Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Institut National Polytechnique de Toulouse
Cotutelle internationale : Université Ouaga 1, Professeur Joseph Ki-Zerbo

Présentée et soutenue par

ABDOULAYE DIAMOUTENE

Le 26 novembre 2018

Contribution de la Théorie des Valeurs Extrêmes à la gestion et à la santé des systèmes

École doctorale : **AA - Aéronautique, Astronautique**

Spécialité : **Mathématiques Appliquées**

Unité de recherche :
LGP - Laboratoire Génie de Production de Tarbes (EA 1905)

Thèse dirigée par
Bernard KAMSU-FOGUEM et Diakarya BARRO

Jury

M. Mohamed LEMDANI, Professeur des Universités, Rapporteur
Mme. Sophie Mercier, Professeur des Universités, Rapporteur
M. Bisso SALEY, Professeur titulaire, Examinateur
M. Bernard KAMSU-FOGUEM, Maître de Conférences, Directeur de thèse
M. Diakarya BARRO, Professeur titulaire, Co-directeur de thèse
M. Farid NOUREDDINE, Maître de Conférences, Co-encadreur de thèse

Contribution de la Théorie des Valeurs Extrêmes à la gestion et à la santé des systèmes

Thèse présentée et soutenue par Abdoulaye DIAMOUTENE
Domaine : Mathématiques Appliquées
Spécialité : STATISTIQUE

Institut National Polytechnique de Toulouse (France)
Laboratoire Génie de Production de Tarbes
Université Professeur Joseph Ki-Zerbo de Ouagadougou (Burkina Faso)
Laboratoire d'Analyse Numérique d'Informatique et de Biomathématiques de Ouagadougou

Cotutelle Internationale

15 décembre 2018

Dédicace

Je dédie ce travail à ma famille, à mes amis et à tous ceux qui me sont chers.

Remerciements

Je tiens à exprimer ma profonde reconnaissance tout d'abord à mon directeur de thèse Monsieur Bernard Kamsu-Foguem, Habilité à Diriger des Recherches (HDR) à l'École Nationale d'Ingénieurs de Tarbes (ENIT), pour avoir accepté d'encadrer cette thèse et de m'avoir accueilli dans son laboratoire de recherche. Je remercie mon co-directeur de thèse professeur Diakarya Barro de l'Université Ouaga 2 au Burkina Faso et le co-encadreur Monsieur Farid Noureddine, Maître de conférences à l'École Nationale d'Ingénieurs de Tarbes. Je leur adresse mes plus sincères remerciements pour leur disponibilité, la confiance placée en moi, leurs précieux conseils et d'avoir permis à ce travail de voir le jour. Merci infiniment.

Je remercie Monsieur le professeur Mohamed Lemdani de l'Université de Lille 2, Madame le professeur Sophie Mercier de l'Université de Pau, pour avoir accepté de rapporter cette thèse et pour le temps précieux pris pour l'instruire. Leurs remarques, critiques et suggestions permettront d'améliorer la qualité scientifique de ce document.

Ma profonde gratitude va également à l'endroit de Monsieur le professeur Bisso Saley de l'Université Abdou Moumouni de Niamey (Niger), pour l'intérêt qu'il a porté à ce travail en acceptant de siéger dans le jury.

L'occasion est venue pour moi de remercier Madame Annie Noureddine, Professeur certifiée de langues, pour les efforts consentis lors de la rédaction des différents articles en anglais.

Je remercie Monsieur le Professeur Blaise Somé, directeur du Laboratoire d'Analyse Numérique d'Informatique et de Biomathématiques (LANIBIO) et Monsieur Longin Somé, Maître de conférences, directeur adjoint du LANIBIO, pour m'avoir accueilli dans leur laboratoire.

Mes remerciements s'addressent à Monsieur Youssouf Paré, directeur adjoint de l'Unité de Formation et de Recherche en Sciences Exactes et Appliquées (UFR-SEA) de l'Université Professeur Joseph Ki-Zerbo de Ouagadougou pour son soutien dans les procédures administratives.

Je remercie la direction et tout le personnel du Laboratoire Génie de Production de l'ENIT pour leur accueil et leurs encouragements pendant les différents séjours à Tarbes et particulièrement l'assistante administrative du laboratoire, Madame Éliane Casta pour son soutien lors du traitement des dossiers d'inscription.

