une loi de Pareto Généralisée. Or, si cette loi est valable pour un seuil u_0, alors elle est aussi valable pour tous les seuils $u > u_0$, puisqu'ils sont alors tous suffisamment grands. On a donc :

$$\begin{aligned}\mathbb{E}(X - u | X > u) &= \frac{\sigma_u}{1-\xi} \\ &= \frac{\sigma_{u_0} + \xi u}{1-\xi}\end{aligned}$$

L'espérance ci-dessus est donc une fonction linéaire de u, pour tout seuil $u > u_0$, avec u_0 le seuil à partir duquel le modèle de la distribution Pareto Généralisée est valable. Si on représente cette espérance pour différentes valeurs de u, et que l'on repère les valeurs de seuils pour lesquels elle est linéaire par rapport à u, alors ces valeurs fonctionnent. En pratique, on approche l'espérance par la moyenne empirique des excès de seuils. On représente

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{max} \right\}$$

où n_u est le nombre d'observations qui dépassent le seuil u. Ce graphique est appelé *mean residual life plot*. Il est important d'ajouter les intervalles de confiance, car ce graphique peut être difficile à interpréter : la linéarité avec u n'est pas toujours évidente.

La seconde méthode pour estimer une valeur convenable du seuil consiste à ajuster le modèle de la loi de Pareto Généralisée sur les excès de seuil pour différentes valeurs de seuil. L'objectif est ici d'observer les valeurs de seuils pour lesquelles les paramètres sont constants.

En effet, si un seuil u_0 est valable, alors tous les seuils u supérieurs à u_0 sont également valables. Le paramètre de forme ξ de ces distributions, qui correspond au paramètre de forme de la distribution GEV des maximums de blocs, est alors constant. De même, pour le paramètre d'échelle, on a $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$. Avec la reparamétrisation $\sigma^* =$

$\sigma_u - \xi u = \sigma_{u_0} - \xi u_0$, on obtient un paramètre d'échelle qui ne change pas avec u. Ces deux paramètres, ξ et σ^* sont donc censés être constants à partir de la valeur de seuil u_0. L'idée est donc d'ajuster différents modèles et de représenter l'évolution de ces paramètres en fonction de u. Le seuil à partir duquel les paramètres apparaissent constants sur le graphique correspond à u_0, et tout seuil supérieur à u_0 est valable et peut être sélectionné.

INFÉRENCE DES PARAMÈTRES De façon habituelle, les paramètres seront inférés par estimation du maximum de vraisemblance. On rappelle l'expression de la fonction de répartition d'un loi de Pareto Généralisée, dans le cas où $\xi \neq 0$, $H(y) = 1 - (1 + \frac{\xi y}{\tilde{\sigma}})^{-\frac{1}{\xi}}$. On calcule la densité et on obtient :

$$f(y) = \frac{1}{\tilde{\sigma}}(1 + \frac{\xi y}{\tilde{\sigma}})^{-\frac{1+\xi}{\xi}}.$$

On en déduit l'expression de la log-vraisemblance, en notant n le nombre d'observations y_i :

$$l(\tilde{\sigma}, \xi) = -n\log(\tilde{\sigma}) - (\frac{1+\xi}{\xi}) \sum_{i=1}^{n} \log(1 + \frac{\xi y_i}{\tilde{\sigma}}).$$

Cette log-vraisemblance est finie seulement si $(1 + \frac{\xi y_i}{\tilde{\sigma}})$ est strictement positif.
Avec un raisonnement similaire dans le cas où $\xi = 0$, on obtient :

$$l(\tilde{\sigma}) = -n\log(\tilde{\sigma}) - \frac{1}{\tilde{\sigma}} \sum_{i=1}^{n} y_i.$$

RETURNS LEVELS Comme dans le cas des distributions GEV, on utilise les quantiles pour interpréter les résultats. Ici, en supposant que X suit une loi de Pareto Généralisée pour une valeur de seuil u, et pour $\xi \neq 0$, on a $P(X > x|X > u) = (1 + \xi(\frac{x-u}{\tilde{\sigma}}))^{-\frac{1}{\xi}}$. Par le théorème de Bayes, on obtient $P(X > x) = n_u(1 + \xi(\frac{x-u}{\tilde{\sigma}}))^{-\frac{1}{\xi}}$ où n_u est l'estimation empirique de la probabilité que X soit supérieur à u, c'est-à-dire n_u = nombre d'observations au-dessus de u / nombre d'observations. Ainsi, le *Return level* x_m dépassé en moyenne une fois toutes les m observations est défini tel que

$$P(X > x_m) = n_u(1 + \xi(\frac{x_m - u}{\tilde{\sigma}}))^{-\frac{1}{\xi}} = \frac{1}{m}.$$

x_m est appelé le *m-observation return level*. Il s'agit du quantile d'ordre $\frac{m-1}{m}$ de la distribution de Pareto Généralisée. A ne pas confondre avec l'interprétation des *Return levels* dans le cas des maximums par blocs, puisque le niveau de retour z_p était alors le quantile d'ordre 1-p. On a donc, d'une certaine façon, p "petit" et m "grand".

VALIDATION DU MODÈLE Comme dans le cas de la modélisation des valeurs extrêmes par maximum de blocs, on représente les *Return level plot*, *QQ plot*, *Probability plot* et *density plot*. Dans le premier, on cherche une correspondance entre la courbe et les probabilités estimées. Dans le dernier, on cherche une correspondance entre l'histogramme des données et la courbe de densité. Enfin, dans les deux du milieu, on cherche à observer une linéarité par rapport à la première bissectrice.

1.2 Cas dépendant

INTRODUCTION Dans tous les théorèmes qui ont été énoncés jusqu'à présent, l'une des hypothèses de bases était que les observations de l'évènement étaient indépendantes. En pratique, c'est rarement le cas, surtout lorsqu'on parle d'évènement extrêmes : il y a plus de chance d'observer un évènement extrême à la suite d'un autre qu'à tout autre moment. Nous allons donc dans cette partie aborder le cas des observations dépendantes dans le temps, c'est-à-dire des séries temporelles. Nous allons cependant nous restreindre au cas des séries stationnaires, et plus particulièrement celles dont la dépendance à long terme est limitée (c'est-à-dire que si deux observations sont suffisamment éloignées dans le temps, alors on peut les considérer comme étant totalement indépendantes).

SÉRIES STATIONNAIRES Cette propriété de dépendance à long terme limitée peut être formulée en terme mathématiques comme suit :

Définition : Une série stationnaire $X_1, ..., X_n$ satisfait les conditions $D(u_n)$ si pour tout $i_1 < ... < i_p < j_1 < ... < j_q$ avec $j_1 - i_p > l$,

$$|P(X_{i_1} < u_n, ..., X_{i_p} < u_n, X_{j_1} < u_n, ..., X_{j_q} < u_n) - P(X_{i_1} < u_n, ..., X_{i_p} < u_n)P(X_{j_1} < u_n, ..., X_{j_q} < u_n)| \leqslant \alpha(n, l),$$

avec $\alpha(n, l_n)$ qui tend vers o pour une suite l_n telle que $\frac{l_n}{n}$ tend vers o quand n tend vers l'infini. Ce théorème stipule donc que des séquences suffisamment éloignées dans le temps (indexée par i et j) peuvent être considérées comme étant indépendantes. En pratique, ce résultat est quasi impossible à prouver, étant donnée que l'on n'a généralement pas la loi de X.

Théorème : Soit $X_1, ..., X_n$ une série stationnaire et soit $M_n = max(X_1, ..., X_n)$. Alors si $\{a_n > 0\}$ et b_n sont des suites de constantes telles que $P(\frac{M_n - b_n}{a_n} \leqslant z) \to G(z)$, où G est une fonction de répartition non dégénérée, avec les conditions $D(u_n)$ satisfaites pour $u_n = a_n z + b_n$ pour tout réel z, alors G appartient à la famille des distributions GEV.

Ce théorème stipule donc que la distribution limite des maximums d'une série stationnaire avec une faible dépendance à long-terme est la même que pour des variables indépendantes (attention cependant, les paramètres inférés ne seront pas exactement les mêmes). Plus précisément, si on note M_n le maximum des réalisations d'une série stationnaire et M_n^* le maximum des réalisations de varibales aléatoires indépendantes ayant la même loi marginale que les variables de la série stationnaire, et si on suppose que pour les mêmes suites de constantes a_n et b_n, les fonctions de répartition des maximums convergent respectivement vers $G_2(z)$ et vers $G_1(z)$, alors $G_2(z) = G_1(z)^\theta$, avec $\theta \in]0, 1]$. Cela revient à considérer la distribution limite du maximum des variables indépendantes avec des paramètres modifiés.

Le paramètre θ est appelé *extremal index*.

Il existe une autre interprétation à ce paramètre. Comme dit plus haut, il est probable d'observer un autre évènement extrême à la suite d'un premier. Les évènements

extrêmes auront ainsi tendance à se produire en groupe, en *clusters*. On a alors $\theta = (\text{taille moyenne limite des clusters})^{-1}$, propriété que l'on admet ici.

MODÉLISATION PAR MAXIMUMS DE BLOCS Nous avons expliqué ci-dessus que si une série stationnaire a une dépendance à long-terme limitée, alors la distribution limite de ses maximums appartient à la famille des distributions GEV. Certes, les paramètres ne sont pas les mêmes que dans le cas indépendant, mais comme ils sont directement inférés par maximum de vraisemblance, cela n'a pas d'importance. Ainsi, on peut considérer que rien ne change par rapport au cas dépendant, et les méthodes de modélisation et d'estimation restent les mêmes.

MODÉLISATION PAR EXCÈS DE SEUIL Il est également possible de modéliser les excès de seuil avec des séries stationnaires ayant une dépendance à long-terme limitée. Cependant, il convient ici de faire attention à la tendance des évènements extrêmes à arriver en *clusters*. En effet, dans le cas de la modélisation par blocs maximas, on ne conserve qu'une observation sur une longue période, et du fait de l'hypothèse d'indépendance pour des observations suffisamment séparées dans le temps, tous les extrêmes conservés peuvent être considérés indépendants. Cependant, lorsqu'on modélise les excès de seuil, on garde toutes les observations supérieures au seuil considérées, et ainsi on garde tous les évènements extrêmes ayant lieu consécutivement, ou presque, et qui ne sont donc pas suffisamment éloignés dans le temps pour être considérés indépendants. Si l'on veut pouvoir considérer que les excès de seuil ont pour marginale une loi de Pareto Généralisée, et appliquer les méthodes d'estimation présentées plus haut (expression de la log-vraisemblance), il faut se débarrasser de la dépendance intra-cluster.

