Table des matières

Table des figures

Remerciements

Louange à Dieu le tout puissant d'avoir mis sur mon chemin plusieurs personnes dont l'intervention a constitué une aide considérable pour l'élaboration de ce projet.

Je voudrais d'abord exprimer mes profonds remerciements et ma reconnaissance à, Mme **EZZAKI Fatima**, l'enseignante-chercheur de haut niveau qui, par ses conseils valeureux, ses remarques perspicaces, son temps précieux bref, pour son meilleur encadrement m'a soutenue dans les différentes étapes de la réalisation de ce projet.

J'adresse ensuite mes remerciements les plus sincères aux membres du jury respectable : **Pr. Ammor Ouafae, Pr. El Hilali Alaoui Ahmed, Pr. El Khomssi Mohamed, Pr. Ettaouil Mohamed, Pr. Ezzaki Fatima** d'avoir bien voulu participer à l'évaluation de ce travail .

Mes sincères remerciements s'adressent à M. **Mohammed Berkkia**, directeur régional de l'ONEE centre nord Fès de m'avoir donné la chance d'effectuer le stage au sein de l'ONEE ainsi qu'à M.**Zitoun Khalid** pour son aide précieuse et ses remarques pertinentes.
J'aimerais encore adresser mes vifs remerciements à mon encadrant professionnel M.**Aissouk** qui par son expertise, ses conseils valeureux et ses commentaires judicieux, m'a aidé pour l'élaboration de ce projet.

Je souhaite aussi manifester mes remerciements à tout le département des mathématiques appliquées et en particulier, le responsable du Master Mathématiques appliquées
M.Hilali Alaoui Ahmed pour les énormes efforts qu'il ne cesse d'épargner à l'égard de tous les étudiants. Mes remerciements vont également à **M.Mohammed Benzakour Amine** pour son aide et sa disponibilité.

Je suis digne à mes parents **Karmouda Jamal** et **Nassib Jamila** qui m'ont toujours encouragée et aidée dans mes études. Ils ont su me donner toutes les chances pour réussir. J'espère qu'ils trouvent, dans la réalisation de ce travail, l'aboutissement de leurs efforts ainsi que l'expression de ma plus affectueuse gratitude.

Je remercie toute ma famille et en particulier, mes deux sœurs **Karmouda Firdaouss** et **Karmouda Faiza** pour leurs soutien et encouragements .

Je remercie enfin toutes les personnes qui ont participé de près ou de loin à la réussite de ce projet .

Introduction

Les risques extrêmes (inondations, sécheresses intenses, tremblements de terre, chocs pétroliers, crises financières, etc) sont généralement des événements rares (ie la probabilité d'apparition est très faible) qui captent l'attention par leurs caractères inattendus et récurrents.

Souvent négligés par certains non-statisticiens et traités comme des observations aberrantes, les extrêmes ont fini par être sous le feu des projecteurs notamment pour l'importance de leurs impacts sociaux et économiques. Mais comment peut on estimer le caractère rare d'une variable aléatoire ?

Les méthodes statistiques classiques utilisent toutes les données observées sans discerner celles qui se rapportent aux événements extrêmes fournissant ainsi des modèles appropriés pour décrire le comportement central, et n'apportant guère d'informations spécifiques aux valeurs extrêmes.

Les outils probabilistes traditionnels développés dans un cadre gaussien sont inadaptés pour l'étude de tels événements extrêmes où les valeurs sont loin d'être concentrées autour de leur moyenne.

L'étude des risques extrêmes vise la protection des systèmes (industriels, économiques, financiers, etc) contre les éventuelles catastrophes susceptibles de les détruire . Dans le cas de notre projet, il s'agit de la gestion du risque hydrologique des crues pour protéger la station de pré-traitement d'eau potable de la ville de Fès contre les inondations.

En effet, la station de pré-traitement d'eau potable de la ville de Fès a connu le 10/10/2008 de graves inondations causées par l'arrivée d'une crue torrentielle d'Oued Sebou, en raison des lâchers exceptionnels du barrage Allal El Fassi suite aux pluies extrêmes qui se sont abattues sur le bassin versant en cette période. Le débit du Oued a été alors estimé à $2600m^3/s$. Cet incident a induit d'importants dégâts au niveau des équipements hydro-électro-mécaniques et électriques de la station et a causé l'arrêt total de la station, ce qui a engendré une rupture provisoire dans l'approvisionnement en eau potable de certains secteurs de la ville de Fès.

Afin de permettre aux décideurs de trancher sur la solution adéquate pour protéger la station contre les inondations (construction d'un mur de protection, digue, etc), il va falloir répondre aux questions suivantes :

1. Quelle est la hauteur d'eau qui est atteinte ou dépassée pour une période donnée ?

2. Dans combien de temps, on peut s'attendre en moyenne à une "grande" hauteur d'eau fixée ?

Statistiquement parlant, la question (1) se rapporte à l'estimation d'un quantile extrême ou niveau de retour en hydrologie. Quant à la question (2), elle se rapporte à l'estimation d'une "petite probabilité", ou de façon équivalente en hydrologie à une période de retour.

Cependant, l'étude des risques extrêmes est souvent confrontée à un problème de manque d'informations, notamment lors d'une analyse statistique. En effet, il n'est pas raisonnable de pouvoir extraire des quantiles de périodes de retour très élevées (à savoir 100 ans voir 1000 ans) à partir de quelques cinquantaines d'observations seulement. Comment peut on alors s'affranchir de ce problème ?

Résumé

Une des voies empruntées pour remédier au problème des valeurs extrêmes est la théorie des valeurs extrêmes (TVE). Celle ci est basée sur l'approximation asymptotique des maxima (ou minima)d'un grand nombre de variables aléatoires supposées *i.i.d.* L'étude des valeurs extrêmes se ramène alors à l'analyse des queues des lois de probabilité, en cherchant à estimer les quantiles extrêmes dépassés avec une probabilité très faible.

Pour s'affranchir du problème du manque d'observations due à la rareté du risque ainsi que des problèmes d'échantillonnage, une méthode basée sur la simulation stochastique a été étudiée : Il s'agit de la méthode SHYPRE. Cette dernière traite le risque des crues par une analyse fine de sa cause : les pluies intenses. En se basant sur l'information pluviométrique, cette méthode consiste à chercher une modélisation probabiliste fine du processus de la pluie et à partir d'une transformation pluie-débit, les débits extrêmes correspondants sont estimés et ceci pour des grandes périodes de retour. L'avantage de cette approche, par rapport à une approche statistique classique, est le fait de disposer de l'intégralité de l'information temporelle des pluies et des crues. De plus, la large prise en compte de l'information des pluies rend la méthode moins sensible aux problèmes d'échantillonnage. Ce qui donne aux aménageurs une idée claire sur les hauteurs d'eau maximales pouvant se reproduire.

Organisation du rapport

Ce rapport est organisé en 3 chapitres :

• Le premier chapitre présente les fondements théoriques et les résultats principaux de la théorie des valeurs extrêmes. En premier lieu, l'approche des lois des maximas basée sur le théorème fondamental de Fisher-Tippet [8] est présentée. Ensuite et pour des limites de la première approche discutée dans le chapitre, une deuxième approche a été étudiée , appelée approche par dépassement de seuil justifiée par le théorème de Pickands [9].Pour chaque approche, les quantiles extrêmes souhaités ont été calculés après avoir déterminé les paramètres de chaque loi considérée en utilisant différentes méthodes d'estimation paramétrique.

• Avant d'aborder la méthode SHYPRE basée sur une simulation stochastique, le deuxième chapitre a été introduit pour présenter différentes méthodes de simulation stochastique à savoir la méthode d'inversion et la méthode d'acceptation-rejet. Dans une deuxième partie, et pour déterminer la probabilité d'occurrence des risques rares, la méthode de Monte Carlo "naive" ne peut être utilisée à cause de la rareté des événements. Pour cela, d'autres méthodes de Monte Carlo ont été présentées principalement : la méthode d'échantillonnage préférentiel.

• Le troisième chapitre de ce projet a été consacré à la présentation d'une deuxième méthode pour la prédétermination des crues. Cette méthode vise à estimer les quantiles extrêmes des débits par un procédé de simulation stochastique. Ce dernier est une combinaison de deux modèles. Le premier est un modèle probabiliste dont l'objectif est la description de l'aléa pluviométrique tandis que le deuxième est un modèle hydrologique pluie-débit dont le but est de simuler la réponse du bassin à des précipitations de tout ordre pour estimer les quantiles extrêmes de différentes périodes de retour.

Présentation du lieu du stage : L'ONEE

Crée en 1972, l'ONEE est un établissement public à caractère industriel commercial. Doté de la personnalité civile et de l'autonomie financière, il assure la production et la gestion de l'eau potable au Maroc.

0.1 Missions principales de L'ONEE

Parmi les missions de l'ONEE :
⋆ **Planifier** l'approvisionnement en eau potable à l'échelle nationale.
⋆ **Étudier** l'approvisionnement en eau potable et assurer l'exécution des travaux des unités de production et de distribution.
⋆ **Gérer** la production de l'eau potable et contrôler la qualité des eaux produites et distribuées.
⋆ **Participer** aux études, projets nécessaires à l'accomplissement de ses missions.

0.2 Les ressources de l'ONEE de Fès

Les ressources utilisées par l'ONEE de Fès, pour la production de l'eau potable sont :
Ressources souterraines et qui sont principalement les forages situés dans la plaine de SAIS.
Ressources superficielles et qui sont les eaux de OUED SEBOU.

0.3 Complexe de production d'OUED SEBOU

Le complexe de production de l'eau potable comprend deux stations :

⋆ **Station de pré-traitement :** Les eaux brutes doivent généralement subir, avant leur traitement proprement dit, un pré-traitement qui comporte un certain nombre d'opérations généralement physiques ou mécaniques. Le rôle de cette station est d'extraire de l'eau brute la plus grande quantité possible d'éléments dont la nature ou la dimension constituerait une gêne pour les traitements ultérieurs.

Elle est située sur la rive gauche de l'Oued Sebou à la sortie de la ville à environ 8 km. Elle fonctionne quand les matières en suspension sont comprises entre $2g/l$ et $50g/l$ notamment lors des crues.

⋆ **Station de traitement** : Située à Ain Noukbi, elle assure :
∗ le traitement des eaux reçues de la station de pré-traitement selon une série d'étapes.
∗ Le contrôle de la qualité des eaux traitées.
∗ Refoulement des eaux vers le réservoir BAB HAMERA.

0.4 Description de la station de pré-traitement

Le rôle de la station de pré-traitement est de diminuer la charge d'eau brute de la matière en suspension à une valeur inférieure à $2g/l$, selon un certain nombre d'opérations.

Dégrillage est destiné à retenir les matières volumineuses et déchets de toutes sortes contenus dans l'eau. Il permet de protéger les ouvrages en aval contre l'arrivée des gros objets susceptibles de provoquer des bouchages dans les différentes unités de traitement.

Figure 1 – Dégrilleur

Relevage Cette opération s'effectue grâce à des vis d'Archimmède, un moyen de relevage particulièrement efficace qui permet le pompage de l'eau vers les dessableurs.

Figure 2 – Vis d'Archimède

Dessablage Un pré-traitement purement physique débarrasse les eaux brutes des sables et des graviers pour éviter les dépôts dans les canalisations et protéger les pompes.

Figure 3 – Dessableur

Débourbage Dernier pré-traitement, consiste à éliminer la boue par une décantation préliminaire dans les débourbeurs.

Figure 4 – Débourbeur

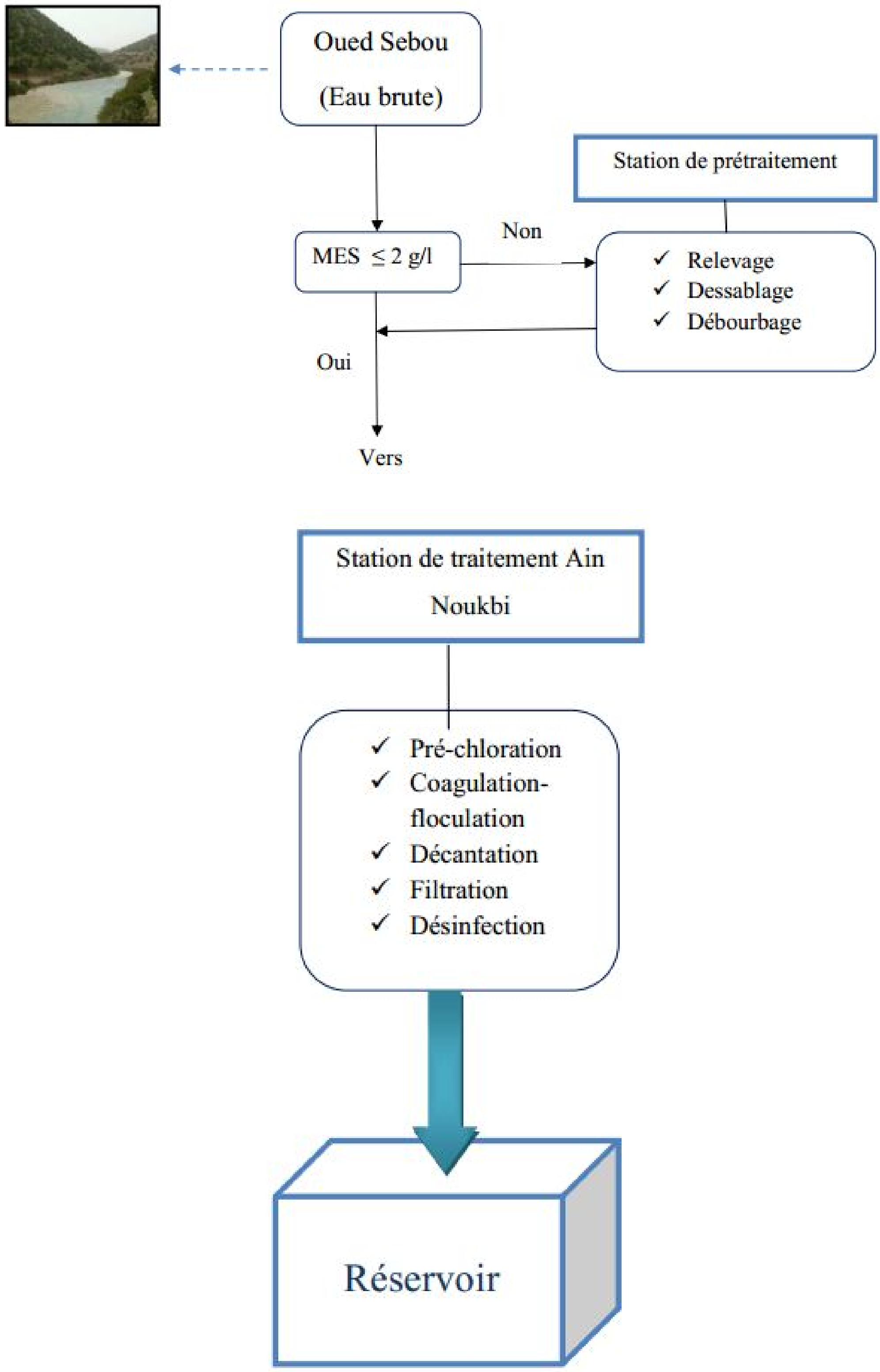


FIGURE 5 – Filière de traitement des eaux de surface

Terminologie

$*$Un **Événement extrême** est un événement qui a une faible probabilité de se produire, mais lorsqu'il se produit, prend de très petites ou de très grandes valeurs et a un grand impact.

$*$Un **Événement rare** est un événement dont la probabilité d'occurrence est faible. Le fait qu'un événement soit rare n'implique pas qu'il soit extrême : il est dépourvu de la notion de quantifiablité (petites ou grandes valeurs). A l'inverse, tout événement extrême est rare dans le sens où il a une faible probabilité de se produire.

$*$Une **Crue** est une forte augmentation, un accroissement du débit / de la hauteur d'eau en écoulement d'un fleuve, d'une rivière, d'un cours d'eau.

$*$**Inondation** est une submersion temporaire par l'eau de terres qui ne sont pas submergées en temps normal. La crue correspond à la montée des eaux d'un cours d'eau, l'inondation au phénomène qui en résulte, l'eau débordant, se répandant sur les terrains alentours.

$*$Le **débit** est la quantité d'une grandeur qui traverse une surface donnée par unité de temps. Il permet de quantifier un déplacement de matière ou d'énergie.

$*$Un **hyétogramme** est la représentation, sous la forme d'un histogramme, de l'intensité de la pluie en fonction du temps.

$*$ Un **hydrogramme** est le graphique de la variation temporelle du débit d'écoulement d'eau, mesurée au sol.

$*$**Prédétermination des crues** Par le terme prédétermination, nous entendons l'annonce d'un événement futur, avec spécification de son intensité et la probabilité d'occurrence, sans en définir précisément une date.

Soit Y une variable aléatoire continue,
$*$Une **Période de retour** est une fonction donnée par :

$$T(y) = \frac{1}{\bar{F}(y)}.$$

où $\bar{F}(y) = \mathcal{P}(Y \geq y)$.
Elle représente le nombre d'observations tel que , en moyenne, il y ait une observation égale ou supérieure à y. Il est évident que la période de retour augmente lorsque y augmente.

On peut alors définir la fonction niveau de retour comme l'inverse de la période de retour.

$$y(T) = \bar{F}^{\leftarrow}\left(\frac{1}{T}\right) = q\left(\frac{1}{T}\right).$$

$*$Un **Niveau de retour** représente le niveau (d'eau par exemple) qui sera atteint ou dépassé pour une certaine période de retour T (de probabilité $\frac{1}{T}$).

Chapitre 1

Étude probabiliste des Valeurs Extrêmes

1.1 Principe de la théorie des extrêmes

Il s'agit dans l'étude des valeurs extrêmes d'analyser l'épaisseur des queues de distributions, ou encore d'étudier les plus grandes observations d'un échantillon pour caractériser sa loi initiale. Ainsi, la théorie des extrêmes vient en complément de la théorie statistique classique où il est plus commun d'étudier le comportement d'une distribution autour de sa moyenne plutôt que dans le domaine des observations extrêmes souvent appelées événements rares. La théorie des extrêmes est fondée sur un équivalent au théorème central limite mais pour les queues de distribution.

On s'intéresse au comportement du maximum d'un échantillon$(X_1, .., X_n)$, variable aléatoire définie par

$$M_n = \max(X_1, ..., X_n),$$

On commence par des définitions essentielles pour l'étude du comportement asymptotique du maximum d'un échantillon.

Définition 1.1. *Soient F une fonction de répartition et $\bar{F} = 1 - F$, la fonction de survie associée et $\alpha \in [0, 1]$ alors, le quantile d'ordre $1 - \alpha$ de F est défini par :*

$$q(\alpha) = \bar{F}^{\leftarrow}(\alpha) = inf\{y : \bar{F}(y) \leq \alpha\}.$$

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

Un quantile sera dit extrême si l'on remplace son ordre α par une suite $\alpha_n \to 0$ quand $n \to \infty$. Le fait que $\alpha_n \to 0$ quand $n \to \infty$ indique que l'information la plus importante pour estimer les quantiles extrêmes est contenue dans la queue de distribution.(voir figure 1.1)

Définition 1.2. *Statistique d'ordre*
La statistique d'ordre de l'échantillon $(X_1, ..., X_n)$ est le réarrangement croissant de $(X_1, ..., X_n)$. On la note par $(X_{1,n}, ..., X_{n,n})$.

Le vecteur $(X_{1,1}, ..., X_{n,n})$ est appelé l'échantillon ordonné associé à l'échantillon $(X_1, ..., X_n)$, et $X_{k,n}$ étant la $k^{ième}$ statistique d'ordre.