Qu'il me soit permis de remercier très sincèrement les camarades Yacouba Simporé pour son hospitalité lors des différents séjours à Ouagadougou et Moussa Ba-

gayoko pour la diligence des documents administratifs.

Un remerciement particulier aux camarades Moumouni Diallo et Sinaly Dissa pour leur soutien indéfectible durant cette thèse.

Je suis très reconnaissant aux doctorants Tamba Camara, Boukaye Traoré et Lassana Coulibaly pour leur soutien lors des différents séjours en France.

Je remercie la direction de l'Institut Polytechnique Rural de Formation et de Recherche Appliquée (IPR/IFRA) de Katibougou au Mali pour son soutien et plus particulièrement toute ma gratitude à l'endroit du chef de département des sciences de base de l'IPR/IFRA, Dr Mamadou N. Sangaré, pour son accompagnement durant cette thèse.

Résumé

Le fonctionnement d'un système, de façon générale, peut être affecté par un incident imprévu. Lorsque cet incident a de lourdes conséquences tant sur l'intégrité du système que sur la qualité de ses produits, on dit alors qu'il se situe dans le cadre des événements dits extrêmes. Ainsi, de plus en plus les chercheurs portent un intérêt particulier à la modélisation des événements extrêmes pour diverses études telles que la fiabilité des systèmes et la prédiction des différents risques pouvant entraver le bon fonctionnement d'un système en général. C'est dans cette optique que s'inscrit la présente thèse. Nous utilisons la Théorie des Valeurs Extrêmes (TVE) et les statistiques d'ordre extrême comme outil d'aide à la décision dans la modélisation et la gestion des risques dans l'usinage et l'aviation. Plus précisément, nous modélisons la surface de rugosité de pièces usinées et la fiabilité de l'outil de coupe associé par les statistiques d'ordre extrême. Nous avons aussi fait une modélisation à l'aide de l'approche dite du « Peaks-Over Threshold, POT » permettant de faire des prédictions sur les éventuelles victimes dans l'Aviation Générale Américaine (AGA) à la suite d'accidents extrêmes. Par ailleurs, la modélisation des systèmes soumis à des facteurs d'environnement ou covariables passent le plus souvent par les modèles à risque proportionnel basés sur la fonction de risque. Dans les modèles à risque proportionnel, la fonction de risque de base est généralement de type Weibull, qui est une fonction monotone ; l'analyse du fonctionnement de certains systèmes comme l'outil de coupe dans l'industrie a montré qu'un système peut avoir un mauvais fonctionnement sur une phase et s'améliorer sur la phase suivante. De ce fait, des modifications ont été apportées à la distribution de Weibull afin d'avoir des fonctions de risque de base non monotones, plus particulièrement les fonctions de risque croissantes puis décroissantes. En dépit de ces modifications, la prise en compte des conditions d'opérations extrêmes et la surestimation des risques s'avèrent problématiques. Nous avons donc, à partir de la loi standard de Gumbel, proposé une fonction de risque de base croissante puis décroissante permettant de prendre en compte les conditions extrêmes d'opérations, puis établi les preuves mathématiques y afférant. En outre, un exemple d'application dans le domaine de l'industrie a été proposé.

Cette thèse est divisée en quatre chapitres auxquels s'ajoutent une introduction et une conclusion générales. Dans le premier chapitre, nous rappelons quelques notions de base sur la théorie des valeurs extrêmes. Le deuxième chapitre s'intéresse aux concepts de base de l'analyse de survie, particulièrement à ceux relatifs à l'analyse de fiabilité, en proposant une fonction de risque croissante-décroissante dans le modèle

à risques proportionnels. En ce qui concerne le troisième chapitre, il porte sur l'utilisation des statistiques d'ordre extrême dans l'usinage, notamment dans la détection de pièces défectueuses par lots, la fiabilité de l'outil de coupe et la modélisation des meilleures surfaces de rugosité. Le dernier chapitre porte sur la prédiction d'éventuelles victimes dans l'Aviation Générale Américaine à partir des données historiques en utilisant l'approche « Peaks-Over Threshold ».

Mots clés : Théorie des Valeurs Extrêmes, fiabilité des systèmes, analyse de survie, approche du peaks over threshold, fonction de risque non monotone, modèles à risques proportionnels, distribution de Pareto Généralisée, bootstrap non paramétrique, statistique d'ordre.

