Statistiques des valeurs extrêmes dans le cas de lois discrètes

Anis Borchani

Statistiques des valeurs extrêmes dans le cas de lois discrètes

Research Center
ESSEC Working Paper 10009
December 2010

Anis Borchani

ESSEC
BUSINESS SCHOOL

Statistiques des valeurs extrêmes dans le cas de lois discrètes

Anis BORCHANI
Ecole Supérieure de la Statistique et de l'Analyse de l'Information, Tunis.

Résumé

Nous proposons une méthode basée sur la notion de quantiles extrêmes pour générer un système d'alertes pour la détection de clusters temporels d'extrêmes dans une série chronologique. A cette fin, nous développons deux approches, l'une utilisant une approximation du temps de retour d'un événement extrême, quelle que soit la nature des données, et l'autre basée sur la théorie classique des valeurs extrêmes après lissage des données discrètes en données continues.
Cette méthode permet ainsi de définir un système de surveillance et de prévision. Une illustration en est proposée dans le cadre de la prévision pour des applications en finance ainsi que pour la mise en place d'un système de surveillance en épidémiologie sur données réelles.

Abstract : We propose a method to generate a warning system for the early detection of time clusters in discrete time series. Two approaches are developed, one using an approximation of the return period of an extreme event, independently of the nature of the data, the other using an estimation of the return period via standard EVT tools after a smoothing of our discrete data into continuous ones.
This method allows us to define a surveillance and prediction system which is applied to finance and public health surveillance.

Mots-clé : applications en assurance et finance, clusters, détection d'événements extrêmes, épidémiologie, niveau de retour, quantiles extrêmes, temps de retour, Théorie des Valeurs Extrêmes, surveillance.
Keywords : applications in insurance and finance, clusters, epidemiology, Extreme Value Theory, extreme quantile, outbreak detection, return level, return period, surveillance.

1. Rapport de stage de fin d'étude effectué à l'ESSEC entre mars et mai 2008 sous la direction de Marie Kratz ; une version a été imprimée à l'ESSAI (Ecole Supérieure de la Statistique et de l'Analyse de l'Information)-Tunis- en Juin 2008.

Table des matières

1 Introduction générale

Ce travail s'inscrit dans le cadre d'un projet de recherche sur la statistique des valeurs extrêmes dans le cas de lois discrètes entre A. Guillou [2], M. Kratz [3] et Y. Le Strat [4] (pour les applications en épidémiologie) et en développe sa partie empirique.
Dans cette étude, nous nous intéressons, non pas au comportement en moyenne de données, mais à celui d'événements à valeurs beaucoup plus fortes (ou plus faibles) que celles observées autour de la moyenne, appelés valeurs extrêmes (V.E.).

Etudier l'occurrence d'événements extrêmes tels les catastrophes naturelles est de première importance pour les assureurs, ou les crises boursières pour les financiers ou encore, les épidémies pour un état (population comme représentants). Ce sont des événements rares aux conséquences désastreuses. C'est pourquoi est-il important de prévoir l'occurence de tels événements.

C'est ce que nous proposons dans ce travail, en estimant les temps de retour de quantiles extrêmes.
Rappelons ce qu'on appelle quantile extrême. Soit un n-échantillon $(X_1, \cdots, X_n)$ associé à une variable aléatoire (v.a.) parente X de loi F, le quantile noté x_{p_t} est défini par :

$$F(x_{p_t}) = p_t = 1 - 1/t$$

où p_t est définie comme l'ordre du quantile x_{p_t}.
Si t est tel que $p_t \longrightarrow 1$, un tel quantile est dit extrême et t représente le temps de retour associé.
Lorsque F est continue, l'estimation d'un quantile extrême fait appel à la théorie classique des valeurs extrêmes (T.V.E.).
Dans le cas discret, une méthode alternative serait de procéder par «approximation» des données discrètes en données continues afin d'utiliser les techniques classiques de la T.V.E, ou/et de mettre en place de nouvelles boîtes à outils adaptées au cas discret.

La première partie est une présentation générale des deux approches de la T.V.E. : la première repose sur les k plus grandes observations d'un échantillon et détermine les trois lois possibles des extrema (synthétisées sous le nom de loi *GEV*) et la seconde utilise les observations au delà d'un seuil déterministe, nous parlons alors de méthode *Peaks Over Thresholds* (P.O.T.).

2. Univ. Louis Pasteur, Strasbourg ; guillou@math.u-strasbg.fr
3. ESSEC, Cergy-Pontoise ; kratz@essec.fr
4. Institut de Veille Sanitaire, St Maurice ; y.lestrat@invs.sante.fr

Dans la section suivante, nous présentons un outil mis en place via une méthode qui s'applique à tout type de données, continues comme discrètes et qui fournit des bornes (inférieure et supérieure) de quantiles extrêmes. Nous choisissons d'approcher le quantile extrême par une valeur : la borne supérieure du niveau de retour b_t. Cette quantité pourrait être utilisée par les praticiens en la considérant comme une valeur critique à ne pas dépasser dans l'horizon de t unités de temps, t représentant le temps de retour associé.
Ainsi, nous proposons une approximation du temps de retour que nous pouvons utiliser tant dans le cadre de prévision pour des applications en finance que pour la mise en place d'un système de surveillance par exemple en épidémiologie.

Dans la troisième partie, nous revenons à la T.V.E. pour l'estimation des queues de distribution. Pour utiliser cette méthode, nous transformons par lissage les données discrètes en continues. Nous pouvons alors estimer les indices puis définir les lois des V.E pour déterminer les quantiles extrêmes. Il s'agira de choisir un type de lissage qui préserve l'information au sein de l'échantillon initial. Elle permet ainsi de définir un système de surveillance et de prévision.

Il existe un package sous *R* appelé surveillance, qui implémente les méthodes de détection temporelle d'événements inhabituels utilisées par les agences de surveillance américaines, anglaises et allemandes.
Une fois testé sur de nombreux jeux de données et s'il s'avère que ces méthodes donnent de bons résultats, nous pourrons demander par la suite à ce qu'elles soient implémentées dans ce package, c'est pourquoi nous avons écrit tout les programme en langage *R* figurant en annexe.

2 La théorie des valeurs extrêmes

Nous disposons de n observations $(x_1, \cdots, x_n)$ associées à un n-échantillon $(X_1, \cdots, X_n)$ de v.a. i.i.d, de fonction de répartition F continue $(F(x) = \mathbb{P}[X \leq x])$.
Nous notons Q la fonction quantile associée définie par

$$Q(p) = \inf\{x : F(x) \geq p\} \tag{1}$$

Soit $X_{1,n} \leq \cdots \leq X_{n,n}$ la statistique d'ordre relative à cet échantillon.
Nous nous intéressons au comportement des V.E. et à l'estimation des quantiles extrêmes. Tout d'abord, nous voulons savoir comment noter la présence de V.E. d'un point de vue descriptif.

En effet, de nombreuses méthodes algébriques, graphiques existent pour détecter les V.E. et étudier la fonction F.
Algébriquement, la détection se fait à partir de la distance relative à une unité au centre de la distribution.
Soient m le paramètre de tendance centrale et s le paramètre d'échelle.
Nous définissons alors la distance relative d'une unité au centre comme suit :

$$d_i = \frac{|x_i - m|}{s}$$

le couple (m, s) peut prendre plusieurs formes, par exemple (moyenne,variance), (médiane, *Median Absolute Deviation* ou MAD) et (médiane,intervalle interquartile) telle que

$$MAD = \text{median}_i\{|x_i - \text{median}_j(x_j)|\}$$

Nous pouvons aussi repérer les événements extrêmes à l'aide d'un graphique simple et facilement compréhensible, le *Boxplot*.
Mais la détermination d'un seuil ne repose pas sur la forme de la queue de la distribution, c'est-à-dire la distance d_i et la distance à l'extrémité de la boîte du *Boxplot* ne suffisent pas à déclarer une observation comme extrême.
Cependant, ces techniques ne sont pas capables de rapporter des informations qui peuvent nous aider à résoudre le problème de description des V.E. Donc, il est nécessaire de développer d'autres techniques dans le but de se focaliser sur les V.E. de quantile très élevé et d'une queue très fine. Ainsi, nous avons recours à la méthode probabiliste, connue dans la littérature par la théorie des valeurs extrêmes.
L'objectif de cette théorie est l'estimation de la probabilité de la queue de distribution, pour cela, on a recours à la fonction de répartition empirique définie par

$$\widehat{F}_n(x) = \frac{i}{n} \qquad si \qquad x \in [x_{i,n}, x_{i+1,n}) \tag{2}$$

telle que $x_{i,n}$ est la $i-me$ valeur de l'échantillon ordonné, ce qui nous mène à estimer la fonction empirique de quantile

$$\widehat{Q}_n(p) = \inf\{x : \widehat{F}_n(x) \geq p\} \tag{3}$$

2.1 La loi du maximum

La théorie des V.E. a été développée pour l'estimation de la probabilité d'occurrence d'événements extrêmes. Elle permet d'extrapoler le comportement de la queue de distribution des données à partir des plus grandes observations.
La fonction de répartition de la loi de $X_{n,n} = \max(X_1, \ldots, X_n)$ est la puissance n-ième F^n de F et pour tout $x < \sup\{x : F(x) < 1\}$, $F^n(x) \to 0$ quand $n \to \infty$. Le théorème de Gnedenko (1943) sur la loi des valeurs extrêmes est pour le maximum de n observations, un analogue du théorème limite central pour la moyenne. Il introduit alors des suites numériques $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ de normalisation permettant d'obtenir une loi limite pour $a_n(X_{n,n} - b_n)$. Ce théorème fait appel aux notions de fonction à variation régulière et normalisée, nous rappellons tout d'abord la définition de telles fonctions.

Définition 1 *Une fonction G est dite à variation régulière (à l'infinie) d'indice $\rho \in \mathbb{R}$ si G est positive à l'infini (ie : s'il existe A tel que pour tout $x \geq A$, $G(x) > 0$) et si pour tout $t > 0$*

$$\lim_{x\to\infty} \frac{G(tx)}{G(x)} = t^\rho \tag{4}$$

Dans le cas particulier où $\rho = 0$, on dit que G est une fonction à variation lente.

En remarquant que si G est à variation régulière d'indice ρ alors $\frac{G(x)}{x^\rho}$ est à variation lente, il est facile de montrer qu'une fonction à variation régulière d'indice ρ peut toujours s'écrire sous la forme $x^\rho \ell(x)$, où ℓ est à variation lente.

Théorème 1 *[16]*

ℓ est une fonction à variation lente si et seulement si pour tout $x > 0$

$$\ell(x) = c(x)\exp\{\int_1^x t^{-1}\epsilon(t)\} \tag{5}$$

où c et ϵ sont des fonctions positives telles que

$$\lim_{x\to\infty} c(x) = c \in]0, +\infty[\quad et \quad \lim_{t\to\infty} \epsilon(t) = 0$$

Remarque 1 *Si la fonction c est constante, on dit que ℓ est normalisée.*

Remarque 2 *Soit G une fonction à variation régulière d'indice ρ.*
En utilisant le fait que $G(x) = x^{\rho}\ell(x)$*, on déduit facilement que pour tout* $x > 0$

$$G(x) = c(x)\exp\left\{\int_1^x t^{-1}\rho(t)\right\} \tag{6}$$

où c et ρ sont des fonctions positives telles que

$$\lim_{x\to\infty} c(x) = c \in]0, +\infty[\quad et \quad \lim_{t\to\infty}\rho(t) = \rho$$

Théorème 2 ***Gnedenko****[5]*

Sous certaines conditions de régularité sur la fonction de répartition F, il existe $\xi \in \mathbb{R}$ *et deux suites normalisantes réelles* $(a_n)_{n\geq 1}$ *et* $(b_n)_{n\geq 1}$ *tels que pour tout* $x \in \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{X_{n,n} - b_n}{a_n} \leq x\right] = H_{\xi}(x), \tag{7}$$

avec, si $\xi > 0$

$$H_{\xi}(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \exp\left[-(x)^{-\frac{1}{\xi}}\right] & \text{si } x > 0 \end{cases} \tag{8}$$

si $\xi < 0$

$$H_{\xi}(x) = \begin{cases} \exp\left[-(-x)^{-\frac{1}{\xi}}\right] & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases} \tag{9}$$

si $\xi = 0$

$$H_0(x) = \exp\left[-\exp(-x)\right] \quad pour\ tout\ x \in \mathbb{R} \tag{10}$$

Remarque 3
- *La fonction de répartition* H_{ξ} *est appelée loi des valeurs extrêmes (que l'on note EVD pour Extreme Value Distribution). Nous disons ainsi que F apppartient au domaine d'attraction de* H_{ξ}*.*
- *Ces lois sont indexées par un paramètre* ξ *appelé indice de valeur extrême et selon son signe, nous distinguons trois domaines d'attraction. On dit que la fonction de répartition F est dans le domaine d'attraction de Fréchet, de Gumbel ou de Weibull selon que* $\xi > 0$*,* $\xi = 0$ *et* $\xi < 0$*.*

Voici ci-dessus le graphe de ces distributions et le graphe de leurs densités pour certaines valeurs de ξ.

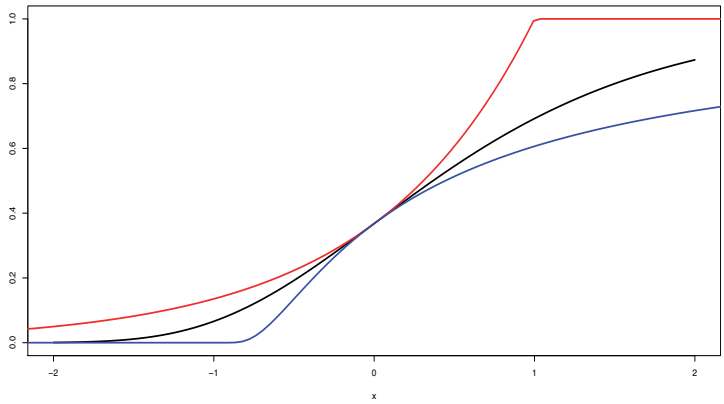

FIGURE 1 – La fonction de répartition de la loi *EVD*. Les courbes en rouge, en noir foncé et en bleu correspondent respectivement aux lois *EVD* dans le domaine d'attraction de *Weibull*, de *Gumbel* et de *Fréchet* d'indices $\xi = -1$, $\xi = 0$ et $\xi = 1$.

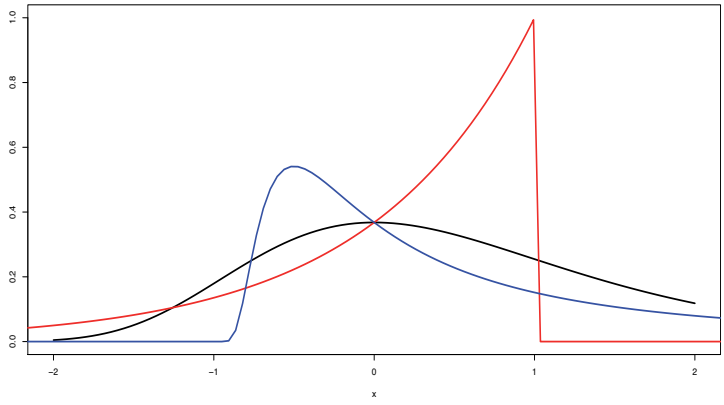

FIGURE 2 – La densité de la distribution *EVD*. Les courbes en rouge, en noir foncé et en bleu correspondent respectivement aux densités de la loi des valeurs extrêmes d'indices $\xi = -1$, $\xi = 0$ et $\xi = 1$.

Remarque 4 *La distribution GEV (Generalized Extreme Value) est une paramétrisation de ces trois lois, en une seule formule :*

$$H_\xi(x) = \exp\left[-(1+\xi x)^{-\frac{1}{\xi}}\right] \quad , \quad \forall \xi \in \mathbb{R} \tag{11}$$

définie pour tout $x > -1/\xi$

C'est seulement par ces trois lois que nous pouvons exprimer la loi limite du maximum normalisé. C'est dans ce sens que ce théorème nous fournit une analogie avec le théorème limite central.
Nous allons donner par la suite des conditions sur la fonction de répartition F pour qu'elle appartienne à l'un des trois domaines d'attraction. Dans la suite, on note

$$x_F = \sup \{x : F(x) < 1\}$$

le point terminal de F et

$$F^{\leftarrow}(y) = \inf \{x \in \mathbb{R} : F(x) \geq y\}$$

où $F^{\leftarrow}$ désigne l'inverse généralisée de F.

Théorème 3 *1. F appartient au domaine d'attraction de Fréchet avec un indice de valeur extrême $\xi > 0$ si et seulement si $x_F = +\infty$ et $1 - F$ est une fonction à variation régulière d'indice $-\frac{1}{\xi}$. Dans ce cas, un choix possible pour les suites $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ est :*

$$a_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right) \quad et \quad b_n = 0 \quad , \quad \forall n > 0 \tag{12}$$

2. F appartient au domaine d'attraction de Weibull avec un indice de valeur extrême $\xi < 0$ si et seulement si $x_F < +\infty$ et $1 - F^*$ est une fonction à variations régulières d'indice $\frac{1}{\xi}$ avec

$$F^*(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ F\left(x_F - x\right)^{-1} & \text{si } x > 0 \end{cases}$$

Dans ce cas, un choix possible pour les suites $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ est :

$$a_n = x_F - F^{\leftarrow}\left(1 - \frac{1}{n}\right) \quad et \quad b_n = x_F \tag{13}$$

Définition 2 *Soit F une fonction de répartition de point terminal x_F (fini ou infini). S'il existe $z < x_F$ tel que*

$$1 - F(x) = c \exp\left\{-\int_z^x \frac{1}{a(t)} dt\right\} \tag{14}$$

*où $c > 0$ et a est une fonction positive absolument continue de densité a' vérifiant $\lim_{x \uparrow x_F} a'(x) = 0$, alors F est une fonction de **Von-Mises** et a est sa fonction auxiliaire.*

Corollaire 1 – *F appartient au domaine d'attraction de Fréchet avec un indice de valeur extrême $\xi > 0$ si et seulement si*

$$x_F = +\infty \quad et \quad 1 - F(x) = x^{-\frac{1}{\xi}} \ell(x) \tag{15}$$

où ℓ est une fonction à variation lente.

– *F appartient au domaine d'attraction de Weibull avec un indice de valeurs extrêmes $\xi > 0$ si et seulement si*

$$x_F < +\infty \, et \, 1 - F(x) = (x_F - x)^{-\frac{1}{\xi}} \ell\left[(x_F - x)^{-1}\right]. \tag{16}$$

où ℓ est une fonction à variation lente.

– *F appartient au domaine d'attraction de Gumbel si et seulement s'il existe une fonction de Von-Mises F^* telle que pour $z < x < x_F$*

$$1 - F(x) = c(x)\left[1 - F^*(x)\right] = c(x) \exp\left\{-\int_z^x \frac{1}{a(t)} dt\right\} \tag{17}$$

où $c(x) \to c > 0$ lorsque $x \to x_F$

2.2 La loi des excès

L'approche basée sur les distributions *GEV* peut être réductrice du fait que l'utilisation d'un seul maxima conduit à une perte d'information continue dans les autres grandes valeurs de l'échantillon. La solution est de considérer plusieurs grandes valeurs au lieu de la plus grande.
La nouvelle approche de la théorie des valeurs extrêmes appelée *POT* consiste à utiliser les observations qui dépassent un certain seuil, plus particulièrement les différences entre ces observations et le seuil, appelées *excès*.
Il est clair que cette méthode nécessite la détermination d'un seuil ni trop faible pour ne pas prendre en considération des valeurs non extrêmes, ni trop élevé pour avoir suffisamment d'observations. Notons le seuil par u.

Définition 3 *Soit X une variable aléatoire de fonction de répartition F et de point terminal x_F. Pour tout $u < x_F$, la fonction*

$$F_u(x) = \mathbb{P}\left[X - u \leq x | X > u\right], \; x \geq 0 \tag{18}$$

est appelée fonction de répartition des excès au dessus du seuil u.

Remarque 5 *Par définition des probabilités conditionnelles, F_u peut être également définie par*

$$F_u(x) = \frac{F(u+x) - F(u)}{1 - F(u)} \quad si \quad x \geq 0 \quad et \quad 0 \quad sinon. \tag{19}$$

Nottons $Y = X - u$ pour $X > u$ et pour n v.a. observées $X_1, \ldots, X_n$, nous pouvons écrire $Y_j = X_i - u$ telle que i est l'indice du $j - me$ excès et $j = 1, \ldots, N_u$.
Le *Mean Excess Plot* appelé aussi le *Mean Residual life Plot* est un outil spécifique pour retenir le seuil performant.
De plus, nous approchons la loi des excès $(Y_1, \ldots, Y_{N_u})$ par une loi de Pareto Généralisée $GPD_{\xi,\sigma}$ (*Generalized Pareto Distribution*) que nous présentons ci-dessous.

2.2.1 La loi de Pareto Généralisée

Le théorème de Pickands est très utile lorsqu'on travaille avec des observations qui dépassent un seuil fixé puisqu'il assure que la loi des excès peut être approchée par une loi de Pareto généralisée.

Définition 4 *Soient $\sigma(u)$ une fonction strictement positive et $\xi \in \mathbb{R}$. La loi de Pareto généralisée a pour fonction de répartition $G_{\xi,\sigma}$:*

$$G_{\xi,\sigma(u)}(y) = \begin{cases} 1 - \left(1 + \xi \frac{y}{\sigma(u)}\right)^{-\frac{1}{\xi}} & \text{si} \quad \xi \neq 0 \\ 1 - \exp\left(-\frac{y}{\sigma(u)}\right) & \text{si} \quad \xi = 0 \end{cases} \tag{20}$$

où $y \geq 0$ si $\xi \geq 0$ et $0 \leq y \leq -\frac{\sigma(u)}{\xi}$ si $\xi < 0$.

Théorème 4 ***Pickands**[14]*
Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes (Fréchet, Gumbel ou Weibull), alors il existe une fonction $\sigma(u)$ strictement positive et un réel ξ tels que

$$\lim_{u \uparrow x_F} \sup_{0 \leq y \leq x_F - u} \left| F_u(y) - G_{\xi,\sigma(u)}(y) \right| = 0 \tag{21}$$

où $G_{\xi,\sigma(u)}$ est la fonction de répartition de la loi de Pareto Généralisée et F_u est la fonction de répartition des excès au delà du seuil u.

Ainsi, pour u grand, la loi des excès est approchée par une loi Pareto généralisée.

$$F_u \approx G_{\xi,\sigma(u)} \tag{22}$$

Voici une représentation de la $GPD_{\xi,\sigma}$ pour certaines valeurs de ξ.

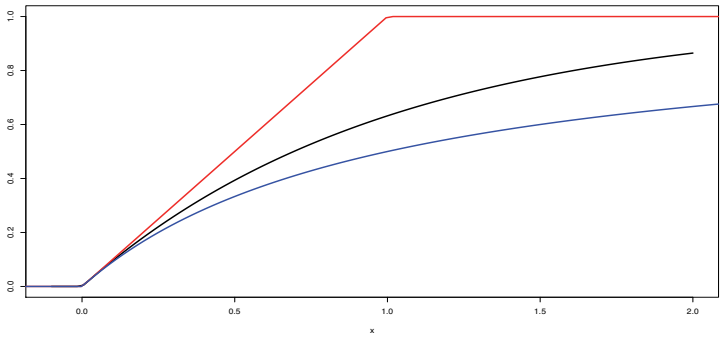

FIGURE 3 – La fonction de répartition de la loi $GPD_{\xi,1}$.

Les courbes rouge, noire foncée et bleue correspondent respectivement aux lois de Pareto généralisée d'indices $\xi = -1$, $\xi = 0$ et $\xi = 1$, avec $\sigma = 1$ fixe.
Nous remarquons que plus les valeurs extrêmes sont extrêmes, plus le paramètre ξ est élevé ; ça se voit très bien dans le graphe de la fonction de répartition.

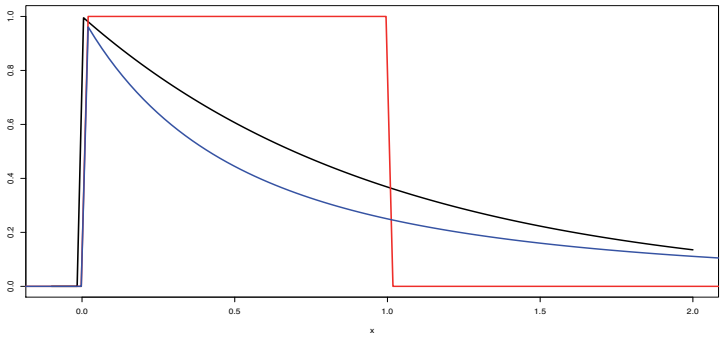

FIGURE 4 – La densité de la loi $GPD_{\xi,1}$.

2.2.2 Le Mean Excess Plot

Le *Mean Excess Plot* est le graphe des points $(u, e(u))$ où $e(u)$ est la moyenne des excès au delà du seuil u, définie par :

$$e(u) = \mathbb{E}\left(Y | X > u\right) \tag{23}$$

Elle correspond à une fonction de u qui s'exprime à l'aide de $1 - F$. Nous supposons que pour ce modèle proposé, $E\left(X\right) < \infty$.
En pratique, la fonction des moyennes des excès e est estimée par $\hat{e}_n$:

$$\hat{e}_n\left(u\right) = \frac{\sum_{i=1}^{n} x_i I_{(u,\infty)}\left(x_i\right)}{\sum_{i=1}^{n} I_{(u,\infty)}\left(x_i\right)} - u$$

telle que $I_{(u,\infty)}$ est égale à 1 si $x_i > u$ et 0 sinon, et nous choisissons la $(k+1) - me$ grande observation :$u = x_{n-k,n}$ comme seuil, d'où l'estimation

$$\hat{e}_n\left(x_{n-k,n}\right) = \frac{1}{k}\sum_{j=1}^{k} x_{n-j+1,n} - x_{n-k,n} \tag{24}$$

Proposition 1 – *Si* $(Y_1, \ldots, Y_{N_u})$ *suivent une loi* $GPD_{\xi,\sigma}$, *alors pour* $\xi < 1$,

$$\mathbb{E}\left(Y|X > u\right) = \frac{\xi}{1-\xi}u + \frac{\sigma}{1-\xi} \tag{25}$$

Dans ce cas, le seuil à retenir u, est celui pour lequel la moyenne des excès est approximativement linéaire.

– *Plus les queues de la distribution sont épaisses, plus la fonction* $e(u)$ *tend rapidement vers l'infini.*

En effet, la fonction moyenne des excès relative à une distribution de loi *Pareto* de paramètre $\alpha > 0$, appartenant au domaine d'attraction de Fréchet d'indice des valeurs extrêmes $\xi = \frac{1}{\alpha}$, s'écrit comme suit : $e(u) = \frac{k+u}{\alpha-1}$
Alors que pour une distribution de queue moins épaisse, par exemple la distribution Weibull de paramètres $\lambda > 0$ et $\tau > 0$, appartenant au domaine d'attraction de Gumbel, la fonction moyenne des excès relatif à cette loi s'écrit comme suit : $e(u) = \frac{u^{1-\tau}}{\lambda\tau}$.