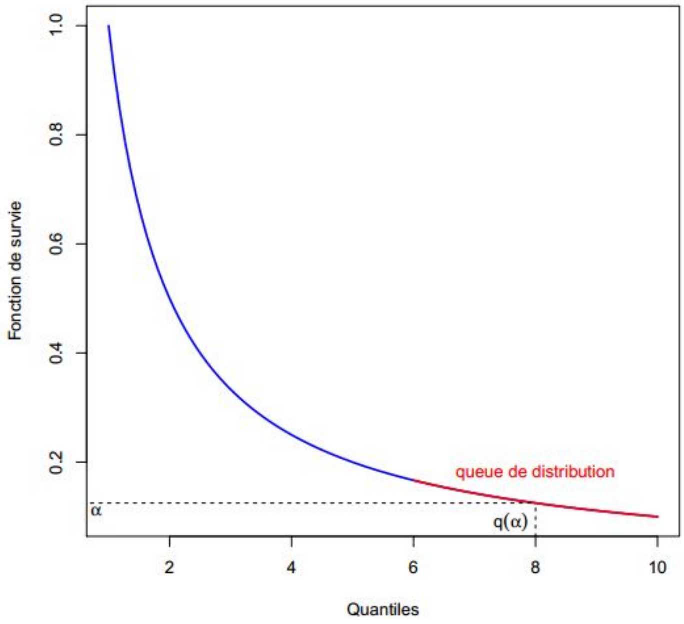


FIGURE 1.1 – Quantiles extrêmes et queue de distribution.

Dans un échantillon de taille n, une statistique d'ordre particulièrement intéressante pour les événements extrêmes est

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

Définition 1.3. *Point extrême*
Avec les notations de la définition précédente ,
On note par x_F le point extrême supérieur de la distribution F (la plus grande valeur possible pour $X_{k,n}, 1 \leq k \leq n$) et est défini par

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

Soit$(X_1, ..., X_n)$ une suite de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition F et $(X_{1,n}, ..., X_{n,n})$, la statistique d'ordre associée, alors la loi du maximum M_n est donnée par :

$$\begin{aligned} F_{M_n}(x) &= \mathbb{P}[\max(X_1, ..., X_n) \leq x] \\ &= \mathbb{P}\left[\cap_{i=1}^{n}(X_i \leq x)\right] \\ &= \prod_{i=1}^{n} P[X_i \leq x] \\ &= \prod_{i=1}^{n} F(x). \end{aligned}$$

Alors,

$$F_{M_n}(x) = [F_X(x)]^n.$$

La fonction de répartition de X n'étant souvent pas connue, il n'est généralement pas possible de déterminer la distribution du maximum à partir de ce résultat. On s'intéresse alors à la distribution asymptotique du maximum en faisant tendre n vers l'infini.On a

$$\lim_{n\to\infty} F_{M_n}(x) = \lim_{n\to+\infty}[F(x)]^n = \begin{cases} 1 \ si \ F(x) = 1, \\ 0 \ si \ F(x) < 1. \end{cases}$$

On constate que la distribution asymptotique, déterminée en faisant tendre n vers l'infini, donne une loi dégénérée (elle prend des valeurs de 0 et 1 seulement).

L'idée est de procéder à une transformation. La plus connue en statistiques est la normalisation illustrée à travers l'exemple du théorème central qui après normalisation, donne la loi asymptotique (non dégénérée) de la moyenne d'un grand nombre de variables aléatoires.

Dans la partie suivante, on commencera par énoncer le résultat fondamental de la théorie des valeurs extrêmes connu sous le nom de Fisher-Tippet-Gnedenko. Il établit la loi asymptotique du maximum de l'échantillon $X_{n,n}$ convenablement renormalisé.

1.2 Loi des extrêmes généralisées, approche block maxima

le théorème suivant montre que la seule distributions limite possible du maximum d'une variable aléatoire convenablement renormalisé est la loi généralisée des extrêmes.

Théorème 1.4. ***(Fisher-Tippet-Gnedenko-Von Mises-Jenkinson)[8]***
Soit $(X_n)_n$ une suite de variables aléatoires i.i.d.
S'il existe un réel γ et deux suites (a_n) et (b_n), $n \in \mathbb{N}$, avec $a_n > 0$ et $b_n \in \mathbb{R}$ telles que :

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

ou

$$\lim_{n\to\infty} F^n(a_n x + b_n) = H_\gamma(x).$$

*pour tout $x \in \mathbb{R}$, où H_γ est une fonction non dégénérée, alors H_γ appartient à une famille de loi appelée **Loi généralisée des valeurs extrêmes** qu'on note GEV.*

F^n est la fonction de répartition de M_n, tandis que H_γ est la fonction de répartition limite de M_n correctement renormalisée par a_n et b_n.On dira que F^n est dans le domaine d'attraction D_γ.

Remarques :

• Ce théorème présente un grand intérêt, car si l'ensemble des distributions est grand, l'ensemble des distributions des valeurs extrêmes est très petit.Ce théorème n'est valable que si les suites (a_n) et (b_n) existent et admettent des limites.
• Ce théorème est un résultat important car il n'est pas nécessaire de faire d'hypothèses sur la loi des X_i. La valeur de γ détermine le comportement de la queue de distribution.
• Le paramètre γ est un paramètre de forme encore appelé indice des valeurs extrêmes ou indice de queue, a_n est un paramètre de position et b_n, paramètre d'échelle.
• les suites de normalisation (a_n) et (b_n) ne sont pas uniques.
• Plus l'indice γ est élevé en valeur absolue, plus le poids des extrêmes dans la distribution initiale est important. On parle alors de distributions à "queues épaisses".

1.2.1 Loi des valeurs extrêmes généralisée

La forme de la distribution des valeurs extrêmes est donnée pour tout $x \in \mathbb{R}$ par

$$H_\gamma(x) = \begin{cases} \exp\{-(1+\gamma x)^{-1/\gamma}\}\mathbb{1}_{\{1+\gamma x>0\}} & si \quad \gamma \neq 0. \\ \exp\{-e^{-x}\} & sinon. \end{cases} \tag{1.1}$$

La fonction densité correspondante h_γ est donnée pour tout $x \in \mathbb{R}$ par

$$h_\gamma(x) = \begin{cases} H_\gamma(x)(1+\gamma)^{-1/\gamma-1}\mathbb{1}_{\{1+\gamma x>0\}} & si \gamma \neq 0, \\ \exp(-x-e^{-x}) & sinon. \end{cases} \tag{1.2}$$

Les lois limites possibles

Le comportement de la queue de distribution d'une suite de variables aléatoires sera complètement caractérisé par le paramètre γ. Une partie sera consacrée aux méthodes d'estimation de ce paramètre. Le signe de γ a une forte influence sur les distributions des extrêmes, et on distingue trois cas.

Domaine d'attraction de Gumbel : Lorsque $\gamma = 0$, la distribution H_0 est appelée distribution de Gumbel. Le support de cette loi est $\mathbb{R}$ et dans ce cas les queues de distribution sont légères et décroissent de manière exponentielle. H_0 sera parfois notée Λ.

Domaine d'attraction de Fréchet : Lorsque $\gamma > 0$, on a $\phi(x) := \exp(-x^{-1/\gamma})\mathbb{1}_{\{x>0\}}$. Il contient les lois dont la fonction de survie est à décroissance polynomiale. De telles distributions possèdent des queues lourdes et la convergence de F^n vers la loi limite se fait très lentement.

Domaine d'attraction de Weibull : Lorsque $\gamma < 0$, on pose $\alpha = -1/\gamma > 0$ et on note

$$\psi_\alpha(x) = \begin{cases} \exp[-(-x)^\alpha] & si\ x \text{ est négatif} \\ 1 & sinon. \end{cases}$$

Toutes les lois de ce domaine d'attraction ont un point extrême fini et la queue de distribution sera très mince.

Exemple 1.5. *Plusieurs auteurs se sont intéressés à la détermination des suites de normalisation $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$. Ainsi, Embretchs et al ont proposé dans [18] une estimation de ces constantes de normalisation dans le cas où la variable d'intérêt a pour loi, la loi normale centrée réduite. Elles sont ainsi définies,*

$$a_n = (2\log n)^{-1/2} \ et\ b_n = (2\log n)^{1/2} - \frac{\log\log n + \log 4\pi}{2(2\log n)^{1/2}} + o((\log n)^{-1/2}).$$

1.2.2 Estimation des quantiles extrêmes par l'approche de la loi GEV

On a, d'après le théorème 1.4, l'approximation suivante :

$$\mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = F^n(a_n x + b_n) \approx H_\gamma(x), \tag{1.3}$$

ou de façon équivalente, en posant $z = a_n x + b_n$

$$\mathbb{P}(M_n \leq z) \approx H_\gamma\left(\frac{z-b_n}{a_n}\right). \tag{1.4}$$

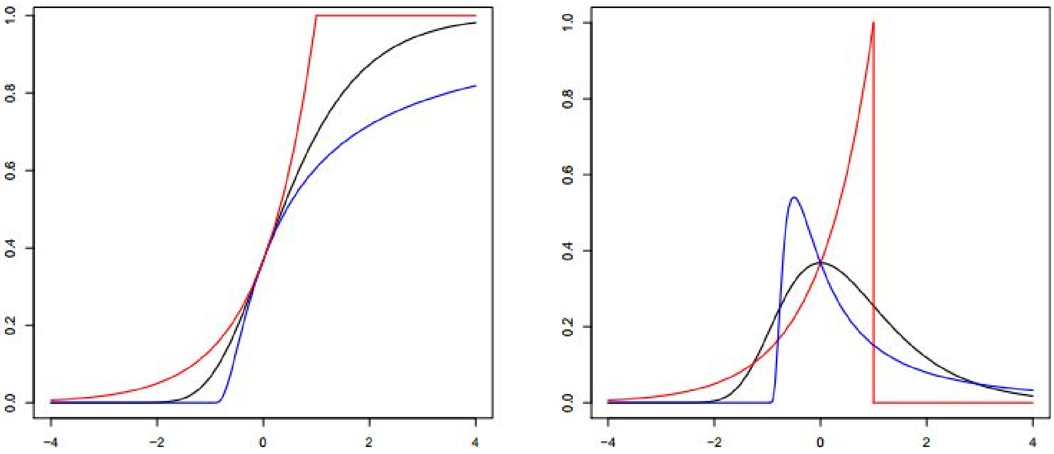


FIGURE 1.2 – Loi des extrêmes généralisée.
A gauche les fonctions de répartition. A droite les densités
(noir :$\gamma = 0$,bleu $\gamma = 1$, rouge $\gamma = -1$)

Dans la suite, on introduit la notation :

$$H_{\gamma,a_n,b_n} = H_\gamma\left(\frac{z - b_n}{a_n}\right).$$

L'équation 1.4 montre que la loi du maximum M_n peut être approchée par H_{γ,a_n,b_n} .
On peut réécrire l'équation 1.3 comme suit :

$$\lim_{n\to+\infty} n\ \log(F(a_n x + b)) = \lim_{n\to+\infty} n\ \log(1 - \bar{F}(a_n x + b)) = \log(H_\gamma(x)).$$

On peut montrer que $a_n x + b_n \to x_F$ quand $n \to \infty$ donc $\bar{F}(a_n x + b_n) \to 0$. On peut alors faire un développement limité de $\log(1 + u)$ au premier ordre ce qui donne :

$$\bar{F}(a_n x + b_n) \simeq -\frac{1}{n}\log(H_\gamma(x)).$$

ce qui est équivalent à :

$$\bar{F}(x) \simeq -\frac{1}{n}\log\left(H_\gamma\left(\frac{x - b_n}{a_n}\right)\right) = -\frac{1}{n}\log(H_{\gamma,a_n,b_n}).$$

à l'aide de l'expression de H_γ donnée dans l'expression 1.1, on obtient ainsi une approximation de la fonction de survie en queue :

$$\bar{F}(x) \simeq \frac{1}{n}\left[1 + \gamma\left(\frac{x - b_n}{a_n}\right)\right]^{-1/\gamma}, \tag{1.5}$$

que l'on peut prolonger à $\gamma = 0$ en faisant tendre $\gamma \to 0$ dans l'équation 1.5 ce qui donne

$$\bar{F}(x) \simeq \frac{1}{n}\exp\left(-\frac{x - b_n}{a_n}\right).$$

On souhaite estimer des quantiles, or par définition du quantile (voir Définition 1.1) , il nous faut inverser la fonction de survie de l'équation 1.5, ce qui nous permet d'approcher le quantile $q(\alpha)$par :

$$q(\alpha) \simeq b_n + \frac{a_n}{\gamma}\left[\left(\frac{1}{n\alpha}\right)^{\gamma} - 1\right]. \tag{1.6}$$

De même que précédemment , le cas $\gamma = 0$ dans l'expression 1.6 peut être vue comme le cas limite lorsque $\gamma \to 0$, on a alors :

$$q(\alpha) \simeq b_n - a_n \log(n\alpha). \tag{1.7}$$

On obtient alors un estimateur du quantile extrême donné dans la définition suivante.

Définition 1.6. *L'estimateur du quantile extrême de la loi GEV est défini par*

$$\widehat{q}_n(\alpha) = \widehat{b}_n + \frac{\widehat{a}_n}{\widehat{\gamma}_n}\left[\left(\frac{1}{n\alpha}\right)^{\widehat{\gamma}_n} - 1\right].$$

où $(\widehat{a}_n, \widehat{b}_n, \widehat{\gamma}_n)$ *sont respectivement des estimateurs des paramètres* (a_n, b_n, γ_n).

De même qu'auparavant, comme le cas $\gamma = 0$ peut être vu comme le cas limite lorsque $\gamma \to 0$. On a d'après l'équation 1.7

$$\widehat{q}_n(\alpha) \simeq \widehat{b}_n - \widehat{a}_n \log(n\alpha).$$

Pour mettre en pratique cette approche basée sur la convergence en loi du maximum (convenablement normalisé) d'un échantillon vers une loi GEV, Jules Emile Gumbel a introduit l'approche des maxima par bloc, en anglais "Block maxima approach".

Ainsi, si l'on dispose d'un échantillon *i.i.d* $X_1, ..., X_n$, il nous faut d'abord obtenir des maxima.L'approche des maxima par bloc consiste à séparer l'échantillon en m sous échantillons (blocs) disjoints, choisis arbitrairement assez grands. On extraira ainsi le maximum de chaque bloc, on dispose alors d'un échantillon de maxima noté $Z_1, ..., Z_m$.
La loi de ces maxima est alors approchée, pour une taille de chaque bloc assez grande, par la loi généralisée des valeurs extrêmes.

Une fois cet échantillon de maxima obtenu, on peut alors s'en servir pour estimer de diverses façons les paramètres (γ, a_n, b_n) de la loi GEV. On présentera dans le paragraphe suivant la méthode du maximum de vraisemblance et la méthode des moments pondérés.

Pour cela, considérons un échantillon $Z_1, ..., Z_m$ de m maxima *i.i.d* de loi H_{γ,a_n,b_n}.

1.2.3 Estimation des paramètres de la loi GEV

1.2.3.0.1 Méthode du Maximum de Vraisemblance (EMV)

Dans la théorie d'estimation paramétrique, la méthode du maximum de vraisemblance est la technique la plus populaire. Dans cette méthode, on retient les paramètres qui maximisent la fonction de vraisemblance ou, plus souvent le logarithme de la fonction de vraisemblance en fonction des paramètres de la famille de la loi choisie pour l'ajustement.

Si H_θ est une fonction de répartition et h_θ , la fonction de densité associée alors la fonction de vraisemblance basée sur les données $Z_1, ..., Z_n$ est définie comme suit

$$L(\theta; Z_1, ..., Z_n) := \prod_{i=1}^{n} h_\theta(Z_i) \mathbb{1}_{\{Z_i>0\}}.$$

et la fonction log-vraisemblance est

$$l(\theta; Z_1, ..., Z_n) := \log L(\theta; Z_1, ..., Z_n).$$

L'estimateur du maximum de vraisemblance de θ est définie comme suit

$$\widehat{\theta}_n = \widehat{\theta}_n(Z_1, ..., Z_n) = \underset{\theta\in\Theta}{argmax}\ l(\theta; Z_1, ..., Z_n),$$

i.e $\widehat{\theta}_n$ maximise $l(\theta; Z_1, ..., Z_n)$ sur l'espace des paramètres Θ :Une approche pour déterminer $\widehat{\theta}_n$ est résoudre le système de vraisemblance en annulant les dérivées partielles de $l(\theta; Z_1, ..., Z_n)$ en ce qui concerne les paramètres γ, a_n, b_n.

la fonction de log-vraisemblance obtenue à partir de l'expression 1.2 s'écrit :

$$\log(\mathcal{L}(\gamma, a_n, b_n)) = -m\ \log(a_n) - \left(1 + \frac{1}{\gamma}\right) \sum_{i=1}^{m} \log\left(1 + \gamma\left(\frac{Z_i - b_n}{a_n}\right)\right) - \sum_{i=1}^{m}\left(1 + \gamma\left(\frac{Z_i - b_n}{a_n}\right)\right)^{-1/\gamma},$$

avec

$$1 + \gamma\left(\frac{Z_i - b_n}{a_n}\right) > 0 \quad \forall i \in \{1, ..., m\}.$$

Le problème de maximisation de la log-vraisemblance nécessite d'avoir recours à des méthodes numériques, type algorithme de Newton-Raphson car l'estimateur du maximum de vraisemblance n'est pas explicite.Un algorithme itératif de maximisation de la fonction de vraisemblance est donné dans [2].

Les propriétés de consistance et la normalité asymptotique sont vérifiées lorsque $\gamma > -0.5$ (voir [5]).

Les résultats fournis dans [6] permettent de former un intervalle de confiance asymptotique.

Si m est le nombre de blocs de maxima , alors on aura la normalité asymptotique suivante pour $m \to \infty$:

$$\sqrt{m}((\widehat{a}_n, \widehat{\gamma}, \widehat{b}_n) - (a_n, \gamma, b_n)) \to \mathcal{N}(0, I^{-1}) \quad , \gamma > 0.5$$

où I est la matrice d'information de Fisher estimée par :

$$I(\Theta) = -\mathbb{E}\left(\frac{\partial^2 \mathcal{L}(Z; \Theta)}{\partial \Theta^2}\right).$$

où $\mathcal{L}(Z; \Theta)$ est la fonction de log-vraisemblance associée à la loi de la variable aléatoire Z, paramétrée par un ensemble de paramètres Θ. Les calculs des éléments de la matrice I pourront être trouvés dans [7].

1.2.3.0.2 Estimation des paramètres de la loi GEV par la méthode des moments pondérés

On définit le moment pondéré d'ordre r par

$$\mu_r = \mathbb{E}(ZH^r_{\gamma,a_n,b_n}(Z)).$$

Cette quantité existe pour $\gamma < 1$ et est donnée par :

$$\mu_r = \frac{1}{r+1}\left(b_n - \frac{a_n}{\gamma}(1-(r+1)^\gamma)\Gamma(1-\gamma)\right).$$

où γ est la fonction gamma définie pour tout $t > 0$ par :

$$\Gamma(t) = \int_0^\infty u^{t-1}\exp(-u)du.$$

En utilisant la formule précédente, trois moments pondérés suffisent pour calculer a_n, b_n et γ. En effet, on a :

$$\begin{aligned}
\mu_0 &= b - \frac{a_n}{\gamma}(1-\Gamma(1-\gamma)),\\
2\mu_1 - \mu_0 &= -\frac{a_n}{\gamma}(1-2^\gamma)\Gamma(1-\gamma),\\
\frac{3\mu_2-\mu_0}{2\mu_1-\mu_0} &= \frac{3^\gamma-1}{2^\gamma-1}.
\end{aligned}$$

Ainsi en remplaçant respectivement μ_r, $r \in \{0,1,2\}$ par son estimateur empirique

$$\widehat{\mu}_{r,n} = \frac{1}{m}\sum_{i=1}^m Z_{i,m}\left(\frac{i-1}{m}\right)^r.$$

où les $Z_{1,m}, ..., Z_{m,m}$ sont les statistiques ordonnées associées à l'échantillon $Z_1, ..., Z_m$ et en résolvant le système précédent, on obtient les estimateurs des moments pondérés des paramètres a_n, b_n et γ.