Abstract

The operation of a system in general may at any time be affected by an unforeseen incident. When this incident has major consequences on the system integrity and the quality of system products, then it is said to be in the context of extreme events. Thus, increasingly researchers have a particular interest in modeling such events with studies on the reliability of systems and the prediction of the different risks that can hinder the proper functioning of a system. This thesis takes place in this very perspective. We use Extreme Value Theory (EVT) and extreme order statistics as a decision support tool in modeling and risk management in industry and aviation. Specifically, we model the surface roughness of machined parts and the reliability of the associated cutting tool with the extreme order statistics. We also did a modeling using the "Peaks-Over Threshold, POT" approach to make predictions about the potential victims in the American General Aviation (AGA) following extreme accidents. In addition, the modeling of systems subjected to environmental factors or covariates is most often carried out by proportional hazard models based on the hazard function. In proportional hazard models, the baseline risk function is typically Weibull distribution, which is a monotonic function. The analysis of the operation of some systems like the cutting tool in the industry has shown that a system can deteriorated on one phase and improving on the next phase. Hence, some modifications have been made in the Weibull distribution in order to have non-monotonic basic risk functions, more specifically, the increasing-decreasing risk function. Despite these changes, taking into account extreme operating conditions and overestimating risks are problematics. We have therefore proposed from Gumbel's standard distribution, an increasing-decreasing risk function to take into account extreme conditions, and established mathematical proofs. Furthermore, an example of the application in the field of industry was proposed.

This thesis is organized in four chapters and to this must be added a general introduction and a general conclusion. In the first chapter, we recall some basic notions about the Extreme Values Theory. The second chapter focuses on the basic concepts of survival analysis, particularly those relating to reliability analysis by proposing a function of increasing-decreasing hazard function in the proportional hazard model. Regarding the third chapter, it deals with the use of extreme order statistics in industry, particularly in the detection of defective parts, the reliability of the cutting tool and the modeling of the best roughness surfaces. The last chapter focuses on the prediction of potential victims in AGA from historical data using the

Peaks-Over Threshold approach.

Keywords : Extreme Values Theory, reliability of systems, survival analysis, approach of peaks over threshold, non monotonous hazard function, proportional hazards models, Generalized Pareto Distribution, non parametric bootstrap, order statistics.

Table des matières

Dédicace i

Remerciements ii

Résumé iv

Table des matières viii

Table des figures xii

Liste des tableaux xiv

Introduction Générale 1

1 Eléments de la Théorie des Valeurs Extrêmes 7
1.1 Introduction . . . 7
1.2 Approche des maxima ou des minima par blocs . . . 11
1.2.1 Comportement asymptotique des extrêmes . . . 12
1.2.2 Domaines d'attraction des lois GEV . . . 15
1.2.2.1 Notion de fonction à variation régulière . . . 16
1.2.2.2 Domaine d'attraction de la loi de Gumbel . . . 17
1.2.2.3 Domaine d'attraction de la loi de Weibull . . . 17
1.2.2.4 Domaine d'attraction de la loi de Fréchet . . . 17
1.2.3 Méthodes d'estimation des paramètres des lois GEV . . . 19
1.2.3.1 Méthodes d'estimation pour les lois du domaine d'attraction de Gumbel . . . 19
1.2.3.2 Méthodes d'estimation pour les lois des domaines d'attraction de Fréchet et Weibull . . . 21
1.3 Approche des dépassements de seuil : loi GPD . . . 22
1.3.1 Modélisation des excès . . . 23
1.3.2 Méthodes pour la détermination du seuil . . . 24
1.3.2.1 Méthodes graphiques . . . 25
1.3.2.2 Une approche de compromis . . . 28
1.3.3 Estimation des paramètres du modèle GPD . . . 28

1.3.3.1 Méthode du maximum de vraisemblance 28
1.3.3.2 Méthode des moments 29
1.3.3.3 Méthode des moments de probabilité pondérés . . . 29
1.3.4 Concepts de quantiles extrêmes 30
1.3.4.1 Quelques éléments théoriques sur les quantiles extrêmes 30
1.3.4.2 Estimation de quantiles extrêmes pour la loi GPD . . 32
1.4 Adéquation des modèles des valeurs extrêmes 33
1.4.1 Introduction 33
1.4.2 Probabilité et Quantile Plot 34
1.4.3 Tests d'adéquation de Kolmogorov Smirnov 34
1.4.4 Quantification de l'incertitude par la méthode du bootstrap dans la théorie des valeurs extrêmes 34
1.4.4.1 Approche non paramétrique du bootstrap 35
1.4.4.2 Bootstrap pour l'estimation d'une erreur standard . . 35
1.4.4.3 Quantification des incertitudes sur les estimations dans l'approche POT 37
1.5 Conclusion 37