La méthode la plus couramment utilisée est appelée *declustering*. Elle consiste à filtrer les évènements extrêmes pour qu'ils puissent tous être considérés indépendants. Elle consiste à définir une règle pour déterminer la taille des clusters des excès, puis à ne conserver que le maximum des excès de seuil au sein de chaque cluster. On suppose alors tous les maximums de clusters indépendant (car suffisamment éloignés si la règle choisie est adaptée) et on ajuste une distribution de Pareto Généralisée sur les excès de seuil conservés.

Une règle pour définir les clusters peut par exemple être de dire que les évènements extrêmes sur une certaine période de temps appartiennent au même cluster.

Les résultats sont malheureusement sensibles au choix de la taille des clusters.

2 APPLICATIONS

2.1 Finance : calcul de la VaR

VALUE AT RISK La *Value at Risk*, plus simplement appelée VaR, est un outil pour évaluer le risque de marché d'un portefeuille d'instruments financiers (actions, obligations, options, monnaies,...). La VaR définit la plus grande perte possible sur une période donnée, avec une probabilité donnée. En d'autres mots, pour une probabilité (aussi appelé niveau de confiance) α et une durée t, la VaR est la perte excédée sur la durée t avec la probabilité $1-\alpha$. La VaR peut ainsi être vue comme un quantile : si on note X la variable aléatoire des profits négatives (c'est-à-dire que X négatif correspond à un profit et X positif correspond à une perte), alors on a :

$$\begin{aligned} &P(X \leqslant VaR) = \alpha \\ &\Leftrightarrow F(VaR) = \alpha \\ &\Leftrightarrow VaR_\alpha = F^{-1}(\alpha) \end{aligned}$$

De nombreuses méthodes existent pour calculer la VaR : la méthode de Variance Covariance, la méthode historique et la simulation de Monte-Carlo.

La première consiste à supposer une distribution au rendement de l'asset considéré, et à prendre pour VaR le quantile (de niveau α ou $1-\alpha$ selon les cas) de la distribution. Dans la plupart des cas, la distribution des rendements sera supposée normale. En réalité, la queue de distribution des rendements est souvent plus lourde que dans une distribution normale, et conduisent à une sous-estimation de la VaR.

La méthode historique permet de ne faire aucune supposition quant à la distribution des rendements. Elle consiste à représenter sous la forme d'un histogramme les données des rendements passés, et à prendre pour VaR, à un niveau de confiance α l'observation passée telle que $1-\alpha\%$ des observations soient plus grandes ou plus petites (selon les cas). Cette méthode a pour défaut de supposer que les conditions du marché ne changeront pas par rapport au passé, et ne permet pas de prédire des évènements pires que ce qui n'a jamais été observé dans le passé.

Enfin, la simulation de Monte-Carlo consiste à modéliser et à générer des rendements. Un certain nombre de rendements sont simulés selon le modèle proposé, puis, comme pour la méthode historique, on regarde la $(1-\alpha)\%$ observations plus grandes ou plus petites. Cette méthode peut s'avérer très coûteuse en fonction de la complexité du portefeuille.

La VaR étant une mesure de risque extrême, localisée dans les queues de distributions, l'utilisation de la Théorie des Valeurs Extrêmes pour son calcul semble naturelle, et permet de plus de contourner certains défauts des méthodes présentées plus haut. Dans son livre, Coles explique d'ailleurs que dans un contexte financier, les quantiles extrêmes de la distribution du rendement d'un instrument financier correspond à la VaR. Ainsi, le *Return Level Plot* est en réalité un graphe de la VaR en fonction du risque. En effet, ce qu'on appelle *Return level* d'ordre z_p est en réalité le quantile d'ordre $1-p$, tel que $P(X \leqslant z_p) = 1-p$. Avec $1-p = \alpha$, c'est exactement la définition donnée plus haut

(mais ici p est "petit" là où α était "grand"). Ainsi en prenant $VaR = z_p$, la probabilité d'excéder la VaR a pour probabilité $p = 1 - \alpha$. Dans le *Return level plot*, on représente z_p en fonction de la *Return period*, c'est-à-dire $\frac{1}{p}$. Plus p est petit, plus $\frac{1}{p}$ est grand, et plus la VaR considérée est élevée et la probabilité de la dépasser faible. C'est ce qu'on peut directement lire sur un tel graphique.

MODÈLE ET DONNÉES On dispose du taux d'échange journalier EUR/USD du 4 janvier 1999 au 27 janvier 2023, soit 6167 données (certains jours sont manquants). Le graphe ci-dessous permet de visualiser l'évolution du taux d'échange EUR/USD en fonction du temps.

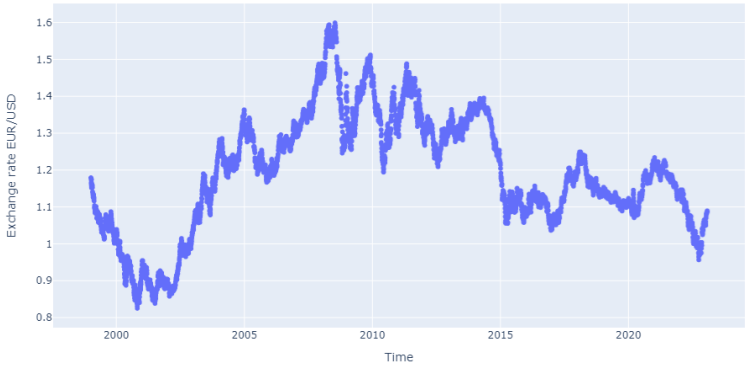


Figure 1: Evolution du taux d'échange EUR/USD en fonction du temps

Il s'agit d'une série temporelle non stationnaire. Comme expliqué plus haut, il est possible d'étendre la Théorie des Valeurs Extrêmes aux séries temporelles, à la condition qu'elles soient stationnaires et que des observations suffisamment éloignées dans le temps puissent être considérées indépendantes.

Afin de travailler sur une série temporelle stationnaire, nous allons nous intéresser au logarithme du rendement, comme c'est usuel en finance. Cela correspond à calculer :

$$r_t = \log(\frac{X_t}{X_{t-1}}) = \log(X_t) - \log(X_{t-1})$$

où X_t est la valeur du taux d'échange au jour t.

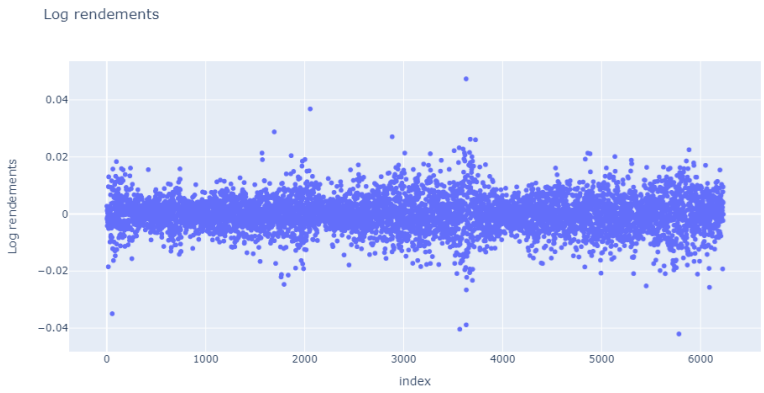


Figure 2: Log rendement EUR/USD

La série est désormais stationnaire. Observons, via un histogramme, la distribution des log rendements.

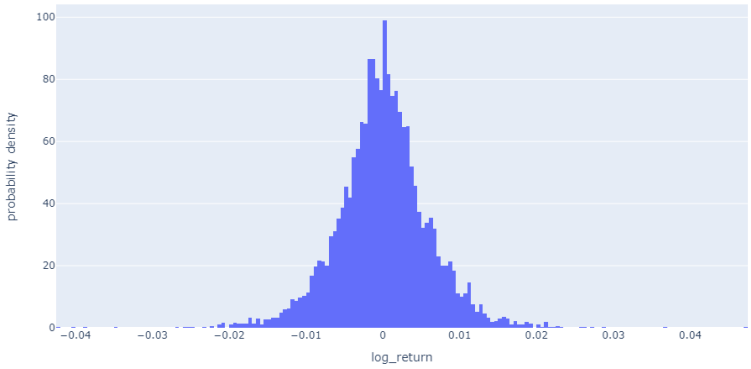


Figure 3: Distribution des log rendements

Vérifions si la distribution est normale.

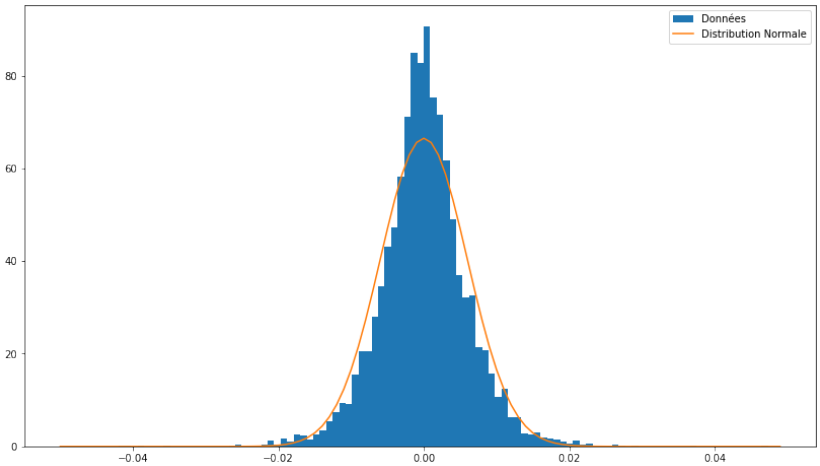


Figure 4

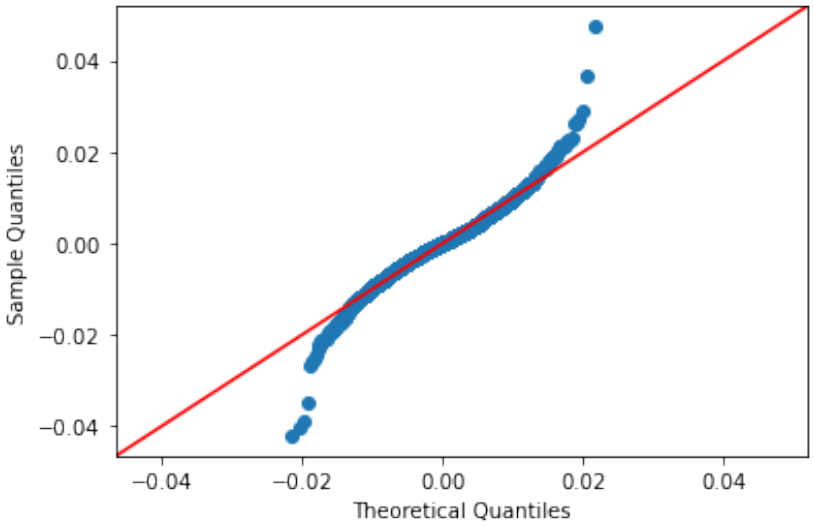


Figure 5

Le premier graphique compare la densité d'une loi normale centrée en o et de variance la variance des données. Le second graphique est un QQ-plot. On observe que bien que les données semblent être normales à priori, le QQ-plot n'est pas linéaire, et les queues de distribution des données semblent être plus lourdes que pour une distribution Gaussienne.