2.3 Estimation des indices extrêmes

Nous avons vu que la loi asymptotique du maximum normalisé de v.a. continues est de trois types possibles (Fréchet, Weibull et Gumbel), cette loi étant indexée par un paramètre noté ξ. De même, la loi asymptotique des excès au delà d'un seuil donné, est une loi de Pareto généralisée indexée par le paramètre ξ et un deuxième paramètre σ. Ces paramètres, appelés respectivements indice de V.E. et paramètre d'échelle, apportent une information sur la forme de la queue de distribution de F. Notamment, selon que $\xi > 0$, $\xi < 0$ ou $\xi = 0$. C'est pourquoi est-il nécessaire d'estimer ces indices. De nombreux estimateurs ont été proposés dans la littérature.

2.3.1 Estimation de queue de distribution sous le modèle de type *Pareto*

Dans cette section, nous considèrons l'estimateur de l'indice extrême ξ dans le cas d'une loi de Pareto, de distribution F telle que :

$$1 - F(x) = x^{-\frac{1}{\xi}}\ell_F(x) \qquad avec \qquad \xi > 0. \tag{26}$$

et

$$Q\left(1 - \frac{1}{x}\right) = U(x) = x^{\xi}\ell_U(x) \tag{27}$$

où ℓ_F et ℓ_U sont deux fonctions à variation lente à l'infini et U est la fonction définie par

$$U(x) = \inf\left\{y : F(y) \geq 1 - \frac{1}{x}\right\}$$

Proposition 2 *Pour toute fonction à variation lente ℓ, on a :*

$$\lim_{x\to\infty} \frac{\log[\ell(x)]}{\log(x)} = 0$$

Remarque 6 *Nous remarquons que lorsque $x \to \infty$*

$$\log U(x) = \xi \log(x) + \log(\ell_U(x)) \sim \xi \log(x)$$

En remplaçant la fonction U par sa version empirique

$$\hat{U}_n(x) = \hat{Q}_n\left(1 - \frac{1}{x}\right),$$

nous obtenons

$$\log\left(\hat{U}_n\left(\frac{n}{k}\right)\right) \sim \xi \log\left(\frac{n}{k}\right)$$

En remarquant que $\hat{U}_n\left(\frac{n+1}{j}\right) = X_{n-j+1,n}$, on obtient finalement l'équivalence suivante :

$$\log X_{n-j+1,n} \sim \xi \log\left(\frac{n+1}{j}\right) \; quand \left(\frac{n+1}{j}\right) \to \infty \tag{28}$$

a) Le Pareto Quantile Plot

C'est le graphe de $\left(\log\left(\frac{n+1}{j}\right), \log X_{n-j+1,n}\right)$. C'est une représentation très utile pour visualiser graphiquement si les observations sont distribuées selon une loi du domaine de Fréchet. Dans ce domaine, le *Pareto quantile plot* serait approximativement linéaire avec une pente ξ, pour les petites valeurs de j, c'est-à-dire les points extrêmes. Deux approches sont possible pour construire un tel estimateur :
soit en forçant la droite à passer par le point $\left(\log\left(\frac{n+1}{k}\right), \log X_{n-k+1,n}\right)$, soit par la méthode des moindres carrés.
Dans le premier cas, Csörgo *et al.* (1985) [3] ont proposé les estimateurs à noyau $K_{k,n}$ définis de la façon suivante :

$$K_{k,n} = \frac{\sum_{j=1}^{k} \frac{j}{k} K\left(\frac{j}{k}\right) (\log X_{n-j+1,n} - \log X_{n-j,n})}{\sum_{j=1}^{k} \frac{1}{k} K\left(\frac{j}{k}\right)} \tag{29}$$

où K représente un noyau d'intégrale égale à 1.
Dans le second cas, en utilisant une approche par moindres carrés, Schultze et Steinebach (1996) [3] et Kratz et Resnick (1996) [12] ont introduit l'estimateur $Zipf$ défini par

$$Z_{k,n} = \frac{\sum_{j=1}^{k} \log \frac{k+1}{j} \log X_{n-j+1,n} - \frac{1}{k} \sum_{j=1}^{k} \log \frac{k+1}{j} \sum_{j=1}^{k} \log X_{n-j+1,n}}{\sum_{j=1}^{k} \log^2 \frac{k+1}{j} - \frac{1}{k} \left(\sum_{j=1}^{k} \log \frac{k+1}{j} \right)^2} \tag{30}$$

b) L'estimateur de ***Hill***
La définition de queue de distribution de type *Pareto* peut être écrite comme

$$\frac{1 - F(ux)}{1 - F(u)} \to x^{-\frac{1}{\xi}} \quad quand \quad u \to \infty \quad pour \quad tout\ x > 1$$

En effet, cette définition découle du fait que

$$\mathbb{P}\left(\frac{X}{u} > x | X > u \right) \sim x^{-\frac{1}{\xi}} \quad pour \quad u \quad trs \quad grand \quad et \quad x > 1$$

Soit $Z_j = \frac{X_i}{u}$, pour $X_i > u$ tel que i est l'indice du $j - me$ excès et $j = 1 \cdots N_u$.
Ainsi, nous écrivons la log vraisemblance, $\log L$ selon

$$\log L(Z_1 \cdots Z_{N_u}) = -N_u \log \xi - \left(1 + \frac{1}{\xi} \right) \sum_{j=1}^{N_u} \log Z_j$$

Alors,

$$\frac{d \log L}{d\xi} = -\frac{N_u}{\xi} + \frac{1}{\xi^2} \sum_{j=1}^{N_u} \log Z_j = 0 \quad \Leftrightarrow \quad \xi = \frac{1}{N_u} \sum_{j=1}^{N_u} \log Z_j \tag{31}$$

Choisissons pour le niveau d'excès u la valeur de la statistique d'ordre $X_{n-k,n}$ et remplaçant N_u par k, on obtient l'estimateur de **Hill** (1975) défini par :

$$\hat{\xi}_{k,n}^{Hill} = \frac{1}{k} \sum_{j=1}^{k} \log X_{n-j+1,n} - \log X_{n-k,n} \tag{32}$$

Nous pouvons retrouver cet estimateur à l'aide de la moyenne des excès. En effet, sous le modèle Pareto, et en faisant une transformation logarithmique nous aboutissons à l'équation :

$$\mathbb{E}(\log X - \log x | X > x) = \int_x^{\infty} \frac{1 - F(t)}{1 - F(x)} \frac{dt}{t} \to \xi \ quand\ x \to \infty.$$

Remplacons la distribution F par sa version empirique $\hat{F}_n$ et x par $X_{n-k,n} \to \infty$.
Ainsi, nous retrouvons

$$\hat{\xi}_{k,n}^{Hill} = \int_{X_{n-k,n}}^{\infty} \frac{1 - \hat{F}_n(t)}{1 - \hat{F}_n(X_{n-k,n})} \frac{dt}{t} \tag{33}$$

2.3.2 Estimation de queue de distribution pour les différents domaines d'attractions

Le cas général $\xi \in \mathbb{R}$ a été moins étudié. Nous recensons dans la littérature trois méthodes permettant d'estimer $\xi \in \mathbb{R}$. La première est proposée par **Pickands** (1975) [14], la deuxième consiste à utiliser les k plus grandes observations où k est un entier positif strictement inférieur à n. Comme exemple d'estimateurs basés sur cette méthode, on cite l'estimateur des moments proposé par Dekkers, Einmahl et de Haan (1989) [4] et enfin la méthode basée sur le *Generalized Quantile Plot* proposée par Beirlant *et al.* (1996) [1].
L'estimateur de **Pickands** est de la forme :

$$\hat{\xi}_{k,n}^{Pickands} = \frac{1}{\log 2} \log \left(\frac{X_{n-\lceil \frac{k}{4} \rceil + 1,n} - X_{n-\lceil \frac{k}{2} \rceil + 1,n}}{X_{n-\lceil \frac{k}{2} \rceil + 1,n} - X_{n-k+1,n}} \right) \tag{34}$$

L'estimateur des moments a été introduit comme une généralisation de l'estimateur de **Hill** et il est défini par

$$\hat{\xi}_{k,n}^{Moment} = \hat{\xi}_{k,n}^{Hill} + 1 - \frac{1}{2} \left(1 - \frac{(\hat{\xi}_{k,n}^{Hill})^2}{S_{k,n}} \right)^{-1} \tag{35}$$

avec

$$S_{k,n} = \frac{1}{k} \sum_{j=1}^{k} \left(\log X_{n-j+1,n} - \log X_{n-k,n} \right)^2$$

Une autre approche permettant d'éviter le choix a priori du domaine du domaine d'attraction consiste à utiliser un *quantile plot* généralisé défini comme le graphe $\left(\log \frac{n+1}{j+1}, \log \hat{\xi}_{j,n}^{UH} \right)$ avec

$$\hat{\xi}_{k,n}^{UH} = X_{n-k,n} \qquad \hat{\xi}_{k,n}^{Hill} \tag{36}$$

On obtient alors une courbe qui ressemble à l'une des trois ci-dessus. Si la courbe ressemble à la courbe bleue (resp. noire, rouge) alors l'indice est positif (resp. nul, négatif)

2.3.3 Estimation de queue de distribution sous le modèle *GPD*

La loi *GPD* est indexée par deux paramètres ξ et σ que nous pouvons estimer par deux méthodes : la méthode de maximum de vraisemblance et la méthode des moments.
Si Y_j désigne la $j - me$ variable d'excès au delà du seuil u, $j = 1, \ldots, N_u$, dans le cas où $u = X_{n-k,n}$, la statistique d'ordre des excès est de la forme $Y_{j,k} = X_{n-k+j} - X_{n-k,n}$ avec $j = 1, \cdots, N_u$.

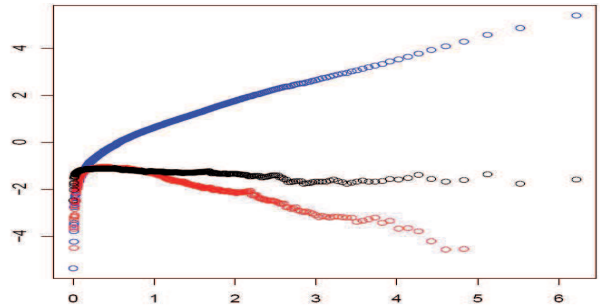

FIGURE 5 – Le quantile plot généralisé $\left(\log \frac{n+1}{j+1}, \log \hat{\xi}_{j,n}^{UH}\right)$.

En posant $\tau = \frac{\xi}{\sigma}$, pour $1 + \xi \frac{Y_i}{\sigma} > 0$ et $i = 1 \cdots N_u$ nous pouvons écrire la log-vraisemblance selon :

$$\log L(\tau, \xi) = -N_u \log \xi + N_u \log \tau - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{N_u} \log (1 + \tau Y_i).$$

Les estimateurs $\left(\hat{\tau}^{MV}, \hat{\xi}^{MV}\right)$ satisfont alors (37) et (38) :

$$\frac{1}{\hat{\tau}^{MV}} - \left(\frac{1}{\hat{\xi}^{MV}} + 1\right) \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{Y_i}{1 + \hat{\tau}^{MV} Y_i} = 0 \tag{37}$$

et

$$\hat{\xi}^{MV} = \frac{1}{N_u} \sum_{i=1}^{N_u} \log \left(1 + \hat{\tau}^{MV} Y_i\right) \tag{38}$$

L'estimation de ces indices est également possible via la méthode des moments, elle a été introduite par Hosking et Wallis (1987) [9].

Proposition 3 *Le r-ème moment de la distribution GPD existe si* $r < \frac{1}{\xi}$.

Sous cette condition, la moyenne et la variance de la *GPD* sont respectivement :

$$E(Y) = \frac{\sigma}{1 - \xi}$$

et

$$Var(Y) = \frac{\sigma^2}{(1 - \xi)^2 (1 - 2\xi)}$$

En remplaçant $E(Y)$ par

$$\overline{Y} = \frac{\sum_{i=1}^{N_u} Y_i}{N_u}$$

et $Var\,(Y)$ par

$$S_Y^2 = \frac{1}{N_u - 1} \sum_{i=1}^{N_u} \left(Y_i - \overline{Y}\right)^2$$

Nous déduisons ainsi les estimateurs, $\widetilde{\xi}$ et $\widetilde{\sigma}$ des paramètre d'une loi *GPD* par la méthode des moments :

$$\widetilde{\xi} = \frac{1}{2}\left(1 - \frac{\overline{Y}^2}{S_Y^2}\right) \tag{39}$$

$$\widetilde{\sigma} = \overline{Y}\left(\frac{1}{2} + \frac{\overline{Y}^2}{S_Y^2}\right) \tag{40}$$

2.4 Estimation de quantiles extrêmes et de périodes de retour

Le choix de la loi pour les valeurs extrêmes et l'estimation de la fonction de répartition en déterminant l'indice de queue de distribution ne sont souvent qu'un objectif intermédiaire, l'objectif réel étant plutôt l'estimation d'un quantile extrême ou d'un niveau de retour ainsi que l'estimation d'une période de retour.

2.4.1 Estimation de quantile extrême ou de niveau de retour

Nous estimerons le quantile d'ordre $p_t = 1 - 1/t$ de F, c'est à dire le nombre x_{p_t} défini par $1 - F\left(x_{p_t}\right) = 1/t$. Si t est grand, un tel quantile est dit quantile extrême.
Nous définissons aussi le niveau de retour comme la valeur x_t telle que nous espérons détecter en moyenne un seul dépassement de cette quantité au bout de t périodes, c'est à dire

$$\begin{aligned} \mathbb{E}\left(\sum_{i=1}^{t} \mathbf{1}_{(X_i > x_t)}\right) = 1 \quad &\Leftrightarrow \quad \mathbb{P}\left[X_i > x_t\right] = 1/t \quad avec \quad i = 1, \ldots, t \\ &\Leftrightarrow \quad 1 - F\left(x_t\right) = 1/t \end{aligned} \tag{41}$$

où **1** représente la fonction indicatrice.
Nous remarquons que l'estimation d'un niveau de retour d'ordre t revient à l'estimation d'un quantile extrême d'ordre $p_t = 1 - 1/t$.
Comme il est situé au dessus de l'observation maximale avec une probabilité qui tend vers 1 lorsque $t \to \infty$, on ne peut pas, comme pour l'estimation de quantiles classiques, inverser tout simplement la fonction de répartition empirique.
Il faut donc, à l'aide des plus grandes observations, estimer la fonction de répartition au-delà de l'observation maximale.
La première méthode utilise un résultat donnant l'expression de la loi asymptotique

du maximum d'un échantillon. Cette loi limite est la loi des valeurs extrêmes*(GEV)*. Nous estimons alors le quantile extrême en inversant la fonction de répartition de la loi des valeurs extrêmes et en estimant les paramètres de cette loi. L'estimateur de quantile extrême obtenu par cette méthode s'écrit sous la forme :

$$\hat{x}_{p_t}^{GEV} = \begin{cases} \hat{\mu} - \frac{\hat{\gamma}}{\hat{\xi}} \left\{ 1 - \left[-\log\left(p_t\right) \right]^{-\hat{\xi}} \right\} & \text{si } \xi \neq 0 \\ \hat{\mu} - \hat{\gamma} \log\left[-\log\left(p_t\right) \right] & \text{si } \xi = 0 \end{cases} \tag{42}$$

où $\hat{\xi}$, $\hat{\mu}$ et $\hat{\gamma}$ sont les estimateurs des paramètres de la loi des valeurs extrêmes H_ξ

$$H_\xi\left(x\right) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\gamma} \right) \right]^{-\frac{1}{\xi}} \right\}, \quad \forall \xi \in \mathbb{R} \quad , \quad \forall \mu \in \mathbb{R} \quad et \quad \forall \gamma > 0 \tag{43}$$

définie pour tout $x > \mu - \gamma/\xi$
La deuxième méthode, appelée méthode des excès ou *P.O.T.*, assure que la loi des observations qui dépassent un seuil u peut être approchée, pour u grand, par une loi de Pareto généralisée *GPD*. Le quantile extrême est alors estimé en inversant la fonction de répartition de la loi de Pareto généralisée et en estimant les paramètres de cette loi à l'aide des observations supérieures au seuil u.
Comme défini en (19),

$$\begin{aligned} F_u\left(x\right) = \frac{F\left(u+x\right) - F\left(u\right)}{1 - F\left(u\right)} \quad si \quad x \geq 0 \quad &\Leftrightarrow \quad F_u\left(x-u\right) = 1 - \frac{1 - F\left(x\right)}{1 - F\left(u\right)} \quad si \quad x \geq u \\ &\Leftrightarrow \quad \overline{F}\left(x\right) = \overline{F}\left(u\right) \overline{F}_u\left(x-u\right) \quad si \quad x \geq u \\ \left(avec \quad \overline{F} = 1 - F\right) & \end{aligned}$$

sachant que

$$\overline{F}\left(u\right) = 1 - \mathbb{P}[X \leq u] = \mathbb{P}[X > u] = \frac{N_u}{n}$$

et

$$\widehat{\overline{F}}_u \approx 1 - G_{\hat{\xi},\hat{\sigma}} \qquad pour \quad u \quad assez \quad grand$$

où $G_{\xi,\sigma}$ est la *GPD* définie en (20), N_u est le nombre de observations au-dessus du seuil u et $\hat{\xi}$ et $\hat{\sigma}$ sont les estimateurs des paramètres de la loi *GPD*,
l'estimateur de $\overline{F}$ peut alors s'écrire

$$\widehat{\overline{F}}\left(x\right) = \frac{N_u}{n}\left(1 + \hat{\xi}\frac{x-u}{\hat{\sigma}} \right)^{-\frac{1}{\hat{\xi}}}, \quad \forall \hat{\xi} \neq 0.$$

Par inversion, l'estimateur obtenu par cette méthode s'écrit sous la forme :

$$\hat{x}_{p_t}^{GPD} = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(\frac{n}{N_u}\left(1 - p_t\right) \right)^{-\hat{\xi}} - 1 \right], \ \forall \hat{\xi} \neq 0 \tag{44}$$

2.4.2 Estimation de la période de retour

Supposons maintenant que nous disposions d'un quantile x_{p_T} et que nous désirions estimer la période de retour T, associée à la probabilité d'excès au delà de x_{p_T}. Cette période s'interprète comme une moyenne d'unités de temps séparant un événement de grandeur donnée x_{p_T} d'un second événement d'une grandeur égale ou supérieure.
Ainsi, estimer une période de retour T d'un quantile x_{p_T} revient à estimer l'ordre de ce quantile $\hat{p}_T = 1 - 1/\widehat{T}$.
Selon (42), nous estimons l'ordre $\hat{p}_T^{GEV}$ du quantile x_{p_T} pour la loi des valeurs extrêmes *GEV* par :

$$\hat{p}_T^{GEV} = \begin{cases} \exp\left\{-\left[1+\hat{\xi}\left(\frac{x_{p_T}-\hat{\mu}}{\hat{\gamma}}\right)\right]^{-\frac{1}{\hat{\xi}}}\right\} & \text{si } \hat{\xi} \neq 0 \\ \exp\left\{-\exp\left(-\frac{x_{p_T}-\hat{\mu}}{\hat{\gamma}}\right)\right\} & \text{si } \hat{\xi} = 0 \end{cases} \tag{45}$$

$\forall \hat{\mu} \in \mathbb{R} \quad et \quad \forall \hat{\gamma} > 0$
d'où

$$\begin{aligned} \widehat{T}^{GEV} &= \begin{cases} \left(1-\exp\left\{-\left[1+\hat{\xi}\left(\frac{x_{p_T}-\hat{\mu}}{\hat{\gamma}}\right)\right]^{-\frac{1}{\hat{\xi}}}\right\}\right)^{-1} & \text{si } \hat{\xi} \neq 0 \\ \left(1-\exp\left\{-\exp\left(-\frac{x_{p_T}-\hat{\mu}}{\hat{\gamma}}\right)\right\}\right)^{-1} & \text{si } \hat{\xi} = 0 \end{cases} \\ &= \frac{1}{1-H_{\hat{\xi}}\left(x_{p_T}\right)} \qquad \forall \hat{\xi} \in \mathbb{R} \end{aligned} \tag{46}$$

$\forall \hat{\mu} \in \mathbb{R} \quad et \quad \forall \hat{\gamma} > 0$
De même, selon (44) et (20), nous estimons la période de retour pour la loi des valeurs extrêmes *GPD* par :

$$\widehat{T}^{GPD} = \frac{n}{N_u}\frac{1}{1-G_{\hat{\xi},\hat{\sigma}}\left(x_{p_T}\right)}, \ \forall \hat{\xi} \in \mathbb{R} \quad et \quad \forall \hat{\sigma} > 0 \tag{47}$$

3 Approximations du niveau et du temps de retour

Il est important de pouvoir déterminer le niveau et le temps de retour (définis dans §2.4) liés au comportement de queue de distribution dans de nombreux domaines. Cela permet, par exemple pour un assureur de prévoir l'occurrence des catastrophes naturelles, pour un financier, d'envisager les scénarios de crises boursières et pour un gouvernement, de prévenir d'éventuelles épidémies.
La T.V.E. suppose la continuité de la loi de distribution pour estimer le niveau et le temps de retour, elle ne semble donc pas judicieuse dans le cas de données discrètes. Comme nous nous intéressons à des données discrètes, nous allons dans un premier temps proposer une estimation de bornes du niveau de retour et une estimation d'un temps de retour au lieu de les calculer exactement. Notons que la mise en place de ces bornes utilise une méthode indépendante de la nature des données.
Soit un n-échantillon $(X_1, \cdots, X_n)$ de n v.a. discrètes positives i.i.d. de v.a. parente X de fonction de répartition F.
Le problème est le suivant : quelle valeur du niveau de retour $\hat{x}_t$ choisir, pour qu'au bout de t-unités de temps, nous espérions un dépassement de $\hat{x}_t$, i.e
$\hat{x}_t$? tel que

$$\mathbb{E}\left(\sum_{i=1}^{t} \mathbf{1}_{(X_i > \hat{x}_t)}\right) = 1 \quad \Leftrightarrow \quad \mathbb{P}\left[X_i > \hat{x}_t\right] = 1/t, \quad \forall i = 1, \cdots, t \tag{48}$$

$$\Leftrightarrow \quad 1 - F\left(\hat{x}_t\right) = 1/t \tag{49}$$

A défaut de calculer explicitement $\hat{x}_t$, nous proposons d'après Guillou et *al.* (2007, [7]), une borne supérieure b_t du niveau de retour tel que pour un t donné

$$\mathbb{P}\left[X_i > b_t\right] \leq \mathbb{P}\left[X_i > \hat{x}_t\right] = 1/t, \quad i = 1, \cdots, t.$$

Notons que nous pouvons proposer également une borne inférieure l_t du niveau de retour tel que pour un t donné

$$\mathbb{P}\left[X_i > \hat{x}_t\right] = 1/t \leq \mathbb{P}\left[X_i > l_t\right], \quad i = 1, \ldots, t$$

3.1 Bornes du niveau de retour

3.1.1 Borne supérieure du niveau de retour

La méthode développée par Guillou et *al.* (2007, [7]) pour obtenir la borne supérieure repose sur l'inégalité de Markov :

$$1 - F\left(x\right) \leq \frac{\mathbb{E}\left(h\left(X\right)\right)}{h\left(x\right)} \quad pour \quad x > 0, \tag{50}$$

avec pour choix de fonction h : $h\left(x\right) = u\left(x\right) v\left(F\left(x\right)\right)$ telle que u et v sont deux fonctions positives et croissantes définies respectivement sur $[0, +\infty]$ et $[0, 1]$.

Pour un quantile $\hat{x}_t$ d'ordre $p_t = 1 - 1/t$, la fonction de survie $1 - F(\hat{x}_t)$ est égale à $1/t$. En remplaçant x par $\hat{x}_t$ (et donc $F(\hat{x}_t)$ par p_t) dans l'inégalité de Markov, nous obtenons

$$u(\hat{x}_t) \leq \frac{t\theta(u,v)}{v(p_t)} \quad avec \quad \theta(u,v) = \mathbb{E}[u(X)\, v(F(X))]$$

d'où

$$\hat{x}_t \leq u^{\leftarrow}\left[\frac{t\theta(u,v)}{v(p_t)}\right] := b_t(u,v), \tag{51}$$

$u^{\leftarrow}$ désignant la fonction inverse de u.
Ainsi,

$$\hat{x}_t \leq \inf\{b_t(u,v) : u \quad et \quad v \quad fonctions \quad croissantes \quad positives\}, \tag{52}$$

et, pour un échantillon ordonné $(X_{1,n}, \cdots, X_{n,n})$, nous considérons l'estimateur naturel de $\theta(u,v)$:

$$\hat{\theta}_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} u(X_{i,n})\, v\left(\frac{i}{n}\right). \tag{53}$$

Rappelons alors les propriétés satisfaites par cet estimateur et l'estimateur de b_t ([7]).

Proposition 4 *Soit X une v.a. positive de loi F.*
Si $v(.)$ est une fonction Lipshitzienne d'ordre 1 sur $[0,1]$ et $u(.)$ une fonction telle que les intégrales $\int |u(x)|\, dF(x)$ et $\int v^2(F(x))\, u^2(x)\, dF(x)$ sont finies, alors

1. *la v.a. $\sqrt{n}\left(\hat{\theta}_n(u,v) - \theta(u,v)\right)$ converge en loi vers une loi Gaussienne centrée de variance*

$$\sigma^2 := \mathbb{E}\left(-v(U)\, v(F^{\leftarrow}(U)) + \theta(u,v) - \int_0^1 \left(\mathbf{1}_{(U\leq t)} - t\right) v'(t)\, u(F^{\leftarrow}(t))\, dt\right)^2 \tag{54}$$

   où U est la fonction de répartition d'une loi uniforme sur $(0,1)$;

2. *la v.a. $\sqrt{n}\left(\hat{b}_t(u,v) - b_t(u,v)\right)$ (où $b_t(u,v)$ est définie en (51)) converge en loi vers une loi normale centrée de variance*

$$\psi^2 := \frac{t^2\sigma^2}{v^2(p_t)}\left[(u^{\leftarrow})'\left(\frac{t\theta(u,v)}{v(p_t)}\right)\right]^2. \tag{55}$$

Proposition 5 *Soit X une v.a. positive de loi F telle que $F(0) = 0$.*
Si $v(.)$ est une fonction positive, strictement croissante et Lipshitzienne d'ordre 1 sur $[0,1]$ et $u(.)$ est une fonction positive, continue et strictement croissante telle que $\mathbb{E}\left(u^3(X)\right) < \infty$, alors :
si σ^2, défini en (54), est non nul,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\hat{\theta}_n(u,v)}{u(x)\,v(F(x))} < 1 - F(x)\right) = 0$$

Sous ces conditions, nous pouvons remplacer $\theta(u,v)$ par $\hat{\theta}_n(u,v)$ dans l'inégalité de Markov (51) et estimer la borne supérieure du niveau de retour par

$$\hat{b}_t(u,v) := u^{\leftarrow}\left(\frac{t\hat{\theta}_n(u,v)}{v(p_t)}\right). \tag{56}$$

Sous les conditions des deux propositions 4 et 5, nous déduisons l'intervalle de confiance IC_{b_t} de la borne supérieure du quantile extrême b_t selon

$$IC_{b_t} = \left[\hat{b}_t(u,v) \pm q_{1-\alpha/2}\frac{t\hat{\sigma}}{\sqrt{n}v(p_t)}\left(u^{\leftarrow}\right)'\left(\frac{t\hat{\theta}_n(u,v)}{v(p_t)}\right)\right] \tag{57}$$

où $\hat{\sigma}$ est un estimateur de σ et $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi normale centrée et réduite.