2 Analyse de Survie ou de Fiabilité des systèmes 39
2.1 Introduction 39
2.2 Concepts de base de l'analyse de fiabilité 39
2.2.1 Fonctions décrivant l'évolution d'un système 40
2.2.1.1 Fonction de survie S ou de fiabilité R 40
2.2.1.2 Fonction de répartition F 40
2.2.1.3 Fonction de densité de probabilité f 41
2.2.1.4 Fonction de risque instantané 41
2.2.1.5 Fonction de risque cumulé 41
2.2.2 Fonction de fiabilité d'un système 41
2.2.3 Taux de défaillance instantané ou fonction de risque instantané 43
2.3 Quelques lois usuelles de fiabilité 44
2.3.1 Loi normale 44
2.3.2 Loi log-Normale 44
2.3.3 Loi exponentielle 45
2.3.4 Loi de Weibull 45
2.3.5 Distribution des valeurs extrêmes 45
2.4 Modèles à risques proportionnels pour la fiabilité des systèmes 46
2.4.1 Modèles à risques proportionnels 46
2.4.1.1 Présentation du modèle à risques proportionnels de Cox 46
2.4.1.2 Facteurs d'influence ou variables d'environnement . . 47
2.4.1.3 Modèles paramétriques classiques à risques proportionnels 48
2.4.2 Modèles GEV à risques proportionnels 48

2.4.2.1 Modélisation de la fonction de risque par la Théorie des Valeurs Extrêmes généralisées 49
2.4.2.2 Fonction de risque croissante-décroissante 49
2.4.2.3 Modélisation d'une nouvelle fonction de risque de base 51
2.4.3 Application du modèle en usinage 52
2.4.3.1 Modélisation de la rugosité 52
2.4.3.2 Étude de la performance d'un outil de coupe 53

3 Statistiques d'ordre et contrôle de qualité 57
3.1 Statistiques d'ordres et contrôle de qualité dans l'industrie 57
3.1.1 Statistiques d'ordre 58
3.1.2 Statistiques d'ordre extrêmes 58
3.1.3 Contrôle statistique de qualité dans l'industrie manufacturière 61
3.1.3.1 Contrôle de fabrication et contrôle en cours de fabrication 62
3.1.3.2 Le contrôle de réception 62
3.2 Contrôle de qualité en usinage 62
3.2.1 Introduction 62
3.2.2 Présentation des données 63
3.2.3 Méthodologie 63
3.2.4 Résultats 64
3.2.4.1 Estimation des paramètres 64
3.2.4.2 Distribution des observations ordonnées pour les cinq lots 66
3.2.4.3 Distribution de minima et maxima 67
3.2.4.4 Diagramme des différentes étapes de la procédure . . 68
3.3 Conclusion 70

4 Modélisation des accidents extrêmes 71
4.1 Gestion des risques d'accidents dans l'aviation 71
4.2 Modélisation dans l'aviation par la méthode POT et application aux données de l'Aviation Générale Américaine 71
4.2.1 Introduction 71
4.2.2 Présentation des données 73
4.2.3 Méthodologie 75
4.2.4 Résultats 75
4.2.4.1 Détermination graphique du seuil 75
4.2.4.2 Graphe de stabilité des paramètres de forme et d'échelle 76
4.2.4.3 Estimation des paramètres de la GPD pour différents seuils 77
4.2.4.4 Estimation du niveau de retour 79
4.2.4.5 Quantification des incertitudes 80
4.2.4.6 Comparaison entre quantile extrême observé et simulé 84