Cependant, pour pouvoir réellement appliquer la Théorie des Valeurs Extrêmes à cette série temporelle, il faut s'assurer qu'elle vérifie la condition $D(u_n)$ définie plus haut. Plus simplement, un ACF permettrait d'observer la corrélation entre X_t et X_{t+h}.

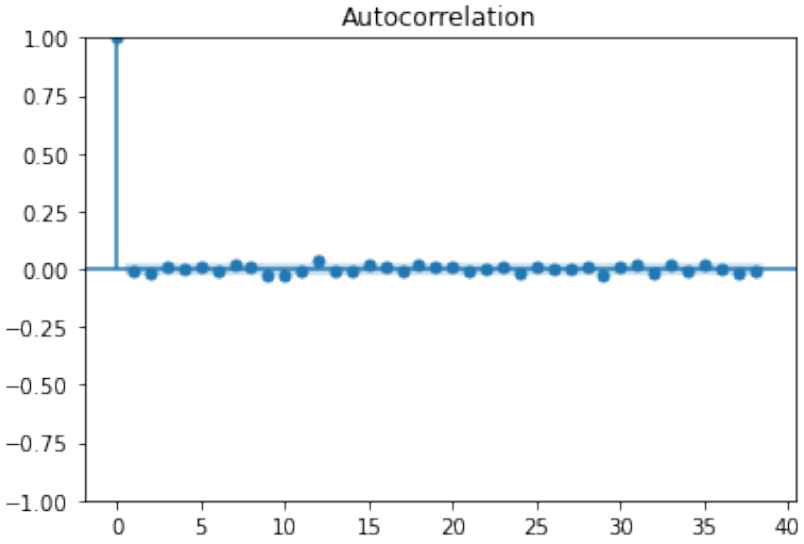


Figure 6

La corrélation entre X_t et X_{t+h} pour $h \neq 0$ est très faible. Il semble donc peu probable que la valeur au jour t influence celle au jour t+h, et on suppose donc que l'on peut appliquer la Théorie des Valeurs Extrêmes à la série des log rendements.

On suppose dans la suite que l'on possède de l'Euro et donc que l'on est en position *long*. Si le rendement est négatif, cela signifie que le taux d'échange a diminué par rapport au jour d'avant et donc que l'Euro est plus faible. La VaR est donc ici, pour un intervalle de confiance donné, et pour une période de 1 jour, la plus grande perte possible. En Théorie des Valeurs Extrêmes, cela correspond à chercher les minimums. Par confort, nous travaillerons donc sur la série des log rendements négatifs, et ainsi nous chercherons à déterminer la distributions des maximums.

Dans la suite, pour cette application, nous utiliserons le package *evd* de R pour construire des modèles et inférer les paramètres.

BLOCK MAXIMAS Dans un premier temps, nous allons appliquer la méthode par Blocs, qui ne change pas par rapport au cas indépendant.

Nous devons d'abord choisir la taille des blocs afin d'en extraire le maximum. Comme nous possédons des données de 1999 à 2023, nous pouvons donc sélectionner un maximum par an (l'année 2023 étant incomplète, elle n'est pas comparable aux autres et est par conséquent exclue des données). Ci-dessous sont représentés les extrêmes pour chaque blocs, c'est-à-dire les variables $Z_1, ..., Z_m$ telles que $Z_i = max(r_1, ..., r_n)$ où n est le nombre de données par bloc et m le nombre de blocs.

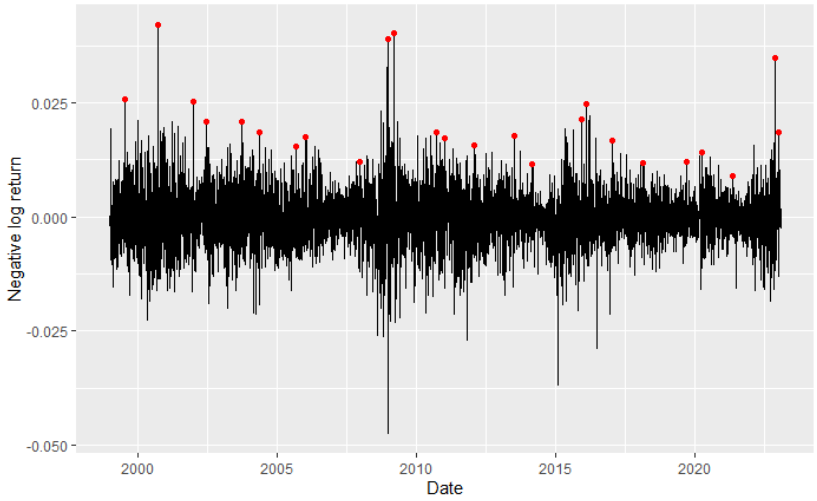


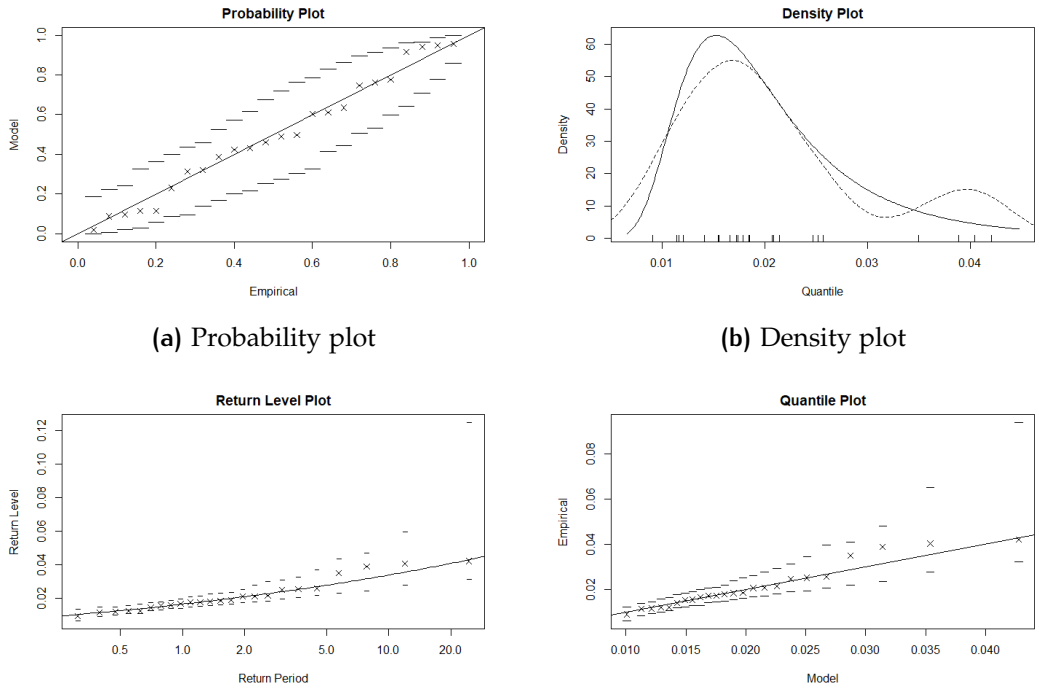
Figure 7: Valeurs extrêmes de la série des Log rendements

Le modèle est ensuite ajusté sur les extrêmes, les paramètres étant déterminés par maximum de vraisemblance. On obtient $\hat{\mu} = 0.016327$ (std($\hat{\mu}$) = 0.0013965), $\hat{\sigma} = 0.005972$ (std($\hat{\sigma}$) = 0.0009465) et $\hat{\xi} = 0.193631$ (std($\hat{\xi}$) = 0.1972221). Les variables Z_i semblent donc appartenir à la famille de distributions de Fréchet, comme c'est habituellement le cas avec les données financières.

```
Call: fgev(x = max_annuels_red[, 2])
Deviance: -165.232

Estimates
     loc     scale     shape
0.016327  0.005972  0.193631

Standard Errors
      loc      scale      shape
0.0013965  0.0009465  0.1972221

Optimization Information
  Convergence: successful
  Function Evaluations: 69
  Gradient Evaluations: 7
```

Figure 8: Résumé du modèle et de ses paramètres estimés

Le diagnostic du modèle semble confirmer que le modèle est adapté aux données. Le *Probability plot* est linéaire. Les points du *Return level plot* et du *Quantile plot* sont tous compris dans les intervalles de confiance, et suivent plutôt correctement la courbe théorique. Le *Density plot* est correct. On peut donc valider ce modèle.



(a) Probability plot

(b) Density plot

(c) Return level plot

(d) Quantile plot

Figure 9: Diagnostic du modèle

Cependant, dans ce modèle, l'erreur standard sur le paramètre de forme est presque égale à lui-même. Si on calcule l'intervalle de confiance pour le paramètre $\hat{\xi}$ en considérant sa propriété de normalité asymptotique, on obtient que $\hat{\xi} \in [-0.1929243, 0.5801864]$.

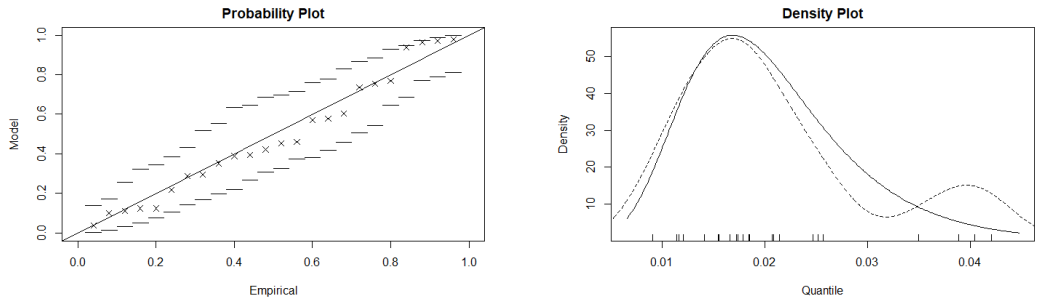
0 appartient donc à cet intervalle. Serait-il possible d'utiliser une distribution de type Gumbel à la place de la distribution de type Fréchet ? Nous pouvons vérifier cela en créant un second modèle dans lequel le paramètre de forme est fixé à 0, et, puisque que le modèle type Gumbel est inclu dans le modèle type Fréchet, en réalisant un test du rapport de vraisemblance de H_0 : le modèle Gumbel est suffisant contre H_1 : le modèle Gumbel n'est pas suffisant.