Considérons $u(x) = x^{\alpha}$ et $v(x) = x^{\beta}$ avec α et β deux réels strictement positifs. Dans ce cas, nous estimons $\hat{\theta}_n$, défini en (53), et $\hat{b}_t$ en (56) selon

$$\hat{\theta}_n(u,v) = \hat{\theta}_n(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n}(x_{i,n})^{\alpha}\left(\frac{i}{n}\right)^{\beta}$$

et

$$\hat{b}_t(u,v) = \hat{b}_t(\alpha,\beta) = \left(\frac{t\hat{\theta}_n(\alpha,\beta)}{(1-1/t)^{\beta}}\right)^{1/\alpha}. \tag{58}$$

Nous calculons

$$\inf\left\{\hat{b}_t(\alpha,\beta) \quad : \quad \alpha \in \ ]0,\alpha_{max}] \quad et \quad \beta \in \ ]0,\beta_{max}] \quad \right\}$$

où α_{max} et β_{max} sont deux constantes données (choisies selon le jeu de données).
Cette valeur correspond à un $\hat{\alpha}_t$ optimum et à un $\hat{\beta}_t$ optimum telle que

$$(\hat{\alpha}_t,\hat{\beta}_t) = \arg\min\left\{\hat{b}_t(\alpha,\beta) \quad : \quad \alpha \in \ ]0,\alpha_{max}] \quad et \quad \beta \in \ ]0,\beta_{max}] \quad \right\} \tag{59}$$

3.1.2 Borne inférieure du niveau de retour

De même, une borne inférieure du niveau de retour de $\hat{x}_t$ a été estimée dans [7] par $\hat{l}$ selon :

$$\hat{x}_t \geq \hat{l}_t(u, w, q) := u^{\leftarrow}\left[\frac{\hat{\theta}^*(u, w) - t^{1/q-1}\left(\hat{\theta}^*(u, w)\right)^{1/q}}{w(1/t)\, p_t}\right] \tag{60}$$

avec

$$\hat{\theta}^*(u, w) = \frac{1}{n}\sum_{i=1}^{n} u(X_{i,n})\, w\left(1 - \frac{i}{n}\right),$$

où u est la fonction définie dans la section précédente (positive et croissante définie sur $[0, +\infty]$), w est une fonction positive et décroissante définie sur $[0, 1]$ et $q > 1$.

Considérons à nouveau $u(x) = x^{\alpha}$ et $w(x) = x^{-\nu}$ avec α et ν deux réels strictement positifs.
Dans ce cas, nous estimons $\hat{l}_t$ définie en (60) selon

$$\hat{l}_t(u, w, q) = \hat{l}_t(\alpha, \nu, q) = \left(\frac{\hat{\theta}^*(\alpha, \nu) - t^{1/q-1}\left(\hat{\theta}^*(\alpha, \nu)\right)^{1/q}}{(1/t)^{-\nu} p_t}\right)^{1/\alpha} \tag{61}$$

avec

$$\hat{\theta}^*(\alpha, \nu) = \frac{1}{n}\sum_{i=1}^{n} (x_{i,n})^{\alpha}\left(1 - \frac{i}{n}\right)^{-\nu}.$$

Nous calculons

$$\max\left\{\hat{l}_t(\alpha, \nu, q) \quad : \quad \alpha \in \ ]0, \alpha_{max}] \quad , \quad \nu \in \ ]0, \nu_{max}] \quad et \quad q > 1\right\}$$

où α_{max} et ν_{max} sont deux constantes données (choisies selon le jeu de données).
Cette valeur correspond à un $\hat{\alpha}_t^*$ optimum, à un $\hat{\nu}_t$ optimum et à un $\hat{q}$ optimum telle que

$$(\hat{\alpha}_t^*, \hat{\nu}_t, \hat{q}_t) = \arg\max\left\{\hat{l}_t(\alpha, \nu, q) \quad : \quad \alpha \in \ ]0, \alpha_{max}] \quad , \quad \nu \in \ ]0, \nu_{max}] \quad et \quad q > 1\right\}. \tag{62}$$

3.2 Approximation du temps de retour

A défaut de calculer explicitement le temps de retour d'un quantile théorique $\hat{x}_t$ défini en (49), nous proposons une démarche pour estimer ce temps de retour T en deux étapes :

i- Nous utilisons les estimateurs des deux bornes ($\hat{b}_t$ définie en (56) et $\hat{l}_t$ définie en (60)) pour tracer le graphe $\left(t, \hat{l}_t, \hat{b}_t\right)$ (figure.6) avec $\hat{l}_t \leq \hat{x}_t \leq \hat{b}_t$ pour tout t.

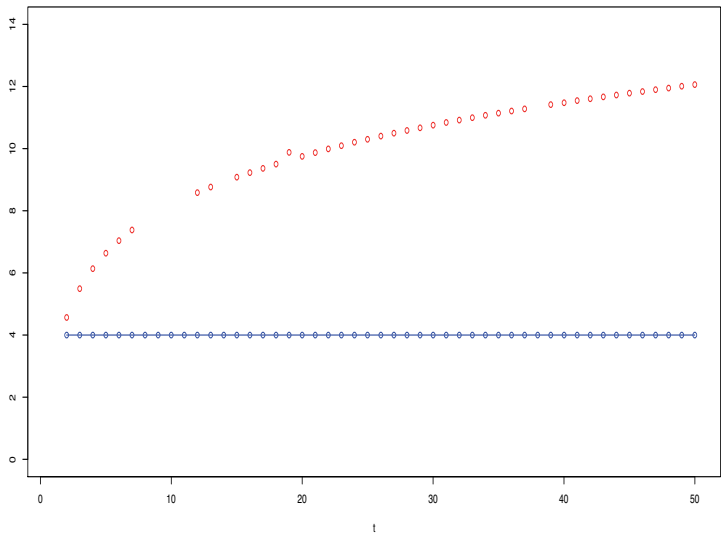

FIGURE 6 – Bornes du niveau de retour. L'axe des abscisses représente les valeurs de t de 2 à 500 et l'axe des ordonnées correspond aux bornes du niveau de retour ou du quantile. Les points rouges correspondent à la borne supérieure du niveau de retour $\hat{b}_t$ définie en (56) et les points bleus correspondent à la borne inférieure du niveau de retour $\hat{l}_t$ définie en (60) avec $\alpha_{max} = 5$, $\beta_{max} = 0.2$, $\nu_{max} = 4$ et $q = 2$; ce choix de α_{max}, β_{max}, ν_{max} et q est obtenu après différentes étapes d'optimisation.

Remarque : nous pouvons aussi représenter les intervalles de confiances des bornes du niveau de retour (par exemple nous avons représenté l'intervalle de confiance de la borne supérieure du niveau de retour pour le jeu de données Derby (figure.11)).

La valeur de $\hat{l}_t$ étant constante pour tout t, nous utiliserons la borne supérieure du niveau de retour pour estimer approximativement le temps de retour de $\hat{x}_t$.

ii- Soit une nouvelle observation x_{t_0} à temps t_0.
- Posons $x_{t_0} = \hat{b}_T$
- Nous lisons sur le graphe précédent la valeur T correspondante, associée au niveau de retour ou au quantile (inconnu) $\hat{x}_T$ avec $\hat{x}_T \leq \hat{b}_T$.

D'où

$$x_{t_0} \geq \hat{x}_T \Leftrightarrow \exists\, \delta t \geq 0 \; tel \; que \; x_{t_0} = \hat{x}_{T+\delta t}. \tag{63}$$

Nous remarquons que le temps de retour théorique $(T + \delta t)$ de l'observation x_{t_0} est

supérieur au temps de retour T de $\hat{b}_T$.
Il serait naturel de procéder de même avec la borne inférieure $\hat{l}_t$ puisque $\hat{l}_t \leq \hat{x}_t \leq \hat{b}_t$ et

$$T_l := \frac{1}{1 - F\left(\hat{l}_t\right)} \leq T = \frac{1}{1 - F\left(\hat{x}_t\right)} \leq T_b := \frac{1}{1 - F\left(\hat{b}_t\right)},$$

où T_l, T et T_b désignent les temps de retour associées à $\hat{l}_t$, $\hat{x}_t$ et $\hat{b}_t$. Ainsi, nous aurions une idée de la qualité de l'approximation du temps de retour.
Cependant, la borne inférieure n'est pas assez fine et nous donne une valeur quasi constante au cours de temps, aussi choisissons nous désormais d'approcher le temps de retour d'un quantile par le temps de retour de sa borne supérieure.
Nous allons utiliser ce temps de retour dans le cadre de prévision pour les applications en finance ainsi que pour la mise en place d'un système de surveillance en épidémiologie.
Notons que du fait de l'hypothèse d'indépendance des $(X_i)_{i=1,\ldots,n}$, nous avons :

$$\mathbb{E}\left(\sum_{i=1}^{T} \mathbf{1}_{\left(X_i \geq \hat{x}_{t_0}\right)}\right) = 1 \Leftrightarrow \mathbb{E}\left(\sum_{i \in I(T)} \mathbf{1}_{\left(X_i \geq \hat{x}_{t_0}\right)}\right) = 1 \tag{64}$$

pour tout intervalle de temps $I(T)$ de longueur T. Nous allons donc considérer $I(T) = [t_0 - T, t_0]$ pour l'application en épidémiologie et $I(T) = [t_0, t_0 + T]$ pour l'application en finance où t_0 représente le temps présent.
Il va de soi que nous aurions pu mettre en place un système de surveillance/détection pour des applications en finance, mais nous avons souhaité illustrer la méthode en nous intéressant à deux types de problèmes.

3.3 Applications

Nous allons appliquer l'approche décrite précédemment à des jeux de données de nature discrète, de type épidémiologiques et financières.

3.3.1 Application en épidémiologie

Nous disposons de nombreux jeux de données réelles de nature discrète, fournies par l'InVS (Institut de Veille Sanitaire) et provenant du CNR (Centre National de Référence des salmonelles de l'Institut Pasteur) sur le nombre de souches revues au CNR pour différents sérotypes de salmonelles. Nous disposons également du nombre de cas de grippe et syndromes grippaux estimé par le réseau sentinelles (INSERM).
La transmission des salmonelles s'effectue par l'intermédiaire de volailles contaminées, d'oeufs et de certains autres aliments. Un type d'intoxication alimentaire est caractérisé par des douleurs abdominales, de la fièvre, des nausées, des vomissements et de la diarrhée.
Ces intoxications nécessitent parfois une intervention rapide de l'état.
Les salmonelles présentent 1400 formes antigéniquement différentes ; nous prendrons comme exemple deux d'entres elles : la salmonelle Derby et la salmonelle Agona.
Nous traiterons aussi le cas de la grippe, caractérisée par son démarrage brutal avec céphalées, fièvre, sensation de froid, toux, mal de gorge, malaises, douleurs musculaires ; parfois, des risques mortels apparaissent chez les personnes âgées.
Dans de telles situations, il est recommandé d'envisager des scénarios en estimant un niveau d'alarme permettant de signaler la présence d'une augmentation anormale du nombre de cas.

Il sera peut-être nécessaire au préalable de transformer les séries d'observations. En effet, parmi les cycles spécifiques identifiés par les épidémiologistes, il en existe de particuliers liés au phénomènes climatiques.
Ces cycles sont qualifiés de saisonniers et pour certaines épidémies, telle la grippe, cette saisonnalité est à prendre en considération.
De même, il peut exister une tendance pour certaines séries d'observations, dont il faudra tenir compte.

D'une façon générale, nous pouvons schématiser le traitement des données selon :

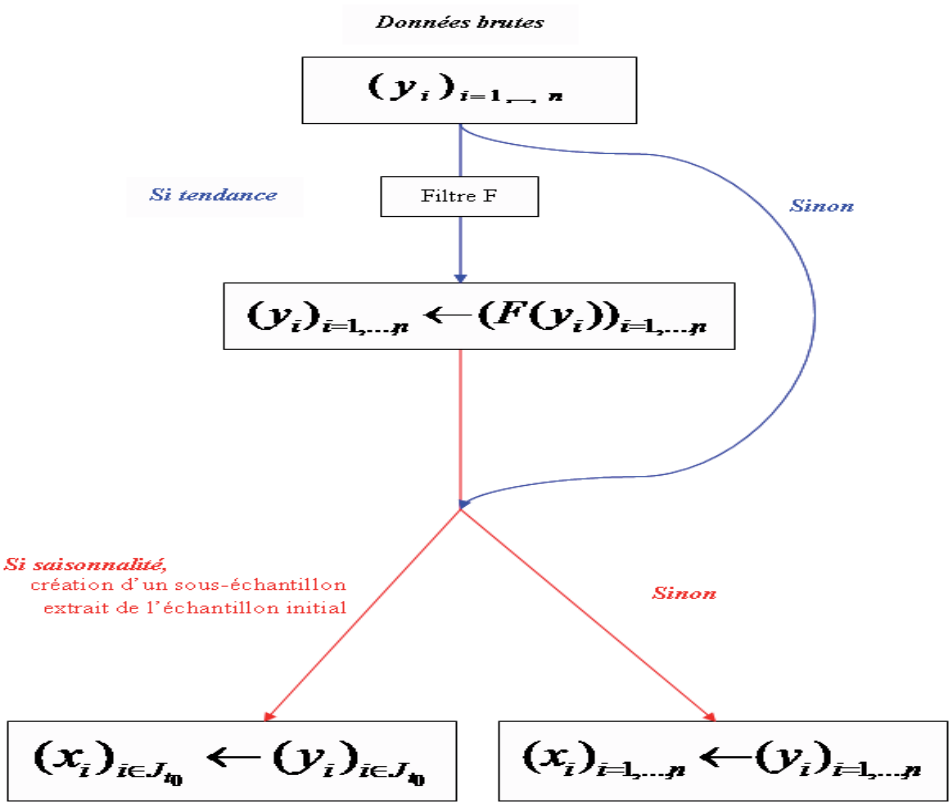

FIGURE 7 – *Schéma de traitement de données.* $J_{t_0} = \{i : i \in [t_0 - t \cdot seq \pm k]\,;\ t = 1, \dots, p\}$ *où* p *correspond au nombre de séquences dans l'échantillon total,* $2k+1$ *représente la longueur du cycle (nombre de semaines considérées par cycle),* t *représente l'ordre du cycle (l'année* t *dans les exemples traités ici),* seq *représente la séquence du cycle (qui correspond à* 52 *semaines si nous avons des données hebdomadaires) et* t_0 *le temps présent.*

Désormais, nous travaillerons sur les données transformées, si nécessaire.

Remarque : si nous cherchons à appliquer les méthodes existantes de la théorie des valeurs extrêmes pour déterminer le temps de retour (décrites en §2.4.2), sans tenir compte de la nature discrète des données (i.e. en les considérant grossièrement continues), nous remarquons en général que nous ne sommes pas en mesure de déterminer la valeur exacte de l'indice de queue de distribution ξ comme nous le verrons sur quelques exemples.
En utilisant (64) avec $I(T) = [t_0 - T, t_0]$ et l'approximation du temps de retour (en §3.2), nous pouvons chercher le temps de retour T associé à la valeur x_{t_0}, observée au

temps t_0 en posant $x_{t_0} = \hat{b}_T$.
Ainsi, s'il existe déjà une valeur supérieure à $\hat{b}_T$ sur l'intervalle $[t_0 - T, t_0[$, une alarme $\mathbb{A}_{t_0}$ sera générée en t_0 tel que

$$\mathbb{A}_{t_0} = \mathbf{1}_{\left\{\mathbb{E}\left(\sum_{i=t_0-T}^{t_0} \mathbf{1}_{\left(X_i \geq x_{t_0-1}\right)}\right) > 1\right\}} \tag{65}$$

Résumons la démarche suivie (en §3.2) :

1. estimer la borne supérieure $\hat{b}_t$ pour différentes valeurs de t et créer le graphe $\left(t, \hat{b}_t\right)$,
2. lire ensuite sur le graphe de $\hat{b}_t$ le temps de retour T de la dernière observation x_{t_0} tel que $x_{t_0} = \hat{b}_T$,
3. tester la présence d'une observation x_t supérieure ou égale à $\hat{b}_T$ pour tout $t \in (t_0 - T, t_0)$ et déduire $\mathbb{A}_{t_0}$.

Si $\mathbb{A}_{t_0} = 1$, nous sonnons l'alarme.
Appliquons cette démarche à quelques exemples de jeux de données.

a) Salmonelle Derby

Le premier jeu de données étudié ici concerne le nombre de souches de Salmonelle Derby reçues par le CNR. Nous cherchons à déterminer si le dernier enregistrement du nombre de souches est anormalement élevé, afin de déclencher une alarme.
La figure.8 et le tableau.1 ci-dessous introduisent et decrivent notre échantillon de 598 observations hebdomadaires du nombre de souches.

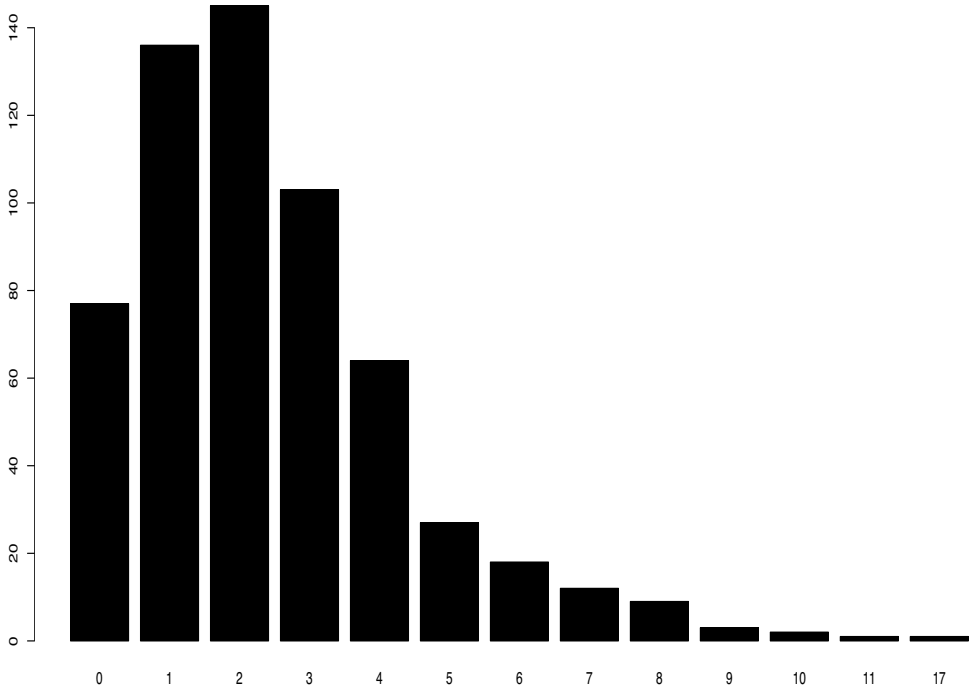

FIGURE 8 – *Diagramme en bâtons du jeu de données : Salmonelle Derby. L'axe des abscisses représente le nombre hebdomadaire de souches et l'axe des ordonnées correspond à leur fréquence.*

Minimum	Quantile 25%	Médiane	Moyenne	Quantile 75%	Maximum
0.00	1.00	2.00	2.45	3.00	17.00

TABLE 1 – *Description du jeu de données : Salmonelle Derby*

Notons la présence de valeurs extrêmes, qui se situent au dessus de 5 ; elles représentent 5.68% du nombre total des observations.

Représentons dans la figure.9 la série temporelle associée.

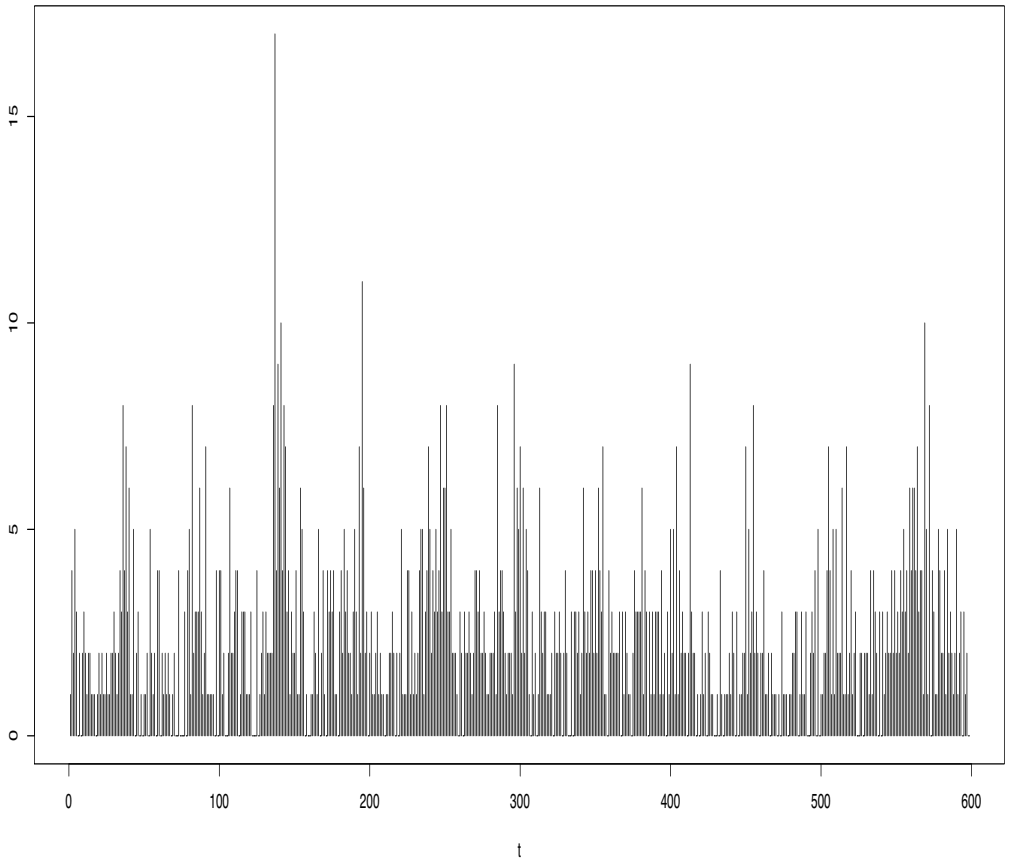

FIGURE 9 – *Le nombre hebdomadaire de souches de Salmonelle Derby. L'axe des abscisses représente les semaines et l'axe des ordonnées correspond au nombre de contaminations.*

Si nous ne tenons pas compte de la nature discrète des données et nous plaçons dans le cadre classique de la théorie des valeurs extrêmes, nous remarquons sur la figure.10 qu'il n'est pas possible de déterminer la valeur exacte de l'indice de queue de distribution ξ.

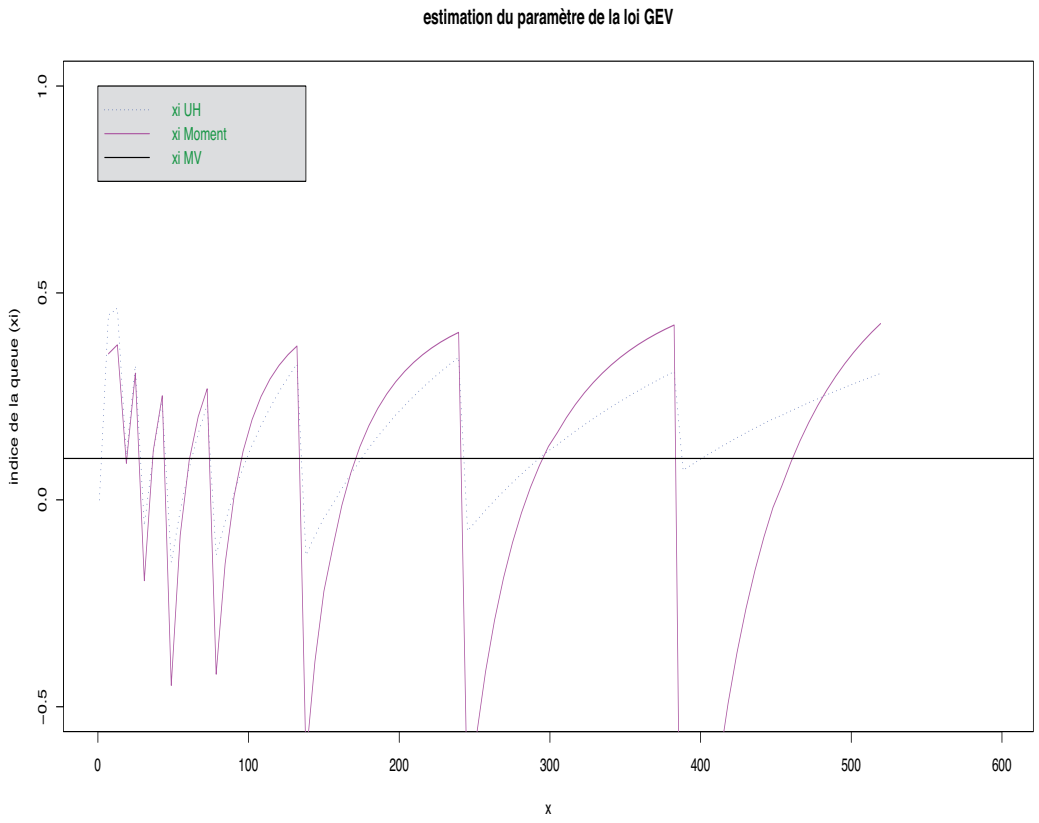

FIGURE 10 – *Estimation de l'indice ξ de la queue de distribution de la loi GEV. L'axe des abscisses représente l'indice k de $\xi_{k,n}$ et l'axe des ordonnées correspond à l'estimateur de l'indice de la queue de distribution $\hat{\xi}_{k,n}$. Les courbes rose, bleue et noir correspondent respectivement aux estimateurs $\hat{\xi}^{Moment}$ défini en (35), $\hat{\xi}^{UH}$ défini en (36) et $\hat{\xi}^{mv}$ obtenu par la méthode de maximum de vraisemblance.*

Il est clair ici, de par l'allure de ces courbes, que l'indice de queue est indéterminable. C'est pourquoi nous utilisons dans un premier temps l'approche décrite en §3.2 dans le but d'avoir au moins une approximation du niveau d'alarme à déclencher.

La figure.11 ci-dessous affiche la courbe de la borne supérieure estimée $\hat{b}_t$ donnée en (58), avec un choix de $\alpha_{max} = 5$ et $\beta_{max} = 0.2$ pour ce jeu de données (obtenu après différentes étapes d'optimisation). Pour chaque valeur de t, nous calculons via (59) les $\hat{\alpha}_t$ et $\hat{\beta}_t$ optimaux puis $\hat{b}_t$.
Rappelons qu'un niveau de retour d'ordre t correspond pour t grand à un quantile extrême d'ordre $1 - 1/t$. L'estimation de la borne supérieure du niveau de retour correspond à une estimation de la borne supérieure du quantile extrême.