Dans le cas d'échantillons de petite ou de moyenne taille, la méthode des moments pondérés donne des meilleurs résultats que la méthode du maximum de vraisemblance. De plus les estimateurs des moments pondérés des paramètres sont plus simples à calculer.

1.2.3.0.3 Discussion sur la méthode block maxima

Le formalisme de la loi des extrêmes généralisée ne tient compte que d'une seule observation, la plus grande alors que seul le maximum de l'échantillon ne permet pas de modéliser le comportement des valeurs extrêmes.
L'ajustement de la loi limite sera toutefois très influencé par la taille des blocs formés à partir de l'échantillon initial.
Cette approche mène aussi à une perte d'informations sur les observations extrêmes qui sont par définition très rares. Par exemple, on peut avoir plusieurs observations extrêmes au sein du même bloc, mais seule la plus grande d'entre elles sera prise en compte.
Une alternative à la loi GEV dans la modélisation du comportement du maximum d'un échantillon est l'approche par dépassement de seuil se basant sur les "grandes valeurs" de l'échantillon.

1.3 Loi des excès, approche POT

L'approche par dépassement de seuil, en anglais "Peaks-Over-Threshold approach " notée (POT), repose sur l'utilisation des statistiques d'ordre supérieur de l'échantillon. Elle consiste à ne conserver que les observations dépassant un certain seuil. L'excès au delà du seuil est défini comme l'écart entre l'observation et le seuil.

Plus précisément, soit un échantillon de variables aléatoires $i.i.d$ $Y_1, ..., Y_n$.
Soit u un seuil fixé (non aléatoire) tel que $u < y_F$. Considérons les N_u observations $Y_{i_1}, ..Y_{i_{N_u}}$ dépassant le seuil u. On appelle excès au delà du seuil u les $Z_j := Y_{i_j} - u$, où $j = 1, ..., N_u$.

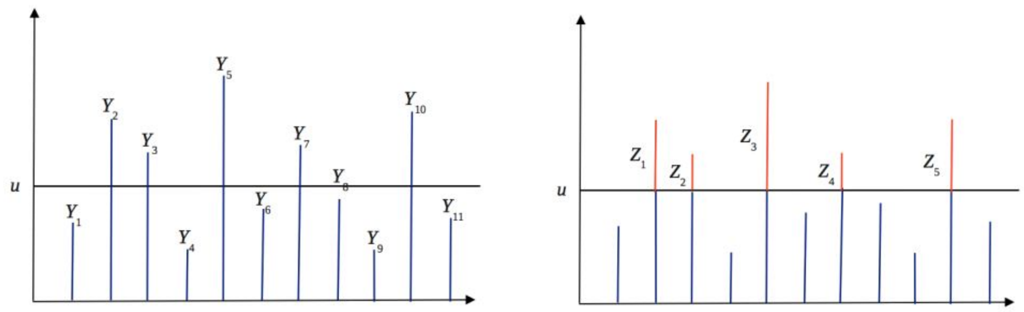


FIGURE 1.3 – Excès au delà du seuil u.

On notera F_u la fonction de répartition de l'excès Z au delà du seuil u. La loi des excès est celle de variables aléatoires $i.i.d$ admettant pour fonction de répartition $F_u(y) = \mathbb{P}(Z \leq y | Y > u)$ représentant la probabilité que la variable aléatoire Y ne dépasse pas le seuil u d'au moins une quantité y sachant qu'elle dépasse u. F_u décrit ainsi la loi de Z sachant que $Y > u$. On peut la réécrire en fonction de F à l'aide du résultat suivant.

On a pour $y \geq 0$:

$$F_u(y) = \mathbb{P}[Z \leq y | Y > u] = \mathbb{P}[Y - u \leq y | Y > u] = \frac{F(u+y) - F(u)}{1 - F(u)}, \tag{1.8}$$

ou de manière équivalente pour la fonction de survie :

$$\bar{F}_u(y) := \mathbb{P}[Z > y | Y > u] = 1 - F_u(y) = \frac{\bar{F}(u+y)}{\bar{F}(u)}. \tag{1.9}$$

L'idée de l'existence d'une équivalence entre la loi GEV et la loi des excès est la suivante.
A l'aide du résultat donné dans le théorème 1.4 pour n assez grand on a ,

$$F^n(u) \approx \exp\left(- \left(1 + \gamma \left(\frac{u - b_n}{a_n} \right) \right)^{-1/\gamma} \right),$$

avec $a_n > 0$ et $(b_n, \gamma) \in \mathbb{R}^2$. Ainsi,

$$n \log(F(u)) \approx - \left(1 + \gamma \left(\frac{u - b_n}{a_n} \right) \right)^{-1/\gamma}. \tag{1.10}$$

Si u est assez grand alors un développement limité donne :

$$\log(F(u)) \approx -(1 - F(u)).$$

En remplaçant dans l'équation 1.10 on obtient pour u assez grand

$$1 - F(u) \approx \frac{1}{n}\left(1 + \gamma\left(\frac{u - b_n}{a_n}\right)\right)^{-1/\gamma}.$$

De même pour $y > 0$ on a,

$$1 - F(u + y) \approx \frac{1}{n}\left(1 + \gamma\left(\frac{u + y - b_n}{a_n}\right)\right)^{-1/\gamma}.$$

En remplaçant dans l'équation 1.9 on obtient :

$$F_u(y) \approx 1 - \left(1 + \gamma\frac{y}{\sigma}\right)^{-1/\gamma},$$

avec

$$\sigma = a_n + \gamma(u - b_n). \tag{1.11}$$

Le théorème suivant donne un résultat très précis sur l'approximation de la fonction de répartition F_u lorsque le seuil u est proche du point extrême y_F.

1.3.1 Théorème de Pickands

Ce théorème montre que la distribution des excès au delà d'un certain seuil u peut être approchée par une loi de Pareto généralisée qu'on notera GPD (Generalized Pareto Distribution) :

Théorème 1.7. *Pickands [9] Soient $X_1, X_2, ..., X_n$ un échantillon de variables aléatoires i.i.d suivant la loi F. F appartient au domaine d'attraction de H_γ si et seulement si il existe une fonction positive $\sigma(.)$ positive telle que*

$$\lim_{u \to x^*} \underset{0<x<x^*}{sup} \{|\mathbb{P}(X - u \leq x | X > u) - \mathcal{G}_{\gamma,\sigma}(x)|\} = 0 \tag{1.12}$$

où $x^* = sup\{x; F_u < 1\}$ et $\mathcal{G}_{\gamma,\sigma}$ est la distribution de Pareto généralisée (GPD) définie par :

$$\mathcal{G}_{\gamma,\sigma}(x) = \begin{cases} 1 - \left(1 + \gamma\frac{x}{\sigma}\right)^{-1/\gamma} & si \gamma \neq 0, x \geq 0;\ et\ x < -\frac{\sigma}{\gamma}\ si\ \gamma < 0 \\ 1 - \exp\{-\frac{x}{\sigma}\} & si\ \gamma = 0, x \geq 0 \end{cases}$$

Remarque :

• Le théorème précédent établit l'équivalence entre la convergence en loi du maximum (convenablement normalisé) d'un échantillon vers une loi des valeurs extrêmes H_γ et la convergence en loi des excès au-delà d'un seuil vers une loi de Pareto généralisée $\mathcal{G}_{\gamma,\sigma}$, lorsque le seuil tend vers la limite supérieure du support de F. Ainsi, les paramètres de la loi de Pareto généralisée ont un lien avec ceux de la loi du GEV.Le paramètre de forme γ est le même que dans la loi GEV, il caractérise donc la lourdeur de la queue de la loi .
Le paramètre d'échelle σ de la GPD est tel que $\sigma = a_n + \gamma(u - b_n)$ (voir l'expression 1.11).

1.3.2 Estimation des quantiles extrêmes par l'approche de la loi des excès

l'approche par la loi des excès est basée sur l'idée suivante. On a d'après l'équation 1.9, pour tout $y \geq 0$ la relation

$$\bar{F}(u+y) = \bar{F}(u)\bar{F}_u(y)$$

Si on effectue le changement de variable $z = u + y$, alors l'approximation de la queue de distribution donne

$$\bar{F}(z) = \bar{F}(u)\bar{F}_u(z-u) \approx \bar{F}(u)\bar{\mathcal{G}}_{\gamma,\sigma}(z-u),$$

où $\bar{\mathcal{G}}_{\gamma,\sigma}$ est la fonction de survie de la loi de Pareto généralisée $\mathcal{G}_{\gamma,\sigma}$ donnée dans le théorème 1.7. On introduit alors la probabilité p que Y dépasse le seuil u,

$$p = \mathbb{P}(Y > u) = \bar{F}(u)$$

d'où :

$$\bar{F}(z) \simeq p\,\bar{\mathcal{G}}_{\gamma,\sigma}(z - \bar{F}^{\leftarrow}(p)).$$

On obtient ainsi pour $\gamma \in \mathbb{R}$ une approximation de la fonction de survie en queue :

$$\bar{F}(z) \simeq p\left(1 + \gamma\left(\frac{z - \bar{F}^{\leftarrow}(p)}{\sigma}\right)\right)^{-1/\gamma}. \tag{1.13}$$

Le cas $\gamma = 0$ peut être vu comme le cas limite $\gamma \to 0$ dans l'équation 1.13 :

$$\bar{F}(z) \simeq p\,exp\left(-\frac{z - \bar{F}^{\leftarrow}(p)}{\sigma}\right).$$

On souhaite estimer des quantiles, or par définition du quantile (voir Définition 1.1), il nous faut inverser la fonction de survie donnée dans l'équation 1.13, ce qui donne :

$$q(\alpha) \simeq \bar{F}^{\leftarrow}(p) + \frac{\sigma}{\gamma}\left(\left(\frac{\alpha}{p}\right)^{-\gamma} - 1\right), \tag{1.14}$$

et dans le cas $\gamma = 0$, on fait tendre l'équation 1.14 ce qui donne :

$$q(\alpha) \simeq \bar{F}^{\leftarrow}(p) - \sigma log\left(\frac{\sigma}{p}\right). \tag{1.15}$$

1.3.2.0.1 Lien entre l'approche block maxima et l'approche par dépassement de seuil On notera la similitude entre l'expression du quantile de la loi GEV (voir l'équation 1.6) et celle du quantile de la loi GPD (voir l'équation 1.14). Il y a trois paramètres inconnus dans chacune d'entre elles :

$\ast$ L'indice des valeurs extrêmes γ qui est le même dans les deux expressions soulignant son importance dans le comportement de la queue de distribution et donc celui des valeurs extrêmes.
$\ast$ Le paramètre d'échelle σ joue le rôle de a_n dans l'approche GEV.
$\ast$ Le seuil $u = \bar{F}^{\leftarrow}(p)$ joue le rôle de b_n l'approche GEV.
Pour pouvoir estimer les quantiles extrêmes, il nous faut donc estimer ces paramètres.

Définition 1.8. *L'estimateur du quantile extrêmes de la loi GPD est défini par :*

$$\widehat{q}(\alpha_n) \simeq u + \frac{\widehat{\sigma}}{\widehat{\gamma}} \left(\left(\frac{\alpha_n}{p} \right)^{-\widehat{\gamma}_n} - 1 \right),$$

où, $\widehat{\gamma}_n$ *et* $\widehat{\sigma}_n$ *sont des estimateurs des paramètres de forme et d'échelle.*

Remarque. Différentes valeurs du seuil u donneront différents échantillons d'excès plus ou moins grands, ce qui influencera l'estimation des quantiles extrêmes. Ce seuil doit être suffisamment grand pour vérifier le caractère asymptotique du modèle et que l'on puisse appliquer le théorème 1.7, mais pas trop grand pour garder un nombre suffisant d'observations qui dépassent ce seuil afin de pouvoir estimer les paramètres du modèle car sinon on utilisera peu d'excès donc peu d'informations.

Choix du seuil en pratique

$*$ **Première méthode** : On fixe le nombre d'excès, noté k_n qui doit tendre vers l'infini avec la taille n de l'échantillon mais rester petit devant n pour que le seuil u_n soit suffisamment grand, $i.e$:

$$\lim_{n\to+\infty} k_n = \infty \; et \; \lim_{n\to+\infty} \frac{k_n}{n} = 0.$$

Ainsi, si $p_n = \dfrac{k_n}{n}$, alors on choisit le seuil u comme étant le quantile d'ordre $1 - p_n$ de la loi des données ($i.e$ $u_n = \bar{F}^{\leftarrow}(p_n)$) qu'on pourra estimer par la statistique d'ordre Y_{n-k_n+1}.

$*$ **Deuxième méthode** :Généralement, le seuil est déterminé graphiquement par la fonction moyenne des excès qui permet de décrire la prédiction du dépassement du seuil u lorsqu'un excès se produit et qui est définie par :

$$e_n(u) = \mathbb{E}(Y - u | Y > u).$$

Cette fonction moyenne des excès est estimée par la somme des excès dépassant un certain seuil élevé u, divisé par le nombre d'observations qui dépassent ce seuil (FME empirique).

$$\widehat{e}_n(u) = \frac{\sum_{i=1}^{n} (X_i - u)^+}{\sum_{i=1}^{n} \mathbb{1}_{\{X_i > u\}}}. \tag{1.16}$$

où $(X_i - u)^+ = sup(X_i - u, 0)$.

Cette approche pratique visant à choisir u consiste à tracer l'estimateur empirique de l'espérance résiduelle de X et à choisir u de manière à ce que $\widehat{e}_n(u)$ soit approximativement linéaire pour tout $x > u$.

le graphique

$$\{(u_n, \widehat{e}_n(u)); u \leq Y_{k_n,k_n}\} \; avec; Y_{k_n,k_n} = \max_{i\in\{1,\ldots,k_n\}} Y_i \; .$$

est appelé en anglais "Mean residual life plot".

Une fois le seuil u choisi, il nous reste à estimer les paramètres γ et σ afin d'obtenir un estimateur du quantile extrêmes $q(\alpha_n)$, ce qui fera l'objet de la partie suivante.

1.4 Estimation des paramètres de la loi des excès

1.4.0.0.1 Méthode du Maximum de Vraisemblance (EMV)

La fonction de log-vraisemblance est obtenue à partir de la loi GPD $\mathcal{G}_{\gamma,\sigma}$ (voir Théorème 1.7) ce qui donne :

$$log(\mathcal{L}(\gamma,\sigma)) = -k_n log(\sigma) - \left(1+\frac{1}{\gamma}\right)\sum_{i=1}^{k_n} log\left(1+\frac{\gamma}{\sigma}Z_i\right).$$

avec $1+\gamma Z_i/\sigma > 0$ pour $i = 1, ..., k_n$ sinon $\mathcal{L}(\gamma,\sigma) = -\infty$

Il n'existe pas d'expression explicite des estimateurs solutions des équations de vraisemblance. Dans la pratique, les méthodes de résolution numérique telles que l'algorithme de Newton-Raphson sont utilisées pour approcher les estimateurs voir [2].

Remarques. $*$ D'après [5] Sous l'hypothèse $\gamma > -1/2$, les estimateurs du maximum de vraisemblance sont asymptotiquement gaussiens et efficaces.
On a le résultat suivant pour $\gamma > -1/2$ et $k_n \to \infty$:

$$\sqrt{n}((\widehat{\gamma},\widehat{\sigma}) - (\gamma,\sigma)) \to \mathcal{N}(0,\Delta).$$

avec

$$\Delta = (1+\gamma)\begin{bmatrix} 1+\gamma & -\sigma \\ -\sigma & 2\sigma^2 \end{bmatrix}$$

On peut alors construire les intervalles de confiance de niveau $(1-\alpha)$ suivant pour les paramètres γ et σ :

$$IC_{1-\alpha}(\gamma) = \left[\widehat{\gamma} - \phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{1+\widehat{\gamma}}{\sqrt{k_n}};\widehat{\gamma} + \phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{1+\widehat{\gamma}}{\sqrt{k_n}}\right].$$

$$IC_{1-\alpha}(\sigma) = \left[\widehat{\sigma} - \phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{2(1+\widehat{\gamma})}{k_n}}\widehat{\sigma};\widehat{\sigma} + \phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{2(1+\widehat{\gamma})}{k_n}}\widehat{\sigma}\right].$$

ϕ^{-1} désigne la fonction quantile d'une loi normale centrée réduite.

$*$ L'estimateur du maximum de vraisemblance est cependant peu utilisé en pratique car il pose des problèmes numériques et il est peu performant sur des échantillons de petite taille ($n < 500$). Il est préférable alors dans ce cas d'estimer les estimateurs par la méthode suivante.

1.4.0.0.2 Méthode des moments

L'espérance et la variance d'une variables aléatoire Z de la loi GPD $\mathcal{G}_{\gamma,\sigma}$ existent si $\gamma < 1/2$. Dans ce cas, on a :

$$\mathbb{E}(Z) = \frac{\sigma}{1-\gamma} \; et \; Var(Z) = \frac{\sigma^2}{(1-\gamma)^2(1-2\gamma)}.$$

On peut alors exprimer les paramètres de la loi GPD γ et σ en fonction de l'espérance et de la variance de Z, soit :

$$\gamma = \frac{1}{2}\left(1-\frac{\mathbb{E}(Z)^2}{Var(Z)}\right) \; et \; \sigma = \frac{\mathbb{E}(Z)}{2}\left(1+\frac{\mathbb{E}(Z)^2}{Var(Z)}\right).$$

Ainsi, en remplaçant $\mathbb{E}(Z)$ et $Var(Z)$ par leurs estimations empiriques :

$$\bar{Z} := \frac{1}{k_n}\sum_{i=1}^{k_n} Z_i \;\; et \;\; s^2(Z) := \frac{1}{k_n-1}\sum_{i=1}^{k_n}(Z_i-\bar{Z})^2.$$

On obtient les estimateurs des moments γ et σ, soit :

$$\widehat{\gamma}_n = \frac{1}{2}\left(1 - \frac{\bar{Z}^2}{s^2(Z)}\right) \quad et \quad \widehat{\sigma}_n = \frac{\bar{Z}}{2}\left(1 + \frac{\bar{Z}^2}{s^2(Z)}\right).$$

Ces estimateurs sont asymptotiquement gaussiens si $\gamma < 1/4$ voir[7] .

1.5 Analyse exploratoire des données

L'information préliminaire utile sur un ensemble de données à analyser, peut être obtenue par plusieurs résultats graphiques et analytiques assez faciles. On commence habituellement l'analyse statistique des données par la détermination des statistiques de base (moyenne, variance,...) et dessin de nuages points, histogrammes, box-plots...en outre, dans l'analyse des extrêmes, la première tâche consiste en étudiant le poids de queue des données.

Définition 1.9. *QQ-plot*
Le QQ-plot est un graphique qui oppose les quantiles de la distribution empirique aux quantiles de la distribution théorique envisagée . Si l'échantillon provient bien de cette distribution théorique, alors le QQ-plot sera linéaire.

Dans la théorie des extrêmes, le QQ-plot sous l'hypothèse d'une distribution exponentielle est la représentation des quantiles de la distribution empirique sur l'axe des X contre les quantiles de la fonction de distribution exponentielle sur l'axe des Y.

Le graphique est l'ensemble des points tel que :

$$\{\left(X_{k,n}, G_{0,1}^{\leftarrow}\left(\frac{n-k+1}{n+1}\right)\right), k = 1, ..., n\},$$

$X_{k,n}$:Représente le k-ème ordre statistique et $G_{0,1}^{\leftarrow}$ est la fonction inverse de la distribution exponentielle.
L'intérêt de ce graphique est de nous permettre d'obtenir la forme de la queue de distribution. Trois cas de figure sont possibles :

— Les données suivent la loi exponentielle : La distribution présente une queue très légère, les points du graphique présentent une forme linéaire.