On réalise ce test à l'aide de la fonction *anova* de R. On obtient une p-valeur de 0.2438. On ne peut donc pas rejeter H_0 pour des niveaux standards. On va donc pouvoir considérer le modèle de type Gumbel dans la suite. Les paramètres de ce modèle sont donc $\hat{\xi} = 0$ (fixé), $\hat{\mu} = 0.016902$ (std($\hat{\mu}$) = 0.001409), et $\hat{\sigma} = 0.006589$ (std($\hat{\sigma}$) = 0.001027).

```
Call: fgev(x = max_annuels_red[, 2], shape = 0)
Deviance: -163.8733

Estimates
     loc     scale
0.016902  0.006589

Standard Errors
     loc     scale
0.001409  0.001027

Optimization Information
  Convergence: successful
  Function Evaluations: 44
  Gradient Evaluations: 6
```

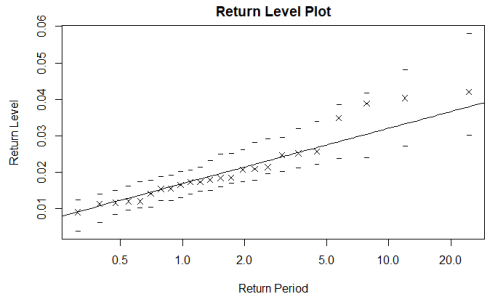
Figure 10: Résumé du modèle final et de ses paramètres estimés

Observons les grahiques de validation de ce modèle. Globalement, il y a peu de différences par rapport au diagnostic précédent, et les mêmes remarques s'appliquent.

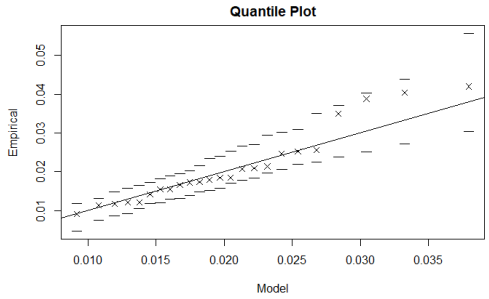


(a) Probability plot

(b) Density plot



(c) Return level plot



(d) Quantile plot

Figure 11: Diagnostic du modèle final

Nous pouvons à présent obtenir la VaR pour différents niveaux de confiance. Pour cela, prenons les *Return levels* d'ordre $p = 0.05, 0.025, 0.01, 0.001$ correspondant aux *Return period* $20, 40, 100$ et 1000 respectivement. Pour obtenir ces valeurs, on prend plus simplement les quantiles d'ordre $1 - p$ de la distribution Gumbel correspondante. Le calcul des intervalles de confiance se base sur la normalité asymptotique d'une fonction des estimateurs du maximum de vraisemblance et de la delta méthode, comme indiqué dans le livre de Coles. On obtient les *Value at Risk* suivantes :

Return period	Valeur	IC inférieur	IC supérieur
20	0.03647153	0.03373994	0.03920313
40	0.04112349	0.03837732	0.04386966
100	0.04721078	0.04445562	0.04996595
1000	0.06241157	0.05965093	0.06517221

Comparons avec les valeurs obtenues par la méthode historique. Pour cela, on suit la méthodologie présentée plus haut et on obtient les résultats suivants :

- pour p = 0.05, VaR = 0.009475322437306116 ;
- pour p = 0.025, VaR = 0.011958186845189178 ;
- pour p = 0.01, VaR = 0.01614900707183195 ;
- pour p = 0.001, VaR = 0.02568734348237925.

Les valeurs obtenues par la méthode historiques sont sous-estimées par rapport à celles obtenues par l'application de la théorie des Valeurs Extrêmes. Cela peut notamment s'expliquer par le fait que la méthode historique se base uniquement sur ce qui a déjà été observé, là où la seconde méthode est capable d'extrapoler.

SEUIL Nous allons à présente appliquer la méthode du seuil sur la même série des log rendements négatifs. Comme expliqué précédemment, pour les séries stationnaires, il convient de réaliser un *declustering* des valeurs extrêmes, afin de s'assurer que celles-ci puissant bien être considérées comme étant indépendantes. Cependant, l'ACF a montré que la corrélation entre X_t et X_{t+h} était très faible : une valeur élevée pour X_t n'indique pas nécessairement que X_{t+h}, avec h petit, soit également élevée.

Avant cela, il convient de sélectionner un seuil suffisamment grand à partir duquel on peut considérer que les excès de seuils suivent une loi de Pareto généralisée. On utilise pour cela le *mean residual life plot*. On a multiplié les données par 100 pour rendre le graphique plus lisible. Pour le calcul des intervalles de confiance, on utilise le TCL, comme suggéré par Coles. En notant $\bar{Y}$ la moyenne empirique des excès de seuil, on a :

$$\frac{\sqrt{n}(\bar{Y} - \mathbb{E}(Y))}{\sqrt{\mathbb{V}(Y)}} \sim \mathrm{Normal}(0, 1)$$

D'où, pour un intervalle de confiance à 95

$$\bar{Y} \in [\mathbb{E}(Y) - 1.96\frac{\sigma_Y}{\sqrt{n}}, \mathbb{E}(Y) + 1.96\frac{\sigma_Y}{\sqrt{n}}]$$

On approximera la moyenne de Y par la moyenne empirique et l'écart-type par l'écart-type empirique.

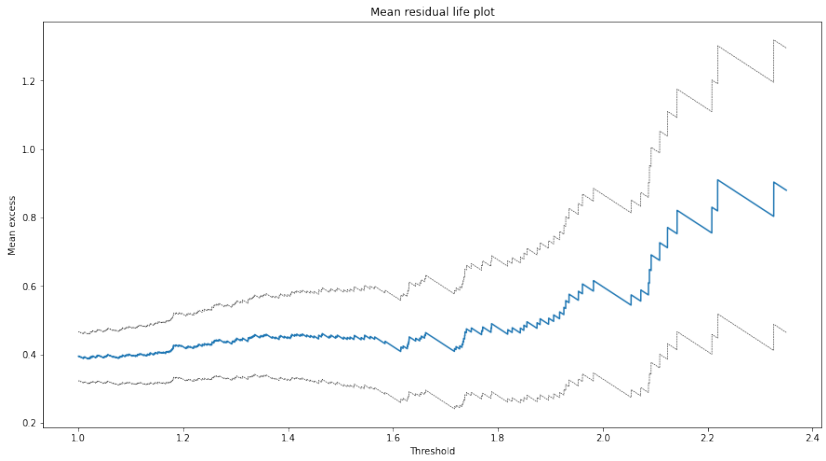
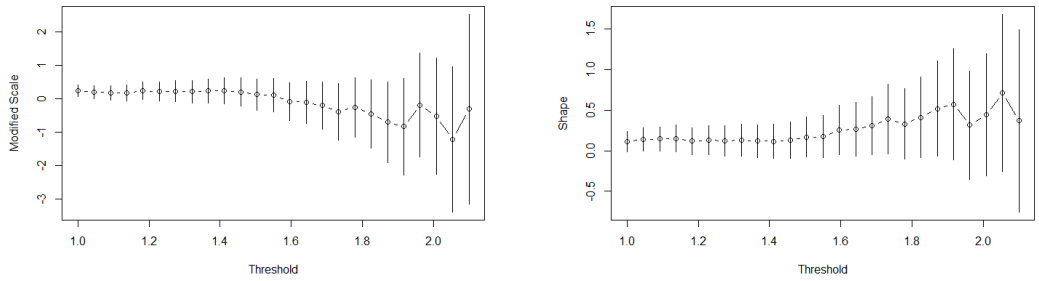


Figure 12: Mean Residual Life plot

Le seuil $u = \frac{1.5}{100} = 0.015$ semble pouvoir fonctionner, cependant il est difficile d'interpréter correctement ce graphique. Réalisons également la deuxième méthode, en ajustant le modèle sur plusieurs valeurs de seuils.



(a) Scale stability plot **(b)** Shape stability plot

Figure 13: Stabilité des paramètres pour différentes valeurs de seuil

Les paramètres semblent stables relativement aux intervalles de confiance. On utilisera la valeur de seuil $u = 0.015$ dans la suite.

Dans un premier temps, pour tester l'hypothèse sur le *declustering* évoquée plus haut, observons les valeurs extrêmes pour le seuil u ci-dessus et sans *declustering*, puis les extrêmes pour ce même seuil et avec des clusters de longueur 24 heures.