Par exemple, $\hat{b}_{10}$ est égale à 8, sur ce jeu de données. C'est une borne supérieure

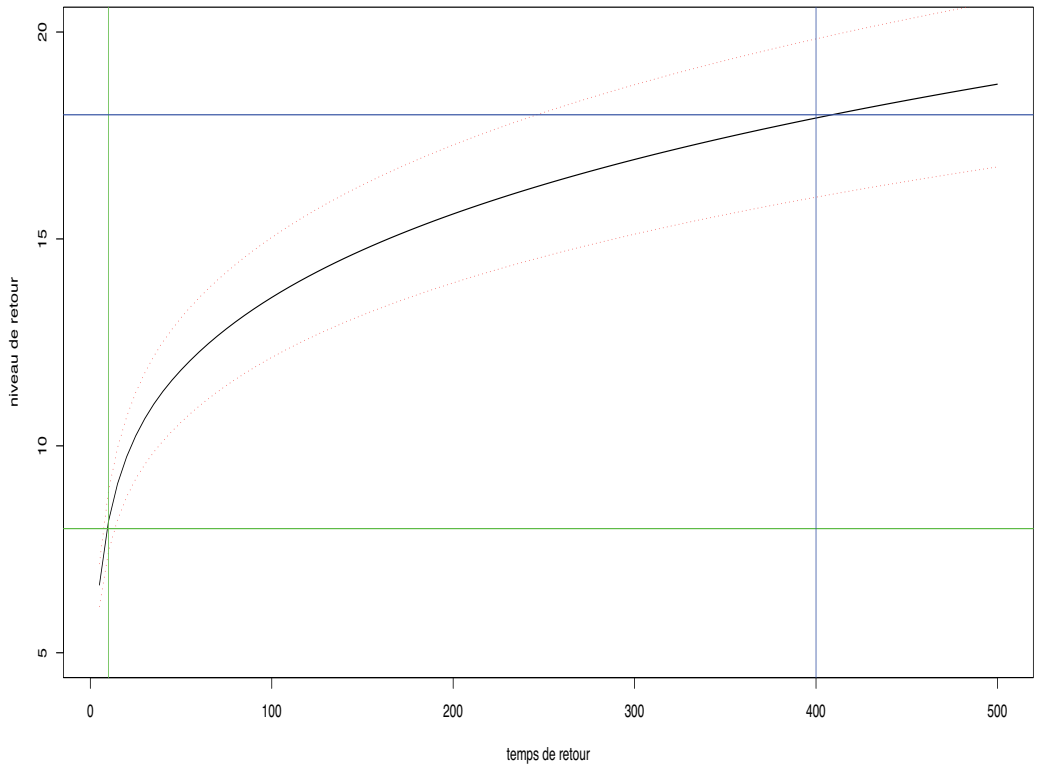

FIGURE 11 – *Le nombre maximal de souches se Salmonelle Derby. L'axe des abscisses représente les valeurs de t de* 5 *à* 500 *semaines, correspondant aussi aux ordres des quantiles extrêmes de* 80% *à* 99.80% *et l'axe des ordonnées correspond à* $\hat{b}_t$ *défini en (58). Les deux courbes rouges pointillées indiquent les bornes de l'intervalle de confiance à 95% de la borne supérieure, défini en (57). (Remarque : du fait de l'échelle, les courbes paraissent être continues mais il s'agit en fait de points juxtaposés.)*

du niveau de retour dans un intervalle de temps de 10 observations successives et du quantile extrême d'ordre 90%. De même, $\hat{b}_{400} = 18$ est une borne supérieure du quantile extrême d'ordre 99.75%. Au vu de ces résultats, nous ne nous attendons donc pas à enregistrer plus de 8 souches par semaine (respectivement 18) pendant 10 semaines successives (respectivement 400), sinon, nous déclenchons l'alarme.

En suivant la démarche décrite dans l'introduction de §3.3.1, si nous enregistrons à la semaine 599, 10 cas d'infections, la valeur de T correspondante, lue sur la figure.11 est égale à 35 (c'est à dire $\hat{b}_{35} = 10$).

Puisque sur la figure.12 nous détectons une obsevation supérieure ou égale à 10

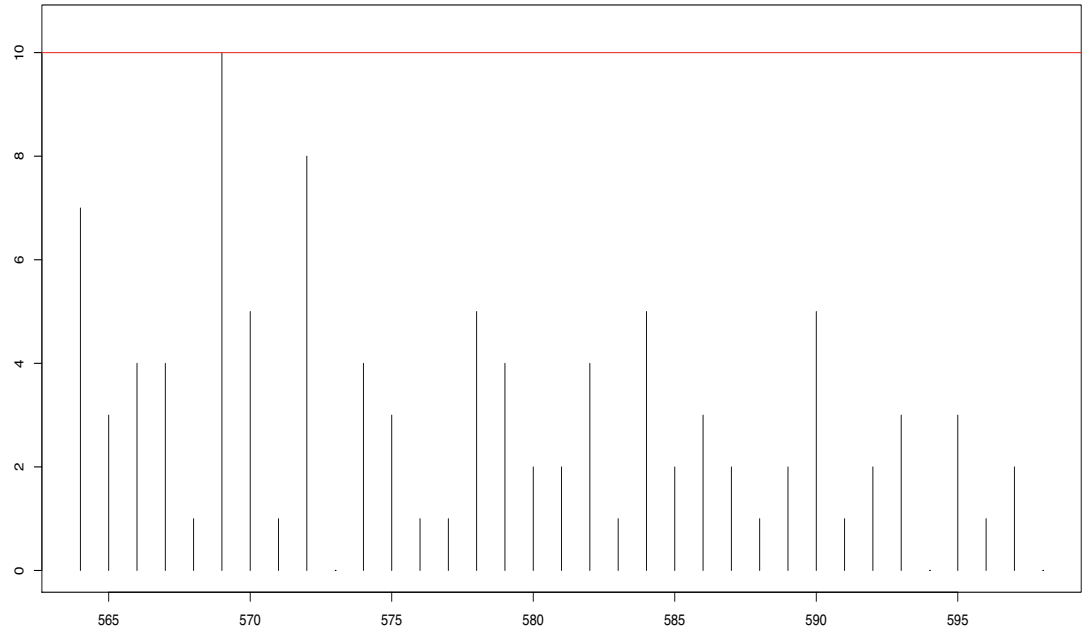

FIGURE 12 – *Graphe de surveillance : Salmonelle Derby. L'axe des abscisses représente l'intervalle de temps* $[564, 599[$*, l'axe des ordonnées correspond au nombre de contaminations et la droite rouge représente le nombre de souches enregistrées au cours de la dernière semaine.*

dans l'intervalle de temps $[564, 599[$, de longueur 35, alors $\mathbb{A}_{599} = 1$ et nous déclenchons une alarme en $t_0 = 599$.

b) Salmonelle Agona

Le jeu de données étudié ici concerne le nombre de souches de Salmonelle Agona. Nous disposons de 683 observations hebdomadaires de l'année 1995 à 2008 du nombre de souches reçues par le CNR.
Nous voulons tester si les observations, dès la première semaine de l'année 2000, c'est à dire de la semaine 260 à la semaine 683, présentent une alarme.
Comme précédemment, nous utilisons la démarche décrite en §3.2 sur des sous-échantillons tels que $(J_{t_0})_{t_0=260,\dots,683}$ obtenus selon la méthode schématisée en figure.7 (avec $k = 3$, $t = 1,\dots,5$ et $seq = 52$: pour une observation à la semaine t_0, le sous-échantillon associé est composé des observations des semaines $(t_0 - 52t - 3,\dots,t_0 - 52t + 3)_{t=1,\dots,5}$) pour mettre en évidence les différentes alarmes générées via la méthode de la borne supérieure.
L'InVS utilise en général l'algorithme de Farrington, une méthode basée sur la prédiction à partir d'un modéle linéaire généralisé (GLM), connue sous le nom de la méthode anglaise.
Représentons sur le même graphe (figure.13) les alarmes générées par les deux méthodes (Farrington et la méthode de la borne supérieure) en travaillant sur les mêmes sous-échantillons constitués en tenant compte de la saisonnalité.

La méthode de la borne supérieure est appliquée sur la série temporelle représentant le nombre hebdomadaire de souches reçu par le CNR, pour la salmonelle Agona. Entre 2000 et 2007, la méthode génère au total 50 alarmes (représentées par les triangles noirs) contre 22 alarmes pour la méthode anglaise (triangles rouges).
En février et mars 2005, 15 alarmes sont générées. Elles correspondent à un nombre anormalement élevé de souches reçues par le CNR (169 souches en 2 mois) qui a fait l'objet, à l'époque, d'un signalement par le CNR pour excès de cas. Ces 15 alarmes correspondent donc à un événement épidémiologique réel et peuvent être considérées comme des vrais positifs. A la fin de l'année 2003, les deux méthodes sont concordantes en générant également 5 alarmes qui n'ont pas fait l'objet à l'époque d'un signalement spécifique.
Il est à noter que la méthode de la borne supérieure génère deux fois plus d'alarmes que la méthode anglaise (50 versus 22). La raison principale est le choix, dans la méthode anglaise, de ne pas générer d'alarme si moins de 5 souches ont été reçues au cours des 4 dernières semaines au CNR. En appliquant ce critère supplémentaire à la méthode de la borne supérieure, 13 alarmes seraient supprimées. Sur 414 semaines, les deux méthodes généreraient alors 20 alarmes simultanément.
Alors que la méthode anglaise génére 2 alarmes sans que la méthode de la borne supérieure en génére, la méthode de la borne supérieure génére 17 alarmes non générées par la méthode anglaise. Les deux méthodes ont une concordance globale de 95%, ce

qui laisse à penser que la méthode de la borne supérieure pourrait être une nouvelle méthode à prendre en compte dans des analyses réelles réalisées par les systèmes de surveillance en France. Le comportement de cette méthode doit être testé de manière exploratoire en routine sur plusieurs sérotypes pour permettre de mesurer ses performances et de décider de l'intégrer ou non dans les méthodes utilisées en routine pour la détection d'événements inhabituels.

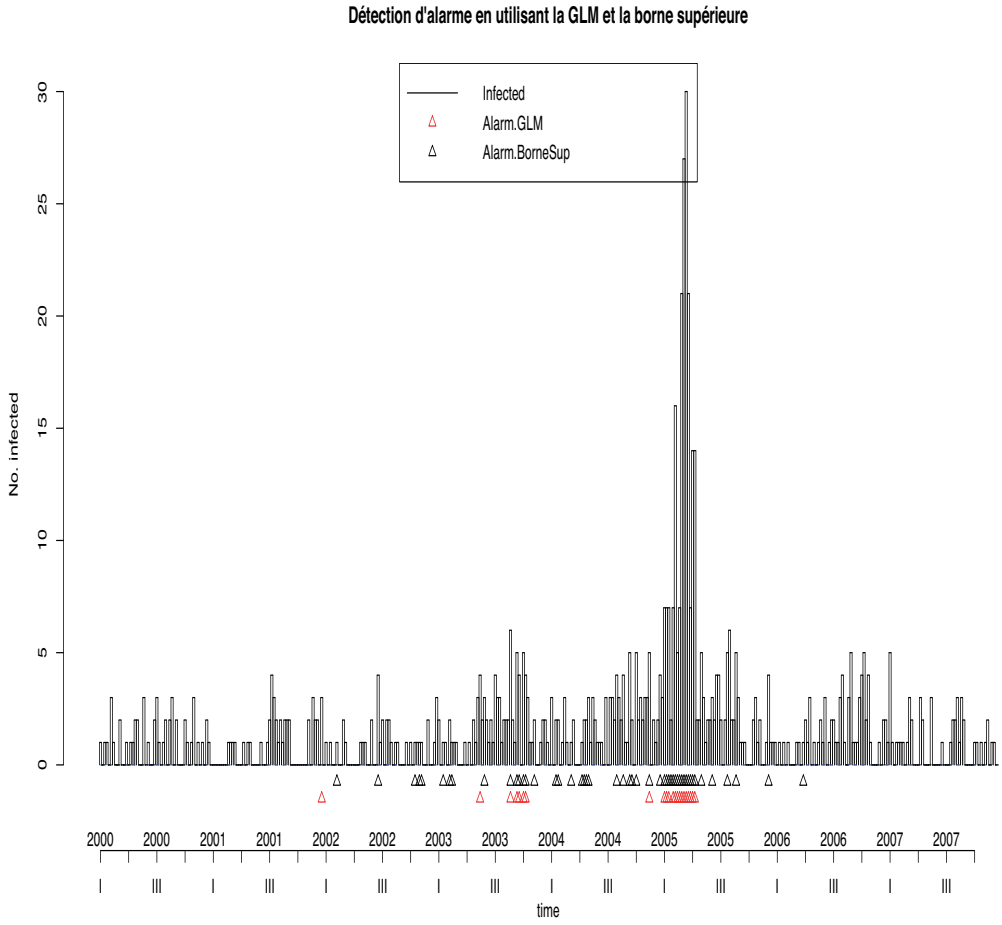

FIGURE 13 – *Détection des alarmes générées selon la GLM et la borne supérieure du niveau de retour. Les triangles rouges représentent les alarmes générées selon l'algorithme de Farrington et les triangles noirs représentent les alarmes générées selon la méthode de la borne supérieure du niveau de retour (figurant à des temps t tel que $\mathbb{A}_t = 1$).*

c) Grippe

Ce jeu de données est un sous-échantillon du jeu de données grippe construit selon le traitement présenté dans la figure.7 (avec $k = 3$, $t = 1, \ldots, 20$ et $seq = 52$). Il concerne le nombre hebdomadaire de cas de grippes et syndromes grippaux estimé en France par le réseau sentinelles de l'INSERM (voir figure.15).
La figure.14 et le tableau.2 décrivent notre sous-échantillon de 140 observations hebdomadaires du nombre de cas d'infections par le virus Grippe.

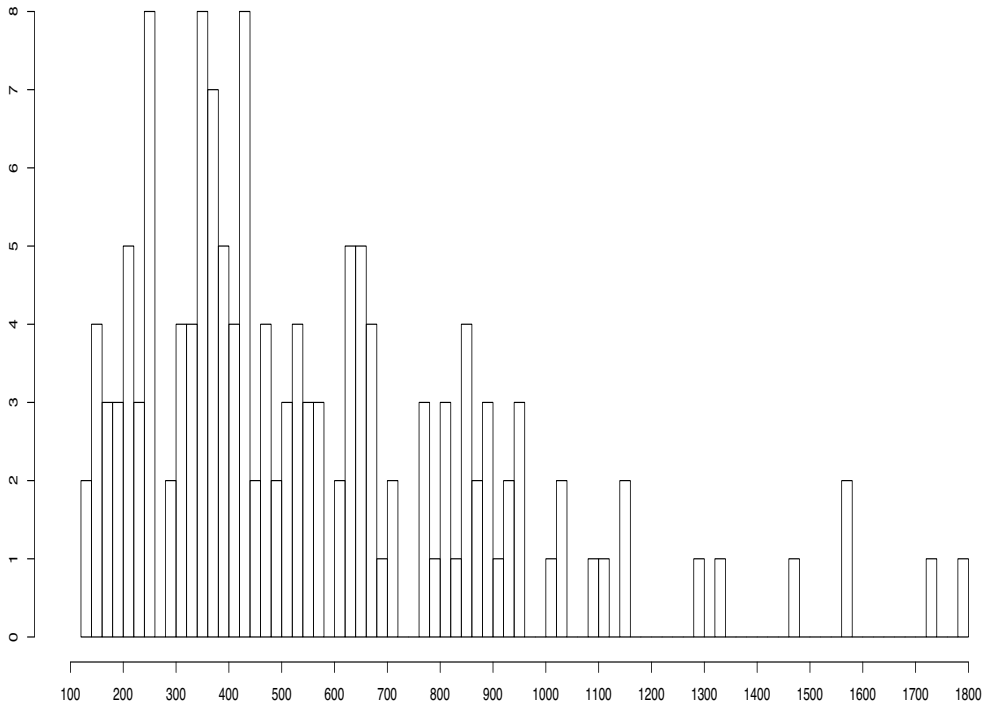

FIGURE 14 – *Histogramme relatif au sous-échantillons du jeu de données : Grippe. L'axe des abscisses représente le nombre d'infections, présenté par classes d'amplitude* 20*, et l'axe des ordonnées correspond à leur fréquence.*

Minimum	Quantile 25%	Médiane	Moyenne	Quantile 75%	Maximum
129.00	337.50	465.00	561.70	734.50	1793.00

TABLE 2 – *Description du sous-échantillon du jeu de données : Grippe*

Nous notons la présence d'événements rares ou de valeurs extrêmes dans ce sous-échantillon. Les valeurs au dessus de 1000 représentent une faible proportion de 9.52% du nombre total des observations.
Nous représentons dans la figure.15 la série temporelle associée au jeu de données Grippe ainsi que celle extraite associée au sous-échantillon notifiée par une couleur foncée.

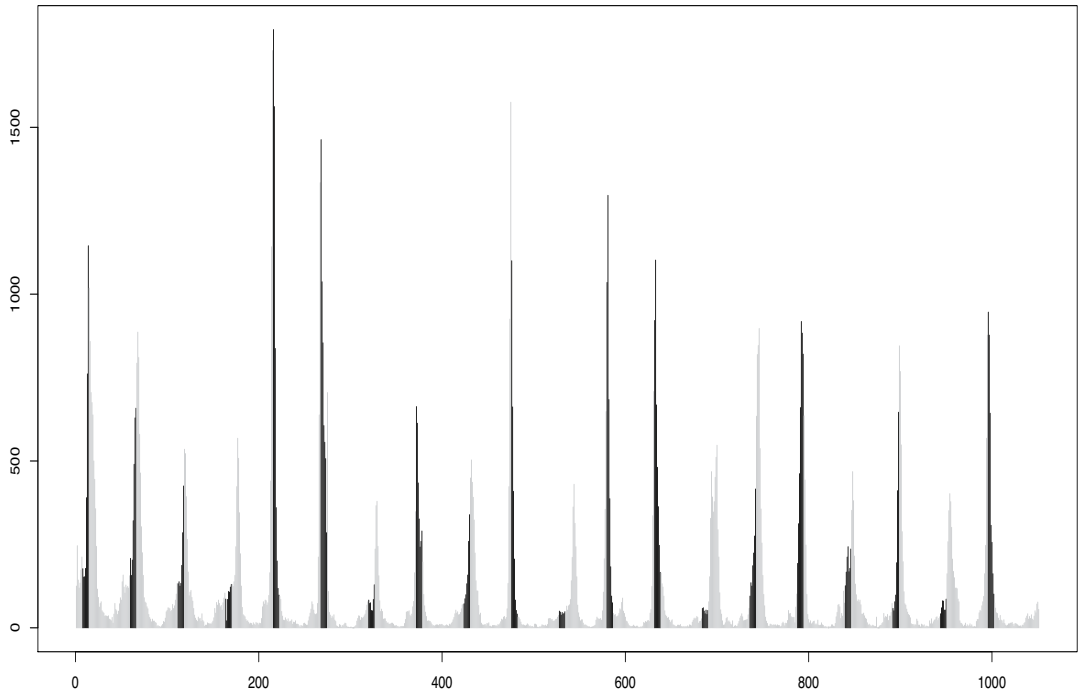

FIGURE 15 – *Le nombre d'infections hebdomadaires par la Grippe. L'axe des abscisses représente les semaines et l'axe des ordonnées correspond au nombre de cas de grippes et de syndromes grippaux. La série temporelle en noire représente le sous-échantillon décrit précédemment.*

A nouveau, la détermination d'une valeur exacte de l'indice de queue de distribution ξ n'est pas possible (lorsque nous considérons les données grossièrement continues), aussi nous utilisons l'approximation du temps de retour de la borne supérieure du niveau de retour pour percevoir la présence d'une alerte.
La figure.16 affiche la courbe de la borne supérieure estimée $\hat{b}_t$ (figurant en (58)) avec $\alpha_{max} = 10$ et $\beta_{max} = 0.3$ pour ce jeu de données. Pour chaque valeur de t, nous calculons via (59) les $\hat{\alpha}_t$ et $\hat{\beta}_t$ optimaux puis $\hat{b}_t$.

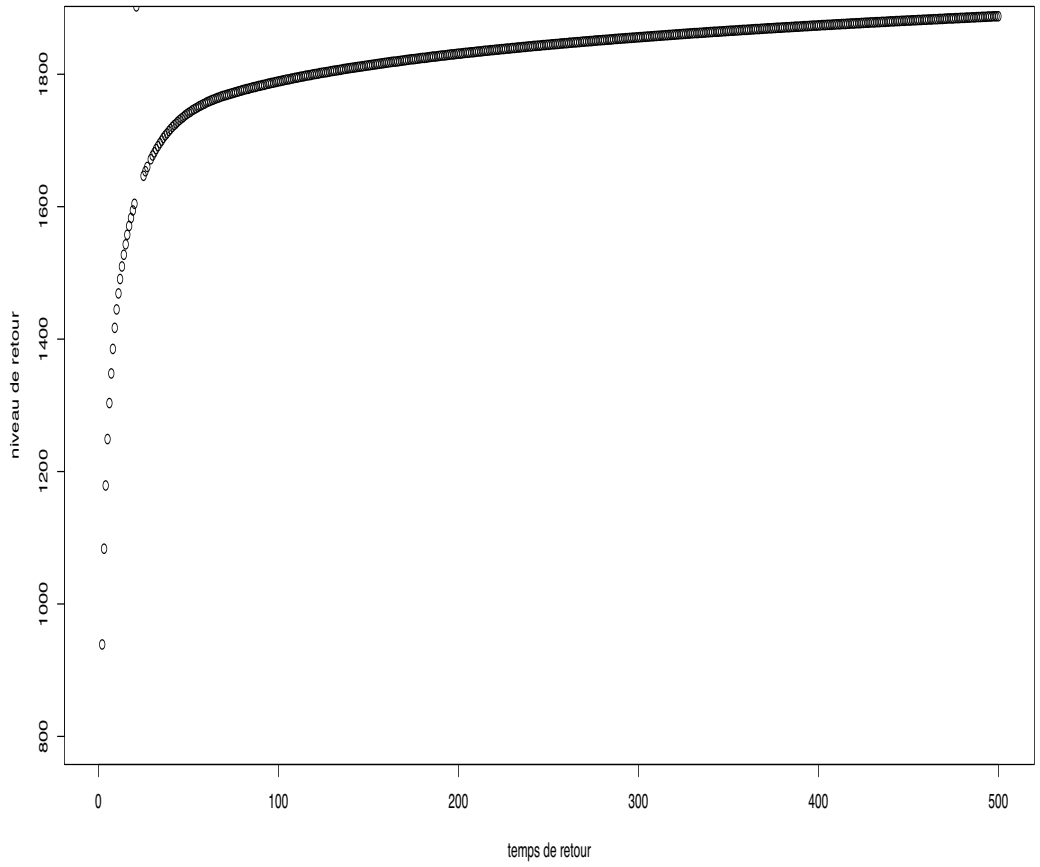

FIGURE 16 – *Le nombre maximal de cas de grippes. L'axe des abscisses représente les valeurs de t, de* 2 *à* 500 *semaines et l'axe des ordonnées correspond à* $\hat{b}_t$ *défini en (58).*

Nous souhaitons pour une nouvelle semaine savoir si le nombre de cas de grippes présente une alerte.
Nous procédons comme pour le jeu de données précédent et nous calculons le temps de retour T relatif à cet enregistrement.
Si nous enregistrons pour une nouvelle semaine 1300 cas de grippe et syndromes grippaux, le temps de retour correspondant, lu sur la figure.16, est égal à 6 et le graphe ci-dessus nous montre qu'il n'existe aucun dépassement dans l'intervalle de temps $[142, 148[$.

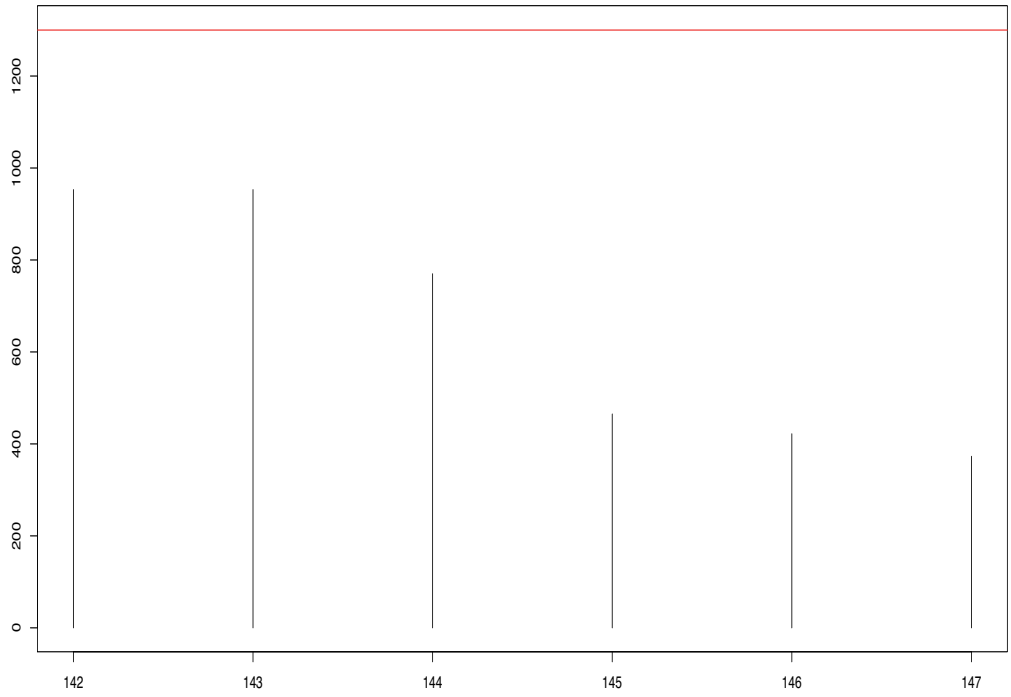

FIGURE 17 – *Graphe de surveillance. L'axe des abscisses représente l'intervalle de temps* $[142, 148[$*, l'axe des ordonnées correspond au nombre de contaminations et la droite rouge représente le nombre de cas de grippes et syndromes grippaux enregistrés au cours de la dernière semaine (*$x_{t_0} = 1300$*).*

Ainsi, $\mathbb{A}_{149} = 0$ et nous ne déclenchons donc pas d'alarme.