— Les données suivent une distribution à queue épaisse : Le graphique QQ-plot est concave.

— Les données suivent une distribution de queue légère :Le graphique QQ-plot a une forme convexe.

Définition 1.10. *ME-plot Le graphique ME-plot permet de déterminer le seuil (à partir duquel , les valeurs sont considérées comme extrêmes) nécessaire pour la modélisation de l'approche GPD .*
LE ME-plot est défini de la manière suivante :

$$\{(X_{k,n}, e_n(X_{k,n})) : k = 1, ..., n\}.$$

Avec $e_n(u)$ est l'estimatoire empirique de la fonction moyenne des dépassements :

$$e(u) = E[X - u/X > u].$$

1.5.1 Application : Estimation probabiliste des débits maximums annuels

Pour appliquer l'approche block maxima de la théorie des valeurs extrêmes ainsi que la méthode par dépassement de seuil, on a utilisé les débits maximums annuels de Oued Sebou depuis 1957 jusqu'à 2008 représentés dans le tableau 1.1.

On commence par explorer les données de la variable aléatoire qui nous intéresse à savoir : Le débit. Dans un premier temps, on présente dans le tableau 1.3 les valeurs des statistiques générales à savoir : la médiane, la moyenne, le premier et le 3 ème quantile , le min et le max de la série d'observations.

Ensuite, on représente dans la figure 3.3 différents graphes permettant d'explorer les valeurs extrêmes des débits comme le QQ-plot et le Mean excess plot.

Les graphes de la figure 3.3 représentent les débits estimés par la théorie des valeurs extrêmes. A gauche, les débits extrêmes sont estimés en utilisant la loi des excès et à droite, en utilisant la loi généralisée des extrêmes.

Pour estimer les quantiles extrêmes des débits, on a utilisé le package fExtremes du logiciel R qui contient les fonctions gev et gpd dédiées respectivement à la loi généralisée des extrêmes et à la loi des excès.

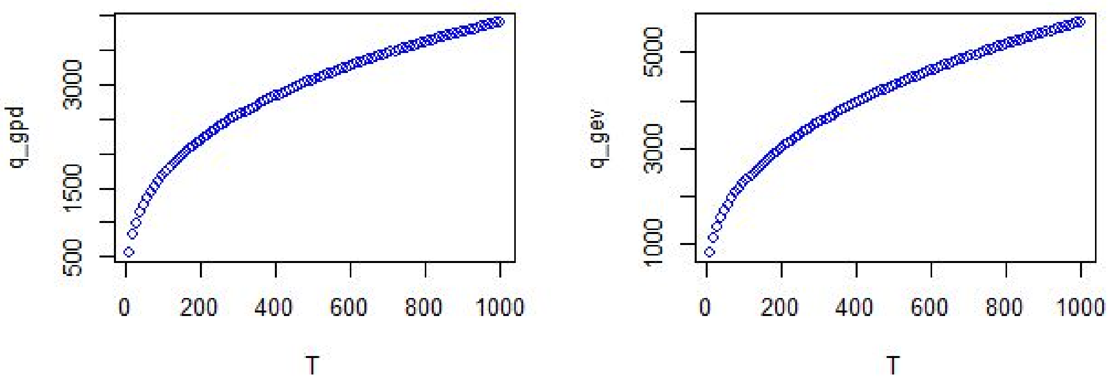


Figure 1.4 – Estimation des débits maximums en fonction des périodes de retour fixées . Àgauche :par l'approche GPD, Àdroite :par l'approche GEV

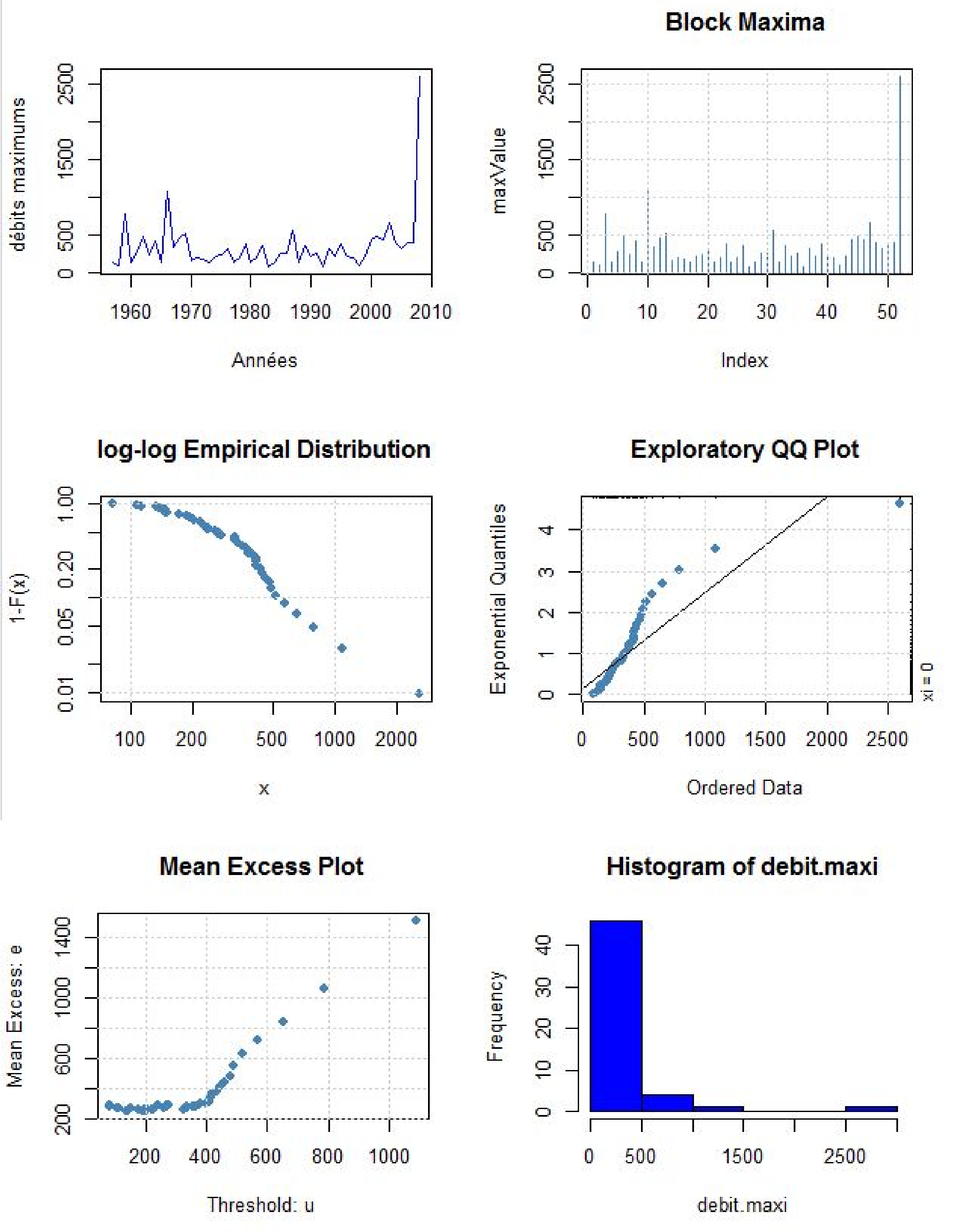


FIGURE 1.5 – Analyse exploratoire des données des débits maximums annuels .

Année	Débits (m^3/s)
1957	148
1958	113
1959	787
1960	152
1961	278
1962	481
1963	239
1964	415
1965	141
1966	1090
1967	335
1968	461
1969	517
1970	174
1971	203
1972	189
1973	139
1974	232
1975	238
1976	326
1977	148
1978	205
1979	380
1980	149
1981	195
1982	367
1983	81.6
1984	135
1985	261
1986	269
1987	570
1988	149
1989	356
1990	221
1991	270
1992	82.3
1993	324
1994	224
1995	376
1996	221
1997	195
1998	107
1999	229
2000	435
2001	490
2002	445

Table 1.1 – Les débits maximums annuels de Oued Sebou depuis 1957 jusqu'à 2002

Année	Débits (m^3/s)
2003	654
2004	414
2005	333
2006	411
2007	400
2008	2600

Table 1.2 – Les débits maximums annuels de Oued Sebou depuis 2003 jusqu'à 2008

Min	1er quantile	Médiane	Moyenne	3e quantile	Max
81.6	185.2	265.0	353	411.8	2600

Table 1.3 – Statistiques générales sur les données des débits maximums annuels

Commentaires sur la figure 3.3 :

• La fonction de répartition empirique attribue la probabilité $\frac{1}{n}$ à chacun des n nombres dans un échantillon.
La fonction de distribution empirique $F_n(x)$ basée sur l'échantillon $X_1, ..., X_n$ est définie par : Pour un x fixé,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

• Le graphique QQ-plot présente une concavité, ce qui montre que les données suivent une distribution à queue épaisse.

• D'après le ME-plot, à partir de la valeur de l'abscisse 420, le graphe devient linéaire. Alors, le seuil à partir duquel les valeurs sont considérées extrêmes peut être estimé à 420.

Chapitre 2

Méthodes de Monte Carlo

2.1 Principe de la méthode

Les méthodes de Monte Carlo permettent d'estimer des quantités en utilisant la simulation de variables aléatoires.

Les problèmes pouvant être rencontrés comprennent le calcul d'intégrales, les problèmes d'optimisation, la résolution des systèmes linéaires...La simplicité, la flexibilité et l'efficacité de la méthode pour les problèmes en grande dimension en font un outil intéressant, pouvant servir d'alternative ou de référence pour d'autres méthodes numériques.

La première étape de la méthode consiste à écrire le problème sous la forme d'une espérance.

Soit une variable aléatoire $\mathbf{X} = (X_1, ..., X_d)$ de loi ν sur $\mathbb{R}^d$ (on abrégera cela par $X \sim \nu$) et une fonction $h : \mathbb{R}^d \longrightarrow \mathbb{R}$. Le problème traité par les méthodes de Monte Carlo est l'estimation de

$$\mathbb{E}_\nu[h(X)] = \int_{\mathbb{R}^d} h(x)\nu(dx)$$

La solution standard à ce problème est de simuler une suite $(\mathbf{X}_n)_{n\geq 1} = (X_{1,1}, ..., X_{d,n})$ de variables aléatoires indépendantes identiquement distribuées (i.i.d) suivant la loi ν, puis d'estimer l'espérance $\mathbb{E}_\nu[h(X)]$ par la moyenne empirique, ie :

$$\mathbb{E}_\nu[h(X)] \approx \frac{1}{n}\sum_{k=1}^{n} h(\mathbf{X}_k) := \frac{1}{n}\sum_{k=1}^{n} Y_k := \bar{Y_n} \tag{2.1}$$

Remarque. Lorsqu'il n'y a pas d'ambiguïté sur la loi de $\mathbf{X}$, on omettra ν dans les notations.

Exemple 2.1. ***Calcul d'une intégrale*** *Soit $h : [a, b]^d \longrightarrow \mathbb{R}^d$. On cherche à calculer*

$$I = \int_{\mathbb{R}^d} h(x_1, ...x_d)dx_1...dx_d.$$

On peut réécrire I sous la forme

$$I = (b-a)^d \int_{\mathbb{R}^d} h(x_1, ...x_d)\frac{1}{(b-a)^d}dx_1...dx_d.$$

Si on pose $\boldsymbol{X} = (X_1, .., X_d)$ un $d-$uplet de variables i.i.d suivant la loi uniforme sur $[a, b]$, on a alors

$$I = (b-a)^d\mathbb{E}[h(X)].$$

Si nous cherchons à évaluer une intégrale de la forme

$$I = \int_{\mathbb{R}^d} h(x)f(x)dx.$$

avec f une densité de probabilité sur $\mathbb{R}^d$ et h, une fonction borélienne, alors nous pouvons écrire, sous les hypothèses d'existence de I, $I = \mathbb{E}[h(X)]$ et l'estimer en utilisant la méthode de Monte Carlo.

2.2 Validité et comportement de la méthode

2.2.1 Convergence de la méthode

La convergence de la méthode de Monte Carlo est assurée par la loi forte des grands nombres, sous l'hypothèse que h est intégrable par rapport à la mesure ν .

Théorème 2.2. *Loi forte des grands nombres*

Soit $(Y_n)_{n\geq 1}$ une suite de variables aléatoires réelles i.i.d intégrables, i.e $\mathbb{E}[|Y_1|] < \infty$.Alors on a la convergence presque sure p.s de la moyenne empirique

$$\bar{Y_n} = \frac{1}{n}\sum_{k=1}^{n} Y_k \xrightarrow[n\to+\infty]{p.s} \mathbb{E}[Y_1].$$

Exemple 2.3. ***Estimation de*** π

Supposons que (X, Y) suive la loi uniforme sur le carré $C = [0, 1] \times [0, 1]$ et que $\phi(x, y) = \mathbb{1}_{x^2+y^2\leq 1}$. En notant $D = \left\{(x, y) \in {\mathbb{R}_+}^2, x^2 + y^2 \leq 1\right\}$ le quart du disque unité, on a donc :

$$I = \int\int_{C} \mathbb{1}_D(x, y)dxdy = \frac{\pi}{4}. \tag{2.2}$$

En simulant des points (X_i, Y_i) uniformément dans C, la propriété précédente assure donc que

$$\widehat{I_n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_D(X_i, Y_i) \xrightarrow[n\to+\infty]{p.s} \frac{\pi}{4} \Leftrightarrow 4 \times \widehat{I_n} \xrightarrow[n\to+\infty]{p.s} \pi.$$

On dispose donc d'un estimateur Monte Carlo pour la constante π.

2.2.2 Vitesse de convergence et erreur d'estimation

Sous l'hypothèse que h est de carré intégrable par rapport à la mesure ν, la vitesse de convergence de la méthode de Monte Carlo est donnée par le théorème suivant.

Théorème 2.4. *Théorème central limite*

Soit $(Y_n)_{n\geq 1}$ une suite de variables aléatoires réelles i.i.d de carré intégrables, i.e, $\mathbb{E}[Y_1^2] < \infty$.
Notons $\sigma^2 = \mathbb{V}ar[Y_1]$,alors

$$\sqrt{n}(\bar{Y_n} - \mathbb{E}[Y_1]) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

Rappel :

Le symbole $\xrightarrow{\mathcal{L}}$ signifie "converge en loi". On dit que la suite $(Y_n)_{n\geq 1}$ converge en loi vers Y si pour toute fonction continue bornée ψ, on a

$$\mathbb{E}[\psi(Y_n)] \xrightarrow{n\to\infty} \mathbb{E}[\psi(Y)].$$

Remarque.

Si nous cherchons à approcher $\mathbb{E}[h(X)]$ par une méthode de Monte Carlo, Le théorème 2.4 renseigne que l'erreur $e_n = \bar{Y_n} - \mathbb{E}[Y_1]$ est d'ordre $\dfrac{\sigma}{\sqrt{n}}$ mais cette erreur est aléatoire .En effet, e_n suit (approximativement) une loi normale $\mathcal{N}(0,\sigma^2)$ donc, elle ne peut être bornée mais elle peut être quantifiée via un intervalle de confiance. Sous les hypothèses du théorème central limite, pour tout $a < b$ réels,

$$\mathbb{P}\left[\bar{Y_n} + a\sqrt{\frac{\sigma^2}{n}} \leq \mathbb{E}[Y_1] \leq \bar{Y_n} + b\sqrt{\frac{\sigma^2}{n}}\right] \underset{n\to\infty}{\longrightarrow} \int_a^b \frac{1}{\sqrt{2\pi}} \exp(\frac{-t^2}{2})dt.$$

Proposition 2.5. ***Intervalles de confiance***

Soit $\alpha \in [0,1]$ fixé. Sous les hypothèses du théorème 2.4,un intervalle de confiance de niveau asymptotique α pour $\bar{Y_n}$ défini en 2.1 est donné par :

$$IC_{\alpha,n} = \left[\bar{Y_n} - \phi^{-1}(1-\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}; \bar{Y_n} + \phi^{-1}(1-\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right].$$

où $\phi^{-1}(1-\dfrac{\alpha}{2})$ désigne le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite.

Remarques.

- Une fois la suite $(X_n)_{n\in\mathbb{N}}$ simulée, donner uniquement la valeur de $\bar{Y_n}$ définie par 2.1 est insuffisant.Cette dernière doit être accompagnée de l'intervalle de confiance $IC_{\alpha,n}$ au niveau de confiance α.
- Néanmoins, dans la pratique, σ^2 n'est pas connue. Pour obtenir une approximation de l'intervalle de confiance, cette variance peut être estimée par une méthode de Monte Carlo comme le montre le lemme suivant.

Lemme 2.6. *Estimateur de la variance*
Si $X_1, X_2, ...$ sont i.i.d avec $\mathbb{E}(X_1^2) < \infty$ et $\sigma^2 = V(X)$, alors

$$\widehat{\sigma}_n^2 = \frac{(X_1-X_2)^2 + (X_3-X_4)^2 + ... + (X_{2n}-X_{2n-1})^2}{2n} \xrightarrow[n\to+\infty]{p.s} \sigma^2.$$

Démonstration. Les variables $(X_1-X_2)^2, (X_3-X_4)^2, ...$ sont $i.i.d$ avec

$\mathbb{E}[(X_1-X_2)^2] = 2\mathbb{E}(X_1^2) - 2\mathbb{E}(X_1)^2 < +\infty$ (car $\mathbb{E}(X_1^2) > E(X_1)^2$ et $\mathbb{E}(X_1^2) < \infty$) donc , le résultat découle de la loi des grands nombres.

□

Lemme 2.7. *Sous les hypothèses du théorème 2.4 , on a :*

$$\sqrt{\frac{n}{\widehat{\sigma_n}^2}}(\bar{Y_n} - \mathbb{E}[Y_1]) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Rappel 2.8. *Le lemme 2.7 est une conséquence du théorème Slutsky.*

Théorème 2.9. *Théorème de Slutsky*
Soient $(Y_n)_{n\in\mathbb{N}}$ et $(Z_n)_{n\in\mathbb{N}}$ deux suites de variables aléatoires. S'il existe une variable aléatoire Y telle que $(Y_n)_{n\in\mathbb{N}}$ converge en loi vers Y, et une constante c telle que $(Z_n)_{n\in\mathbb{N}}$ converge en probabilité vers c, alors $(Y_n, Z_n)_{n\geq 1}$ converge en loi vers (Y, c). En particulier , $Z_nY_n \xrightarrow[n\to+\infty]{\mathcal{L}} cY$.

Il résulte du lemme 2.7 que l'intervalle de confiance au niveau α pour $\mathbb{E}[h(X)]$ définie en 2.1 s'écrit :

$$[\bar{Y_n} - c_\alpha\sqrt{\frac{\widehat{\sigma_n}^2}{n}}; \bar{Y_n} + c_\alpha\sqrt{\frac{\widehat{\sigma_n}^2}{n}}]$$

avec $c_\alpha > 0$ et vérifie $P(|Z| < c_\alpha) = \alpha$ où $Z \sim \mathcal{N}(0,1)$.