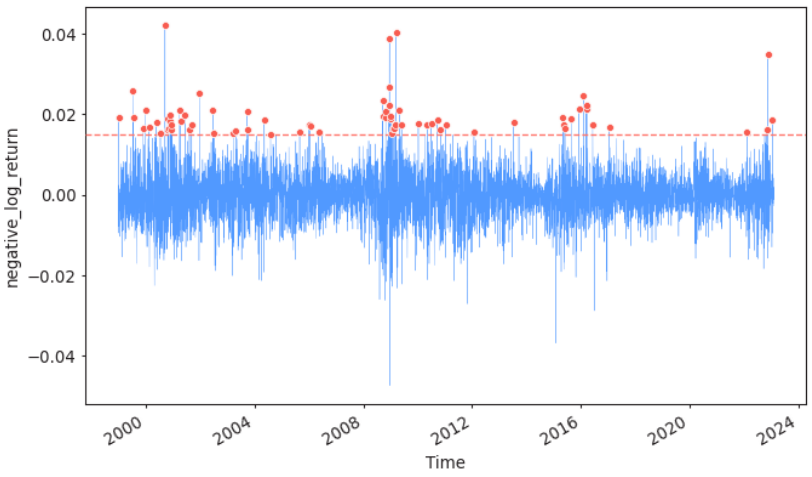


Figure 14: Extrêmes pour u = 0.015 et r = 1H

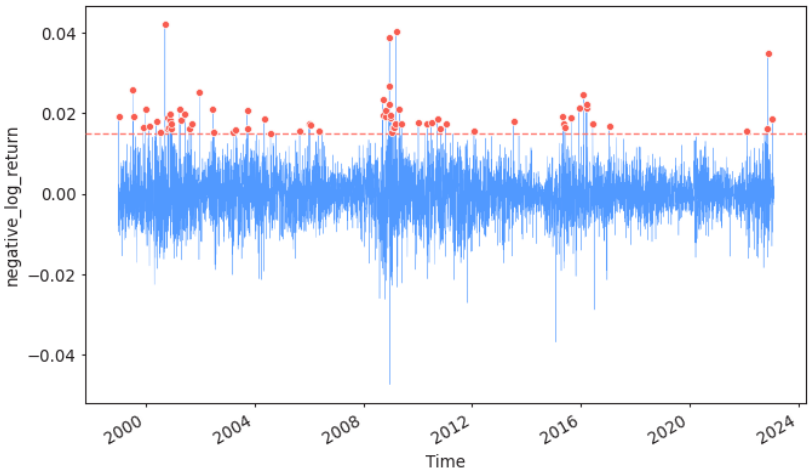


Figure 15: Extrêmes pour u = 0.015 et r = 24H

Il n'y a pas de différence. On va donc garder les extrêmes sans *declustering*.
On ajuste le modèle et on obtient les paramètres suivants :

```
Call: fpot(x = data2$Negative_log_return, threshold = 0.015)
Deviance: -633.7256

Threshold: 0.015
Number Above: 72
Proportion Above: 0.0117

Estimates
    scale      shape
4.513e-03  1.098e-17

Standard Errors
    scale      shape
0.0006241  0.0786075

Optimization Information
  Convergence: successful
  Function Evaluations: 27
  Gradient Evaluations: 1
```

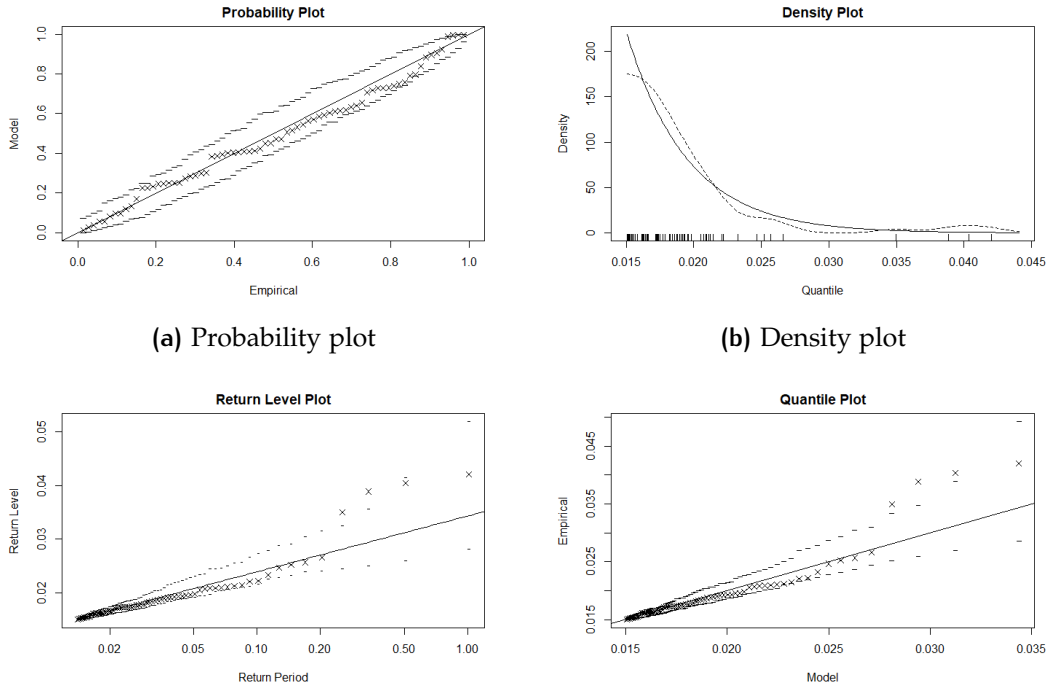
Figure 16: Résultats de l'ajustement

La première chose que l'on constate est la très faible valeur du paramètre de forme, surtout comparée à son erreur standard. Tentons un autre modèle, avec cette même valeur de seuil, en fixant le paramètre de forme à 0, et réalisons un test du rapport de vraisemblance.

```
Analysis of Deviance Table

     M.Df Deviance Df Chisq Pr(>chisq)
fit3    2  -633.73
fit4    1  -633.73  1     0          1
```

Figure 17: Résultats du test de rapport de vraisemblance pour les excès de seuil

La p-valeur vaut 1 : on ne peut pas rejeter H_0 et on peut donc conserver le modèle simplifié avec le paramètre de forme fixé à 0, ce qui correspond à une distribution exponentielle de paramètre $\frac{1}{\hat{\sigma}}$. Observons les graphiques de validation du modèle.



(a) Probability plot

(b) Density plot

(c) Return level plot

(d) Quantile plot

Figure 18: Diagnostic du modèle

Note : l'échelle du Return level plot n'est pas annuelle et il ne faut donc pas l'interpréter comme telle ; je ne suis pas parvenue à la modifier.

Le modèle semble plutôt satisfaisant : le *Probability plot* est presque linéaire, le *Return level plot*, l'est au début, et le décalage sur la fin reste compris dans les intervalles de confiance. De même pour le *QQ plot*. La densité de la distribution ajustée semble visuellement cohérente avec la répartition des données.

Observons à présent les *Return level*. On cherche toujours ici à obtenir $P(Y > x) = p$, avec Y la variable des excès de seuil et $p = 0.05, 0.025, 0.01$ et 0.001. Les *Return period* associées sont donc 20, 40, 100 et 1000 respectivement. Pour cela, on récupère les quantiles de la loi généralisée de Pareto avec les paramètres du modèle ajusté. On n'oublie pas de mettre le paramètre *loc* à 0.015 (autrement, les quantiles ne seront pas supérieur à 0.015 ce qui n'aurait pas de sens ici). Pour les intervalles de confiance, on applique la méthode indiquée par Coles. On utilise donc le théorème de la delta méthode, en n'oubliant pas également de prendre en compte la proportion empirique de données dépassant le seuil, car cette valeur intervient en réalité dans le calcul. Enfin, la valeur de $\frac{1}{m}$ du livre correspond ici à p.

Return period	Valeur	IC inférieur	IC supérieur
20	0.02851867	0.02205626	0.03498107
40	0.03164659	0.02456807	0.03872512
100	0.03578149	0.02788851	0.04367446
1000	0.04617223	0.03623266	0.05611180

Les valeurs obtenues sont plus faibles que dans le cas de la méthode par blocs (c'est souvent le cas).

Dans la réalité, les portefeuilles financiers sont composés de plusieurs instruments ou de plusieurs assets. La prochaine étape consisterait donc à étendre cette application univariée à un calcul multivarié, pour plus de réalisme.

2.2 Environnement: Montant de l'indemnité de l'assurance incendie

2.2.1 Introduction

1. Environnement expérimental

   Cette application est réalisée à l'aide de la bibliothèque PyExtremes.

   - PyExtremes est une bibliothèque Python pour l'analyse statistique des valeurs extrêmes qui fournit une gamme de fonctions et de classes pour la modélisation et l'analyse des événements extrêmes.
   - Sur l'ajustement de modèles des valeurs extrêmes : justement d'une large gamme de distributions telles que la distribution généralisée des valeurs extrêmes (GEV), la distribution de Pareto généralisée (GPD), en utilisant des méthodes telles que MLE, L-Moments, Bayesian MCMC, etc. Distribution exponentielle, etc.
   - Analyse des valeurs extrêmes : calcule les valeurs extrêmes au-delà d'un seuil spécifique et fournit des méthodes courantes d'analyse des valeurs extrêmes telles que Peak Over Threshold (POT), Block Maxima (BM), etc.
   - Quantification de l'incertitude : estimation de l'incertitude des paramètres, de l'incertitude de la distribution, de l'incertitude de la prédiction, etc.
   - Visualisation : méthodes de traçage telles que les histogrammes, les diagrammes de dispersion, les diagrammes QQ, les diagrammes PP, les fonctions de distribution empirique et la génération de graphiques interactifs via les bibliothèques Bokeh et Matplotlib.

2. Méthodes d'extraction des valeurs extrêmes

   - Dans la bibliothèque PyExtremes, POT[1] et BM[2] sont deux méthodes d'analyse des valeurs extrêmes ; la méthode POT (Peak Over Threshold) identifie les valeurs extrêmes en identifiant les pics dans les données qui sont supérieurs à un seuil prédéterminé, tandis que la méthode BM (Block Maxima) divise les données en blocs de taille fixe et utilise la valeur maximale de chaque bloc comme valeur extrême.
   - L'avantage de la méthode POT est qu'elle maximise l'utilisation des données et permet l'identification rapide d'un petit nombre de valeurs extrêmes dans l'ensemble de données. En outre, la méthode POT permet de choisir différents seuils afin de détecter des événements extrêmes de différentes tailles. Toutefois, l'inconvénient de la méthode POT est qu'il est impossible d'identifier les extrêmes lorsqu'il n'y a pas d'observations dans l'ensemble de données qui soient plus grandes que le seuil, et qu'il peut être difficile de choisir le bon seuil.
   - L'avantage de la méthode BM est que les données peuvent être découpées en morceaux, ce qui permet de surmonter les inconvénients de la méthode POT, tout en améliorant l'efficacité des calculs lorsque l'on traite de grandes quantités de données. En outre, la méthode BM ne nécessite pas la sélection d'un seuil, puisque chaque bloc a une valeur maximale, ce qui permet à l'analyste

d'appliquer plus facilement les données à des séries temporelles ou spatiales. Toutefois, les inconvénients de la méthode BM sont la possibilité d'obtenir des résultats trop conservateurs dans les cas extrêmes et la nécessité de choisir la bonne taille de bloc.

3. Algorithmes d'ajustement des distributions de valeurs extrêmes

   Dans la bibliothèque PyExtremes, MLE et MCMC sont deux algorithmes permettant d'ajuster les distributions de valeurs extrêmes.