3.3.2 Application en finance/actuariat

L'assurance est une opération par laquelle une personne, l'assureur, s'engage à indemniser une autre personne, l'assuré, en cas de réalisation d'un risque déterminé, moyennant le paiement préalable d'une prime ou d'une cotisation.
L'assurance occupe une place très importante dans l'économie moderne : son mécanisme contribue à accroître le niveau de protection de l'ensemble des individus, et sa pratique a été rendue obligatoire dans de très nombreux domaines.
Le mécanisme de l'assurance s'appuie sur la compensation des risques : si tous les assurés sont soumis à un risque, la probabilité de voir celui-ci se réaliser pour tous les assurés est faible. Les victimes sont dédommagées grâce aux contributions versées par la collectivité des cotisants. L'assureur doit donc être capable de prévoir les charges des sinistres qu'il aura à supporter du fait des risques qu'il couvre lorsqu'il établit ses polices d'assurance.
Ces évaluations sont de grande importance pour que la compagnie d'assurance évite la ruine et pour assurer la solvabilité de son portefeuille.
Donc, prévoir l'occurence de telles charges de sinistres est très important pour prendre les précautions nécessaires.
Habituellement, ces évaluations sont effectuées grâce au calcul de probabilités réalisé par un actuaire. La T.V.E permet, en effet, d'établir des scénarios de survenance des charges de sinistres anormaux, ce qui permet à la compagnie d'assurance de considérer ces évenements et garder sa solvabilité.
Le problème est que cette méthodologie considère la loi des charges de sinistres comme continue [8] et ne prend pas en considération la nature discrète des charges de sinistres.
Considérons le modèle simple de Cramér-Lundberg où le montant total des sinistres à l'instant $t \in \mathbb{N}$ est donné par $\sum_{k=1}^{N(t)} W_k$, où $(N(t))$ est un processus de Poisson d'intensité λ et $(W_k)_{k\geq 1}$ sont des v.a. i.i.d. discrètes de loi *Log-discrète* de paramètre α (de fonction de répartition notée F) et indépendantes de $(N(t))$.
Une méthode naïve de mettre en place un système d'alarmes est de considérer des scénarios de survenance de sinistres extrêmes, ce qui revient à estimer un temps de retour T associé à des sinistres très élevés (dans le cas de lois n'ayant pas de queue de distribution lourde ; cf. par exemple [13]).
On cherche donc, dans un premier temps, le niveau de retour $\gamma = \gamma(t)$ d'ordre t tel que

$$\mathbb{E}\Big[\sum_{i=1}^{N(t)} 1_{(W_i>\gamma)}\Big] = 1.$$

Le processus (N_t) étant dans ce modèle indépendant des W_i, on a

$$\mathbb{E}\Big[\sum_{i=1}^{N(t)} 1_{(W_i>\gamma)}\Big] = \mathbb{E}\Big(\mathbb{E}\Big[\sum_{i=1}^{N(t)} 1_{(W_i>\gamma)} \mid N(t)\Big]\Big) = \mathbb{E}[N(t)]P[W_i > \gamma] = \lambda t \bar{F}(\gamma)$$

et par conséquent

$$F(\gamma) = 1 - \frac{1}{\lambda t}$$

ce qui signifie que l'estimation de γ correspond à l'estimation du quantile extrême $p_t = 1 - \frac{1}{\lambda t}$.
En utilisant l'approximation du temps de retour développée en §3.2 et en remarquant que nous pouvons écrire (en utilisant les propriétés du processus de Poisson et l'indépendance de N et des W_i)

$$\mathbb{E}\left(\sum_{i=0}^{N(t)} 1_{(W_i>\gamma)}\right) = \mathbb{E}\left(\sum_{i=N(s)}^{N(s+t)} 1_{(W_i>\gamma)}\right),$$

nous pouvons chercher le temps de retour T associé à la valeur x_{t_0}, observée au temps t_0 en posant $x_{t_0} = \hat{b}_T$.
Ainsi, s'il existe une valeur supérieure ou égale à $\hat{b}_T$ sur l'intervalle $(t_0 - T, t_0)$, on sonne une alarme en t_0.

Résumons la démarche suivie en §3.2 :

1. estimer la borne supérieure $\hat{b}_t$ pour différentes valeurs de t et créer le graphe $\left(t, \hat{b}_t\right)$,
2. lire ensuite sur le graphe de $\hat{b}_t$ le temps de retour T de la dernière observation x_{t_0} tel que $x_{t_0} = \hat{b}_T$.

Notre jeu de données est une simulation de n-observations de charges de sinistres $(X_t)_{t=1,\cdots,n}$ définie en (66) de paramètres $\alpha = 0.9$ et $\lambda = 1$.

La figure.18 et le tableau.3 introduisent notre échantillon de 1000 observations simulées et la figure.19 représente la série temporelle associée.

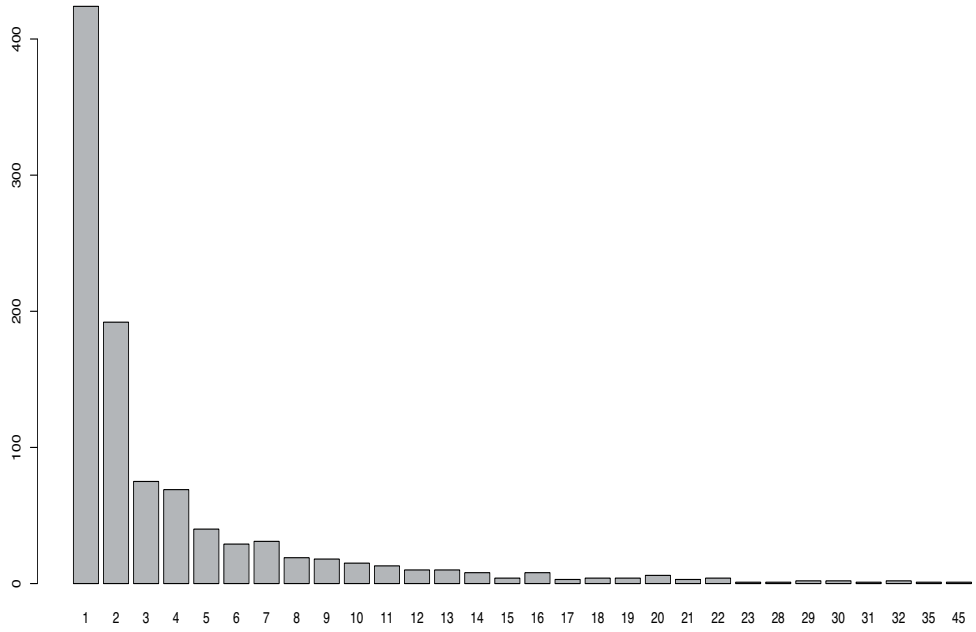

FIGURE 18 – *Diagramme en bâtons du jeu de données : charges des sinistres. L'axe des abscisses représente les charges des sinistres et l'axe des ordonnées correspond à leurs fréquences.*

Minimum	Quantile 25%	Médiane	Moyenne	Quantile 75%	Maximum
1.00	1.00	2.00	3.87	4.00	45.00

TABLE 3 – *Description du jeu de données : charges des sinistres*

Notons la présence de valeurs extrêmes, celles qui se situent au dessus de 8 ; elles représentent 9.5% du nombre total des observations.

Si nous ne tenons pas compte de la nature discrète des données et nous plaçons dans le cadre classique de la théorie des valeurs extrêmes, nous remarquons sur la figure.20 qu'il n'est pas possible de déterminer la valeur exacte de l'indice de queue de distribution ξ.

Il est clair ici, de par l'allure de ces courbes, que l'indice de queue est indéterminable. C'est pourquoi l'utilisation de l'approche de la borne supérieure décrite en §3.2 paraît souhaitable.

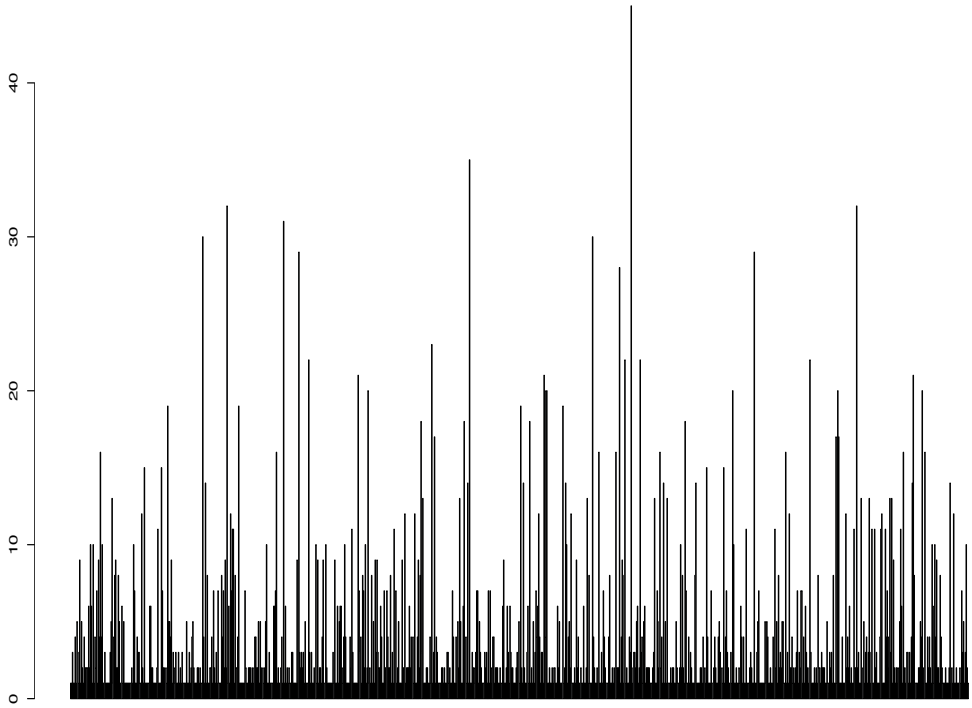

FIGURE 19 – *Les charges de sinistres. L'axe des abscisses représente le temps de* 1 *à* 1000 *et l'axe des ordonnées correspond aux charges de sinistres.*

La figure ci-dessous affiche la courbe de la borne supérieure estimée $\hat{b}_t$ (58) avec $\alpha_{max} = 30$ et $\beta_{max} = 0.2$ pour ce jeu de données (choix obtenu après différentes étapes d'optimisation). Pour chaque valeur de t de 2 à 500, nous calculons via (59) les $\hat{\alpha}_t$ et $\hat{\beta}_t$ optimaux puis $\hat{b}_t$.

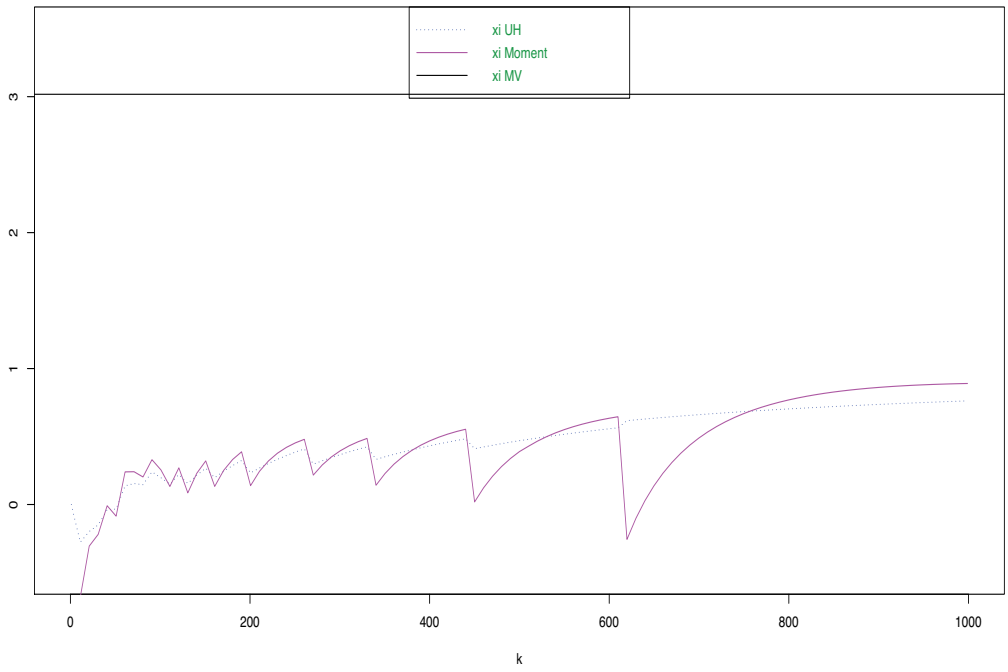

FIGURE 20 – *Estimation de l'indice ξ de la queue de distribution de la loi GEV. L'axe des abscisses représente l'indice k de $\xi_{k,n}$ et l'axe des ordonnées correspond à l'estimateur de l'indice de la queue de distribution $\hat{\xi}_{k,n}$. Les courbes rose, bleue et noire correspondent respectivement aux estimateurs $\hat{\xi}^{Moment}$ défini en (35), $\hat{\xi}^{UH}$ défini en (36) et $\hat{\xi}^{mv}$ obtenu par la méthode de maximum de vraisemblance.*

Pour ce jeu de données, l'actuaire ne s'attend pas à une charge de sinistres de valeur égale ou supérieure à 30 qu'après $T = 40$ (lue sur la figure.21) unités de temps de la dernière charge de sinistres de valeur égale à 30 , c'est à dire à $t = 1008$ sachant que la dernière valeur de charge de sinistres égale à 30 était enregistrée à $t = 968$.

Nous schématisons ce scénario dans la figure ci-dessous.

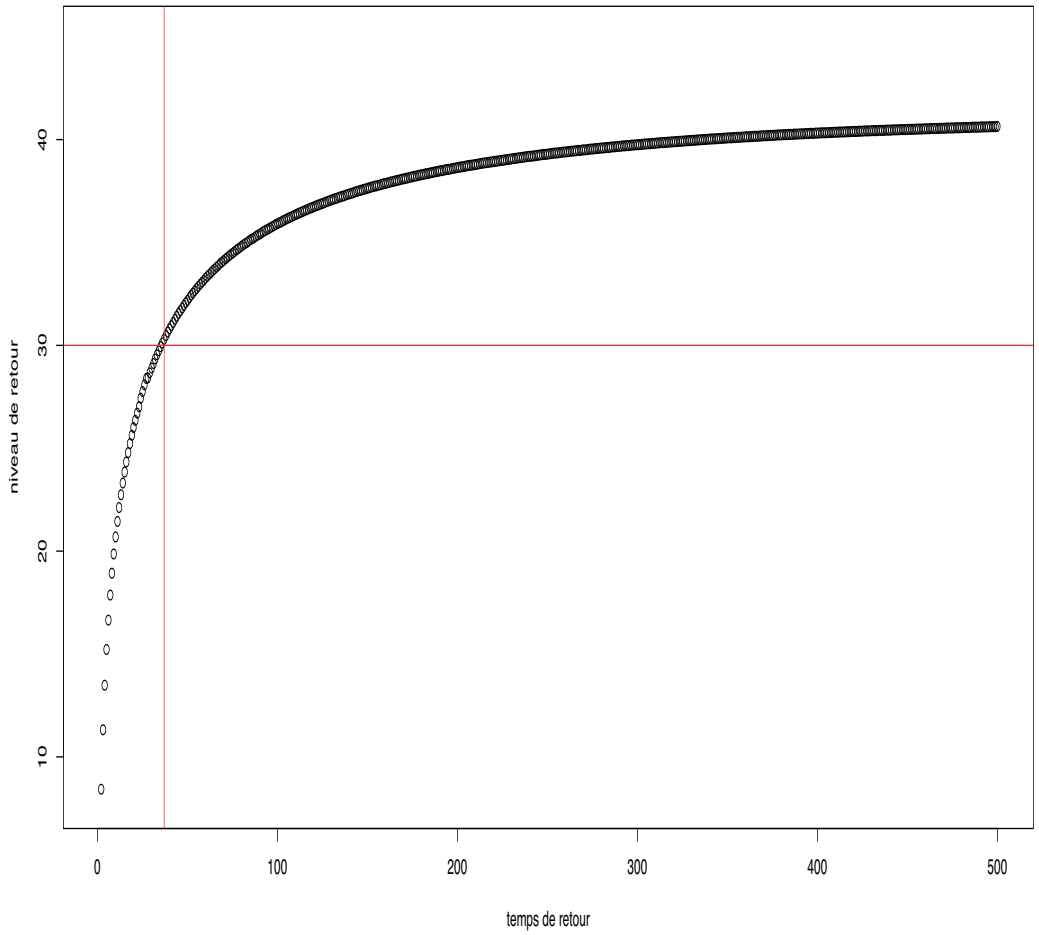

FIGURE 21 – *Les charges de sinistres maximales. L'axe des abscisses représente les valeurs de* t *de* 2 *à* 500 *et l'axe des ordonnées correspond à* $\hat{b}_t$ *défini en (58).*

Selon cette figure, la survenance d'une charge de sinistres supérieure ou égale à 30 serait probable dès l'instant $t = 1009$. L'actuaire devrait donc prendre en considération cet événement et agir afin d'assurer la solvabilité de la compagnie d'assurance.

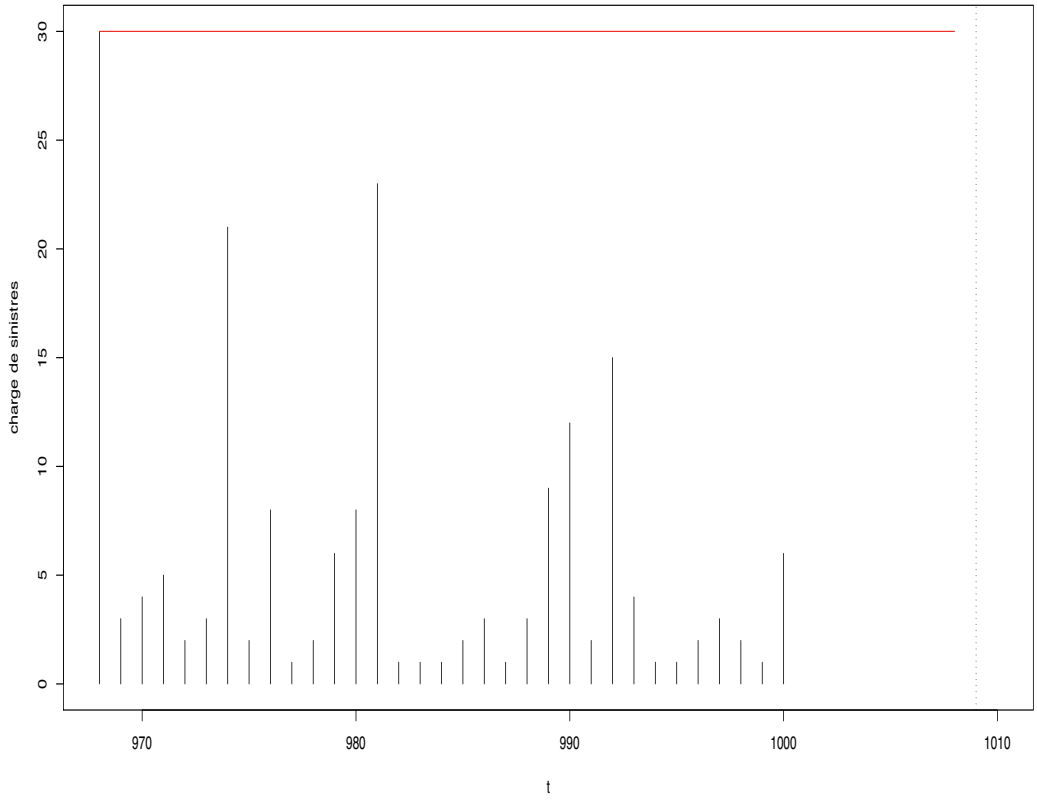

FIGURE 22 – *Temps de retour d'une charge de sinistres de valeur égale à* 30. *L'axe des abscisses représente les valeurs de* t, *l'axe des ordonnées la charge de sinistres correspondante et le segment rouge le niveau de confiance dans un horizon de* $T = 40$.

Une alternative pour détecter les sinistres pouvant mener à la ruine est de considérer la charge de sinistres X_t entre les instants $t-1$ et t $(t \in \mathbb{N}^*)$ définie par

$$X_t = \sum_{k=N(t-1)+1}^{N(t)} W_k. \tag{66}$$

Afin de mettre en place des scénarios de survenance de charges de sinistres extrêmes, nous cherchons dans ce cas à estimer un temps de retour T associé à une charge de sinistres très élevée, en utilisant une nouvelle fois

$$\mathbb{E}\Big[\sum_{k=1}^{t} 1_{(X_k>\gamma)}\Big] = 1$$

ce qui signifie que $\gamma = F_X^{\leftarrow}(1 - 1/t)$, où F_X désigne la fonction de répartition F_X des variables aléatoires i.i.d. X_k, définie par

$$F_X(x) = \sum_{m=0}^{\infty} P[N(1) = m] P[\sum_{i=1}^{m} W_i \leq x].$$

Dans le cas où F_X est discrète, une façon de procéder serait d'adapter la méthode de la borne supérieure au cas conditionnel et de démontrer la convergence de l'estimateur de la borne supérieure dans ce nouveau cadre.

3.3.3 Application en finance (marché financier)

Le marché financier est l'ensemble des lieux où se négocient et s'échangent des capitaux à long terme entre divers intervenants. Dans le vocabulaire courant, l'expression est synonyme de Bourse, ensemble des espaces où se confrontent des offres et des demandes de produits financiers.
Les marchés financiers sont surveillés par une institution qui joue le rôle de régulation et de contrôle en veillant au bon fonctionnement des opérations. Il s'agit de la Commission des Opérations de Bourse (COB), crée en 1967, organisme public indépendant chargé d'assurer la transparence sur les marchés, de sanctionner les éventuelles fraudes constatées et d'envisager des scénarios des crises boursières.
Une quantité importante caractérisant le dynamisme des prix dans les marchés financiers est le volume d'actions échangées. En effet, l'échange et la négociation massive d'actions durant une courte période contribuent à une variation brutale de la valeur des actions.
Estimer un temps de retour du volume d'actions échangées est donc d'une importance primordiale pour la COB.

C'est ce que nous proposons de traiter ici à partir du modèle G.D.M.(*Generalized Discret Model*) de Gontis ([6]) qui généralise le modèle de Kaulakys et Meskauskas ([11]) (en introduisant un $AR\,(2)$ plutôt q'un $AR\,(1)$ dans le modèle ci-dessous).

- Supposons que les transactions se passent aux temps t_k (aléatoires), satisfaisant le modèle :

$$\begin{cases} t_k = t_{k-1} + \tau_k \\ \tau_k = \tau_{k-1} + \alpha\,(\tau_{k-1} - \tau_{k-2}) - \gamma\,(\tau_{k-2} - \bar{\tau}) + \epsilon_k \end{cases}$$

  où $(\epsilon_k)_{k\in\mathbb{N}} : BB\left(0, \sigma^2\right)$, $\sigma^2 > 0$ et où α, $\gamma << 1$, $\bar{\tau}$ (moyenne des τ_k) sont des paramètres donnés ([6]) et τ_k représentent des fluctuations dûes à des perturbations extérieures (aléatoires).
- Nous observons le volume total V d'actions échangées sur une période de longueur N telle que $N = n\tau$.

– Nous observons alors le volume $(V_r)_{r=1,\dots,n}$ d'actions échangées sur chacun des intervalles (de longueur τ), $I_\tau(r) = [r\tau, (r+1)\tau[$, $r = 1, \dots, n$. Nous avons donc :

$$V_r = \sum_{k=1}^{N_r} q_k$$

où N_r est le nombre (aléatoire) de transactions sur l'intervalle $[r\tau, (r+1)\tau[$ et q_k est le volume d'actions par transaction.

– Supposons, comme dans [6], que les q_k peuvent être remplacés par leur moyenne notée $\langle q_k \rangle$, ainsi $V_r = N_r \langle q_k \rangle$.

Nous allons reprendre le même jeu de données que celui de [6] et donc simuler n-observations discrètes du volume d'actions échangées $(V_t)_{t=1,\cdots,n}$ où $n = 1000$, $\alpha = 0.75$, $\gamma = 0.0007$, $\bar{\tau} = 0.01$, $\sigma = 0.00015$, $\tau = 1$ jour et $\langle q_k \rangle = 1$.
La figure.23 et le tableau.4 introduisent notre échantillon de 1000 observations simulées et la figure.24 représente la série temporelle associée.

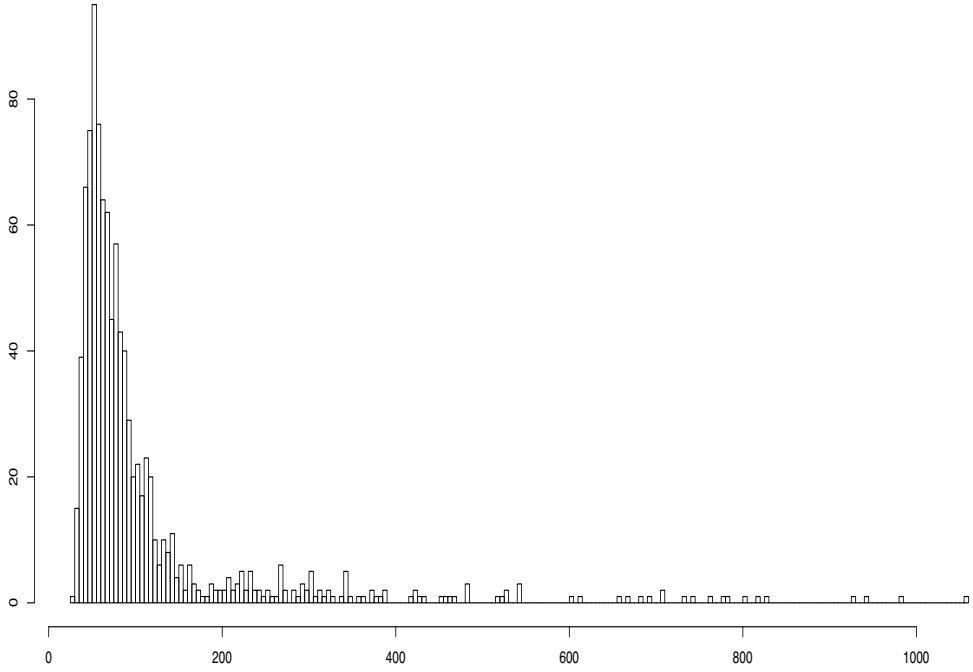

FIGURE 23 – *Diagramme en bâtons du jeu de données : Volume d'actions échangées. L'axe des abscisses représente le volume d'actions échangées et l'axe des ordonnées correspond à leur fréquence.*

Minimum	Quantile 25%	Médiane	Moyenne	Quantile 75%	Maximum
28.00	53.00	71.00	111.60	106.00	1060.00

TABLE 4 – *Description du jeu de données : Volume d'actions échangées*

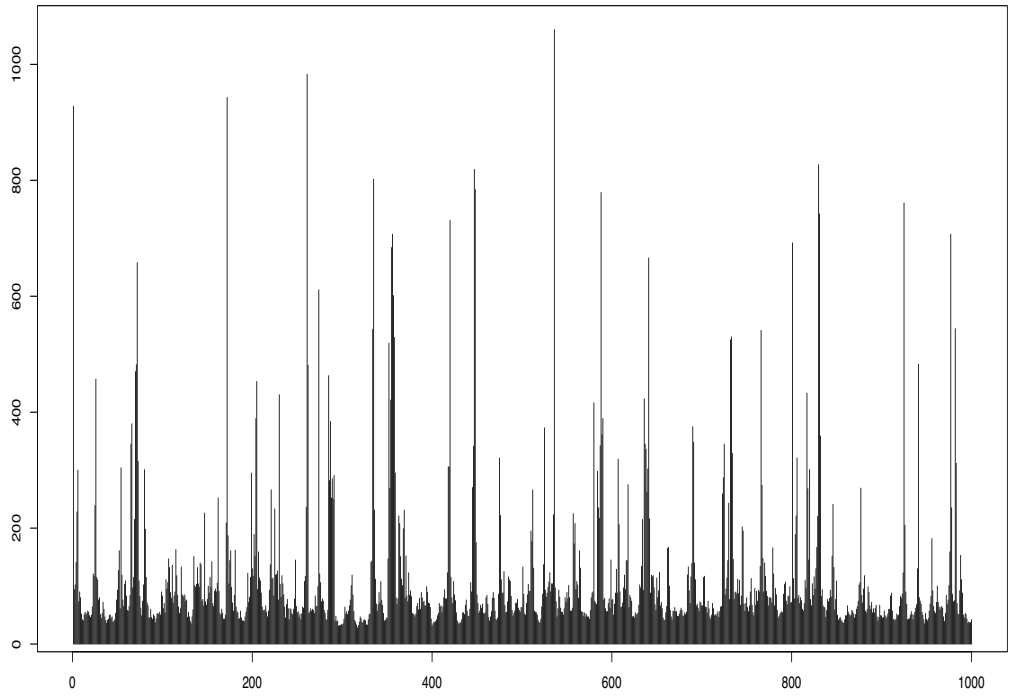

FIGURE 24 – *La série temporelle associée au volume d'actions échangées. L'axe des abscisses représente les journées boursières et l'axe des ordonnées correspond au volume d'actions échangées au jour t.*

Notre objectif est de déterminer dans un premier temps un niveau de retour ou un quantile extrême du volume d'actions échangées d'ordre très élevé en l'estimant par la borne supérieure du niveau de retour et dans un second temps, de calculer un temps de retour T du volume d'actions échangées de valeurs très élevées.
La figure.25 affiche la courbe de la borne supérieure estimée $\hat{b}_t$ (58) (avec $\alpha_{max} = 8$ et $\beta_{max} = 1$ pour ce jeu de données). Pour chaque valeur de t de 5 à 500, nous calculons via (59) les $\hat{\alpha}_t$ et $\hat{\beta}_t$ optimaux puis $\hat{b}_t$ et obtenons le graphe $\left(t, \hat{b}_t\right)$ suivant :

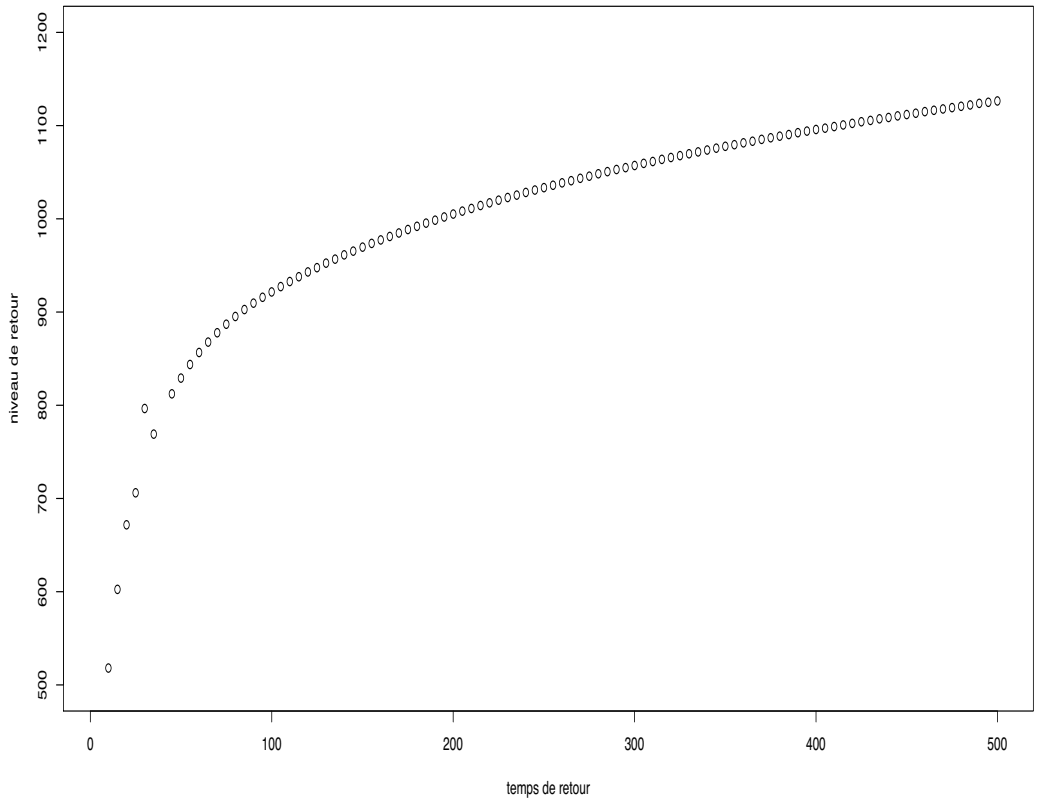

FIGURE 25 – *Le volume maximal d'actions échangées. L'axe des abscisses représente les valeurs de t de* 5 *à* 500 *journées boursières, correspondant aussi aux ordres des quantiles extrêmes de* 80% *à* 99.80% *et l'axe des ordonnées correspond à* $\hat{b}_t$ *défini en (58).*

En suivant la démarche décrite précédemment (§3.2), nous pouvons déduire une approximation du temps de retour T du volume d'actions échangées.