Exemple 2.10. *Soit U une variable aléatoire dans $\mathbb{R}^d$ $d \in \mathbb{N}^*$ et $f : \mathbb{R}^d \to \mathbb{R}$ mesurable.*

On cherche à calculer $p = P(f(U) \leq \lambda)$ pour un certain λ. Soit $X = \mathbb{1}_{f(U)\leq\lambda}$.Alors $p = E(X)$. On peut approcher p en faisant des tirages i.i.d $X_1, X_2, ...$ de même loi que X par

$$p_n = \frac{X_1 + ... + X_n}{n} \approx p.$$

Nous calculons $\sigma^2 = V(X) = p(1-p)$. Sous les hypothèses du théorème 2.4, notons $e_n = p_n - p$. Si nous voulons une erreur $|e_n|$ inférieure à 0.01 avec un niveau de confiance 95%, c'est à dire que l'on veut

$$P(|e_n| \geq 0.01) \leq 0.05.$$

ce qui est équivalent à

$$P(|e_n| \leq 0.01) \geq 0.95.$$

Calculons :

$$\begin{aligned}
P(|e_n| \leq 0.01) =& P\left(-0.01 \leq \frac{X_1 + ... + X_n}{n} - E(X_1) \leq 0.01\right)\\
=& P\left(-0.01\frac{\sqrt{n}}{\sigma} \leq \frac{\sqrt{n}}{\sigma}\left(\frac{X_1 + ... + X_n}{n} - E(X_1)\right) \leq \frac{\sqrt{n}}{\sigma}0.01\right)\\
(\textit{pour } n \textit{ assez grand}) \approx& \int_{-0.01\frac{\sqrt{n}}{\sigma}}^{+0.01\frac{\sqrt{n}}{\sigma}} \frac{e^{\frac{-t^2}{2}}}{\sqrt{2\pi}}dt\\
(\textit{par symtrie de } \mathcal{N}(0,1)) =& 2\int_{-\infty}^{+0.01\frac{\sqrt{n}}{\sigma}} \frac{e^{\frac{-t^2}{2}}}{\sqrt{2\pi}}dt - 1.
\end{aligned}$$

2.3 Comparaison avec l'intégration numérique

Supposons qu'on veut calculer une approximation de l'intégrale

$$I = \int_{[0,1]^d} \phi(x)dx.$$

Si la fonction ϕ est suffisamment régulière, des méthodes déterministes d'intégration numérique permettent également d'approcher cette intégrale.

Commençons par la dimension $d = 1$ et rappelons quelques résultats classiques :

— Dans ce cas, si ϕ est de classe $\mathcal{C}^1$, la méthode des rectangles sur la subdivision régulière $\left\{\frac{1}{n}, ..., \frac{n-1}{n}, 1\right\}$ a une précision $\mathcal{O}(\frac{1}{n})$. Plus précisément, si on convient de noter $M_1 = sup_{x\in[0,1]}|\phi'(x)|$ et R_n l'approximation obtenue, alors

$$|R_n - I| \leq \frac{M_1}{2n}.$$

— Si ϕ est de classe $\mathcal{C}^2$, la méthode des trapèzes sur la même subdivision a une précision en $\mathcal{O}(\frac{1}{n^2})$. Plus précisément, si on note $M_2 = sup_{x\in[0,1]}|\phi''(x)|$ et T_n, l'approximation obtenue, alors

$$|T_n - I| \leq \frac{M_2}{12n^2}.$$

— La méthode de Simpson consiste à approcher f sur chaque segment $\left[\frac{k}{n}, \frac{k+1}{n}\right]$ par un arc de parabole qui coïncide avec f aux deux extrémités et au milieu de ce segment. Si ϕ est de classe $\mathcal{C}^4$, cette méthode a une précision en $\mathcal{O}(\frac{1}{n^4})$. Plus précisément, si on note $M_4 = sup_{x\in[0,1]}|\phi^{(4)}(x)|$ et S_n, l'approximation obtenue, alors

$$|S_n - I| \leq \frac{M_4}{2880n^4}.$$

— De façon générale, si ϕ est de classe $\mathcal{C}^s$, une méthode numérique adaptée à cette régularité permettra d'atteindre une vitesse en $\mathcal{O}(n^{-s/d})$.
Cette vitesse s'effondre avec l'augmentation de la dimension d. Par contre, la méthode de Monte Carlo, avec une vitesse en $\frac{1}{\sqrt{n}}$, est insensible à la dimension et peut donc s'avérer plus avantageuse dès que l'on travaille en dimension grande ou avec une fonction irrégulière.

2.4 Monte Carlo dans un contexte bayésien

Les méthodes de Monte Carlo sont d'usage courant dans le cadre typique où la loi de la variable $\mathbb{X}$ dépend d'un paramètre θ. Dans l'approche fréquentiste, θ est inconnu mais supposé avoir une valeur fixée et les observations $(x_1, ..., x_n)$ permettent de l'estimer , par une méthode donnée par exemple au maximum de vraisemblance.

L'approche bayésienne est différente : Elle consiste à considérer que θ est lui même aléatoire et suit une loi (dite à priori) donnée, les observations $(x_1, ..., x_n)$ permettent d'affiner cette loi via sa mise à jour au vu des observations.

Plus formellement, notons π la densité de la loi à priori de θ et $f(x|\theta)$ la densité conditionnelle de x sachant θ. Par la règle de Bayes, la densité à posteriori de θ sachant x s'écrit alors

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|t)\pi(t)}dt \propto f(x|\theta)\pi(\theta).$$

On pourra s'intéresser à une valeur moyenne par rapport à cette loi à posteriori, laquelle s'écrira donc

$$\mathbb{E}[\phi(\theta)|x] = \int \phi(\theta)\pi(\theta|x)d\theta = \frac{\int \phi(\theta)f(x|\theta)\pi(\theta)d\theta}{\int f(x|t)\pi(t)dt}. \tag{2.3}$$

Le paramètre θ est en général multidimensionnel. Par conséquent, le calcul de l'intégrale 2.3 est, sauf cas particulier, impossible analytiquement et difficile par intégration numérique déterministe.

Supposons cependant que l'on sache simuler suivant la loi à priori $\pi(\theta)$ et, pour tout θ ,évaluer la quantité $\phi(\theta)f(x|\theta)$. On retombe alors exactement dans le cadre d'application des méthodes Monte Carlo d'intégration, puisqu'il suffit de générer des réalisations $i.i.d(\theta_1, ..., \theta_n)$ selon $\pi(\theta)$ pour en déduire que

$$\frac{1}{n}\sum_{i=1}^{n}\phi(\theta_i)f(x|\theta_i) \xrightarrow[n\to+\infty]{p.s} \int \phi(\theta)f(x|\theta)\pi(\theta)d\theta$$

L'estimateur du dénominateur de 2.3 correspond au cas où $\phi = 1$ et se traite de la même façon .Au total, l'estimateur Monte Carlo de $I = \mathbb{E}[\phi(\theta)|x]$ s'écrit :

$$\widehat{I_n} = \frac{\sum_{i=1}^{n}\phi(\theta_i)f(x|\theta_i)}{\sum_{i=1}^{n}f(x[\theta_i)}$$

En particulier, l'estimateur de la moyenne à posteriori de θ est donné par

$$\widehat{\theta}(x) = \int \theta\pi(\theta|x)d\theta.$$

qui admet pour estimateur Monte Carlo

$$\widehat{\theta_n}(x) = \frac{\sum_{i=1}^{n}\theta_i f(x|\theta_i)}{\sum_{i=1}^{n}f(x[\theta_i)}$$

Remarque. Il s'avère bien que les méthodes de Monte Carlo interviennent dans plusieurs domaines d'application . Néanmoins, leur application suppose savoir simuler des variables aléatoires identiques et indépendantes suivant une loi donnée .Pour cela, la section suivante traite une liste non exhaustive des méthodes de simulation de variables aléatoires les plus connues.

2.5 Simulation de variables aléatoires

Introduction

Une méthode de Monte Carlo repose sur la simulation d'une suite de variables aléatoires $(X_n)_{n\geq 1}$ indépendantes et identiquement distribuées $i.i.d$ selon une loi donnée, cette partie expose quelques méthodes pour y parvenir en commençant par la loi uniforme, sur laquelle toutes les autres sont basées.

2.5.1 Simulation suivant la loi uniforme

Dans la suite, on suppose que l'on a un générateur de nombres aléatoires suivant la loi uniforme sur $[0, 1]$.Disposer d'un tel générateur n'est néanmoins pas trivial :Un ordinateur ne dispose d'aucun composant aléatoire. Un générateur de nombres aléatoires est donc un programme déterministe qui produit une suite de valeurs "suffisamment" désordonnées pour ressembler à un échantillon aléatoire, on parle alors de générateur pseudo-aléatoire.

Définition 2.11. *Générateur pseudo-aléatoire*
Un générateur de nombres pseudo-aléatoires est un algorithme qui, à partir d'une valeur initiale u_0, appelée graine, et une transformation D, produit une suite $(u_n)_{n\geq 1} = D^n(u_0)$ dans $[0,1]$. Pour tout n, les valeurs $u_1, ..., u_n$ reproduisent le comportement d'un échantillon i.i.d de loi uniforme.

On suppose disposer d'un générateur pseudo-aléatoire pour la loi uniforme. Dans la suite, on montre comment, partant de la loi uniforme, il est possible de générer d'autres lois.

2.5.2 Méthode d'inversion

Soit F une fonction de répartition bijective, alors si U suit une loi uniforme sur $[0,1]$, la variable aléatoire $X = F^{-1}(U)$ a pour fonction de répartition F. En effet, pour tout réel x,

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

où l'on a appliqué respectivement l'aspect bijectif de F, sa croissance et la forme spécifique de la fonction de répartition de la loi uniforme.

Ce résultat représente un grand intérêt :Si F est facilement inversible, alors pour générer une variable aléatoire de loi F, il suffit de générer une variable uniforme U et de lui appliquer F^{-1}. On peut même généraliser ceci à des fonctions de répartition non bijectives, c'est d'ailleurs l'objectif de ce qui suit.

Définition 2.12. *Inverse généralisé*
Soit X une variable aléatoire de fonction de répartition F. On appelle inverse généralisée de F, ou fonction quantile de X, la fonction F^{-1} définie pour tout $u \in]0,1[$ par

$$F^{-1}(u) = inf\{x \in \mathbb{R} : F(x) \geq u\}.$$

Rappel 2.13. *Le réel $q_{1/2}$: $F^{-1}(1/2)$ est appelé la médiane de F. De façon générale, lorsque $0 < \alpha < 1$, $q_{1-\alpha} := F^{-1}(1-\alpha)$ est appelé quantile d'ordre $1-\alpha$ de F. On le rencontre constamment dans les tests hypothèses, le plus fameux d'entre eux étant celui associé aux intervalles de confiance à 95% de la loi gaussienne centrée réduite : $q_{0.975} := 1.96$ où ϕ est la fonction de répartition de la loi $\mathcal{N}(0,1)$.*

Si F est inversible, l'inverse généralisée de F coïncide avec son inverse classique.

Proposition 2.14. *Méthode d'inversion*
Soit X une variable aléatoire réelle de fonction de répartition F et $U \sim \mathcal{U}([0,1])$. Alors,$F^{-1}(U)$ suit la loi de X.Autrement dit, il suffit de simuler u suivant $\mathcal{U}([0,1])$ puis d'appliquer la transformation $x = F^{-1}(u)$ pour simuler suivant la loi de X.

Démonstration. Il s'agit de montrer que $F^{-1}(U)$ et X ont même fonction de répartition , *i.e*, pour tout y réel

$$\mathbb{P}[F^{-1}(U) \leq y] = P[X \leq y] = F(y) = \mathbb{P}[U \leq F(y)].$$

Il suffit de montrer que pour tout y réel et tout u dans $[0,1]$,

$$\{(u,y) : F^{-1}(u) \leq y\} = \{(u,y) : u \leq F(y)\}.$$

($\subseteq$) Soit (u,y) tel que $F^{-1}(u) \leq y$. Par croissance de F, on a $F\{F^{-1}(u)\} \leq F(y)$. Et par définition de l'inverse généralisée $F\{F^{-1}(u)\} \geq u$. D'où $u \leq F(y)$.

($\supseteq$) Si $u \leq F(x)$,par définition de l'inverse généralisé , $F^{-1}(u)$ est le plus petit réel qui vérifie $u \leq F\{F^{-1}(u)\}$ donc, $F^{-1}(u) \leq x$. □

Exemple 2.15. *Loi discrète*

Soit X une variable aléatoire discrète à valeurs dans l'ensemble fini de support $\Omega = \{x_k \in \mathbb{R}, k \in \mathbb{N}^*\}$.

Notons, pour $k \geq 1$,

$$p_k = \mathbb{P}[X = x_k], s_0 = 0 \text{ et } s_k = \sum_{i=1}^{k} p_k.$$

Pour tout $u \in [0,1]$, l'inverse généralisé est donné par

$$F^{-1}(u) = inf\{x \in \mathbb{R} : \sum_{k=1}^{+\infty} p_k \mathbb{1}_{\{x_k \leq x\}} \geq u\}$$

Alors, pour $U \sim \mathcal{U}([0,1])$, la variable Z définie ci-dessous suit la même loi que X

$$Z = \begin{cases} x_1 & si & u \in [0,1] \\ x_k & si & u \in]s_{k-1}, s_k],\ k \geq 2 \\ 0 & si & sinon. \end{cases} \tag{2.4}$$

En pratique,, il faut trouver, pour une réalisation u suivant la loi $\mathcal{U}([0,1])$, l'unique indice k tel que $s_{k-1} < u \leq s_k$ et ceci, en testant successivement si $u > p_1$, puis si $u > p_1 + p_2$,etc.

Exemple 2.16. *Loi exponentielle*

Soit $X \sim \varepsilon(\lambda)$ une variable aléatoire de loi exponentielle de paramètre $\lambda > 0$.

Rappel *La loi exponentielle a pour support $\mathbb{R}_+$. Sa densité est définie par*

$$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{\{x \geq 0\}}.$$

et sa fonction de répartition par

$$F(x) = (1 - e^{-\lambda x}) \mathbb{1}_{\{x \geq 0\}}.$$

F est bijective et pour tout $u \in [0,1]$,

$$F^{-1}(u) = \frac{-1}{\lambda} ln(1-u).$$

alors, pour tout $U \sim \mathcal{U}([0,1])$,

$$F^{-1}(U) = -\lambda^{-1} ln(1-U) \sim -\lambda^{-1} ln(U) \sim \varepsilon(\lambda).$$

car $1-U$ et U ont même loi sur $[0,1]$.

Exemple 2.17. *Loi de Weibull*

La loi de Weibull est utilisée dans différents domaines(ingénierie, théorie des valeurs extrêmes, hydrologie, assurances,...). On dit qu'une variable aléatoire X suit une loi de Weibull de paramètres λ , $k \in \mathbb{R}_+^*$ lorsque sa fonction de répartition est donnée, pour tout réel, $x \geq 0$, par

$$F(x) = 1 - \exp\{-(x/\lambda)^k\}.$$

Comme pour la loi exponentielle, la fonction de répartition est bijective et

$$F^{-1} = \lambda\{-ln(1-u)\}^{1/k}.$$

Ainsi, pour $U \sim \mathcal{U}([0,1]), \lambda\{-ln(U)\}^{1/k}$ suit une loi de Weibull de paramètres s λ , $k \in \mathbb{R}_+^*$.

Application 2.18. *Simulation d'une loi conditionnelle*
Soit X une variable aléatoire de fonction de répartition F, On peut simuler suivant la loi de X sachant l'événement $\{X > c\}$ i,e, on veut simuler suivant la distribution conditionnelle

$$\mathbb{P}(X \in .|X > c).$$

L'algorithme est le suivant :

1. *Simuler U suivant la loi uniforme.*
2. *Calculer $Z = F^{-1}\left((1 - F(c))U + F(c)\right)$*

En effet,

$$\begin{aligned}\mathbb{P}(Z \leq x) &= \mathbb{P}((1-F(c))U + F(c) \leq F(x)) = \mathbb{P}\left(U \leq \frac{F(x)-F(c)}{1-F(c)}\right) \\ &= \frac{F(x)-F(c)}{1-F(c)} = \frac{c \leq X \leq x}{X > c} = P(X \leq x|X > c).\end{aligned}$$

Remarque.

La méthode d'inversion n'est exacte qu'à condition de connaitre l'expression explicite de F^{-1} comme pour les fonctions de répartition des lois (exponentielle,double exponentielle,Weibull,Cauchy,...).
Il ne s'agit néanmoins que de cas très limités. Un contre exemple classique à cette méthode est la loi normale $\mathcal{N}(0,1)$: Il n'existe pas de formulation simple de sa fonction de répartition ϕ ou de son inverse ϕ^{-1}. Il faut alors utiliser un algorithme d'approximation :Il existe des polynômes donnant une approximation de ϕ ou de son inverse ϕ^{-1}. La méthode d'inversion peut s'appliquer à ces polynômes pour simuler une loi normale.C'est la méthode utilisée par défaut dans $\mathbb{R}$.
Si la méthode d'inversion est indiquée pour de nombreuses lois discrètes, elle est rarement la plus efficace pour les variables aléatoires continues (même lorsque F^{-1} est explicitement connue).D'autres méthodes sont alors à disposition comme on va voir dans ce qui suit.

2.5.3 Méthode d'acceptation-rejet

Dans le cas où il est difficile, voire impossible, de simuler une variable aléatoire par transformation inverse ou d'en obtenir une représentation alternative(facilement utilisable), on utilise les méthodes d'acceptation-rejet.
Pour simuler suivant une densité f, appelée loi cible, ces méthodes utilisent une densité alternative g ,appelé densité

instrumentale, plus simple(d'un point de vue simulation,*e.g*, lois uniformes, exponentielle, normale...)combinée avec une procédure de rejet.

Proposition 2.19. *Méthode d'acceptation-rejet*

Soit $X = (X_1, ..., X_d)$ *une variable aléatoire de densité* f *de* $\mathbb{R}^d$ *et* g *une densité de* $\mathbb{R}^d$ *telle qu'elle existe une constante* $M \geq 1$ *satisfaisant*

$$f(x) \leq Mg(x) \text{ pour tout } x \text{ dans } \mathbb{R}^d.$$

Soit $(U_n)_{n\geq 1}$ *une suite de variables i.i.d de loi* $\mathcal{U}([0,1])$ *et* $(Y_n)_{n\geq 1} = (Y_{1,n}, ..., Y_{d,n})_{n\geq 1}$ *une suite de variables i.i.d de loi de densité* g *telles que ces deux suites sont indépendantes. Alors, pour* T *défini par*

$$T := inf\{n \geq 1 : U_n \leq \alpha(Y_n)\} \text{ avec } \alpha(Y_n) = \frac{f(Y_n)}{Mg(Y_n)},$$

Y_T *suit la loi de densité* f*. Autrement dit, pour simuler* $X \sim f$*, il suffit de simuler*

$$Y \sim g \text{ et } U|Y = y \sim \mathcal{U}([0, Mg(y)]),$$

jusqu'à ce que $u < f(y)$.