   - Algorithme MLE (maximum de vraisemblance)

Algorithm 1 Maximum Likelihood Estimation Algorithm for GPD[3]

Input: sample data $y_1, y_2, ..., y_n$
Initialize: $\theta = (\xi, \sigma)$
Define likelihood function: $L(\theta|y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{GPD}(y_i|\theta)$
Define log-likelihood function: $\ell(\theta|y_1, \ldots, y_n) = \sum_{i=1}^{n} \ln f_{GPD}(y_i|\theta)$
Set convergence criterion: ϵ
repeat
    Compute gradient of log-likelihood: $\nabla\ell(\theta|y_1, \ldots, y_n)$
    Compute Hessian matrix of log-likelihood: $H(\theta|y_1, \ldots, y_n)$
    Update θ using Newton-Raphson method:
    $\theta \leftarrow \theta - H(\theta|y_1, \ldots, y_n)^{-1}\nabla\ell(\theta|y_1, \ldots, y_n)$
until $|\nabla\ell(\theta|y_1, \ldots, y_n)| < \epsilon$
Output: θ

   – MLE est une méthode couramment utilisée pour estimer les paramètres d'une distribution de probabilité à partir d'un échantillon de données. Dans le contexte de l'analyse des extrêmes, MLE peut être utilisé pour ajuster un modèle de distribution généralisée d'extrêmes (GPD), qui est utilisé pour décrire les processus extrêmes dans les données.

   – Dans le modèle GPD, l'algorithme MLE est utilisé pour estimer les paramètres du modèle en maximisant la fonction de vraisemblance maximale (maximum likelihood function). Cette fonction peut être représentée par :

$$L(\theta \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{GPD}(y_i \mid \theta)$$

   où $f_{GPD}(y_i \mid \theta)$ représente la densité de probabilité de la distribution GPD, θ est le paramètre du modèle GPD, et $y_1, \ldots, y_n$ sont les données de l'échantillon. En maximisant cette fonction, on peut obtenir l'estimation optimale des paramètres du modèle GPD.

- Algorithme MCMC (chaînes de Markov Monte Carlo) :

   – MCMC est une méthode utilisée pour générer des échantillons aléatoires à partir de distributions de probabilité complexes. Dans le contexte de

Algorithm 2 Markov Chain Monte Carlo Algorithm for BEVD[4][5]

Input: sample data $y_1, y_2, ..., y_n$
Initialize: model parameters θ, number of iterations N, burn-in period B
Define likelihood function: $L(\theta|y_1, \ldots, y_n) \propto \prod_{i=1}^{n} f_{BEVD}(y_i|\theta)$
Define prior distribution: $\pi(\theta)$
for $t = 1$ to N **do**
  Propose new parameter values: $\theta' \sim q(\theta'|\theta)$
  Compute acceptance ratio: $r = \min\left(1, \frac{L(\theta'|y_1,...,y_n)\pi(\theta')q(\theta|\theta')}{L(\theta|y_1,...,y_n)\pi(\theta)q(\theta'|\theta)}\right)$
  Sample random number $u \sim U[0, 1]$
  if $u < r$ **then**
    Accept new parameter values: $\theta \leftarrow \theta'$
  end if
  if $t > B$ **then**
    Store parameter values: θ_t
  end if
end for
Output: posterior distribution samples $\theta_{t_{t=B+1}}^{N}$

l'analyse des extrêmes, MCMC peut être utilisé pour ajuster un modèle de distribution de probabilité d'extrêmes de Bayes (BEVD), qui est également utilisé pour décrire les processus extrêmes dans les données.

– Dans le modèle BEVD, l'algorithme MCMC est utilisé pour estimer la distribution a posteriori en générant des échantillons aléatoires et en extrayant les paramètres du modèle à partir de ces échantillons. Les échantillons aléatoires générés par MCMC peuvent être utilisés pour calculer la distribution de probabilité a posteriori, ce qui permet d'estimer les paramètres du modèle et l'incertitude associée.

2.2.2 Processus de modélisation

1. Présentation des données

   - L'ensemble de données univariées a été collecté auprès de Copenhagen Reinsurance et comprend 2167 pertes pour cause de sinistres incendie sur la période 1980-1990. Ils ont été corrigés de l'inflation pour refléter les valeurs de 1985 et sont exprimés en millions de couronnes danoises. La distribution des données est représentée dans [Fig. 19] et nous pouvons voir que certaines valeurs semblent plus extrêmes que d'autres.

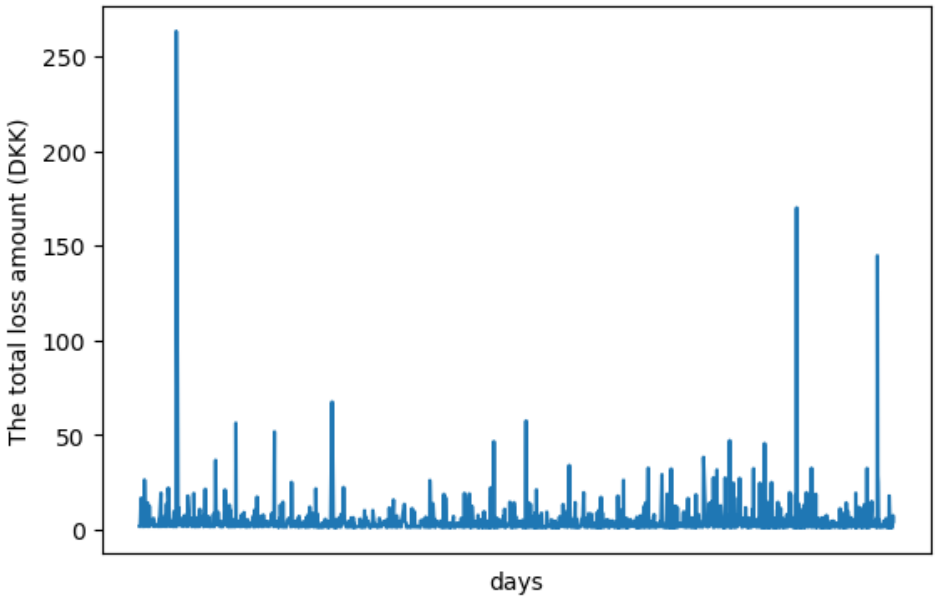


Figure 19: Représentation des données

Comme le montrent les deux graphiques ci-dessous [Fig. 20], la distribution des données présente une forte asymétrie positive et quelques montants de sinistres très élevés, qui peuvent représenter des cas extrêmes.

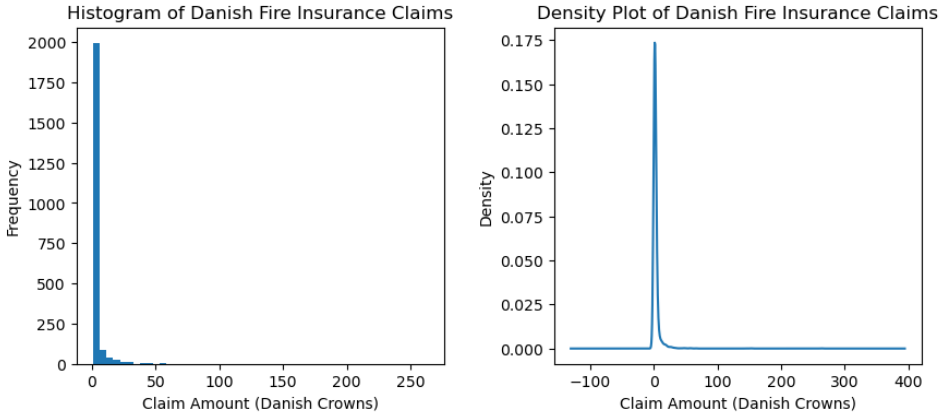
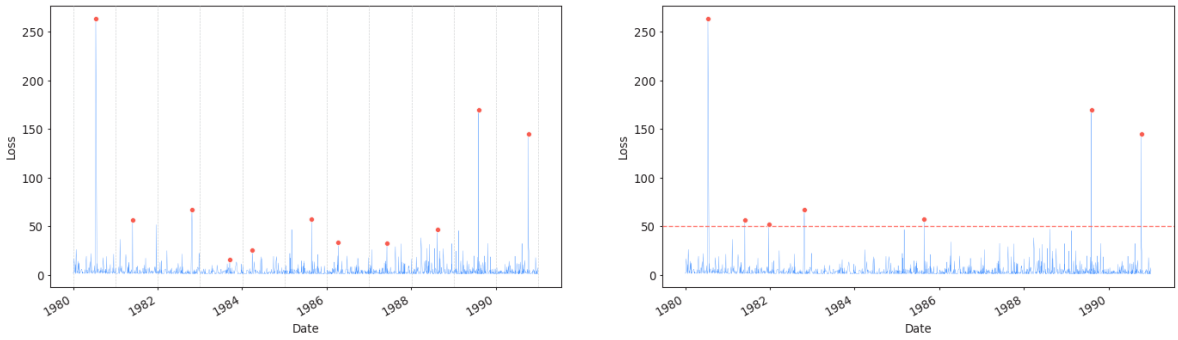


Figure 20: Graphiques de densité et histogrammes de données

2. Analyse du modèle

   Comme le montre [Fig.21], nous avons utilisé deux modèles, BM et POT, respectivement, le modèle POT utilisant 50 comme seuil et les points rouges de la figure représentant les extrêmes.