La figure.26 présente les temps de retour de certaines valeurs extrêmes du volume d'actions échangées, lus sur la figure.25 et représentés par des segments de droite horizontaux.

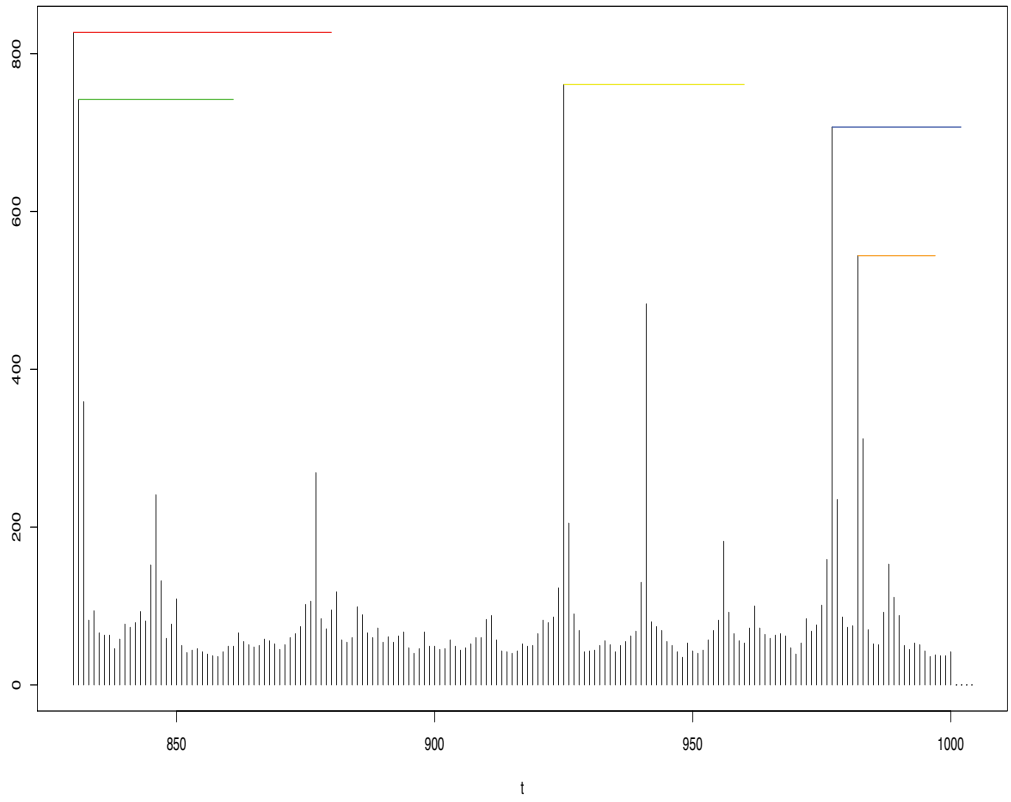

FIGURE 26 – *Temps de retour du volume d'actions échangées. L'axe des abscisses représente les journées boursières* $t = 830, \ldots, 1000$ *et l'axe des ordonnées correspond au volume d'actions échangées au jour* t*. Les segments de droite horizontaux représentent les intervalles de temps* $[t_0, t_0 + T[$ *de longueur* T*, le temps de retour associé au volume d'actions échangées à* t_0*.*

Selon cette figure, un volume d'actions échangées de valeur supérieure ou égale à $x_{977} = 707$ ne serait probable qu'après $T = 25$ jours (ce scénario est représenté par le segment de droite bleu). Ainsi, si la COB enregistre un volume d'actions échangées supérieure ou égale à $x_{977} = 707$ dans l'intervalle de temps $[977, 1002[$, elle devrait être en mesure d'agir et de détecter les éventuelles fraudes.

Remarque : un système de surveillance/détection de fraudes pourrait être mis en place comme cela été traité pour les exemples d'épidémiologie.

4 La T.V.E. dans le cas de lois discrètes

Nous avons proposé dans la section précédente une approximation du quantile extrême d'une distribution de loi F discrète par une borne supérieure du niveau de retour. Cette approche peut être satisfaisante pour certaines applications, mais demeure approximative puisque le quantile est surestimé (étant donné qu'il est estimé par une borne supérieure). Nous allons donc envisager une autre méthode permettant de nous ramener à des variables continues afin de pouvoir appliquer la T.V.E..
Cette méthode repose sur l'estimation de la densité spectrale du modèle discret (via un périodogramme lissé), permettant ainsi de passer d'un caractère discret à un caractère continu pour les données.
Cette estimation induit l'apparition de deux paramètres, un paramètre de lissage et d'une fenêtre spectrale (*spectral window*), tous deux jouant un rôle crucial dans la vitesse de convergence de l'estimateur obtenu.
Après lissage par noyau, nous sommes ramemés à un cadre continu et pouvons alors appliquer les résultats de la T.V.E obtenus dans § 2 afin d'estimer le quantile extrême et le temps de retour.

4.1 Estimation de la densité spectrale

Soit $x := (x_1, \ldots, x_n)' \in \mathbb{C}^n$; le produit scalaire de deux éléments u et v de $\mathbb{C}$ est noté

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i \bar{v}_i$$

où $\bar{v}_i$ est le conjugué de v_i.
Nous pouvons exprimer x selon

$$x = \sum_{j \in D_n} a_j e_j$$

avec $e_j = \frac{1}{\sqrt{n}} \left(\exp\left(i\omega_j\right), \exp\left(i2\omega_j\right), \cdots, \exp\left(in\omega_j\right)\right)'$, $\omega_j = 2\pi j/n$, $-\pi < \omega_j \leq \pi$
(appelées fréquences de Fourrier de la série $(x_1, \ldots, x_n)$), $D_n = \left\{j \in \mathbb{Z} : -\pi < \omega_j = 2\pi j/n \leq \pi\right\} =$
$\left\{-\left[(n-1)/2\right], \cdots, [n/2]\right\}$ et $[x]$ est la partie entière de x.
Les a_j sont les coefficients de x dans la base $\left(e_j\right)_j : a_j = \langle x, e_j \rangle = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t \exp\left(-it\omega_j\right)$.

Nous définissons le périodogramme de x à la fréquence $\omega_j = 2\pi j/n$, $j \in D_n$ selon

$$I\left(\omega_j\right) = \left|a_j\right|^2 = \left|\langle x, e_j \rangle\right|^2 = n^{-1} \left|\sum_{t=1}^{n} x_t \exp\left(-it\omega_j\right)\right|^2. \tag{67}$$

Ainsi, le périodogramme de $(X_t)_{t=1,\cdots,n}$ relatif aux fréquences $\omega_j = 2\pi j/n$, $\omega_j \in [-\pi, \pi]$ est défini par

$$\begin{cases} I_n(0) = n\left|\overline{X}\right|^2 & \text{si } \omega_j = 0, \\ I_n(\omega_j) = \sum_{|k|<n} \hat{\gamma}(k) \exp(-ik\omega) & \text{si } \omega_j \neq 0 \end{cases} \tag{68}$$

où $\hat{\gamma}(k) := n^{-1} \sum_{t=1}^{n-|k|} \left(X_t - \overline{X}\right)\left(X_{t+|k|} - \overline{X}\right)$ $(\star)$ et $\overline{X} = n^{-1} \sum_{t=1}^{n} X_t$.
Le périodogramme est utilisé comme un estimateur de la densité spectrale définie pour un processus stationnaire de moyenne μ et de fonction d'autocovariance γ par :

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} \gamma(k) \exp(-ik\omega), \quad \omega \in [-\pi, \pi]. \tag{69}$$

$I_n(\omega_j)$ est certes un estimateur de $f(\omega_j)/2\pi$, mais il n'est pas convergeant.
Construisons alors un autre estimateur de la densité spectrale de type noyau, en lissant le périodogramme. Cet estimateur a l'avantage d'être asymptotiquement consistant (sous certaines conditions, il converge en moyenne quadratique [15]) et s'exprime selon :

$$\hat{f}_L(\omega) = \frac{1}{2\pi} \sum_{|k|\leq r} w(k/r)\, \hat{\gamma}(k) \exp(-ik\omega) \tag{70}$$

où $\hat{\gamma}(k)$ est définie en $(\star)$ et $w(.)$, appelée *lag window*, est une fonction paire et continue par morceaux satisfaisant les conditions suivantes :

i- $w(0) = 1$,

ii-$|w(x)| \leq 1$ pour tout x,

iii-$w(x) = 0$ pour $|x| > 1$,

et r est le paramètre de lissage que l'on se donne.
Nous définissons la fenêtre spectrale (*spectral window*) W associée à w selon

$$W(\omega) = \frac{1}{2\pi} \sum_{|k|\leq r} w(k/r) \exp(-ik\omega) \tag{71}$$

Le paramètre r gouverne le niveau de lissage ainsi que le compromis entre biais et variance :

- quand r tend vers 0, il n'y a pas de pénalité de lissage et le modèle fournit un ajustement parfait : les valeurs ajustées sont les données elles-mêmes.
- quand r est grand, l'ajustement est un lissage parfait : les valeurs ajustées tombent le long d'une ligne droite forçant ainsi la liaison à être linéaire. Il en résulte une faible variance mais un biais important.

Le paramètre r doit être compris entre ces deux cas extrêmes pour produire le niveau désiré de lissage afin d'avoir un bon ajustement du modèle avec un équilibre biais - variance. Nous choisissons en théorie la plus faible valeur de r qui minimise l'erreur quadratique moyenne (E.Q.M) ($biais^2 + variance$), mais c'est difficilement réalisable ([2], page 193).
En pratique, nous pouvons :

1. choisir une valeur de r selon le niveau de lissage désiré (nous testerons plusieurs valeurs),
2. puis, choisir un type de fenêtre spectrale W parmi les modèles proposés ci-dessous tel que l'E.Q.M de $\hat{f}_L$ soit rendue minimale [10].

Exemple 1 *: la fenêtre rectangulaire*
Cette fenêtre a pour lag window

$$w = \begin{cases} 1 & \text{si } |x| \leq 1, \\ 0 & \text{sinon} \end{cases}$$

et sa fenêtre spectrale correspond au noyau de ***Dirichlet*** *défini selon*

$$W(\omega) = \frac{1}{2\pi} \frac{\sin\left((r+1/2)\,\omega\right)}{\sin\left(\omega/2\right)}. \tag{72}$$

Exemple 2 *: la fenêtre triangulaire ou de Bartlett*
Dans ce cas

$$w = \begin{cases} 1 - |x| & \text{si } |x| \leq 1, \\ 0 & \text{sinon} \end{cases}$$

et la fenêtre spectrale correspondante est donnée par le noyau de ***Fejer***

$$W(\omega) = \frac{1}{2r\pi} \frac{\sin^2\left(r\omega/2\right)}{\sin^2\left(\omega/2\right)}. \tag{73}$$

Exemple 3 *: la fenêtre de Daniell*
La fenêtre spectrale de ***Daniell*** *est définie selon*

$$W = \begin{cases} r/2\pi & \text{si } |\omega| \leq \pi/r, \\ 0 & \text{sinon} \end{cases} \tag{74}$$

relative à ce lag window

$$w = \frac{\sin\left(\pi x\right)}{\pi x}, \quad -1 \leq x \leq 1.$$

4.2 Génération des variables continues via la densité spectrale

Nous pouvons schématiser le processus de génération des variables continues via la densité spectrale selon :

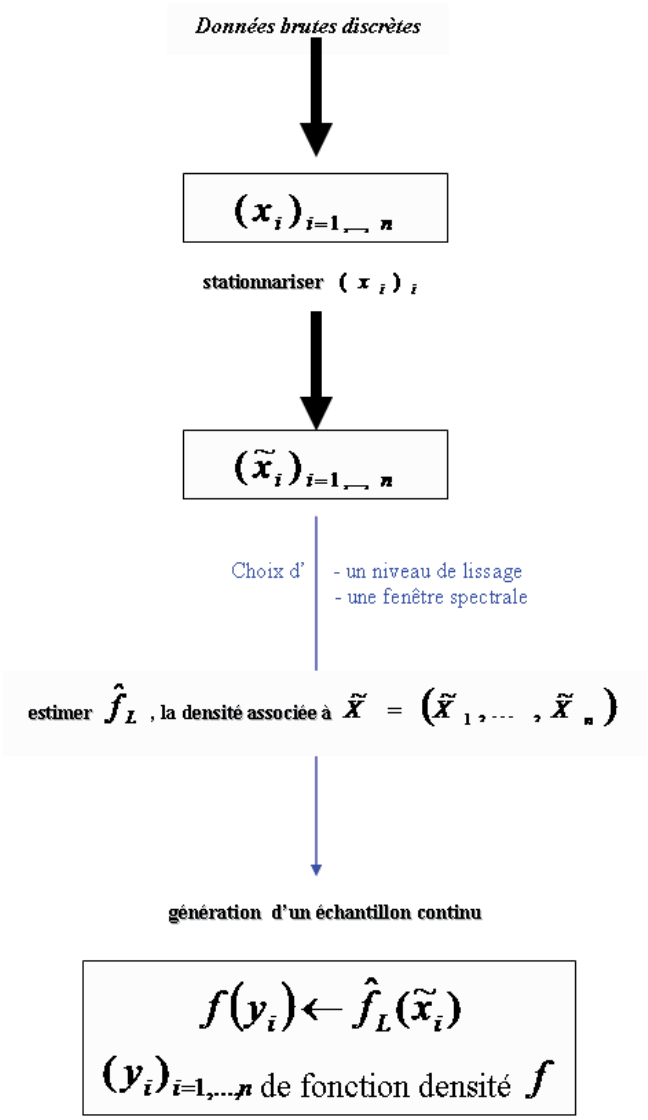

FIGURE 27 – *Procédure de génération des variables continues via la densité spectrale*

Désormais, nous pouvons appliquer les résultats de la T.V.E. sur l'échantillon $(Y_i)_{i=1,\dots,n}$ à n-observations $(y_i)_{i=1,\dots,n}$, méthode décrite en §2.4 pour estimer le niveau et le temps de retour.

4.3 Applications

Nous allons à présent suivre la procédure introduite dans la section précédente aux jeux de données traités dans la section §3.3.
Notons que la méthode de lissage fait apparaître des valeurs mathématiquement acceptables mais à réinterpréter en pratique.

4.3.1 Application en épidémiologie

a) Salmonelle Derby

Ce jeu de données, décrit en §3.3.1.a), concerne le nombre de souches de Salmonelle Derby. Nous voulons savoir si le dernier enregistrement du nombre de souches est anormalement élevé et tester la présence d'une alarme en t_0 définie selon $\mathbb{A}_{t_0}$ en (65).
Nous transformons la série d'observations $(x_t)_t$ en une série continue $(y_t)_t$ selon le schéma donné en figure.27.
Nous choisissons un niveau de lissage $r = 5$ et nous calculons l'E.Q.M. de l'estimateur de la densité spectrale $\hat{f}_L$ selon le type de fenêtre considéré (§4.1).

fenêtre spectrale	fenêtre rectangulaire	fenêtre triangulaire	fenêtre de Daniell
EQM	0.005226491	0.01220235	0.02003714

TABLE 5 – *L'erreur quadratique moyenne de $\hat{f}_L$.*

D'après ce tableau, nous choisissons la fenêtre rectangulaire pour calculer $\hat{f}_L$ et obtenons ainsi un nouvel échantillon d'une série temporelle continue $(Y_t)_{i=t,\ldots,598}$ que nous représentons ci-dessous ainsi que la série discrète initiale (figure.28).

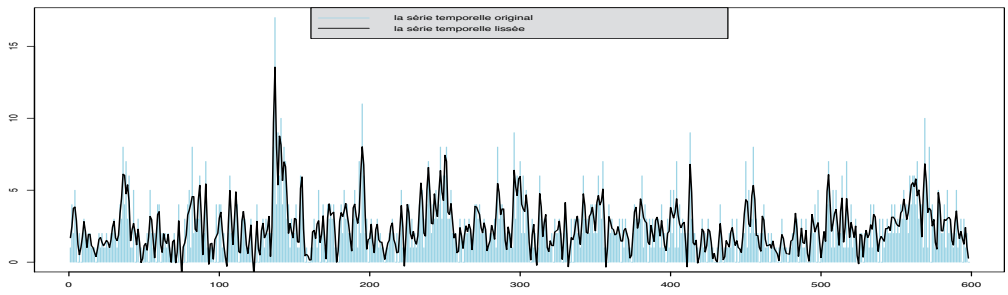

FIGURE 28 – *La série temporelle lissée.L'axe des abscisses représente les semaines et l'axe des ordonnées correspond au nombre hebdomadaire des souches.*

L'échantillon de la série $(Y_t)_{i=t,\dots,598}$ associé au jeu de données Salmonelle Derby est constitué de 598 observations hebdomadaires et est décrit selon :

Minimum	Quantile 25%	Médiane	Moyenne	Quantile 75%	Maximum
-0.8746	1.3670	2.1710	2.4500	3.2940	13.5600

TABLE 6 – *Description du jeu de données : nombre de souches de Salmonelle Derby (série lissée)*

La loi des $(Y_t)_t$ étant continue, nous pouvons calculer le temps de retour selon les deux lois *GEV* et *GPD* définies respectivement en (46) et (47).
Nous nous intéressons dans un premier temps au calcul de $\widehat{T}^{GEV}$.
La figure.29 affiche les estimateurs de queue de distribution $\hat{\xi}^{Moment}$ défini en (35), $\hat{\xi}^{UH}$ défini en (36) et $\hat{\xi}^{MV}$ l'estimateur de maximum de vraisemblance.

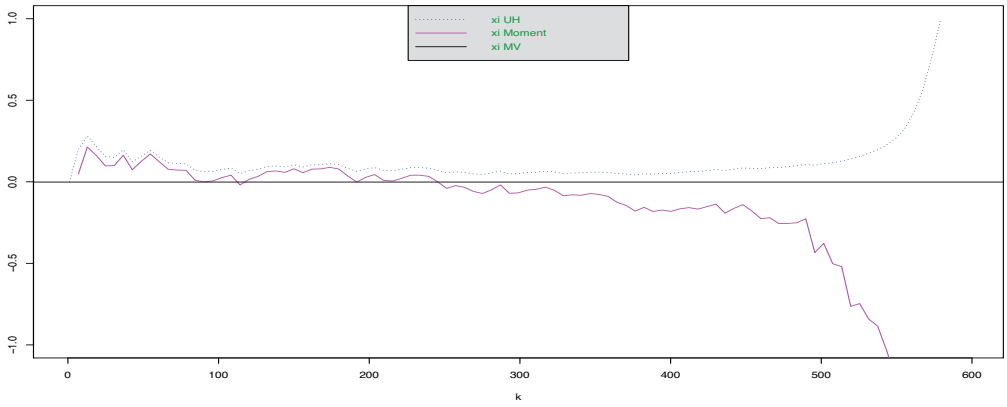

FIGURE 29 – *Estimation de l'indice ξ de la queue de distribution de la loi GEV.L'axe des abscisses représente l'indice k de $\hat{\xi}_{k,n}$ et l'axe des ordonnées correspond à l'estimateur de l'indice de la queue de distribution. Nous pouvons conclure que $\hat{\xi} \approx 0$.*

Ainsi, nous calculons le temps de retour $\widehat{T}^{GEV}$ associé à l'observation lissée y_{t_0} à t_0.
Plaçons nous au temps $t_0 = 599$. Nous observons la valeur $x_{599} = 10$, à laquelle correspond $y_{599} = 9.165$, l'observation lissée à t_0. Nous obtenons alors $\widehat{T}^{GEV} = 621.97 \approx 622$.

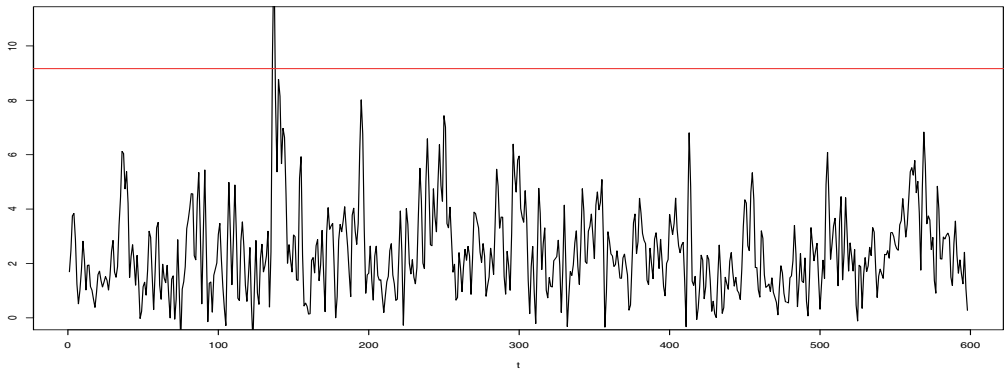

FIGURE 30 – *Graphe de surveillance du nombre de souches de Salmonelle Derby selon la méthode GEV.L'axe des abscisses représente l'intervalle de temps* $[1, 599[$*, l'axe des ordonnées correspond au nombre de contaminations lissé et la droite rouge représente* $y_{599} = 9.165$ *associé au nombre original de souches* x_{599}*, enregistré au cours de la dernière semaine.*

Nous détectons une observation supérieure ou égale à $y_{599} = 9.165$ dans l'intervalle de temps $[1, 599[$. Ainsi, nous devrons générer une alarme en $t_0 = 599$, ie $\mathbb{A}_{599} = 1$.
Reprenons le calcul du temps de retour, mais cette fois selon la loi *GPD*.

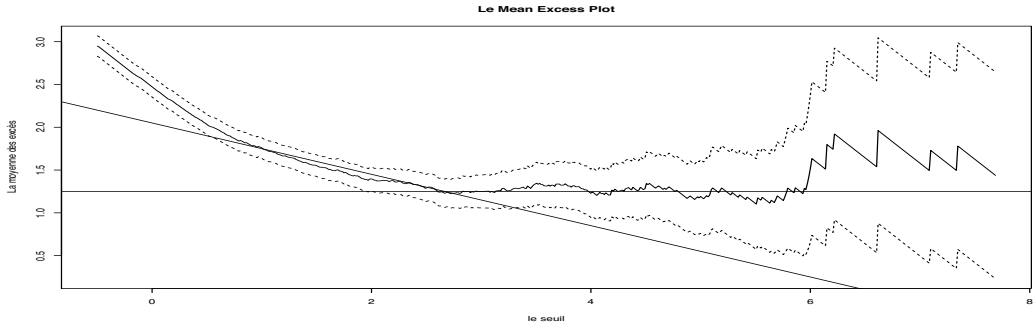

FIGURE 31 – *Mean Excess PLot L'axe des abscisses représente le seuil u du nombre hebdomadaire des souches et l'axe des ordonnées correspond à la moyenne des excès définie en (23).Les deux courbes pointillées indiquent les bornes de l'intervalle de confiance à* 95% *de la moyenne des excès.*

Le graphe du *Mean Excess Plot* (figure.31) nous donne une approximation du seuil u, introduit dans §2.2.1. Nous pouvons déduire qu'il est égal à 3 : c'est le seuil pour lequel la fonction moyenne des excès change de pente.
Les estimateurs des indices de la loi *GPD*, obtenus par la méthode de maximum de vraisemblance et définis en (37) et (38), sont égaux à $\left(\hat{\xi}, \hat{\sigma}\right) = (0.02, 1.34)$. Ainsi, le temps de retour de $y_{599} = 9.165$ selon la loi GPD, calculé via (47), est égal à $\widehat{T}^{GPD} = 412.13 \approx 412$.