Remarque. f et g étant des densités, on a nécessairement $M \geq 1$.En effet,

$$f(x) \leq Mg(x) \Rightarrow \int_{\mathbb{R}^d} f(x)dx \leq M \int_{\mathbb{R}^d} g(x)dx.$$

Démonstration. Montrons que $E[h(Y_T)] = \int_{\mathbb{R}^d} h(y)f(y)dy$ pour toute fonction h borélienne positive sur $\mathbb{R}^d$.

$$\mathbb{E}\left[\sum_{n=1}^{+\infty} h(Y_n)\mathbb{1}_{T=n}\right] = \sum_{n=1}^{+\infty} \mathbb{E}\left[h(Y_n)\mathbb{1}_{T=n}\right].$$

D'après la définition du temps aléatoire T, on a

$$\mathbb{1}_{\{T=n\}} = \mathbb{1}_{\{U_n \leq \alpha(Y_n)\}} \prod_{i=1}^{n-1} \mathbb{1}_{\{U_i > \alpha(Y_i)\}}.$$

Les variables aléatoires (U_i, Y_i), $i \in \mathbb{N}^*$, étant *i.i.d* , on en déduit

$$\mathbb{E}[h(Y_T)] = \sum_{n=1}^{+\infty} \mathbb{E}[h(Y_n)\mathbb{1}_{U_n \leq \alpha(Y_n)}]\{\mathbb{E}[\mathbb{1}_{\{U_1 > \alpha Y_1\}}]\}^{n-1}$$

Pour tout $n \geq 1$, U_n et Y_n étant indépendantes

$$\begin{aligned}
\mathbb{E}[h(Y_n)\mathbb{1}_{U_n \leq \alpha(Y_n)}] &= \int_{\mathbb{R}^d} h(y)g(y) \int_{u\in[0,1]} \mathbb{1}_{u \leq \alpha(y)} du dy && (2.5)\\
&= \int_{\mathbb{R}^d} h(y)g(y)\alpha(y)dy && (2.6)\\
&= \frac{1}{M} \int_{\mathbb{R}^d} h(y)f(y)dy. && (2.7)
\end{aligned}$$

On en déduit en particulier pour la fonction constante $h = 1$,

$$\mathbb{E}[\mathbb{1}_{U_n \leq \alpha(Y_n)}] = \frac{1}{M}\int_{\mathbb{R}^d} f(y)dy = \frac{1}{M}(= 1 - \mathbb{E}[\mathbb{1}_{U_n > \alpha(Y_n)}]). \tag{2.8}$$

En combinant les équations 2.7 et 2.8, on obtient

$$\mathbb{E}[h(Y_T)] = \sum_{n=1}^{+\infty}(1 - \frac{1}{M})^{n-1}\frac{1}{M}\int_{y\in\mathbb{R}^d} h(y)f(y)dy.$$

Sous l'hypothèse $M \geq 1$, la série géométrique $\sum_{n=1}^{+\infty}\left(1 - \frac{1}{M}\right)^{n-1}$ converge et

$$\begin{aligned}\mathbb{E}[h(Y_T)] &= \frac{1}{M}\int_{\mathbb{R}^d} h(y)f(y)dy \times \frac{1}{1-(1-\frac{1}{M})}\\ &= \int_{\mathbb{R}^d} h(y)f(y)dy,\end{aligned}$$

□

Une conséquence de la proposition 2.19 est que la probabilité d'acceptation est exactement $1/M$ (*c.f.*, équation 2.8). De plus, le nombre moyen d'essais moyen jusqu'à ce qu'une variable soit acceptée est M. En effet,

$$\begin{aligned}E(T) &= \sum_{n=1}^{+\infty} n\mathbb{P}[T = n]\\ &= \sum_{n=1}^{+\infty} n\mathbb{P}[U_1 > \alpha(Y_1), ..., U_{n-1} > \alpha(Y_{n-1}), U_n \leq \alpha(Y_n)].\end{aligned}$$

Comme $(U_n, Y_n), n \geq 1$, sont *i.i.d* ,

$$\begin{aligned}E[T] &= \sum_{n=1}^{+\infty} n\{\mathbb{P}[U_1 > \alpha(Y_1)]\}^{n-1}\mathbb{P}[U_n \leq \alpha(Y_n)]\\ &= \sum_{n=1}^{+\infty} n\left(1 - \frac{1}{M}\right)^{n-1}\frac{1}{M}.\end{aligned}$$

Si on introduit la série entière $h(z) = \sum_{n\geq 1} nz^{n-1}$ de rayon de convergence 1 *i.e*, la série converge pour $|z| < 1$. h est la dérivée de la série entière $\sum_{n=1}^{+\infty} z^n = (1-z)^{-1}$. On en déduit $h(z) = (1-z)^{-2}$ et par suite

$$E[T] = \frac{1}{M}h\left(1 - \frac{1}{M}\right) = M.$$

Exemple 2.20. *Pour $x \geq 1$,*
Soit $f(x) = \frac{e^{-x^3}}{Z}\mathbb{1}_{\{x\geq 1\}}$ avec $Z = \int_1^{+\infty} e^{-x^3}dx$.
On veut simuler suivant f en utilisant la méthode d'acceptation-rejet , introduisons alors une fonction g .

Soit $g(x) = \dfrac{\mathbb{1}_{\{x\geq 1\}}}{x^2}$*.Les fonctions* f *et* g *sont des densités de probabilité.*
**Il est facile de simuler suivant la densité* g *en utilisant la technique d'inversion de la fonction de répartition.*
En effet, pour $x \geq 1$*,* $G(x) = \int_1^x g(u)du = \left[-\dfrac{1}{u}\right]_1^x = 1 - \dfrac{1}{x} = 1 - \dfrac{1}{x}$.
et pour $u \in [0;1], G^{-1}(u) = \dfrac{1}{1-u}$.
**Cherchons la constante* k *tel que* $f(x) \leq kg(x)$ *pour tout* $x \geq 1$.
Posons $h = \dfrac{f}{g} : x \geq 1 \to x^2 e^{-x^3}$.
Nous avons $h'(x) = (2x - 3x^4)e^{-x^3} \leq 0$ *pour tout* $x \geq 1$*.Donc, pour tout* $x \geq 1$ $h(x) \leq h(1) = e^{-1}$.
Nous en déduisons que pour tout $x \geq 1$,

$$f(x) \leq \frac{1}{eZ} g(x).$$

Posons $k = 1/(eZ)$ *et soit pour tout* $x \geq 1$,

$$\alpha(x) = \frac{f(x)}{kg(x)} = x^2 e^{-x^3+1}.$$

Nous ne connaissons pas Z *mais nous pouvons calculer* $\alpha(x)$ *pour tout* $x \geq 1$ *donc nous pouvons utiliser l'algorithme de simulation par rejet pour simuler suivant la densité* f.

2.6 Méthodes de réduction de la variance

On se concentre dans cette partie sur le problème d'estimation de

$$I = \mathbb{E}_\nu[h(X)], X \sim \nu,$$

où h est une fonction mesurable telle que :
(H_1) h est intégrable par rapport à la mesure ν (I est bien définie),
(H_2) h est de carré intégrable par rapport à la mesure ν.
Sous l'hypothèse (H_1), pour une suite $(X_n)_{n\geq 1}$ de variables aléatoires $i.i.d$ suivant la loi ν, l'estimateur

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \tag{2.9}$$

converge vers I.

On a vu dans la première partie, que sous l'hypothèse (H_2) théorème central limite), la précision, $\epsilon(\alpha)$, de l'estimateur $\bar{Y}_n$ au niveau de confiance $1-\alpha$ est donné par :

$$\epsilon(\alpha) = 2\phi^{-1}\left(1 - \frac{\alpha}{2}\right)\frac{\widehat{\sigma}_n}{\sqrt{n}} = 2\phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{Var(\bar{Y}_n)}.$$

L'objectif de cette partie est de trouver un estimateur plus efficace (en terme de réduction de variance) que l'estimateur de Monte Carlo classique $\bar{Y}_n$. Il s'agit de trouver une variable aléatoire $\tilde{Y}$, ou de façon équivalente un estimateur $\widehat{\delta}_n$ tel que

$$\mathbb{E}[\tilde{Y}] = E[h(X)] \; et \; Var(\tilde{Y}) < Var[h(X)]$$

ou,
$$\mathbb{E}[\widehat{\delta}_n] = E[h(X)] \; et \; Var(\widehat{\delta}_n) < Var[h(X)]$$
Généralement, $\tilde{Y}$ et $\widehat{\delta}_n$ s'écrivent $\tilde{Y} = g(X)$ et $\widehat{\delta}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_n)$, avec $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ une fonction mesurable. La nouvelle méthode d'estimation implique donc le même co$\widehat{u}$t de simulation.

2.6.1 Échantillonnage préférentiel (*Importance Sampling*)

La méthode d'échantillonnage préférentiel repose sur l'idée que pour calculer $\mathbb{E}_f[h(X)]$ pour X de densité f, il peut être plus intéressant d'échantillonner suivant une autre loi de densité g *i.e* écrire
$$\mathbb{E}_f[h(X)] = \mathbb{E}_g[\tilde{h}(X)],$$
de sorte que l'estimateur Monte Carlo de l'espérance sous g soit plus efficace que l'estimateur Monte Carlo sous f.

Exemple 2.21. *Probabilité d'événement rare*
On souhaite calculer $\mathbb{P}[X > 4.5]$ pour $X \sim \mathcal{N}(0,1)$. Étant donné $(X_n)_{n\geq 1}$ une suite de $n =$ 10000 variables aléatoires i.i.d suivant la loi $\mathcal{N}(0,1)$, on obtient via l'estimateur de Monte Carlo classique
$$P[X > 4.5] \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > 4.5\}} = 0.$$
Dans le cas d'un événement rare, la simulation "naïve" peut s'avérer très inefficace.Une solution est d'avoir recourt à une distribution alternative .Soit $Z \sim \tau\epsilon(4.5, 1)$ une variable aléatoire suivant la loi exponentielle de paramètre 1 et translatée de 4.5 i.e ,la densité de Z est donnée par
$$g(z) = \frac{e^{z-4.5}}{\int_{4.5}^{+\infty} e^{-x}dx}\mathbb{1}_{\{z\geq 4.5\}}.$$
On peut écrire
$$\mathbb{P}[X > 4.5] = \int_{4.5}^{+\infty} \phi(x)dx = \int_{4.5}^{+\infty} \frac{\phi(x)}{g(x)}g(x)dx.$$
Ainsi pour une suite $(Z_n)_{n\in\mathbb{N}}$ de $n = 10000$ variables aléatoires i.i.d suivant la loi de $\mathbb{Z}$, on a l'estimation
$$\mathbb{P}[X > 4.5] \approx \frac{1}{n}\sum_{k=1}^{n} \frac{\phi(X_k)}{g(Z_k)}\mathbb{1}_{\{Z_k > 4.5\}} = 3.36 \times 10^{-6}.$$

Le postulat de base de l'échantillonnage préférentiel est que pour toute densité $g : \mathbb{R}^d \to \mathbb{R}$, $(Z \sim g)$ tel que $supp(f) \subset supp(g)$, on a
$$\mathbb{E}_f[h(X)] = \mathbb{E}_g\left[\frac{h(Z)f(Z)}{g(Z)}\right]. \tag{2.10}$$

Définition 2.22. *Estimateur d'échantillonnage préférentiel*
Soient g une densité telle que $supp(f) \subset supp(g)$ et $(Z_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d suivant la densité g.On définit l'estimateur d'échantillonnage préférentiel par
$$\widehat{\delta}_n(g) = \frac{1}{n}\sum_{k=1}^{n} \frac{f(Z_k)}{g(Z_k)}h(Z_k).$$
La densité g est appelée loi instrumentale (ou loi d'importance) et le rapport $f(Z_k)/(g(Z_k)$ est appelé poids d'importance.

Biais de l'estimateur. Sous l'hypothèse $supp(f) \subset supp(g)$,on obtient en utilisant l'équation 2.10 que l'estimateur est sans biais.
Convergence de l'estimateur. Les variables aléatoires $(h(Z_n)f(Z_n)/g(Z_n))_{n\in\mathbb{N}}$ sont *i.i.d* et d'espérance finie sous g. La loi forte des grands nombre donne

$$\widehat{\delta}_n(g) \xrightarrow[n\to+\infty]{p.s} \mathbb{E}_g\left[\frac{h(Z)f(Z)}{g(Z)}\right] = \mathbb{E}_f[h(X)].$$

Variance de l'estimateur. Les variables aléatoires $(Z_n)_{n\geq 1}$ étant *i.i.d* suivant la densité g, la variance de l'estimateur s'écrit

$$\begin{aligned} Var_g[\widehat{\delta}_n(g)] &= \frac{1}{n}\left\{\mathbb{E}_g\left[\frac{h^2(Z)f^2(Z)}{g^2(Z)}\right] - \mathbb{E}_g\left[\frac{h(Z)f(Z)}{g(Z)}\right]^2\right\} \\ &= \frac{1}{n}\left\{\mathbb{E}_f\left[\frac{h^2(Z)f(Z)}{g(Z)}\right] - \mathbb{E}_f[h(X)]^2\right\} \end{aligned}$$

Proposition 2.23. *L'estimateur d'échantillonnage préférentiel $\widehat{\delta}_n$ est plus efficace en terme de réduction de variance que l'estimateur de Monte Carlo classique si*

$$\mathbb{E}_f\left[h^2(X)\left(\frac{f(X)}{g(X)} - 1\right)\right] < 0.$$

Démonstration. En effet,
Posons

$$\theta = \mathbb{E}_f[h(X)] = \mathbb{E}_g\left[\frac{h(Z)f(Z)}{g(Z)}\right].$$

Calculs de la variance :

$$Var[h(X)] = \int h^2(x)f(x)dx - \theta^2$$

$$Var\left[\frac{h(Z)f(Z)}{g(Z)}\right] = \int \frac{h^2(x)f^2(x)}{g(x)}dx - \theta^2$$

Alors,

$$\begin{aligned} Var\left[\frac{h(X)f(X)}{g(X)}\right] - Var[h(X)] - &= \int h^2(x)\left[\frac{f(x)}{g(x)} - 1\right] f(x)dx \\ &= \mathbb{E}_f\left[h^2(X)\left(\frac{f(X)}{g(X)} - 1\right)\right] < 0. \end{aligned}$$

□

Remarque. Bien que l'estimateur $\widehat{\delta}_n(g)$ converge presque surement vers $\mathbb{E}_f[h(X)]$ pour toute densité g, il n'est de variance finie que lorsque

$$\mathbb{E}_f\left[\frac{h^2(Z)f(Z)}{g(Z)}\right] = \int_{supp(f)} h^2(x)\frac{f^2(x)}{g(x)}dx < \infty.$$

Ainsi les densités instrumentales g telles que f/g n'est pas bornée sont à proscrire. Lorsque $Var[h(X)] < \infty$, une condition suffisante pour garantir que $\widehat{\delta}_n(g)$ soit de variance finie est de choisir une densité g telle que f/g soit bornée.

Parmi les estimateurs d'échantillonnage préférentiel de variance finie, il est possible d'expliciter en fonction de f et h, la loi instrumentale optimale en terme de variance.

Proposition 2.24. *L'estimateur d'échantillonnage préférentiel de variance minimale , $\widehat{\delta}_n(g^*)$, est obtenu pour la densité instrumentale*

$$g^*(z) = \frac{|h(z)|f(z)}{\int_{supp(f)} |h(x)|f(x)dx}, \ z \in \mathbb{R}^d. \tag{2.11}$$

Démonstration. La fonction d'importance optimale en terme de variance est définie

$$g^* = \underset{g}{argmin}\ E_f\left[\frac{h^2(X)f(X)}{g(X)}\right] = \underset{g}{argmin}\ \mathbb{E}_g\left[\frac{h^2(Z)f^2(Z)}{g^2(Z)}\right].$$

□

Or,l'inégalité de Jensen donne

$$E_g\left[\frac{h^2(Z)f^2(Z)}{g^2(Z)}\right] \geq E_g\left[\frac{|h(Z)|f(Z)}{g(Z)}\right]^2 = \mathbb{E}_f[|h(X)|]^2.$$

La borne inférieure est indépendante de g et est atteinte pour g définie par 2.11.

Lorsque $h > 0$, la densité instrumentale optimale est $g^* = hf//\mathbb{E}_f[h(X)]$. Néanmoins, cela requiert de connaitre $\mathbb{E}_f[h(X)]$ qui est justement la quantité d'intérêt . La proposition précédente fournit une stratégie pour choisir g : Un candidat pertinent est tel que $|h|f/g$ soit quasi constant de variance finie.

Remarque. La méthode d'échantillonnage préférentiel est une méthode générale dont l'intérêt majeur , mais aussi la difficulté majeure, réside dans les hypothèses faibles sur le choix de densité instrumentale g.

Il faut également noter que le *coût* de calcul de cet estimateur peut être différent de celui de l'estimateur de Monte Carlo.Ce critère est à prendre compte dans la comparaison des performances des estimateurs.

2.6.2 Variables de contrôle

Définition 2.25. *Méthode de la variable de contrôle*
Soit $h_0 : \mathbb{R}^d \to \mathbb{R}$ telle que son espérance, notée m, soit facilement calculable et $Var[h_0(X)] > 0$.
Pour tout réel b, étant donnée $(X_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d suivant la loi ν, on définit l'estimateur

$$\widehat{\delta}_n = \frac{1}{n}\sum_{k=1}^{n} h(X_k) - b(h_0(X_k) - m). \tag{2.12}$$

Biais de l'estimateur. Comme $m = \mathbb{E}[h_0(X)]$, on a directement $\mathbb{E}[\widehat{\delta}_n] = \mathbb{E}[h(X)]$.

Convergence de l'estimateur. La loi forte des grands nombres pour les suites de variables aléatoires $(h(X_n))_{n\geq 1}$ et $(h_0(X_n))_{n\geq 1}$, donne

$$\left.\begin{array}{l} \displaystyle\frac{1}{n}\sum_{k=1}^{n} h(X_k) \xrightarrow[n\to+\infty]{p.s} \mathbb{E}[h(X)] \\ \displaystyle\frac{1}{n}\sum_{k=1}^{n} h_0(X_k) - m \xrightarrow[n\to+\infty]{p.s} 0 \end{array}\right\} \widehat{\delta}_n \xrightarrow[n\to+\infty]{p.s} \mathbb{E}[h(X)], \text{pour tout réel} b.$$

Intervalle de confiance. Les variables aléatoires $(\tilde{Y}_n(b))_{n\geq 1} = (h(X_n) - b[h_0(X_n) - m])_{n\in\mathbb{N}}$ sont $i.i.d$ et de variance finie ,notée $\sigma^2(b)$.Le théorème central limite donne alors, pour tout réel b,

$$\sqrt{n}(\widehat{\delta}_n(b) - \mathbb{E}[h(X)]) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2(b)).$$

On en déduit l'intervalle de confiance au niveau de confiance $1-\alpha$

$$\begin{aligned} IC_{1-\alpha} &= \left[\widehat{\delta}_n - q_{1-\alpha/2}\frac{\sigma(b)}{\sqrt{n}}, \widehat{\delta}_n + q_{1-\alpha/2}\frac{\sigma(b)}{\sqrt{n}}\right] \\ &= \left[\widehat{\delta}_n - q_{1-\alpha/2}\sqrt{Var[\widehat{\delta}_n(b)]}, \widehat{\delta}_n + q_{1-\alpha/2}\sqrt{Var[\widehat{\delta}_n(b)]}\right], \end{aligned}$$

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

Remarque. On estime la variance de la méthode d'estimation via la variance empirique associée aux réalisations de la variable aléatoire $\tilde{Y} = h(X) - b(h_0(X) - m)$:

$$\sigma^2_{n-1} = \frac{1}{n-1}\sum_{k=1}^{n}[h(X_k) - b(h_0(X_k) - m) - \widehat{\delta}_n(b)]^2.$$

Variance de l'estimateur. En utilisant le fait que les variables $(\tilde{Y}_n(b))_{n\geq 1}$ sont $i.i.d$, on obtient que la variance de l'estimateur 2.12, est pour tout réel b,

$$\begin{aligned} Var[\widehat{\delta}_n(b)] &= \frac{1}{n}\sigma^2(b) = \frac{1}{n}\{Var[h(X)] + b^2 Var[h_0(X)] - 2bCov[h(X), h_0(X)]\} \\ &= Var[\bar{Y}_n] + \frac{1}{n}\{b^2 Var[h_0(X)] - 2bCov[h(X), h_0(X)]\} \end{aligned}$$

On en déduit le résultat suivant :

Proposition 2.26. *L'estimateur $\widehat{\delta}_n(b)$ est de variance plus faible que l'estimateur classique $\bar{Y}_n$ si, et seulement si, on choisit b et h_0 tels que*

$$b^2 Var[h_0(X)] - 2bCov[h(X), h_0(X)] < 0.$$

La variance de l'estimateur $\widehat{\delta}_n(b)$ étant une fonction quadratique et convexe en b, elle admet un unique minimum b^*.On a alors le résultat suivant :

Proposition 2.27. *L'estimateur de variance minimale, $\widehat{\delta}_n(b^*)$, est obtenu pour*

$$b^* = \underset{b\in\mathbb{R}}{argmin} Var[\widehat{\delta}_n(b)] = \frac{Cov[h(X), h_0(X)]}{Var[h_0(X)]}.$$

et

$$Var[\widehat{\delta}_n(b^*)] = Var[\bar{Y}_n](1 - \rho(h(X), h_0(X)))^2,$$

avec,

$$\rho(h(X), h_0(X)) = \frac{Cov[h(X), h_0(X)]}{\sqrt{Var[h(X)]Var[h_0(X)]}}.$$

On en déduit que l'estimateur de variance minimale a une variance inférieure à celle de l'estimateur classique $\bar{Y}_n$.La méthode est d'autant plus efficace que $\rho(h(X), h_0(X))^2$ est proche de 1, $i.e$, les variables $h(X)$ et $h_0(X)$ sont corrélées.