(a) Visualisation du modèle BM

(b) Visualisation du modèle POT

Figure 21: Visualisation du modèle des valeurs extrêmes

Après une comparaison expérimentale, nous avons constaté que les meilleurs résultats étaient obtenus en utilisant BM comme méthode d'extraction des valeurs extrêmes et en utilisant MLE comme algorithme d'ajustement de la distribution des valeurs extrêmes. Les résultats du modèle sont présentés dans [Fig. 24]

```
                        Univariate Extreme Value Analysis
================================================================================================
                                      Source Data
------------------------------------------------------------------------------------------------
Data label:                      Loss      Size:                             1,645
Start:                    January 1980     End:                       December 1990
================================================================================================
                                     Extreme Values
------------------------------------------------------------------------------------------------
Count:                             11      Extraction method:                   BM
Type:                            high      Block size:          365 days 05:49:12
================================================================================================
                                        Model
------------------------------------------------------------------------------------------------
Model:                            MLE      Distribution:                  gumbel_r
Log-likelihood:               -60.369      AIC:                            126.238
------------------------------------------------------------------------------------------------
Free parameters:          loc=52.348       Fixed parameters: All parameters are free
                         scale=44.791
================================================================================================
```

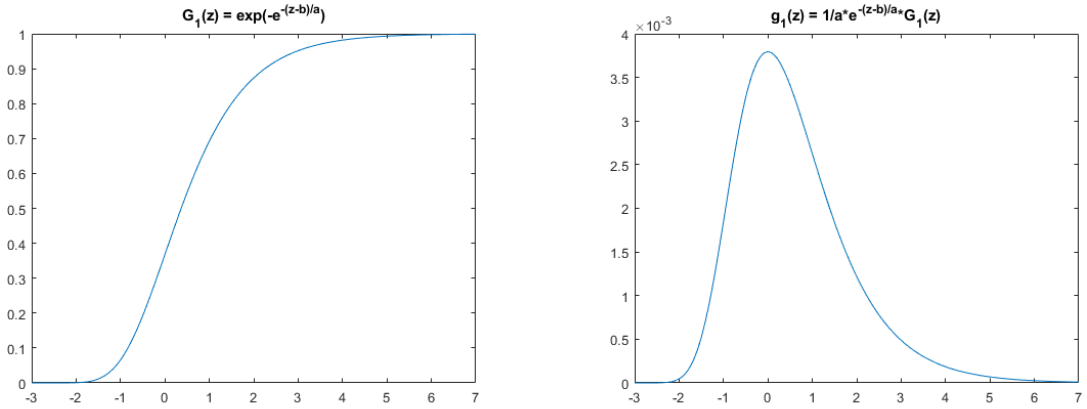
Figure 22: Résultats de la modélisation

Nous pouvons constater que le modèle ajusté correspond à la distribution de Gumbel.

- La distribution de Gumbel est l'une des distributions de probabilité continue utilisée pour décrire la distribution des événements extrêmes.

$$G_1(z) = exp\{-exp(-\frac{z-b}{a})\}, -\infty < z < +\infty$$

est appelée distribution de **Gumbel**, dont la densité est $g_1(z) = \frac{1}{a}exp(-\frac{z-b}{a})G_1(z)$



(a) Fonctions de distribution **(b)** Fonctions de densité

Figure 23: Distribution de Gumbel

- La distribution de Gumbel est largement utilisée en statistique des valeurs extrêmes, par exemple pour décrire la distribution des températures, des précipitations et des vitesses du vent extrêmes.

2.2.3 Présentation des résultats

Les résultats montrent que pour une période de régression donnée, par exemple 2, 5, 10, 25, 50, 100 ans, etc., la compensation maximale des incendies (valeur de retour) prédite par le modèle et la plage de valeurs extrêmes (valeur de retour) sont de 65,94 pour une période de régression de 2 ans, avec un intervalle de confiance de 95% compris entre 59,68 et 99,29.

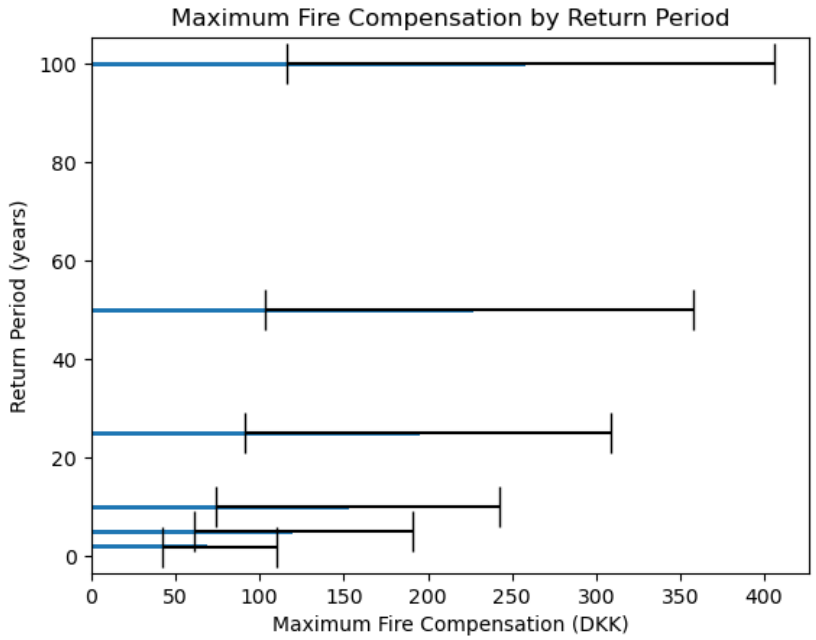


Figure 24: Indemnisation maximale en cas d'incendie par période de retour

- Pour la période de régression de 2 ans, la valeur extrême moyenne (valeur de retour) était de 65,94 avec un intervalle de confiance de 95% entre 59,68 et 99,29.
- Pour la période de régression de 5 ans, la valeur extrême moyenne était de 126,35, avec un intervalle de confiance à 95% allant de 96,37 à 286,05.
- Pour la période de régression de 10 ans, la valeur extrême moyenne était de 172,05, avec des intervalles de confiance à 95% allant de 124,12 à 427,32.
- Pour la période de régression de 25 ans, la valeur extrême moyenne était de 232,45, avec des intervalles de confiance à 95% allant de 160,81 à 614,08.
- Pour la période de régression de 50 ans, la valeur extrême moyenne est de 278,15, avec un intervalle de confiance à 95% compris entre 188,56 et 755,36.
- Pour la période de régression de 100 ans, la valeur extrême moyenne est de 323,84, avec des intervalles de confiance à 95% allant de 216,32 à 896,63.

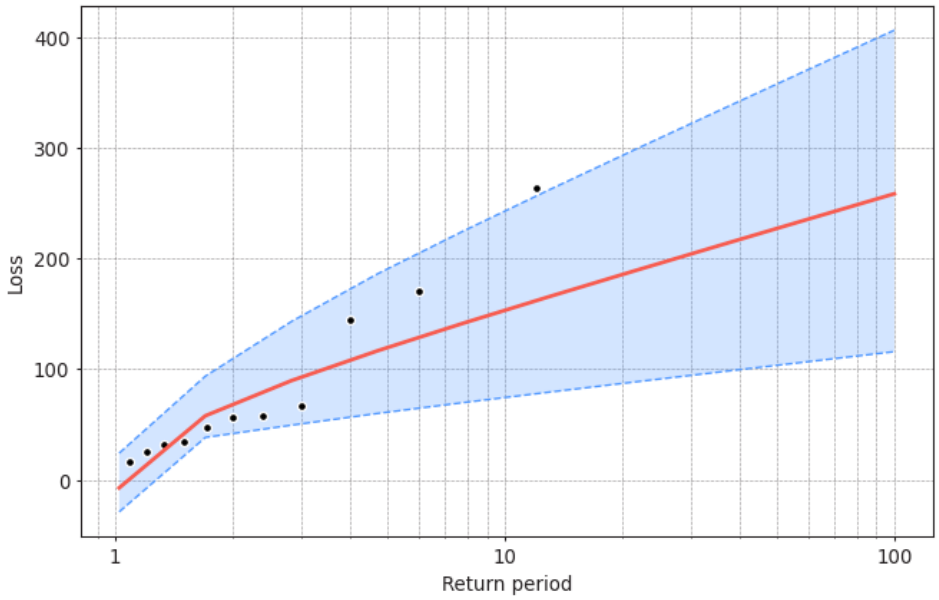


Figure 25: Diagramme de la valeur de retour du modèle

Nous pouvons voir que l'intervalle de confiance dans [Fig. 25] est suffisamment large pour inclure tous les points, ce qui signifie généralement que le seuil utilisé dans l'ajustement du modèle est relativement bas. Les méthodes suivantes peuvent être essayées pour améliorer la précision et réduire la largeur des intervalles de confiance.

- Augmenter la taille de l'échantillon : un plus grand nombre de données peut améliorer la précision de la statistique et donc réduire la largeur de l'intervalle de confiance. Envisager d'augmenter la taille de l'échantillon ou de relever le seuil des valeurs extrêmes afin d'accroître le nombre de valeurs extrêmes.

- Modification de la taille du bloc : la taille du bloc est un paramètre important dans le calcul des intervalles de confiance ; essayer d'ajuster la taille du bloc pour obtenir une estimation plus précise.
- Modification du seuil : le seuil des valeurs extrêmes est un facteur important dans l'ajustement du modèle et le calcul des intervalles de confiance ; essayez d'ajuster le seuil des valeurs extrêmes pour obtenir de meilleurs résultats.
- Essayer d'autres algorithmes : outre la MLE et la MCMC, d'autres algorithmes peuvent être utilisés pour l'analyse des valeurs extrêmes, tels que la modélisation hiérarchique bayésienne, qui peuvent être essayés pour obtenir des estimations plus précises.

Cependant, en raison des fonctionnalités limitées de la bibliothèque PyExtreme, nous conservons ce résultat, bien qu'il pourrait être amélioré.

2.3 Classification binaire dans les régions extrêmes

INTRODUCTION Nous allons travailler sur l'article "On Binary Classification in Extreme Regions" de Hamid Jalalzai, Stephan Clémençon et Anne Sabourin [6].

Précisons le cadre de l'article. L'objectif est de prédire une variable aléatoire Y qui dépend à priori d'un vecteur aléatoire X. Y est une variable aléatoire binaire qui prend la valeur 1 ou -1. (X, Y) définit ainsi une paire aléatoire définie sur un espace de probabilité donné avec pour distribution de probabilité jointe P qui est inconnue.
L'objectif est de construire un classifieur $g : \mathbb{R}^d \mapsto \{-1, 1\}$ à l'aide d'algorithmes classiques de Machine Learning (tel les Random Forests) qui s'entraîne à partir d'un jeu d'entraînement $D_n = \{(X_i, Y_i)_{i=1,\dots,n}\}$ et qui minimise théoriquement la probabilité d'erreur $L_p(g) = \mathbb{P}(Y \neq g(X))$. De manière empirique, on cherchera à trouver $\hat{g_n}$ qui minimise $\widehat{L_n}(g) = (1/n) \sum_{i=1}^{n} \mathbb{1}\{Y_i \neq g(X_i)\}$.
On fixe $t > 0$ arbitrairement qui représente le seuil à partir duquel on considère une valeur comme extrême.
Les X tels que $\|X\| > t$ sont ainsi rares et en minorité dans le jeu d'entraînement, ce qui implique que si l'on prédit mal Y pour X grand, cela n'aura que peu d'importance sur la probabilité d'erreur. Ainsi on a pas de garantie que la prédiction soit bonne pour de telles valeurs extrêmes. On va alors chercher un prédicteur g tel que $L_t(g) := L_{P_t}(g) = \mathbb{P}\{Y \neq g(X) \mid \|X\| > t\}$ soit minimal quand t tend vers $+\infty$.