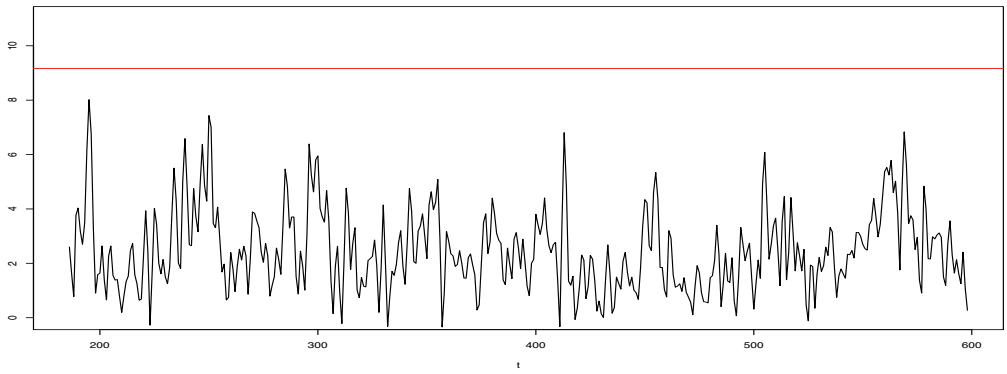

FIGURE 32 – *Graphe de surveillance du nombre de souches de Salmonelle Derby selon la méthode GPD.L'axe des abscisses représente l'intervalle de temps* $[186, 599[$*, l'axe des ordonnées correspond au nombre de contaminations lissé et la droite rouge représente* $y_{599} = 9.165$ *associé au nombre original de souches* x_{599} *enregistré au cours de la dernière semaine.*

Contrairement à ce que nous avons trouvé précédemment, nous ne remarquons aucun dépassement de la valeur $y_{599} = 9.165$ dans l'intervalle de temps $[186, 599[$. Ainsi, nous ne générons pas d'alarme en $t_0 = 599$ (ie $\mathbb{A}_{599} = 0$).

b) Salmonelle Agona

Reprenons le jeu de données décrit dans §3.3.1.b) concernant le nombre de souches de Salmonelle Agona.
Nous voulons tester si les observations, dès la première semaine de l'année 2000, présentent une alarme.
Comme précédemment, nous transformons la série d'observations $(x_t)_t$ en une série continue $(y_t)_t$, selon la figure.27, en choisissant un niveau de lissage $r = 5$ et la fenêtre rectangulaire pour calculer $\hat{f}_L$. Nous calculons ensuite le temps de retour selon la loi *GEV*, défini en (46), et mettons en évidence les différentes alarmes générées sur les sous-échantillons obtenus selon la méthode schématisée en figure.7 (avec $k = 3$, $t = 1, \ldots, 5$ et $seq = 52$).
Représentons sur le même graphe (figure.33) les alarmes générées par les trois méthodes, Farrington, la méthode de la borne supérieure et la méthode GEV, en travaillant sur les mêmes sous-échantillons constitués en tenant compte de la saisonnalité.

La méthode *GEV* génère un nombre plus limité d'alarmes que la méthode de la borne supérieure. En éliminant les alarmes qui sont produites par un nombre limité de souches sans qu'au moins 5 souches n'aient été vues au cours des 4 semaines précédentes, les alarmes générées par la méthode *GEV* sont similaires aux alarmes générées par la méthode anglaise.
Comme pour la méthode de la borne supérieure, les nombres élevés de souches observés durant deux mois en 2005, correspondant à un réel événement épidémiologique inhabituel, sont détectés par cette méthode.
Pour cette série temporelle, les deux méthodes ont une bonne sensibilité et ne génère pas trop d'alarmes, rendant ainsi l'interprétation des alarmes plus aisée.

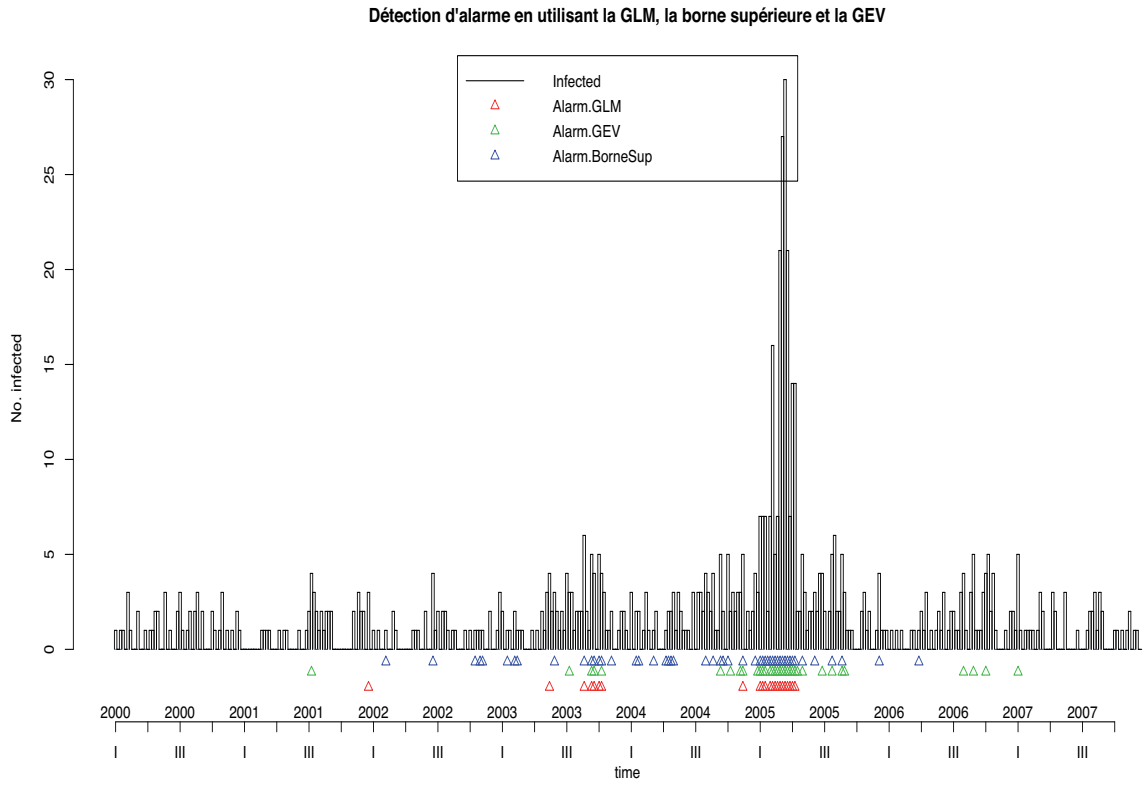

FIGURE 33 – *Détection des alarmes générées selon la GLM, la GEV et la borne supérieure. Les triangles rouges représentent les alarmes générées selon l'algorithme de Farrington, les triangles verts représentent les alarmes générées en calculant le temps de retour selon la loi GEV appliquée à la série lissée et les triangles bleus représentent les alarmes générées selon la méthode de la borne supérieure obtenues en* §3.3.1.*b*).

4.3.2 Application en finance/actuariat

Reprenons le jeu de données décrit en §3.3.2, concernant n-observations simulées de charges de sinistres et rappelons que notre objectif est l'estimation d'un temps de retour associé à une charge de sinistres très élévée.
Nous transformons la série d'observations en une série continue $(y_t)_t$ selon la figure.34. Nous choisissons un niveau de lissage $r = 5$ et nous calculons l'E.Q.M. de l'estimateur de la densité spectrale $\hat{f}_L$ selon le type de fenêtre considéré (§4.1).

fenêtre spectrale	fenêtre rectangulaire	fenêtre triangulaire	fenêtre de Daniell
EQM	0.005359713	0.009655344	0.02052883

TABLE 7 – *L'erreur quadratique moyenne de* $\hat{f}_L$.

D'après ce tableau, nous choisissons la fenêtre rectangulaire pour calculer $\hat{f}_L$ et obtenons ainsi un nouvel échantillon d'une série temporelle continue $(Y_t)_{i=t,\dots,1000}$ que nous représentons ci-dessous ainsi que la série discrète initiale (figure.34).

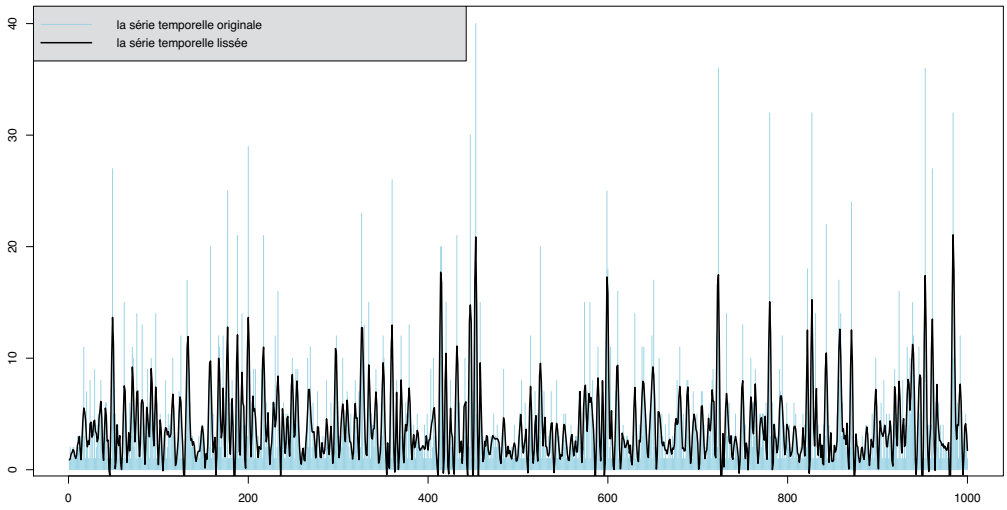

FIGURE 34 – *La série temporelle lissée.L'axe des abscisses représente le temps de* 1 *à* 1000 *et l'axe des ordonnées correspond aux charges de sinistres.*

L'échantillon de la série lissée $(Y_t)_{i=t,\dots,1000}$ est constitué de 1000 observations de charges de sinistres et décrit selon :

Minimum	Quantile 25%	Médiane	Moyenne	Quantile 75%	Maximum
-4.447	1.464	2.940	3.887	5.341	25.710

TABLE 8 – *Description du jeu de données : charges de sinistres (série lissée)*

La loi des $(Y_t)_t$ étant continue, nous pouvons calculer le temps de retour selon les deux lois *GEV* et *GPD* définies respectivement en (46) et (47).
Nous nous intéressons dans un premier temps au calcul de $\widehat{T}^{GEV}$.
La figure.35 affiche les estimateurs de queue de distribution $\hat{\xi}^{Moment}$ défini en (35), $\hat{\xi}^{UH}$ défini en (36) et $\hat{\xi}^{MV}$ l'estimateur de maximum de vraisemblance.

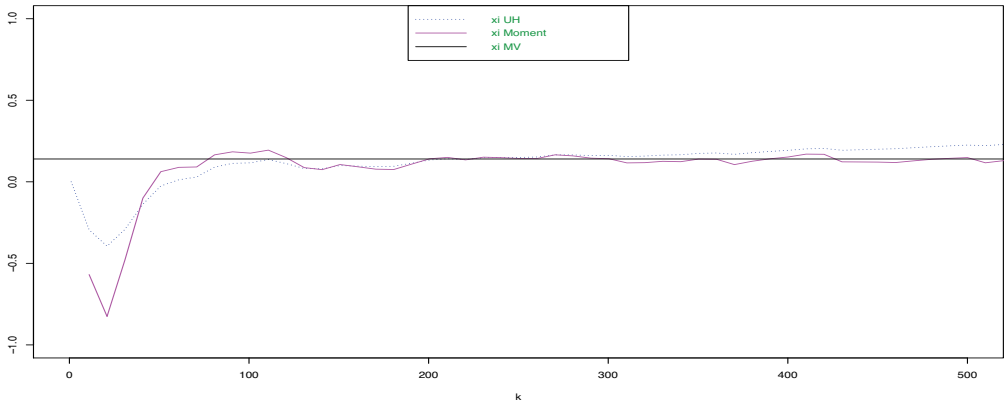

FIGURE 35 – *Estimation de l'indice ξ de la queue de distribution de la loi GEV.L'axe des abscisses représente l'indice k de $\hat{\xi}_{k,n}$ et l'axe des ordonnées correspond à l'estimateur de l'indice de la queue de distribution.*

D'après ce graphe, nous pouvons conclure que $\hat{\xi} = 0.14$.
Calculons alors le temps de retour $\widehat{T}^{GEV}$ associé à l'observation lissée y_{t_0} à t_0.

Plaçons nous au temps $t_0 = 984$. Nous observons la valeur $x_{984} = 32$ à laquelle correspond $y_{984} = 21.06$.
Nous obtenons alors $\widehat{T}^{GEV}_{984} = 147.89 \approx 148$, le temps de retour associé à t_0.

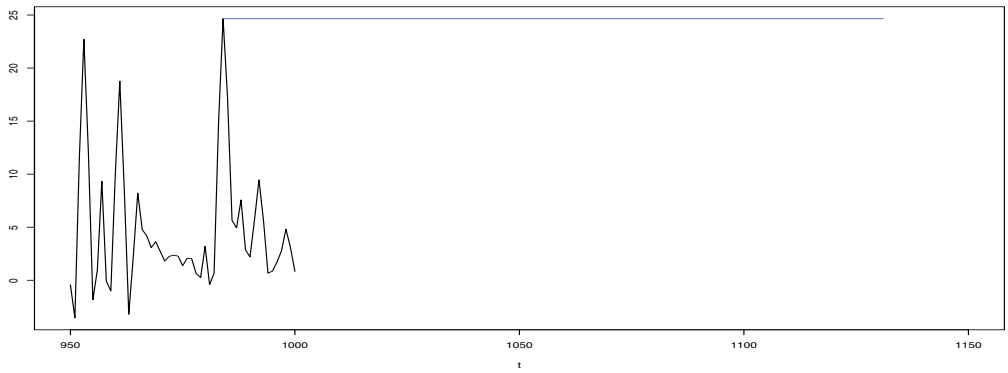

FIGURE 36 – *Temps de retour d'une charge de sinistres selon la méthode GEV.L'axe des abscisses représente l'intervalle de temps* $[950, 1150[$*, l'axe des ordonnées correspond au charge de sinistres lissé et le segment de droite horizontal représente l'intervalle de temps de longueur* $\widehat{T}_{984}^{GEV}$.

Selon cette figure, la survenance d'une charge de sinistre supérieure ou égale à 32 serait probable dès l'instant $t = 1132$, l'actuaire devrait donc prendre en considération cet événement et agir afin d'assurer la solvabilité de la compagnie d'assurance.
Reprenons le calcul du temps de retour, mais cette fois selon la loi *GPD*.

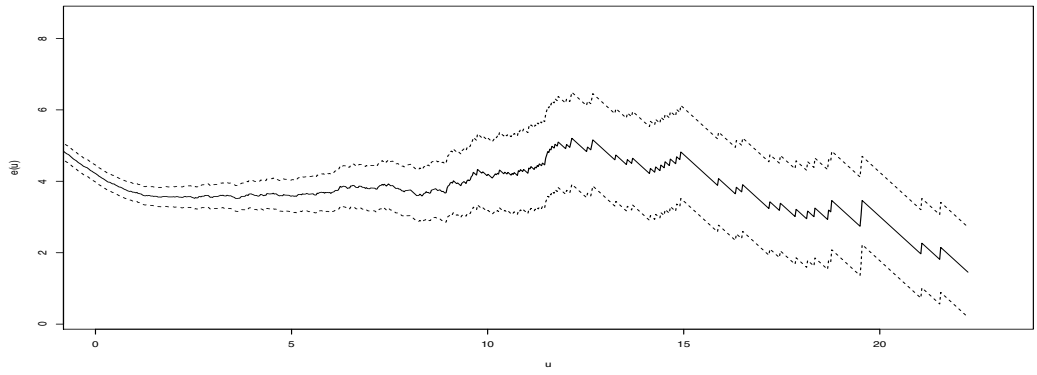

FIGURE 37 – *Mean Excess PLot L'axe des abscisses représente le seuil u du charge de sinistres et l'axe des ordonnées correspond à la moyenne des excès définie en (23).*

Le graphe du *Mean Excess Plot* (figure.37) nous donne une approximation du seuil u, introduit dans §2.2.1. Nous pouvons déduire qu'il est égal à 8 : c'est le seuil pour lequel la fonction moyenne des excès change de pente.
Les estimateurs des indices de la loi *GPD*, obtenus par la méthode de maximum de vraisemblance et définis en (37) et (38), sont égaux à $\left(\hat{\xi}, \hat{\sigma}\right) = (0.0892, 3.4289)$. Ainsi, le temps de retour de $y_{984} = 21.06$ selon la loi GPD, calculé via (47), est égal à $\widehat{T}_{984}^{GPD} = 221.22 \approx 221$.

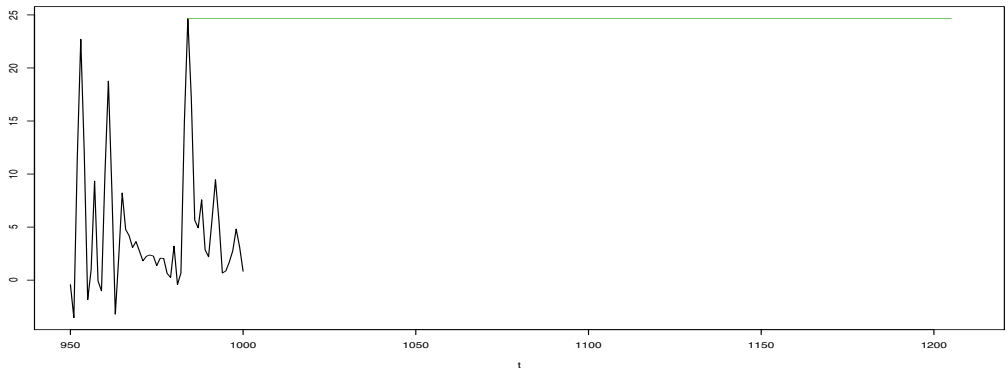

FIGURE 38 – *Temps de retour d'une charge de sinistres selon la méthode GPD. L'axe des abscisses représente l'intervalle de temps* $[950, 1210[$*, l'axe des ordonnées correspond au charge de sinistres lissé et le segment de droite horizontale représente l'intervalle de temps de longueur* $\widehat{T}_{984}^{GPD}$.

Ici, la survenance d'une charge de sinistre supérieure ou égale à 32 serait probable dès l'instant $t = 1205$, un temps qui est supérieur à celui trouvé selon la méthode *GPD* ($t = 1132$). Notons aussi que ces deux temps sont supérieurs au temps obtenu par la méthode de la borne supérieure développée en §3.2, égal à $t = 1091$.

4.3.3 Application en finance (marché financier)

Ce jeu de données décrit en §3.3.3, concerne le volume V_r d'actions échangées, simulé via le modèle G.D.M. de Gontis ([6]).
A nouveau, nous proposons d'estimer le temps de retour du volume d'actions échangées selon les lois *GEV* et *GPD* et de le comparer avec celui obtenu par la méthode de la borne supérieure.

Comme précédemment, nous transformons la série d'observations en une série continue $(y_t)_t$ selon la figure.39.
Nous choisissons un niveau de lissage $r = 5$ et nous calculons l'E.Q.M. de l'estimateur de la densité spectrale $\hat{f}_L$ selon le type de fenêtre considéré (§4.1).

fenêtre spectrale	fenêtre rectangulaire	fenêtre triangulaire	fenêtre de Daniell
EQM	$1.330353e^{-05}$	$9.528359e^{-06}$	$1.74176e^{-05}$

TABLE 9 – *L'erreur quadratique moyenne de* $\hat{f}_L$.

L'EQM étant d'une ordre de grandeur comparable pour les différentes fenêtres, nous choisissons la fenêtre donnant la représentation de la série lissée la plus adaptée à l'échantillon discret, à savoir la fenêtre rectangulaire.

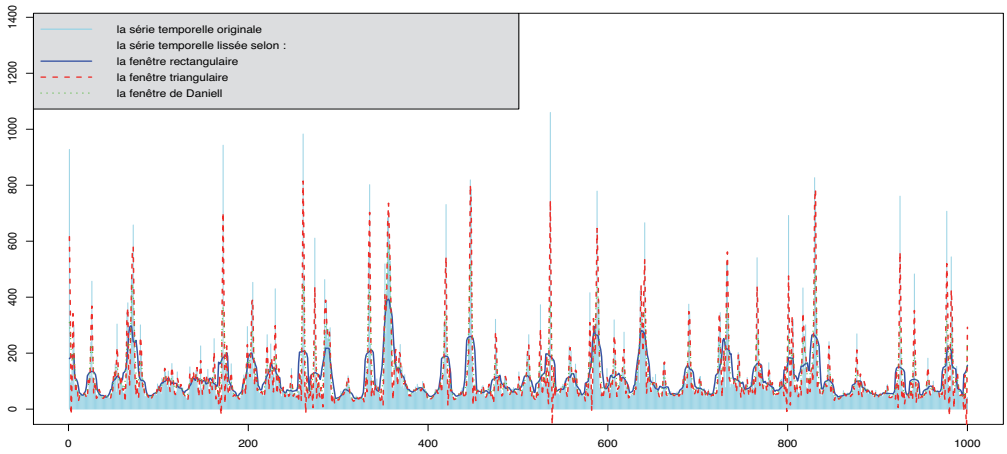

FIGURE 39 – *La série temporelle lissée. L'axe des abscisses représente les journées boursières* $t = 1, \ldots, 1000$ *et l'axe des ordonnées correspond au volume d'actions échangées.*

L'échantillon de la série lissée $(Y_t)_{i=t,\ldots,1000}$ est constitué de 1000 observations de volume d'actions échangées ;

Minimum	Quantile 25%	Médiane	Moyenne	Quantile 75%	Maximum
−75.62	50.27	69.97	111.59	115.63	815.03

TABLE 10 – *Description du jeu de données : volume d'actions échangées (série lissée)*

La loi des $(Y_t)_t$ étant continue, nous calculons le temps de retour selon les deux lois *GEV* et *GPD* définies respectivement en (46) et (47).
La figure.40 affiche les estimateurs de queue de distribution $\hat{\xi}^{Moment}$ défini en (35), $\hat{\xi}^{UH}$ défini en (36) et $\hat{\xi}^{MV}$ l'estimateur de maximum de vraisemblance.

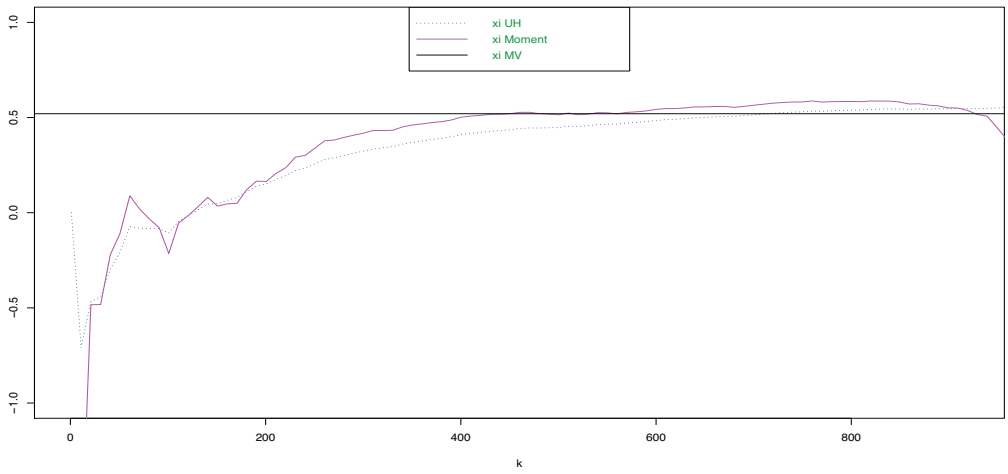

FIGURE 40 – *Estimation de l'indice ξ de la queue de distribution de la loi GEV. L'axe des abscisses représente l'indice k de $\hat{\xi}_{k,n}$ et l'axe des ordonnées correspond à l'estimateur de l'indice de la queue de distribution.*

D'après ce graphe, nous pouvons conclure que l'estimateur de la queue de distribution de la loi *GEV* est égal à $\hat{\xi} = 0.52$.
Nous pouvons aussi approximer le seuil u, paramètre de la loi *GPD*, par la valeur $u = 110$ et obtenons $(\hat{\xi}, \hat{\sigma}) = (0.205, 112.265)$, indices de la loi *GPD* estimés par la méthode de maximum de vraisemblance.
Nous représentons ci-dessous le temps de retour de certaines observations extrêmes, calculé selon les deux lois *GEV* et *GPD* et comparons ce temps de retour avec celui obtenu par la méthode de la borne supérieure.

temps t_0	$\widehat{T}^{GEV}$	$\widehat{T}^{GPD}$	$\widehat{T}^{BS}$
830	58.67	50.09	42
831	59.38	50.90	33
925	22.21	16.02	35
977	24.29	17.57	20
982	17.90	12.94	8

TABLE 11 – *Temps de retour $\widehat{T}^{GEV}$, $\widehat{T}^{GPD}$ et $\widehat{T}^{BS}$ désignent les temps de retour obtenus selon la lois GEV, la loi GPD et la méthode de la borne supérieure.*

Nous remarquons que le temps de retour obtenu via la méthode de la borne supérieure est constamment inférieur aux deux autres temps de retour calculés via la T.V.E.
Les décideurs de la COB pourront choisir l'un des ces trois temps de retour selon leur exigence quant au risque. En effet, les plus précautionneux préfèrent choisir le temps de retour le plus petit car cela leur permettra de prévoir plus tôt un volume d'actions échangées considéré comme élevé.

5 Conclusion et perspectives

L'objectif de cette étude était l'estimation d'un quantile extrême et d'un temps de retour dans le cas de lois discrètes. Pour ce faire, nous avons proposé deux méthodes originales adaptées au cas discret : la méthode de la borne supérieure du niveau de retour et la T.V.E. après lissage des données discrètes.
Nous avons ensuite utilisé ces méthodes pour estimer un quantile extrême ainsi qu'un temps de retour. Les résultats obtenus sont très encourageants, même si cette étude ne fait que débuter. Nous dressons ici une liste (non exhaustive) des directions possibles que peuvent prendre nos recherches futures.

Nous pourrions améliorer la méthode de la borne supérieure en déterminant notamment une borne inférieure plus fine que celle proposée en [7]. Cela permettrait d'affiner la qualité des estimateurs du niveau et du temps de retour et d'avoir ainsi une méthode plus opérationnelle.
De plus, comme nous avons pu en juger sur quelques applications, le choix de α_{max} et β_{max} a une influence importante sur le comportement de la borne supérieure. Il nous semble donc primordial de proposer une méthode permettant de choisir ces paramètres de manière optimale et automatique.

Nous avons aussi proposé une méthode empirique permettant d'utiliser les résultats de la T.V.E. en transformant l'échantillon de variables aléatoires de loi discrète en un échantillon de variables aléatoires de loi continue en procédant par un lissage de type noyau. Cela nous a permis d'estimer le temps de retour via les deux lois GEV et GPD. Il nous parait important à présent de tester d'autres méthodes de lissage (par exemple la méthode des ondelettes).

De plus, nous souhaiterions établir cette étude à des variables aléatoires dépendantes dans le cadre de lois discrètes et savoir quelle est la méthode la plus pertinente, que ce soit en termes de faisabilité (en théorie) comme en termes d'applications.

Nous avons, dans l'application en épidémiologie, créé un système de surveillance en utilisant ces deux méthodes fondées sur les statistiques des valeurs extrêmes et avons comparé les éventuelles alarmes générées par le système de surveillance élaboré et celui utilisé par l'InVS.
Il serait très intéressant d'appliquer ces deux méthodes sur des données financières réelles et les adapter à des problématiques en finance.