Exemple. On cherche à calculer $I = \int_0^1 e^{x^2}dx$ par une méthode de Monte Carlo. Nous avons $I = \mathbb{E}(e^{U^2})$ avec $U \sim \mathcal{U}([0, 1])$.Donc nous pouvons faire l'approximation

$$I \approx \frac{e^{U_1^2} + ...e^{U_n^2}}{n}$$

avec $U_1, U_2, ...i.i.d$,$\sim \mathcal{U}([0, 1])$.Nous avons donc une première méthode de Monte Carlo. Avec les notations ci dessus, nous avons $h(x) = e^{x^2}$ .Posons $h_0 : x \to 1 + x^2$ est proche de h sur $[0, 1]$.Nous savons calculer

$$E[h_0(X)] = \int_0^1 1 + x^2 dx = 1 + \frac{1}{3} = \frac{4}{3}.$$

Nous pouvons faire l'approximation en utilisant la méthode de variable de contrôle pour réduire la variance

$$I \approx \frac{(e^{U_1^2} - 1 - U_1^2) + ... + (e^{U_n^2} - 1 - U_n^2)}{n} - \mathbb{E}(h_0(U)).$$

2.7 Application : Estimation des débits maximums annuels par la méthode de Monte Carlo

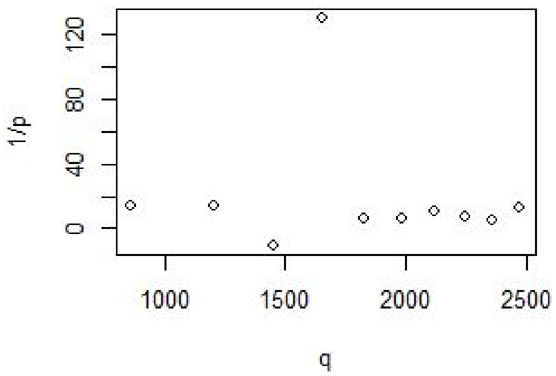
Estimer la période de retour d'un quantile extrême revient à estimer la probabilité d'occurrence de ce quantile.
Considérons la variable aléatoire X représentant le débit maximum annuel et soit q un quantile extrême, on cherche la probabilité de dépassement du quantile q donnée par $p = P[X > q]$. En se basant sur le résultat de l'approche maxima de la théorie des valeurs extrêmes , la variable aléatoire X a comme densité h_γ , avec h_γ la densité de la loi généralisée des extrêmes définie dans le premier chapitre (voir la formule 1.2).

Alors,

$$p = \int_0^q h_\gamma(t)dt \tag{2.13}$$
$$= \int_0^q qh_\gamma(t)\frac{1}{q}dt \tag{2.14}$$
$$= E(qh_\gamma) \tag{2.15}$$
$$\approx \frac{1}{n}\sum_{i=1}^{n} qh_\gamma(u_i). \tag{2.16}$$

Avec n le nombre de simulations et $U \sim \mathcal{U}[0, q]$.

Les paramètres de la densité h_γ sont ajustés par la méthode des moments pondérés sous le logiciel R en utilisant les données du tableau 1.1. En utilisant l'approximation fournie par la formule 2.16, on représente dans la figure ci-dessous la période de retour en fonction des quantiles extrêmes estimés par la loi généralisée des extrêmes.



;

FIGURE 2.1 – Estimation de la période de retour des quantiles extrêmes.

Commentaire :
Pour une période de retour de 130 ans, le débit maximum estimé en utilisant la méthode de Monte Carlo est à peu près $1700m^3/s$ et en utilisant l'approche maxima de la théorie des valeurs extrême, le débit maximum est estimé à peu près à $1500m^3/s$.

Chapitre 3

Prédétermination des crues extrêmes par simulation : Méthode SHYPRE

Introduction La problématique des crues réside dans le besoin d'extrapoler en fréquence un phénomène aléatoire très variable et relativement peu observé par rapport à sa variabilité. Cette problématique est abordée dans ce chapitre par la modélisation et la simulation des phénomènes réalisées à partir de deux outils de modélisation : un générateur stochastique de chroniques de pluies horaires et une modélisation hydrologique conceptuelle permettant de transformer la pluie en événements de crues. Du couplage de ces deux modèles résulte une méthode originale de prédétermination des crues extrêmes, la méthode SHYPRE appartenant à la famille des méthodes dites par simulation.

3.1 État de l'art de la méthode SHYPRE

La méthode SHYPRE(Simulation d'Hydrogrammes pour la PREdtermination des crues) a été conçue pour étudier les distributions de variables hydrologiques (pluies et débit).Elle combine un modèle stochastique pour la pluie horaire avec un modèle de transformation de la pluie en débit. L'extrapolation de la pluie vers les grandes durées de retour est obtenue en générant beaucoup d'événements différents sur une grande période de simulation plutôt qu'en ajustant directement une distribution de probabilité théorique sur des valeurs observées.

3.1.1 Simulation de pluies

La pluie résulte de processus physiques naturels dont les principes fondamentaux sont connus. Les phénomènes énergétiques et mécaniques concernant la physique atmosphérique sont alors à la base des modèles de prévision météorologique. Mais ces modèles sont associés à des échelles souvent trop grandes pour être utilisés en hydrologie. Le phénomène de précipitation est alors plutôt considéré comme un phénomène aléatoire qui peut être étudié de façon statistique par le biais d'approches stochastiques,à la base de générateur de pluies.
Les modèles de génération de pluies sont plus fréquemment associés à un pas de temps journalier.La modélisation des pluies au pas de temps journalier est plus simple car le phénomène est moins complexe qu'aux pas de temps horaire plus fin or l'étude des crues nécessite souvent d'avoir une information à pas de temps infra-journalier.
Les modélisations de chroniques de pluies à pas de temps fin ont fait l'objet de nombreuses recherches à travers différentes approches.Le modèle retenu est basé sur une simulation "directe" des hyétogrammes, définie à partir de l'analyse descriptive du signal de pluie observé. Ce type de modèle part aussi du principe que la pluie peut être assimilée à un processus aléatoire et

intermittent (succession d'états secs et pluvieux).
L'élaboration de ce type de modèle se résume donc à trouver les bonnes variables aléatoires indépendantes décrivant le processus de pluie, ainsi que les lois de probabilité qui les représentent le mieux. La reconstruction du signal s'effectue grâce à une hiérarchisation logique du tirage des différentes variables, afin de parvenir à une représentation la plus fidèle possible du signal de départ.
La validation du modèle consiste à étudier sa capacité à reproduire des variables représentatives du signal simulé, non utilisés lors du calage du modèle.
Sur cette base, différents modèles peuvent êre élaborés, les différences proviennent en particulier du choix des variables descriptives et de leurs lois de probabilité.

3.1.2 Le passage aux débits

Les pluies simulées peuvent être transformées en événements de crues par le biais de modèles hydrologiques. Il existe une multitude de modélisations plus ou moins complexes, s'appuyant sur des approches différentes.Ils peuvent être utilisés en prévision des crues, pour déterminer l'évolution des débits dans un futur proche, à partir d'hypothèses sur les pluies à venir. Une classification de ces modèles hydrologiques est présentée en fonction de la simplification qui est faite dans la modélisation des processus.Trois types de modèles sont différenciés :

⋆ **Les modèles empiriques** en simplifiant au maximum le processus, ils ne s'intéressent qu'à la réponse du système. Cette réponse est modélisée grâce à une équation ou à un opérateur dont les paramètres sont déterminés par les résultats expérimentaux. Ces modèles sont assimilés à une "boite noire" donnant une description purement mathématique du fonctionnement du système, sans faire appel à une notion liée à la physique du processus.Ces modèles sont très dépendants des données et donc difficilement extrapolables au delà de leur condition d'élaboration.

⋆ **Les modèles conceptuels** Ils sont basés sur une conception simple du processus qui ne fait pas vraiment appel aux processus physiques mis en jeu. Ces processus physiques et leur interaction sont modélisés par des opérateurs dont les paramètres n'auront pas de signification physique réelle. Ces modèles sont surement ceux qui présentent le plus de développements, surement parce qu'ils présentent un bon compromis entre leur facilité de mise en œuvre et leurs bonnes performances. Couplés à des générateurs de pluies, il y a par exemple le modèle hydrologique GR3H utilisé dans la méthode SHYPRE (voir [?]).

⋆**Les modèles à base physique** Ils cherchent à représenter et à expliquer le fonctionnement du système par la description de ses mécanismes internes, avec l'utilisation des lois physiques théoriques. Les paramètres de ces modèles sont censés être mesurables .Les difficultés de la mise en place de ces modèles restent liées à a complexité des processus hydrologiques et à la disponibilité des données relatives à la mesure de ses différentes composantes.

3.2 Modèle de génération stochastique de pluies horaires de la méthode SHYPRE

3.2.1 Principe

Le générateur stochastique de pluies horaires vise la simulation directe des hyétogrammes. C'est un modèle événementiel, c'est à dire qu'il crée de façon indépendante et discontinue des événements pluvieux au pas de temps horaire.

Cette mise en œuvre du générateur de pluies horaires implique trois étapes :

1-**Analyse descriptive des hyétogrammes observés** qui déterminera la structure et le calage du générateur. Elle conduit au choix des variables qui définissent au mieux la structure temporelle du phénomène et au choix des variables et au choix des lois de probabilités qui vont décrire ces variables. Les paramètres des lois de probabilité de chaque variable sont ensuite estimés à partir d'un ajustement statistique des distributions de fréquence des valeurs observées.

2-**Simulation d'événements pluvieux** Différentes valeurs des variables définies à l'étape précédente sont générées par simulation stochastique. La génération des variables est réalisée suivant un ordre dicté par la construction des événements pluvieux.L'hypothèse d'indépendance des différentes variables descriptives permet le tirage des valeurs des variables indépendamment les unes des autres.

3-**Validation des résultats obtenus** La validation du modèle se fait alors sur les variables tests (hauteur et durée totale des épisodes).En comparant les premiers moments statistiques de ces variables tests calculés sur les épisodes observés et simulés. Ces variables tests n'ayant pas servi lors de la génération des hyétogrammes, leur bonne restitution lors de la simulation permettra de juger la capacité du modèle à reproduire la structure des hyétogrammes.

3.2.2 Variables du modèle

3.2.2.1 Définition des variables du modèle

Avant de définir les différentes variables du modèle de génération de pluies horaire, il convient de définir quelques termes.
∗ **Une averse** est définie comme une succession de pluies horaires ne présentant qu'un maximum local, avec une décroissance des pluies autour de ce maximum local.
∗ **Une période pluvieuse** est une période de pluies horaires comprenant une ou plusieurs averses.
∗ **Un événement pluvieux** est une succession de pluies journalières supérieures à $4mm$ et comprenant au moins un cumul journalier supérieur à $20mm$. Chaque événement pluvieux sélectionné suivant ce critère est alors analysé au pas de temps horaire.
∗ **Un épisode pluvieux** est l'équivalent d'un événement pluvieux, mais représenté au pas de temps horaire.
Dans un épisode pluvieux étudié, on distingue les averses principales des averses ordinaires.
L'averse principale est unique est repérée comme étant celle qui apporte la plus grande quantité d'eau au cours de l'événement. Toutes les autres averses sont alors considérées comme des **averses ordinaires.**

Les variables du modèle de génération de hyétogrammes horaires sont :

Variables	Notation
Nombre d'événements pluvieux par année	NE
Origine de l'événement dans la journée	TSE
Nombre de périodes pluvieuses par événement	NG
Nombre d'averses par période pluvieuse	NA
Durée de l'averse(en heures)	DA
Volume de l'averse(en mm)	VOL
Rapport entre la pluie maximale de l'averse et son volume	RX
Position relative de la pluie horaire maximale dans l'averse	RPX

Prenons l'exemple d'un épisode pluvieux dans la figure 3.1 pour illustrer la détermination des variables descriptives du modèle.

Cet épisode est composé de trois périodes pluvieuses entrecoupées de deux périodes sèches. La seconde période pluvieuse contient trois averses. Au total, il y a 5 averses et la première est l'averse principale. Les 4 autres sont des averses ordinaires. le tableau suivant décrit les valeurs des différentes variables décrivant l'épisode pluvieux de la figure 3.1 qui contient 3 périodes pluvieuses.

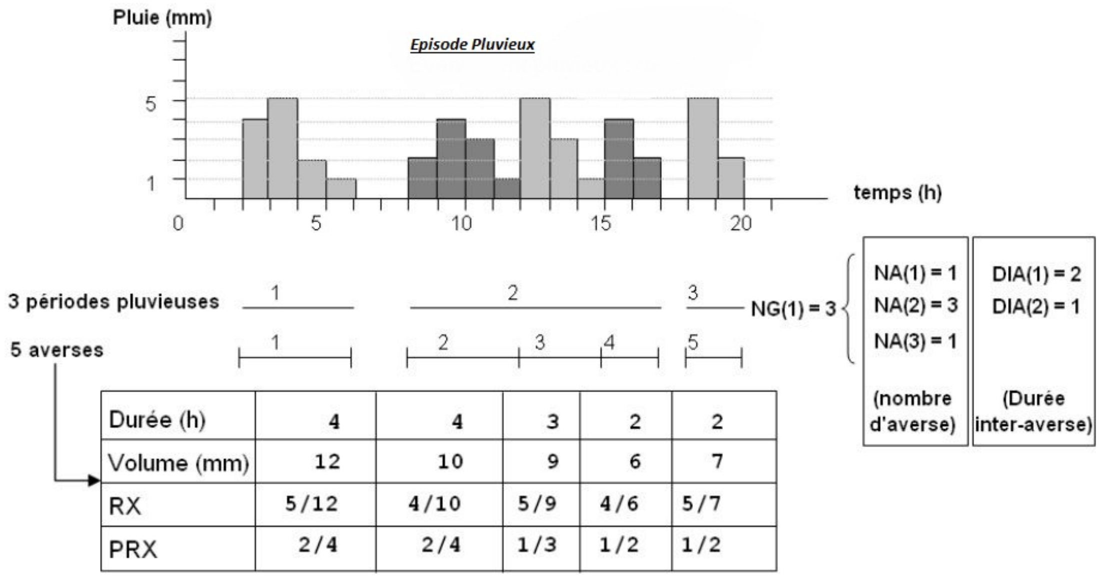


FIGURE 3.1 – Exemple d'épisode pluvieux.

Variables	Averse 1	Averse 2	Averse 3	Averse 4	Averse 5
NA	1	3	•	•	1
DIA	2	0	0	1	•
VOL	12	10	9	6	7
DAP	4	•	•	•	•
DAO	•	4	3	2	2
RX	1.67	1.6	1.67	1.33	1.43
RPX	0.5	0.5	0.33	0.5	0.5

Chaque événement pluvieux, repéré sur les critères des pluies journalières, est analysé de la même façon au pas de temps horaire pour fournir une série de valeurs prises par les différentes variables. Pour chaque variable, ces valeurs forment l'échantillon représentant la population de la variable.Elles vont être utilisées pour l'ajustement de la loi de probabilité de cette variable.

3.2.2.2 Lois de probabilité des variables du modèle

Les valeurs des différentes variables descriptives définies précédemment, sont extraites des pluies horaires issues des séries pluviographiques. La distribution de probabilité de chaque variable va être ajustée par une loi de distribution théorique dont on calculera les paramètres.

Variable	loi de probabilté
NE	loi de poisson
NG,NA,TSE	Géométrique
DIA	Géométrique tronquée
DAP	Poisson tronquée
DAO	Poisson tronquée
RX	Uniforme
RPX	loi normale tronquée
VOL	Exponentielle

Ces lois théoriques, calées sur les distributions empiriques des valeurs observées, vont servir à générer de façon aléatoire chacune des variables nécessaires à la simulation des épisodes pluvieux. Les paramètres des lois statistiques décrivant la distribution des variables du modèle sont alors calculés en utilisant la méthode des moments .

L'étape descriptive étant terminée. Il reste à réaliser l'étape de modélisation. Cette étape consiste à hiérarchiser le tirage des différentes variables utilisées et de les combiner suivant des règles précises pour générer les hyétogrammes au pas de temps horaire.

3.2.2.3 Structure du modèle

Les variables du modèle sont générées suivant un ordre précis, par un tirage aléatoire dans la loi de répartition qui les caractérise :

Pour chaque année à simuler

1. Tirage du nombre d'épisodes pluvieux à générer dans l'année.
   Pour chaque épisode pluvieux
2. Tirage du nombre d'averses : calcul du nombre total d'averses de l'épisode.
3. Tirage de la position de l'averse principale dans une loi uniforme bornée au nombre total d'averses.
   Pour chaque averse à construire on génère
   — Tirage de la durée (DA)
   — Tirage du volume (VOL)
   — Tirage du ratio pluie maximale horaire /VOL (RX)
   — Tirage de la position relative du maximum (RPX)
4. Positionnement des averses avec leur volume.
5. Calcul et positionnement de la pluie maximale horaire des averses.

6. Répartition aléatoire des pluies horaires autour du maximum de l'averse.
7. La durée qui sépare l'averse de la suivante.
8. et enfin, tirage de l'origine de l'épisode dans la journée.

3.3 Modèle de transformation pluie-débit(Modèle GR3H)

le modèle de transformation de la pluie en débit requis par la méthode SHYPRE est le modèle GR3H (GR pour Génie Rural, 3 pour son nombre de paramètres, H pour horaire) Il s'agit d' un modèle hydrologique de type conceptuel. Le signal d'entrée est la pluie horaire sur le bassin et le signal de sortie le débit à l'exutoire.

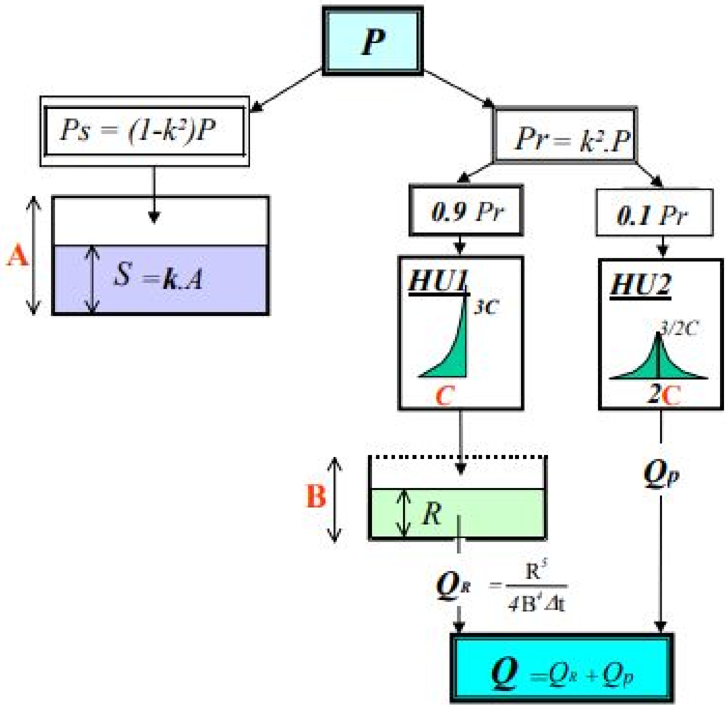


FIGURE 3.2 – Structure du modèle GR3H.