CADRE PROBABILISTE Soit $\alpha > 0$ et X une variable aléatoire. X est dite régulièrement variable (regularly varying en anglais) avec un indice de queue α si : $\mathbb{P}\{X > tx \mid X > t\} \underset{t\to\infty}{\longrightarrow} x^{-\alpha}, \quad x > 1$. On montre que c'est le cas si $b : \mathbb{R}_+ \to \mathbb{R}_+^*$ avec b qui tend vers $+\infty$ telle que pour tout $x > 0$, $t\mathbb{P}\{X/b(t) > x\}$ tends vers une limite $h(x)$ quand $t \to \infty$.
On peut dans cette caractérisation choisir b définit par $b(t) = t^{1/\alpha}$ et $h(x) = cx^{-\alpha}$ pour un certain $c > 0$.
On se sert de se résultat pour étendre le cadre à queue épaisse au cas où $X = \left(X^{(1)}, \dots, X^{(d)}\right)$ à valeurs dans $\mathbb{R}_+^d$ est multivariée.
On dit alors que X varie régulièrement avec un indice de queue α si il existe une mesure positive de Radon non nulle μ sur l'espace $E = [0, \infty]^d \backslash \{0\}$ et une fonction $b(t) \to \infty$ telle que pour tout borélien $A \subset E$ vérifiant $0 \notin \partial A$, et $\mu(\partial A) = 0$, l'on ait

$$t\mathbb{P}\{X/b(t) \in A\} \underset{t\to\infty}{\longrightarrow} \mu(A)$$

On a alors μ qui vérifie la propriété d'homogénéité $\mu(tC) = t^{-\alpha}\mu(C)$ pour tout $t > 0$ et tout borélien $C \subset E$.
Cela incite à décomposition de μ en une composante radiale et une composante angulaire Φ. Pour tout $x = (x_1, \dots, x_d) \in \mathbb{R}_+^d$, on définit ainsi, où S est la sphère unité.

$$\begin{cases} R(x) = \|x\| \\ \Theta(x) = \left(\frac{x_1}{R(x)}, \dots, \frac{x_d}{R(x)}\right) \in S \end{cases}$$

CLASSIFIEURS ET RISQUE On définit le risque de classification pour un classifieur g par $L_t(g) := L_{P_t}(g) = \mathbb{P}\{Y \neq g(X) \mid \|X\| > t\}$.
On cherche à minimiser le risque dans les extrêmes donné par $L_\infty(g) = \limsup_{t\to\infty} L_t(g)$.
On note $L_t^* := \mathbb{E}[\min\{\eta(X), 1-\eta(X)\} \mid \|X\| > t], \quad t > 0$ où $\eta(X) = \mathbb{P}\{Y = 1 \mid X\} = L_t(g^*)$ et l'on a que $L_t(g) \geqslant L_t^*$ pour tout classifieur g.
Sous certaines hypothèses détaillées dans l'article, on a que L_t^* converge lorsque t tend vers $+\infty$, vers L_∞^*, qui vérifie $L_\infty^* = \inf_g L_\infty(g)$.
Le point central est que cet infimum est atteint et qu'un classifieur g le vérifiant ne dépend que de la composante angulaire de l'observation et non de la composante radiale, ce qui fait que l'on peut se ramener à un classifieur qui "travaille" sur la sphère unité.
Enfin l'article montre la pertinence de cette approche lorsque l'on travaille sur des données concrètes et donc avec des quantités empiriques. Etant donné n observations, un réel $\tau > 0$ fixé, $k = \lfloor n\tau \rfloor$, on considère le risque empirique dans les extrêmes d'un classifieur g

$$\widehat{L}_k(g) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\left\{Y_{(i)} \neq g\left(\Theta\left(X_{(i)}\right)\right)\right\}$$

où les $(X_{(i)})$, sont rangés dans l'ordre décroissant en norme (X_i est l'observation la plus extrême, et X_n, l'observation la moins extrême). On note $\widehat{g_k}$, le classifieur qui minimise ce risque.
On montre que lorsque n tend vers $+\infty$ et donc k également au vu de sa défniition, qu'on a la convergence en probabilité $L_\infty(\widehat{g}_k) \to L_\infty^*$. La suite de classifieurs $\widehat{g}_k$ ainsi définie est consistante dans les extrêmes ce qui justifie son utilisation.

APPLICATION Le but de cet article est de pouvoir prédire efficacement en fonction d'observations qui sont extrêmes. En effet, les classifieurs classiques utilisés en machine learing (Random Forests,...), ne sont pas entraînés à traiter de telles données extrêmes qui arrivent très rarement.
On adopte ici une approche série temporelle. Connaissant un certains nombres d'éléments du passé, on cherche à avoir une information binaire sur le présent.
On note $U = (U_i)$ la série temporelle sur laquelle on travaille. Ici, concrétement, en connaissant une observation extrême $X_i = (U_{i-1},...,U_{i-l})$, on cherche à avoir une information binaire sur U_i.
On dit que $(U_{i-1}, ..., U_{i-l})$ est extrême si la norme (que l'on choisit) du vecteur ainsi constitué est supérieure à un certain réel c.
Notre approche a été de tenter de prédire si U_i sera également une valeur extrême comparée aux observations extrêmes précédentes.
Formellement, on regarde si $|U_i| > d \cdot \|X_i\|$ avec d une constante choisie intelligemment. Si tel est le cas, on définit $Y_i = 1$, sinon, $Y_i = -1$.
On utilisera une approche apprentissage ultérieurement, pour prédire Y, d'où la constante d est choisie de manière à ce qu'il y ait un nombre similaire de valeurs de i telles que $Y_i = 1$ et $Y_i = -1$.
L'algorithme utilisé présenté dans l'article et adapté à notre cas est donné ci-après.

Nous avons testé cette approche et l'algorithme précédent sur des données S&P 500 à la cloture différenciées pour enlever la tendance, mais les résultats obtenus ne sont pas

Algorithm 3 Algorithme de classification dans les extrêmes

1: Entrée : Série temporelle $(U_1, ..., U_n)$, famille de classifieurs, nombre l qui donne le nombre d'éléments du passé que l'on regarde pour déterminer la classe
2: Sortie : Classifieur
3: Préparation des données : création d'une liste contenant des vecteurs à l éléments : au rang i, la liste contient le vecteur suivant $X_i = (U_i, ..., U_{i+l-1})$, et la classe associée Y_i
4: Choix de n élements au hasard pour former le jeu d'entraînement
5: (étape est optionnelle) Calcul de la fonction de répartition empirique $\widehat{F_l}$ pour chaque composante des vecteurs (que l'on multiplie par $n/(n-1)$, pour éviter d'obtenir un nombre infini), et création d'une nouvelle liste contenant les vecteurs $\hat{T}(X_i) = \left(1/\left(1 - \hat{F}_j\left(U_{i+j}\right)\right)\right)_{j=1,...,l}$
6: Caclul de la norme (que l'on choisit, dans notre cas norme L_1 ou L_2) de chaque élément de la liste précédente
7: Conservation des k plus grands $\hat{T}(X_i)$ en norme (où des valeurs dépassant un certain seuil)
8: Phase d'entraînement sur les données précédentes (choix du classifieur qui minimise l'erreur)
9: Renvoi du classifieur

probants. Le code produit sera donné en complément.
Les familles de classifieurs utilisées sont les mêmes que mentionnées dans l'article, les Random Forests, les k-NN, et nous avons également tenté d'utiliser des réseaux de neurones (avec une fonction d'activation sigmoid, où l'on attribue la classe 1 aux sorties positives et -1 aux sorties négatives).
L'approche de tester si le présent est encore "extrême" si on le compare aux observations précédentes n'est pas forcément pertinente dans le cas du S&P 500. Nous manquons de temps pour tester cette approche sur d'autres jeux de données ou pour réfléchir à une autre approche pour appliquer l'algorithme de l'article.
Une idée serait de lier des événements extrêmes sur le S&P 500 à un autre événement qui pourrait être corrélé, et ainsi de lier plusieurs sources de données entre elles. Par exemple, on pourrait en gardant des notations similaires à précédemment, regarder, si une évolution récente extrême (sur les deux ou trois derniers jours) du S&P 500 entraîne une évolution sur un autre marché (par exemple sur la paire EUR-GBP ou en cryptomonnaie), et voir si l'on arrive à trouver un classifieur qui va permettre de prédire l'arrivée d'un phénomène sur ces marchés lorsque l'on atteint un extrême sur le S&P 500. Un tel prédicteur peut être de dire si oui ou non on aura un extrême également sur ces marchés.

3 CONCLUSION

Ce projet nous aura permis de découvrir les bases de la théorie des Valeurs Extrêmes et d'en réaliser des applications directes. Nous n'avons pas couvert l'intégralité du livre de Stuart Coles, qui constituait notre ressource principale ; nous n'avons notamment pas abordé les extrêmes de séries non stationnaires ou encore les extrêmes multivariés. L'étude des valeurs extrêmes peut donner lieu à l'utilisation d'autres outils mathématiques utilisés de manière plus classiques, qui sous certaines hypothèses permettent de travailler sur ces valeurs extrêmes. Par exemple, l'utilsiation du machine learning peut être pertinente, comme nous avons pu le voir à travers l'étude de la classification binaire en régions extrêmes.

Les valeurs extrêmes peuvent apparaître dans toutes sortes de problèmes, en santé, en climatologie, en prédiction de phénomènes naturels, en finance, en économie, en industrie (défaillance de pièces) et la connaissance, même basique, de la théorie et de certains exemples classiques, pour les traiter pourra nous être très utile.

RÉFÉRENCES

[1] Fernando J Méndez, Melisa Menéndez, Alberto Luceño, and Inigo J Losada. Estimation of the long-term variability of extreme significant wave height using a time-dependent peak over threshold (pot) model. *Journal of Geophysical Research: Oceans*, 111(C7), 2006.

[2] Ana Ferreira and Laurens De Haan. On the block maxima method in extreme value theory: Pwm estimators. 2015.

[3] Scott D Grimshaw. Computing maximum likelihood estimates for the generalized pareto distribution. *Technometrics*, 35(2):185–191, 1993.

[4] Marcelo Hartmann and Ricardo S Ehlers. Bayesian inference for generalized extreme value distributions via hamiltonian monte carlo. *Communications in Statistics-Simulation and Computation*, 46(7):5285–5302, 2017.

[5] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[6] Stephan Clémençon Hamid Jalalzai and Anne Sabourin. On binary classification in extreme regions. *Neural Information Processing Systems*, 2018.