Références

[1] J Beirlant, P Vynckier, and J.M Teugels. Excess functions and estimation of the extreme value index. *Bernoulli*, 2 :293–318, 1996.

[2] P Bloomfield. *Fourier Analyses of Time Series, An Introduction*. JOHN WILEY & Sons, INC, 2000.

[3] S Csörgö, P Deheuvels, and D.M Mason. Kernel estimates of the tail index of a distribution. *Annals of Statistics*, 13 :1050–1077, 1985.

[4] A.L.M Dekkers, J.H.J Einmahl, and L De Haan. A moment estimator for the index of an extreme value distribution. *Annals of Statistics*, 17 :1833–1855, 1989.

[5] B Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44 :423–453, 1943.

[6] V Gontis. Modelling share volume traded in financial markets. *Lithuanian Journal of Physics*, 41 :551–555, 2001.

[7] A Guillou, P Naveau, J Diebolt, and P Ribereau. Return level bounds for small and moderate sample sizes for discret and continuous random variables. AMS Subject classifications : 62G05, 62G20, 62G30.

[8] E.J Gumbel. *Statistics of extremes*. Columbia University Press, 1985.

[9] J.R.M Hosking and J.R Wallis. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3) :339–349, 1987.

[10] P J.Brockwell and R A.Davis. *Introduction to Time Series and Forecasting*. Springer-Verlag New York, Inc., 2002.

[11] B Kaulakys and T Meskauskas. *Phys. Rev.E*, V 58 :p.7013, 1998.

[12] M Kratz and SI Resnick. The qq-estimator of the index of regular variation. *Communications in Statistics : Stochastic Models*, 12 :699–724, 1996.

[13] T. Mikosch. *Non-life Insurance Mathematics*. Springer, 2004.

[14] J Pickands. Statistical inference using extreme-order statistics. *Annals of Statistics*, 3 :119–131, 1975.

[15] M.B Priestley. *Spectral Analysis and Time Series*. Academic Press, 1981.

[16] S.I Resnick. *Extreme Values, regular Variation, and Point Process*. Springer-Verlag, New-York, 1987.

Remerciements

Le moment est venu d'exprimer toute ma reconnaissance à l'encadreuse de ce projet, Marie KRATZ, pour son enthousiasme, son soutien sans faille, sa gentillesse et sa patience qu'elle a manifesté à mon égard durant ces trois mois.
Je ne sais pas comment exprimer ma gratitude à cette personne autrement qu'en lui promettant d'agir comme elle avec des étudiants dans ma situation, si un jour m'en est donné.

Je tiens également à remercier Yann LE STRAT de Institut de Veille Sanitaire pour avoir trouvé le temps, malgré ses nombreuses occupations, de lire ce rapport et surtout de me rencontrer pour me faire part de ses remarques. Je lui suis très reconnaissant pour l'attention qu'il a porté à ce travail.
J'adresse mes sincères remerciements à Armelle GUILLOU pour l'intérêt qu'elle a manifesté pour ce projet. Malgré son éloignement, elle s'est déplacée à Paris pour me prodiguer ses conseils et elle a été disponible pour m'inculquer sa grande rigueur.
Je tiens aussi à remercier François-Xavier WEILL du Centre National de Référence des Salmonella Laboratoire des Bactéries Pathogènes Entériques Institut Pasteur, pour la mise à ma disposition des données.
Je tiens en particulier à mentionner le plaisir que j'ai eu à travailler au sein de l'ESSEC, et j'en remercie ici tous les membres et tous les doctorants.
J'exprime toute ma reconnaissance à mes enseignants de l'ESSAI sans qui rien n'aurait été possible et je tiens particulièrement à remercier Mokhtar KOUKI.
Enfin, j'exprime ma gratitude à ma famille qui m'a toujours soutenu et encouragé dans la voie je m'étais fixé.
Je remercie mes parents qui m'ont stimulé et encouragé pendant mes études.
Surtout, bonne chance et bon courage à Zied et à Rania pour leurs futurs projets de fin d'étude.

A.BORCHANI

6 Annexe

– **Lois de la T.V.E.**

```
#densité de la distribution GEV
curve(dgev(x, shape=0), lwd=3, xlim=c(-2,2), ylim=c(0,1.5),
      ylab="densité", main="")
for (s in seq(-2,2,by=.1)) {
  curve(dgev(x, shape=s), add=T)
}
curve(dgev(x, shape=-1), lwd=3, add=T, col='red')
curve(dgev(x, shape=1), lwd=3, add=T, col='blue')
#densité de la distribution Fréchet
curve(dfrechet(x, shape=1), lwd=3, xlim=c(-1,2), ylim=c(0,1),
      ylab="", main="densité de la distribution Fréchet")
for (s in seq(.1,2,by=.1)) {
  curve(dfrechet(x, shape=s), add=T)
}
curve(dfrechet(x, shape=2), lwd=3, add=T, col='red')
curve(dfrechet(x, shape=.5), lwd=3, add=T, col='blue')
#densité de la distribution (reverse) Weibull
 curve(drweibull(x, shape=1), lwd=3, xlim=c(-2,1), ylim=c(0,1),
      ylab="", main="densité de la distribution (reverse) Weibull ")
for (s in seq(.1,2,by=.1)) {
  curve(drweibull(x, shape=s), add=T)
}
curve(drweibull(x, shape=2), lwd=3, add=T, col='red')
curve(drweibull(x, shape=.5), lwd=3, add=T, col='blue')
#densité de la distribution Gumbel
curve(dgumbel(x), lwd=3, xlim=c(-2,2), ylim=c(0,1),
ylab="", main="densité de la distribution Gumbel")
#densité de la distribution GPD
curve(dgpd(x, shape=0), lwd=3, xlim=c(-.1,2), ylim=c(0,2),
      xlab="y",ylab="", main="densité de la distribution GPD")
for (s in seq(-2,2,by=.1)) {
  curve(dgpd(x, shape=s), add=T)
}
curve(dgpd(x, shape=-1), lwd=3, add=T, col='red')
curve(dgpd(x, shape=1), lwd=3, add=T, col='blue')
```

– **Paramètres des lois de la T.V.E (*GEV* et *GPD*)**

```
#Mean Excess Plot
x <- sort(x)
e <- rep(NA, length(x))
for (i in seq(along=x)) {
  u <- x[i]
  e[i] <- mean( (x-u)[x>u] )
}
plot(x,e, type='n',xlab='u',ylab='e(u)', main="Mean Excess Plot")
#Estimateur de Hill
x<-sort(x)
n<-length(x)
hill<-rep(NA,(n-1))
for(j in seq(from=1,to=(n-1))){
  hill[j]<-((sum(log(x[(n-j+1):n])))/j)-(log(x[n-j]))
}
plot(hill,main='estimateur de Hill')
#Estimateur de UH
for(j in seq(from=1,to=(n-1))){
  UH[j]<-(z[n-j])*hill[j]
  xi_UH[j]<-((sum(log(UH[1:j])))/j) - log(UH[j])
}
plot(xi_UH,main='estimateur de UH')
#Estimateur des Moments
Moment<-rep(NA,(n-1))
for(j in seq(from=1,to=(n-1))){
  S[j]<-((sum((log(z[(n-j+1):n]))-(log(z[n-j])))^2)/j)
  Moment[j]<-hill[j]+1-(1/2*(1-(((hill[j])^2)/S[j])))
}
plot(Moment,main='estimateur des Moments')
#Generalized Quantile Plot
x<-sort(x)
n<-length(x)
hill<-rep(NA,(n-1))
UH<- rep(NA,(n-1))
logq<-rep(NA,(n-1))
for(j in seq(from=1,to=(n-1))){
  hill[j]<-((sum(log(x[(n-j+1):n])))/j)-(log(x[n-j]))
  UH[j]<-(x[n-j])*hill[j]
```

```
  logq[j]<-log((n+1)/j)
  }
plot(logq,log(UH),main='Generalized Quantile Plot')
```

– **Bornes du niveau de retour**

```
bornes=function(X,amax,bmax,cmax,pmax,t){
X<-sort(X)
n=length(X)
aminimisersup=function(theta){
  a=theta[1]
  b=theta[2]
  thetachapeau=0
  for(i in 1: n){
    thetachapeau=thetachapeau+ (1/n)*(X[i]^a)*((i/n)^b)
  }
  denominateur=(1-1/t)^b
  totale2=(t*thetachapeau/denominateur)^(1/a)
  if(is.na(totale2)){totale2=222222222}
  if(totale2==Inf){totale2=2000000}
  return(totale2)
}
bornesup=optim(par=c(0.1,0.1),fn=aminimisersup, method ="L-BFGS-B",lower=c(0.1,
0.1),upper=c(amax,bmax))$value#la borne supérieure du niveau de retour
R=optim(par=c(0.1,0.1),fn=aminimisersup, method ="L-BFGS-B", lower=c(0.1,
0.1),upper=c(amax,bmax))$par

aminimiserinf=function(thetaetoile){
  a=thetaetoile[1]
  c=thetaetoile[2]
  p=thetaetoile[3]
  thetaetoilechapeau=0
  for(i in 1: n){
    thetaetoilechapeau=thetaetoilechapeau+ ((1/n)*(X[i]^a)*((1-(i/n))^(-c)))
  }
  denominateur2=(1-1/t)*((1/t)^(-c))
  totale21=((thetaetoilechapeau-((t^(1/(p-1)))*(thetaetoilechapeau^
  (1/p))))/denominateur2)^(1/a)
  if(is.na(totale21)){totale21=222222222}
```

```
  if(totale21==Inf){totale21=2000000}
  return(totale21)
}
borneinf=optimize(f = aminimiserinf , lower=c(0.1,0.1,1),upper=c(amax,cmax,
pmax), maximum = TRUE)$maximum#la borne inférieure du niveau de retour

thetachapeau=0
for(i in 1: n){
  thetachapeau=thetachapeau+ (1/n)*(X[i]^a)*((i/n)^b)
}
aint3=function(u){
  aint=function(t){
    b*(-t)*t^(b-1)*(quantile(X,t)^a)
  }
  aint2=function(t){
    b*(1-t)*t^(b-1)*(quantile(X,t)^a)
  }
  return((-u^b*(quantile(X,u))^a+thetachapeau-integrate(aint,0,u,
  stop.on.error=F)$value-integrate(aint2,u,1,stop.on.error=F)$value)^2)
}
variancesup=integrate(aint3,0,1, subdivisions=1000)$value
plus=1.96*t*sqrt(variancesup)/(a*sqrt(n)*(1-1/t)^(b))*
((t*thetachapeau/(1-1/t)^(b))^(1/a-1))
vsi=bornesup-plus#vsi est la borne inférieure de l'intervalle
#de confiance de la borne supérieure à 95%
vss=bornesup+plus#vsi est la borne supérieure de l'intervalle
#de confiance de la borne supérieure à 95%
return(bornesup,borneinf,vsi,vss)
}
```

– **Génération de n_{sim} v.a. i.i.d. de loi Log discrète de paramètre α**

```
n_sim=1000
i_max=100
alpha=0.9
F_log<-rep(NA,i_max)
W<-rep(NA,n_sim)
for(i in seq (from=2, to=i_max)){
  F_log[1]<-(-alpha)/log(1-alpha)
```

```
  F_log[i]<-F_log[i-1]+((-(alpha^i))/(i*log(1-alpha)))
}
i_sim<-0
repeat{
  random<-runif(1)
  if (random<F_log[1]){
    W[i_sim]<-1
  }
  if (random>F_log[i_max]){
    W[i_sim]<-i_max
  }
  for(i in seq (from=1, to=(i_max-1))){
    if((F_log[i]<=random)&&(F_log[i+1]>random)){
      W[i_sim]<-i+1
    }
  }
  i_sim<-i_sim+1
  if(i_sim==n_sim+1)  {break}
}
W#"n_sim" v.a. i.i.d. de loi Log discret (Charge de sinistres)
```

– **Génération de n_{sim} v.a. i.i.d. de loi Generalized Discret Model**

```
n_sim=1000
Vr<-matrix(data = 0,  ncol= 1, nrow = n_sim)
Nr<-matrix(data = 0,  ncol= 1, nrow = n_sim)
alpha=0.75
gamma=0.0007
taubar=0.01
sigma=0.00015
r=1
tau2=taubar
tau1=taubar
temps=1
repeat{
  bb<-rnorm(1, mean = 0, sd = 1)
  tau3=tau2+(alpha*(tau2-tau1))-(gamma*(tau2-taubar))+ (sigma*bb)
  tauinter=tau2
  tau1=tauinter
```

```
  tau2=tau3
  temps=temps+tau2
  for(r in seq(from=1, to=n_sim)){
    if((temps>=r)&&(temps<(r+1))){
      Vr[r]<-Vr[r]+1
    }
  }
  if(temps>=n_sim){break}
}
Vr#"n_sim" v.a. i.i.d. de loi G.D.M (Volume d'actions échangées)
```

– **Génération d'un échantillon continu par lissage de type noyau**

```
#Lissage par noyau
r=5
k1 <- kernel("daniell", r)
k2 <- kernel("dirichlet",r,3)
k3 <- kernel("fejer",r,3)
#série lissée par :
#la fenêtre de Daniell
x1 <- kernapply(x, k1, circular = TRUE)
#la fenêtre rectangulaire
x2 <- kernapply(x, k2, circular = TRUE)
#la fenêtre triangulaire
x3 <- kernapply(x, k3, circular = TRUE)
```

– **Génération d'un sous-échantillon en tenant compte de la saisonnalité**

```
b<-20     #nbre d'année
freq<-52  #L'occurence du cycle :52 semaine
w<-3      #(2*w+1) représente le nbre des observations pour un cycle
k<-length(X)
S_X<-rep(0,k)
for (i in (b:1)) {
  for (j in (((k - i * freq) - w):((k - i * freq) + w))) {
  S_X[j]<-X[j]
  }
}
```

```
S_X#sous-échantillon de X
```

– **Algorithme de surveillance**

```
library(surveillance)
library(evd)
algo.farrington.bornesup.gev.anis=function (disProgObj, control = list(range =
NULL, b = 3, w = 3, reweight = TRUE, verbose = FALSE, alpha = 0.01,
powertrans = "2/3"))
{

    observed <- disProgObj$observed
    k2 <- kernel("dirichlet", 5,r=3)
    observed2 <- kernapply(observed, k2, circular = TRUE)
    a=fgev(observed2)$param[1]
    b=fgev(observed2)$param[2]
    s=fgev(observed2)$param[3]
    freq <- disProgObj$freq
    if (is.null(control$range)) {
        control$range <- (freq * control$b - control$w):length(observed)
    }
    if (is.null(control$b)) {
        control$b = 5
    }
    if (is.null(control$w)) {
        control$w = 3
    }
    if (is.null(control$reweight)) {
        control$reweight = TRUE
    }
    if (is.null(control$verbose)) {
        control$verbose = FALSE
    }
    if (is.null(control$alpha)) {
        control$alpha = 0.05
    }
    if (is.null(control$trend)) {
        control$trend = TRUE
    }
    if (is.null(control$plot)) {
        control$plot = FALSE
    }
```

```
if (is.null(control$limit54)) {
    control$limit54 = c(5, 4)
}
if (is.null(control$powertrans)) {
    control$powertrans = "2/3"
}
if (!((control$limit54[1] >= 0) & (control$limit54[2] > 0))) {
    stop("The limit54 arguments are out of bounds: cases >= 0 and
    perior > 0.")
}
alarm <- matrix(data = 0, nrow = length(control$range), ncol = 3,dimnames
 = list(c(1:length(control$range)),c("glm", "GEV.anis","BS.Anis")))
trend <- matrix(data = 0, nrow = length(control$range), ncol = 1)
T <- matrix(data = 0, nrow = length(control$range), ncol = 1)
upperbound <- matrix(data = 0, nrow = length(control$range), ncol = 1)
n <- control$b * (2 * control$w + 1)
for (k in control$range) {
    if (control$verbose) {
        cat("k=", k, "\n")
    }
    wtime <- NULL
    for (i in control$b:1) {
        wtime <- append(wtime, seq(k - freq * i - control$w, k - freq
         * i + control$w, by = 1))
    }
    response <- NULL
    for (i in (control$b:1)) {
        if (control$verbose) {
            cat("b=", i, "\trange=", ((k - i * freq) - control$w):
            ((k - i * freq) + control$w), "\n")
        }
        for (j in (((k - i * freq) - control$w):((k - i * freq)
        + control$w))) {
            if (j < 1) {
              cat("Warning: Selection index less than 1!\n")
            }
            else {
              response <- append(response, observed[j])
            }
        }
```

```
}
if (control$verbose) {
    print(response)
}
model <- algo.farrington.fitGLM(response, wtime, timeTrend =
control$trend, reweight = control$reweight)
if (is.null(model))
    return(model)
doTrend <- control$trend
if (control$trend) {
    p <- summary.glm(model)$coefficients["wtime", 4]
    significant <- (p < 0.05)
    mu0Hat <- predict.glm(model, data.frame(wtime = c(k)), type =
     "response")
    atLeastThreeYears <- (control$b >= 3)
    noExtrapolation <- mu0Hat <= max(response)
    if (!(atLeastThreeYears && significant && noExtrapolation)) {
        doTrend <- FALSE
        model <- algo.farrington.fitGLM(response, wtime,timeTrend =
        FALSE, reweight = control$reweight)
    }
}
else {
    doTrend <- FALSE
}
pred <- predict.glm(model, data.frame(wtime = c(k)), dispersion
= model$phi, type = "response", se.fit = TRUE)
lu <- algo.farrington.threshold(pred, model$phi, skewness.transform
 = control$powertrans, alpha = control$alpha)
if (control$plot) {
    data <- data.frame(wtime = seq(min(wtime), k, length = 1000))
    preds <- predict(model, data, type = "response",dispersion =
    model$phi)
    plot(c(wtime, k), c(response, observed[k]), ylim = range(
    c(observed[data$wtime],lu)), , xlab = "time", ylab = "No. infected"
    , main = paste("Prediction at time t=", k, " with b=", control$b,
    ",w=", control$w, sep = ""), pch = c(rep(1,length(wtime)), 16))
    lines(data$wtime, preds, col = 1, pch = 2)
    lines(rep(k, 2), lu, col = 3, lty = 2)
}
```

```
enoughCases <- (sum(observed[(k - control$limit54[2] +1):k])
>= control$limit54[1])
X1 <- ifelse(enoughCases, (observed[k] - pred$fit)/(max(lu) -
pred$fit), 0)
alarm[(k - min(control$range) + 1),1] <- (X1 > 1)

response2 <- NULL
for (i in (control$b:1)) {
    if (control$verbose) {
        cat("b=", i, "\trange=", ((k - i * freq) - control$w)
        :((k - i * freq) + control$w), "\n")
    }
    for (j in (((k - i * freq) - control$w):((k - i * freq)
    + control$w))) {
        if (j < 1) {
          cat("Warning: Selection index less than 1!\n")
        }
        else {
          response2 <- append(response2, observed2[j])
        }
    }
}

if(observed2[k]<=(a-(b/s))){
T[k]=0
}
else{
  if((abs(s))<=0.1){
    T[k] = (1-exp(-(exp(-(observed2[k]-a)/b))))^(-1)
  }
  else{
    T[k] = (1-exp(-(1+s*((observed2[k]-a)/b))^(-1/s)))^(-1)
  }
}
if(T[k]>=length(response2)){
  T[k]=length(response2)
  alarm[(k - min(control$range) + 1),2]<- 1
}
if(max(response)<observed2[k]){
```

```
  alarm[(k - min(control$range) + 1),2]<- 1
}
if(T[k]<10){alarm[(k - min(control$range) + 1),2]<- 0}
else{
    for (i in (((length(response2))-T[k]+1):length(response2))){
      if (response2[i]>=observed2[k]){
      alarm[(k - min(control$range) + 1),2] <- 1
      }
    }

}

borne=function(X,amax,bmax,t){
  n=length(X)
  X<-sort(X)
  aminimisersup=function(theta){
    a=theta[1]
    b=theta[2]
    thetachapeau=0
    for(i in 1: n){
      thetachapeau=thetachapeau+ (1/n)*(X[i]^a)*((i/n)^b)
    }
    denominateur=(1-1/t)^b
    totale2=(t*thetachapeau/denominateur)^(1/a)
    if(is.na(totale2)){totale2=222222222}
    if(totale2==Inf){totale2=2000000}
    return(totale2)
  }
  bornesup=optim(par=c(0.01,0.01),fn=aminimisersup, method =
  "L-BFGS-B", lower=c(0.01,0.01),upper=c(amax,bmax))$value
  return(bornesup)
}
Tmax <- 100
BS<-rep(NA,Tmax)
for(t in 2:(1+Tmax)){
B=0
B<-borne(response,1,0.2,t)
BS[t-1]<-B[1]
}
```

```
          T<-0
          for(t in 1:Tmax){
            if((BS[t]) <= observed[k]){
              T=T+1
            }
          }

          if(T>=length(response)){
            T=length(response)
            alarm[(k - min(control$range) + 1),3]<- 1
            }
          if(max(response)<observed[k])
            {
            alarm[(k - min(control$range) + 1),3]<- 1
            }
          if(T==0){alarm[(k - min(control$range) + 1),3]<- 0}
          else{
              for (i in (((length(response))-T+1):length(response))){
                if (response[i]>=observed[k]){
                alarm[(k - min(control$range) + 1),3] <- 1
                }
              }

          }
      }
      control$name <- paste("farrington(", control$w, ",", 0,
      ",", control$b, ")", sep = "")
      control$data <- paste(deparse(substitute(disProgObj)))
      result <- list(alarm = alarm, T=T,observed2=observed2,upperbound
      = upperbound, trend = trend, disProgObj = disProgObj, control = control)
      class(result) <- "survRes"
      return(result)
  }
```

– **Graphe de surveillance**

```
  #x<-algo.farrington.bornesup.gev.anis("data",...)
  plot.GLM.BS.GEV=function (x, method = x$control$name, disease = x$control$data,
      domany = FALSE, ylim = NULL, xaxis.years = TRUE, startyear = 2001,
```

```
    firstweek = 1, xlab = "time", ylab = "No. infected", main = NULL,
    type = "hhs", lty = c(1, 1, 2), col = c(1, 1, 4, 3), outbreak.symbol =
     list(pch = 3,col = 3), alarm.symbol.GLM = list(pch = 24, col = 2),
     alarm.symbol.ANIS = list(pch = 24, col = 4) ,alarm.symbol.ANIS.GEV =
      list(pch = 24, col = 3),legend.opts = list(x = "top",
        legend = c("Infected", "Alarm.GLM","Alarm.GEV", "Alarm.BorneSup"),
        lty = NULL, col = NULL, pch = NULL), ...)
{
    if (is.null(main))
        main = paste("Détection d'alarme en utilisant la GLM, la borne
        supérieure et la GEV")
    if (domany)
        main = ""
    survResObj <- x
    observed <- survResObj$disProgObj$observed[survResObj$control$range]
    state <- survResObj$disProgObj$state[survResObj$control$range]
    tab <- 0.5
    observedxl <- (1:length(observed)) - tab
    observedxr <- (1:length(observed)) + tab
    upperboundx <- (1:length(survResObj$upperbound))
    max <- max(max(observed), max(survResObj$upperbound))
    if (is.null(ylim)) {
        max <- max(max(observed), max(survResObj$upperbound))
        ylim <- c(-1/20 * max, max)
    }
    else {
        max <- ylim[2]
    }
    if (ylim[1] >= 0)
        ylim[1] <- -1/20 * max
    xstuff <- cbind(observedxl, observedxr, upperboundx)
    ystuff <- cbind(observed, observed, survResObj$upperbound)
    matplot(x = xstuff, y = ystuff, xlab = xlab, ylab = ylab,
        main = main, ylim = ylim, axes = !(xaxis.years), type = type,
        lty = lty, col = col, ...)
    for (i in 1:length(observed)) {
        matlines(c(i - tab, i + tab), c(observed[i], observed[i]),
            col = col[1])
        if (survResObj$alarm[i,1] == 1)
            matpoints(i, -1/15 * max, pch = alarm.symbol.GLM$pch,
```

```
                col = alarm.symbol.GLM$col)
        if (survResObj$alarm[i,2] == 1)
            matpoints(i, -1/25 * max, pch = alarm.symbol.ANIS.GEV$pch,
                col = alarm.symbol.ANIS.GEV$col)
        if (survResObj$alarm[i,3] == 1)
            matpoints(i, -1/45 * max, pch = alarm.symbol.ANIS$pch,
                col = alarm.symbol.ANIS$col)
        if (state[i] == 1)
            matpoints(i, -1/20 * max, pch = outbreak.symbol$pch,
                col = outbreak.symbol$col)
    }
    if (max * 2/3 >= max(max(observed[1:floor(1/4 * length(observed))]),
        max(survResObj$upperbound[1:floor(1/4 * length(survResObj$upperbound))
        ]))) {
        xlegpos <- 0
    }
    if (xaxis.years) {
        myat.week <- seq(ceiling((52 - firstweek + 1)/13) * 13 +
            1, length(observed) + (floor((52 - firstweek + 1)/13) *
            13 + 1), by = 13)
        year <- (myat.week - 52)%/%52 + startyear
        quarterFunc <- function(i) {
            switch(i + 1, "I", "II", "III", "IV")
        }
        quarter <- sapply((myat.week - 1)%/%13%%4, quarterFunc)
        myat.week <- myat.week - (52 - firstweek + 1)
        mylabels.week <- paste(year, "\n\n", quarter, sep = "")
        axis(at = myat.week, labels = mylabels.week, side = 1,
            line = 1)
        axis(side = 2)
    }
    if (!is.null(legend.opts) && (class(legend.opts) == "list")) {
        if (is.null(legend.opts$lty))
            legend.opts$lty = c(1, NA, NA,NA)
        if (is.null(legend.opts$col))
            legend.opts$col = c(1, alarm.symbol.GLM$col ,
            alarm.symbol.ANIS.GEV$col, alarm.symbol.ANIS$col)
        if (is.null(legend.opts$pch))
            legend.opts$pch = c(NA, alarm.symbol.GLM$pch ,
            alarm.symbol.ANIS.GEV$pch,alarm.symbol.ANIS$pch)
```

```
            if (is.null(legend.opts$x))
                legend.opts$x = "top"
            if (is.null(legend.opts$legend))
                legend.opts$legend = c("Infected", "Alarm GLM","Alarm GEV",
                "Alarm Borne Sup")
            do.call("legend", legend.opts)
        }
        invisible()
    }
```

ESSEC Business School
Avenue Bernard Hirsch
BP 50105
95021 Cergy-Pontoise Cedex
France

Tél. +33 (0)1 34 43 30 00
Fax +33 (0)1 34 43 30 01

www.essec.fr

ESSEC Executive Education
CNIT BP 230
92053 Paris-La Défense
France

Tél. +33 (0)1 46 92 49 00
Fax +33 (0)1 46 92 49 90

http://formation.essec.fr

ESSEC Business School
Singapore Campus
100 Victoria Street
National Library Building # 13-02
Singapore 188064

essecasia@essec.fr
Tél. +65 6884 9780
Fax +65 6884 9781

www.essec.edu

Informations

+33 (0)1 34 43 30 91
blanchard@essec.fr
research.center@essec.fr
www.essec.fr

ESSEC BUSINESS SCHOOL

CCI Chambre de commerce et d'industrie Versailles Val-d'Oise - Yvelines
À vos côtés, précisément.

AACSB ACCREDITED

EFMD EQUIS ACCREDITED