3.3.1 Architecture du modèle GR3H

L'architecture du modèle GR3H repose sur deux réservoirs, caractérisés par un paramètre chacun et sur deux hydrogrammes unitaires, caractérisés par un même paramètre .

Paramètres du modèle

Paramètre	Signification
S	Niveau de remplissage du réservoir-sol
A	Niveau maximal du réservoir-sol
C	Nombre de pas de l'hydrogramme unitaire
R	Niveau de remplissage du réservoir-eau-gravitaire
B	Niveau maximal du réservoir-eau-gravitaire

Le fonctionnement du bassin versant est décrit par :

1) **La fonction de production** L'étude de la fonction de production se limite à l'étude du réservoir-sol. Cette fonction assure la transformation de la pluie brute P en pluie nette Pr . La différence Ps est stockée définitivement dans le réservoir-sol et ne participe pas à la génération des débits.
La répartition de la pluie P est fonction du taux de remplissage k du réservoir-sol k défini comme le rapport du niveau de remplissage S.

Si le système est soumis à une pluie élémentaire dP, la pluie nette est égale à $dPr = \left(\frac{S}{A}\right)^2 dP$ et son complément est dirigé vers le réservoir-sol. La variation du niveau du réservoir sol est donc $dS = \left[1 - \left(\frac{S}{A}\right)^2\right] dP$.
Cette relation en valeur instantanée peut être intégrée sur le pas considéré. Ainsi, l'accroissement du niveau du réservoir est donné par la formule :

$$\Delta S = S_2 - S_1 = A\,tanh\left(\frac{P}{A}\right)\left[\frac{1-\left(\frac{S_1}{A}\right)^2}{1+(S_1/A)\,tanh(P/A)}\right] = Ps.$$

avec $tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}$. où S_1 est le niveau du réservoir au début du pas de temps pendant lequel il reçoit la pluie P et S_2 le niveau du réservoir à la fin de ce pas de temps.
2) **La fonction de transfert** Elle assure la transformation de la pluie nette Pr en débit Q avec un retard (apporté par C) et un étalement dans le temps (assuré par le réservoir de transfert R, qui stocke partiellement pour assurer la décrue) .Il est composé de :
• **Le premier hydrogramme unitaire :** Il décrit la propagation d'une importante fraction (90%) des pluies sortantes de la fonction de production jusqu'au deuxième réservoir.Elle correspond à l'idée d'un écoulement souterrain.

• **Le deuxième hydrogramme unitaire :** Il décrit la propagation d'une petite fraction (10%) des pluies issues de la fonction de production, et participant à l'écoulement direct sans passer par le second réservoir.Elle correspond à la prise en compte du ruissellement direct.

• **Le réservoir eau-gravitaire** :Il reçoit les débits issus du premier hydrogramme unitaire. Sa vidange, définie par une fonction puissance, fournit la seconde partie du débit à l'exutoire. Le niveau R du réservoir détermine le débit Q_r qu'il relâche selon la relation :

$$Q_r = \frac{R^5}{4 * B^4 * \Delta t}.$$

Pour l'utilisation du modèle GR3H en prévision, les paramètres A, B et C sont fixés, considérés comme des valeurs caractéristiques du bassin versant et obtenus par calage sur un échantillon d'événements pluie débit. Le modèle GR3H fonctionne en mode événementiel, *i.e* discontinu dans le temps, ceci nécessite l'initialisation en début de crue du réservoir de production (le paramètre d'initialisation du modèle $S0/A$ est optimisé à chaque événement de crue).

3.3.2 Calage du modèle GR3H

Le modèle GR3H possède 3 paramètres A,B et C à identifier. De façon générale, le calage des paramètres d'un modèle consiste à déterminer le jeu de paramètres qui permet d'obtenir

une réponse du modèle la plus proche de la réponse observée. Pour cela, on cherche à minimiser une fonction critère qui permet de juger la bonne adéquation entre les réponses simulées(ou calculées) et les réponses observées.
L'optimisation du modèle $GR3H$ est faite par un algorithme qui s'apparente à la technique des gradients. L'optimisation va porter sur la valeur du critère de performance, ou fonction critère. Le critère utilisé est appelé *Critère de Nash* tient compte de l'écart entre les observations et les simulations et est donné par :

$$NASH = 1 - \frac{\sum_{i=1}^{N}(Q_{obs}(i) - Q_{cale}(i))^2}{\sum_{i=1}^{N}(Q_{obs}(i) - \bar{Q}_{obs})^2}.$$

tend vers 1 pour une adéquation parfaite.
Pour chaque valeur prise par les paramètres lors des différentes itérations, le critère de NASH est calculé à partir des débits simulés et observés.

3.3.3 Application : Estimation stochastique des débits maximums annuels de Oued Sebou

Pour appliquer la méthode SHYPRE, on a utilisé les données de pluies horaires enregistrées à Fès et qui sont disponibles dans [19].

On a sélectionné tous les événements pluvieux qui vérifient le critère extrême (toute succession de pluies journalières supérieures ou égales à $4mm$ qui contient au moins un cumul journalier de $20mm$).

50 événements pluvieux ont été sélectionnés. Dans les tableaux 3.1, 3.2 , 3.3, 3.3, on présente les détails de chaque événement (nombre de périodes pluvieuses de chaque événement, nombre d'averses de chaque période pluvieuse, le volume des averses de chaque période pluvieuse, la durée entre les averses, la pluie maximale de chaque averse ainsi que sa position au sein de l'averse).

En utilisant les données des tableaux 3.1, 3.2 , 3.3, 3.3, les paramètres de chaque variable aléatoire sont ainsi ajustés en utilisant la méthode des moments pondérés. Ensuite, et en respectant l'ordre de simulation décrit dans le paragraphe 3.2.2, on simule différentes valeurs de chaque variable sous le logiciel R pour construire des hyétogrammes .

Et pour transformer la pluie simulée en débit, on utilise le modèle GR3H de transformation de pluie en débit. Les paramètres A, B, C de ce modèle sont déterminées en optimisant la fonction de critère NASH décrite dans le paragraphe 3.3.2. Ainsi, les débits sont calculés en utilisant les formules figurantes dans 3.2.

NumE	NG	NA	VOL	DIA
1	4	4	0.7,14,6,4	2,4,4
2	6	7	2,2,7,6,4,8	5,4,3,4,1,0
3	2	2	3,5	3
4	6	9	15,19,5,30,20,20	0,1,0,3,1,1,0,1
5	7	7	8,14,31,0.7,0.1,1,7	2,2,11,11,11,4
6	5	5	4,4,12,3,2	4,2,4,18
7	5	5	4,12,0.5,12,6	1,5,3,3
8	4	4	4,12,0.5,6	1,17,3
9	2	2	0.5,22	1
10	3	3	1,3,13	4,14
11	2	2	3,5	3
12	4	4	0.4,0.8,10,3	11,3,11
13	1	1	18	0
14	8	8	17,7,17,6,0.1,3,2,8	5,2,10,11,5,4
15	1	3	36	0,0
16	3	3	32,9,19	3,1
17	2	2	2,21	1
18	7	8	0.9,2,1,6,28,14,14	5,5,1,0,2,1
19	1	3	31	0,0
20	2	3	10,20	1,0
21	5	6	0.3,4,5,12,0.6	10,4,7,11
22	2	2	6,10	3
23	3	4	0.2,3,13	1,4
24	1	3	27	4,5,6
25	2	3	18,38	1,0
26	1	2	23	0
27	1	3	45	0,0
28	4	4	6,9,11,0.5	3,8,17
29	1	3	35	0,0
30	5	5	15,2,5,7,7	12,9,2,1
31	5	5	26,23,0.8,6,7	1,4,1,1
32	6	6	0.4,0.4,5,2,6,0.9	5,3,11,3,23
33	2	3	0.8,21	1,0
34	4	6	39,11,0.4,1	0,0,3,11,5
35	1	3	32	0,0
36	7	10	7,7,17,3,0.5	4,1,0,5,5,18,5
37	4	4	0.4,7,0.2,11	4,5,3
38	7	7	0.4,4,11,16,0.4,0.9,4	4,14,2,11,5,4
39	4	4	0.4,0.4,13,1	5,2,17
40	1	3	34	0,0
41	1	2	21	0
42	1	2	31	4,4,0
43	3	3	7,6,11	4,16
44	3	4	4,19,30	2,1

TABLE 3.1 – Données d'événements pluvieux pour construire les hyétogrammes.

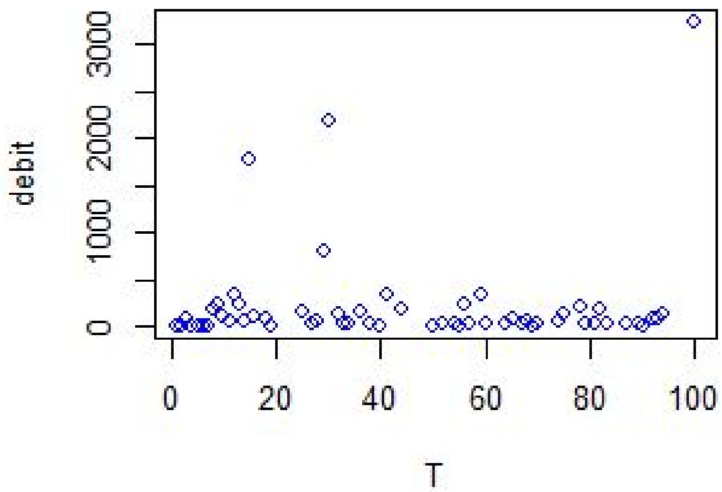


FIGURE 3.3 – Estimation par la méthode SHYPRE des débits maximums en fonction des périodes de retour fixées.

Commentaires :
• Pour une période de retour de moins de 50 ans, Le débit maximum est estimé à $2500m^3/s$ et pour une période de 100 ans, il s'élève à $3500m^3/s$.
• En comparant les résultats fournis par la méthode SHYPRE et la théorie des valeurs extrême et , on constate que cette dernière sous estime le risque des crues à une valeur de $1500m^3/s$ contre $3500m^3/s$ fourni par la méthode SHYPRE et ceci est du aux problèmes d'échantillonnage.

NumE	NG	NA	VOL	DIA
45	2	2	14,17	4
46	3	3	14,14,8	3,10
47	4	4	6,15,0.1,12	3,5,15
48	1	2	25	0
49	2	2	8,13	4
50	9	9	12,5,7,8,2,5,4,13,18	9,23,4,5,5,10,3,3

TABLE 3.2 – Données d'événements pluvieux pour construire les hyétogrammes(suite).

DA	RPX	Pmax
1,4,2,2	1,0.75,1,1	0.7,8,4,3
1,1,2,3,2,3,2	1,1,0.2,2/3,1,1/3	2,2,5,3,3,2,3
2,3	0.5,1/3	2,3
3,2,3,2,3,5,3,2,5	2/3,0.5,2/3,0.5,1,2/5,2/3,0.5,1/5	3,5,7,5,3,9,5,6,9
3,4,4,1,1,1,2	1,0.5,1,1,1,1,0.5	4,5,10,0.7,0.1,1,5
3,2,3,2,1,1,1,1,3,4	1/3,1/2,2/3,1/2,1,1,1,1,2/3,3/4	4,5,5,6,3,0.5,3,0.4,4,4
4,2,4,2,1,1,1,1	1/2,1,1/4,1/2,1,1,1,2/3	6,4,7,4,0.1,3,2,4
1,2,1,3	1,1/2,1,1	0.4,5,0.2,5
2,2,4,2,1	1/2,1,3/4,1/2,1	3,3,5,2,2
2,5,1,3	2/3,2/5,1,1/3,2/3	2,4,0.5,5,3
2,5,1,3	1,1/2,1,2/3	3,4,0.5,3
1,6	1,2/5	0.5,7
1,2,4	1,1/2,1/2	1,2,5
2,1	1/2,2/3	2,3
1,1,3,1	1,1,1,1	0.4,0.8,6,3
5	4/5	6
4,4,3	3/4,1/4,1/3	5,6,4
6,3,5	1/2,1,1/3	10,5,6
1,5	1,1/2	2,7
1,1,1,2,4,1,4,4	1,1,1,1/2,3/4,1,3/4	0.9,2,1,,8,11,6,5
4,3,2	3/4,1/3,1/2	4,5,6
3,3,2	1,2/3,1/2	5,6,5
1,2,2,3,2,1	1,1/2,1,2/3,1/2,1	0.3,3,3,3,5,0.6
3,3	2/3,1/3	3,5
1,2,3,3	1,1,2/3,1/3	0.2,2,4,3
3,2,4	2/3,1/2,1/4	4,5,6
4,2,3	1,1/2,1/3	8,12,11
4,2	3/4,1/2	6,7
5,4,2	3/5,1/4,1/2	5,8,6
2,3,4,1	1,1/3,1/2,1	5,4,4,0.5
3,3,3	2/3,1/3,1/3	6,7,6
4,1,3,1,2	1/2,1,2/3,1,1/2	6,2,3,7,5
8,5,1,2,2	1/3,2/5,1,1/2,1/2	7,9,0.8,4,5
1,1,3,1,3,1	1,1,1,2/3,1,1/3,1	0.4,0.4,3,2,4,0.9
1,3,3	1,1/3,1/3	0.8,5,5
3,3,2,3,1,1	2/3,2/3,1/2	8,7,6,5,0.4,1
4,2,3	3/4,1/2,2/3	5,3,6
1,2,4,4,1,1,2	1,1/2,3/4,3/4,1,1,1	0.4,3,4,6,0.4,0.9,3
1,1,4,1	1,1,3/4,1	0.4,0.4,5,1
2,4,2	1/2,3/4,1/2	6,7,6
3,1	3/4,1	7,5
3,3	2/3,1/3	8,10
2,2,2	1,1/2,1	5,4,7

TABLE 3.3 – Données d'événements pluvieux pour construire les hyétogrammes.

DA	RPX	Pmax
2,4,3,2	1,3/4,2/3,1/2	3,6,8,6
3,3	2/3,1/3	6,7
3,3,2	1/3,2/3,1	6,6,5
2,3,1,3	1/2,2/3,1,2/3	4,6,0.1,5
2,3	1/2,1/3	8,6
2,2	1,1/2	5,8
3,2,2,2,1,1,2,3,3	2/3,1/2,1,1/2,1,1,1/2,1/3,1/3	6,4,4,5,2,5,3,6,9
1,4,2,2	1,0.75,1,1	0.7,8,4,3

Table 3.4 – Données d'événements pluvieux pour construire les hyétogrammes. (suite)

Conclusion

Le risque hydrologique des crues peut être traité statistiquement en utilisant la théorie des valeurs extrêmes de deux manières différentes :L'approche block maxima qui permet d'estimer les quantiles extrêmes des débits d'une période de retour donnée en utilisant les valeurs maximales des blocs choisis pour l'ajustement des paramètres de la loi généralisée des extrêmes .

Néanmoins,cette approche utilise peu d'informations puisque seulement les maximums par bloc sont considérés, pour cela, l'approche par dépassement de seuil fournit un remède et consiste à considérer comme extrêmes toutes les valeurs dépassant un seuil fixé .

Cependant, l'étude statistique des crues nécessite d'extrapoler la distribution des débits au delà du domaine des observations. Il est donc essentiel d'exploiter des informations complémentaires issues du réseau pluviométrique comme c'est le cas dans la méthode SHYPRE qui utilise un modèle de génération stochastique de pluies horaires, couplé à un modèle conceptuel global de transformation de la pluie en débit.

Ainsi, la simulation d'événements de pluie et leur transformation en débit permet de générer différents scénarios de crues de diverses formes pour tester le comportement de l'ouvrage étudié. L'originalité de cette approche est que l'extrapolation des débits ayant une probabilité de dépassement faible, se fait de façon empirique, et non plus sur l'ajustement direct des distributions observées.

Pour affiner l'étude de la méthode SHYPRE, il s'avère important d'utiliser de nouveaux outils mathématiques, tels que les copules, pour affiner la prise en compte des dépendances entre les variables descriptives de la pluie impactant le comportement du modèle vers les valeurs extrêmes.

De plus, les résultats de la méthode SHYPRE en utilisant le pas de temps horaire pour la simulation des pluies sont fortement tributaires de l'initialisation du modèle. C'est à partir de ce constat qu'est venue l'idée de coupler avec le modèle de pluies horaire un modèle de pluie journalier qui va permettre de situer les événements pluvieux dans une chronique simulée en continu qui va servir à l'initialisation du modèle horaire et donc la diminution du nombre de paramètres nécessaires à la modélisation au pas de temps horaire.

Bibliographie

[1] Agnès Lagnoux Renaudie, Analyse des modèles de branchement avec duplication des trajectoires pour l'étude des événements rares. Th.Doc.Université Paul Sabatier Toulouse III, 2007.

[2] HOSKING,J(1985).Algorithm as 215 :Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution.*Journal of the Royal Stastical Society.Series C(Applied Statistics*,34(3) :301-310.

[3] Arnaud, P., and J. Lavabre (2002), Coupled rainfall model and discharge model for flood frequency estimation.Water Resources Research, 38(6).

[4] Hosking, J. et Wallis, J. (1987). Parameter and quantile estimation for the generalized Pareto distribution. Technometrics, 29(3) :339–349. 37, 38, 80

[5] R.L Smith. Maximum likelihood estimation in a class of nonregular cases.*Biometrika*, 72(1) : 67,1985.

[6] M.G. Kendall and S. Alan. The advanced theory of statistics. Vols. I and II 1961.

[7] E.Castillo and A.S.HADI. Extreme value & related models with applications in engineering & science. *Recherche*, 67 :02, 2004.

[8] Fisher, R.A. et Tippett, L.H.C. (1928). Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. Proceedings of the Cambridge Philosophical Society 24, 180-190.

[9] Pickands J.Pickands .Statistical inference using estreme order statistics. *The Annals of Statistics*, 3(1) : 119-131,1975.

[10] Arnaud P.,Lavabre J.(2000) Using a stochastic model for generating hourly hyetographs to study extreme rainfalls. Hydrological Scoences Journal,44(3) 443-446.

[11] Arnaud P,1997, Modèle de prédétermination de crues basé sur la simulation,Extension de sa zone de validité, paramétrisation horaire par l'information journalière et couplage des deux pas de temps.Th.Doc.Univ.Montpellier II ,286 p.

[12] Federico Garavaglia, Méthode SCHADEX de prédetermination des crues extrêmes : Méthodologie, application, études de sensibilité.Th.Doc.Université de Grenoble, 2011.

[13] Jean Philippe Gatarayiha, Méthode de Simulation avec les variables antithétiques Mémoire en Statistiques. Univ de Montréal, 2007.

[14] Sylvain Rubenthaler, Méthodes de Monte Carlo M1 IM, Univ Nice Sophia Antipolis.

[15] BATEKA Samah, Détermination du nombre de statistiques d'ordre extrême, Mémoire de Magister en Mathématiques, Univ. Mohamed Khider BISKRA, 2010.

[16] Pierre Brigode, Zoran Micovic, Emmanuel Paquet, Pietro Bernardara, Méthodes probabilistes et déterministes d'estimation des débits extrêmes.

[17] Benlagha Nourredine, Michel Grun-Réhomme, Olga Vasechko (2009), Les sinistres graves en aussurance automobile : Une nouvelle approche par la théorie des valeurs extrêmes.

[18] Embretchs,P.,Kluppelberg, C., Mikosch ,T.*Modelling Extremal Events for Insurance and Finance.* New York, 1997.

[19] https ://www.infoclimat.fr/climatologie-mensuelle/60141/janvier/2001/fes-sais.html