4.3 Conclusion 85

Conclusion Générale et perspectives **86**

Bibliographie **87**

Bibliographie **88**

Table des figures

0.0.1 Statistique classique et statistique des valeurs extrêmes 2

1.1.1 Maxima annuels du niveau de la mer au Port Pirie, sud de l'Australie Coles (2001) 8
1.1.2 Températures journalières minimales en Ohio, Coles (2001) 9
1.1.3 Cumul des précipitations journalières (sud-ouest de l'Angleterre), Coles (2001) 10
1.1.4 Histogramme de résistance à la rupture de quelques fibres de verre, Coles (2001) 11
1.2.1 Approche des blocs de maxima (minima) 12
1.2.2 À gauche fonctions de répartition, à droite fonctions de densité des lois des valeurs extrêmes 15
1.3.1 Excès au-dessus du seuil 23
1.3.2 À gauche : fonctions de répartition et à droite : fonctions de densité des lois de Pareto Généralisée 24
1.3.3 Graphe de la fonction moyenne des excès (MRlplot) pour des données de pluie anglaises, (Coles, 2001) 27
1.3.4 Graphe de la stabilité des paramètres (tcplot) 27
1.3.5 Quantiles extrêmes et queue de distribution 31
1.4.1 Bootstrap pour l'estimation d'une erreur standard 36

2.2.1 Fonction de fiabilité et de défaillance 42
2.4.1 Fonction de risque de base 51
2.4.2 Comparaison entre la fonction de risque de base et le modèle à risque proportionnel GEV 54
2.4.3 Comparaison entre la fonction de risque proportionnel GEV et la fonction flexible de Weibull 56

3.1.1 Fonction de répartition à gauche et de densité à droite du maximum pour un échantillon de taille 10, 100 et 1000 pour une loi de Gumbel standard 60
3.1.2 Fonction de répartition à gauche et de densité à droite du minimum pour un échantillon de taille 10 (verte), 100 (rouge) et 100 (bleue) pour une loi de Gumbel standard. 61
3.2.1 Fonction de densité à gauche et de répartition à droite de Weibull pour les rugosités de surface des pièces usinées 65
3.2.2 Distribution des minimums pour les cinq séries de mesure 67

3.2.3 Distribution des maximums pour les cinq séries de mesure 68
3.2.4 Diagramme des différentes étapes 69

4.2.1 Victimes annuelles 74
4.2.2 Concentration des accidents de l'Aviation Générale Américaine 75
4.2.3 Fonction moyenne des excès 76
4.2.4 Graphe de stabilité des paramètres de forme et d'échelle 77
4.2.5 QQ-plot et PP-plot par MV (droite) et MM (gauche) avec le seuil $u = 40$ 81
4.2.6 QQ-plot et PP-plot par MV (droite) et MM (gauche) avec le seuil $u = 45$ 81
4.2.7 QQ-plot et PP-plot par MV (droite) et MM (gauche) avec le seuil $u = 50$ 82

Liste des tableaux

1.2.1 Domaines d'attraction et quelques lois associées 18

2.3.1 Fonctions de densité de quelques lois usuelles de fiabilité 46

3.2.1 Paramètres de Weibull pour les cinq lots 65
3.2.2 Probabilité d'avoir la dixième mesure inférieure ou égale à 4μm et fiabilité correspondante associée à l'outil de coupe 66
3.2.3 Probabilité d'obtention d'au moins huit mesures inférieures ou égales à 4μm et fiabilité correspondante associée à l'outil de coupe 67

4.2.1 Estimation des paramètres par la méthode du maximum de vraisemblance 78
4.2.2 Estimation des paramètres par la méthode des moments 78
4.2.3 Estimation des paramètres pour la méthode des probabilités de moments pondérés 79
4.2.4 Estimation possible de quantiles pouvant être attendus dans 5, 10 et 25 ans 79
4.2.5 Récapitulatif des statistiques par bootstrap pour le seuil 40 83
4.2.6 Récapitulatif des statistiques par bootstrap pour le seuil 45 83
4.2.7 Récapitulatif des statistiques par bootstrap pour le seuil 50 83
4.2.8 Erreurs quadratiques moyennes pour les niveaux de retour, les paramètres de forme et d'échelle 84
4.2.9 Estimation des niveaux de retour attendus pour les différents seuils avec un intervalle de confiance à 95% 84

Sigles, Abréviations et Notations

Sigles	Significations
AG :	Aviation Générale
MV :	Méthode du Maximum de Vraisemblance
MM :	Méthode des Moments
MOM :	Method Of Moment
PWM :	Méthode des Moments de Probabilités Pondérés
NTS :	National Transportation Statistics
AGA :	Aviation Générale Américaine
rmse :	root mean standard error
GEV :	Generalized Extreme Value
EVT :	Extreme Value Theory
GPD :	Generalized Pareto Distribution
POT :	Peaks Over Threshold
TVE :	Thérie des Valeurs Extrêmes
PHM :	Proportional Hazard Models
iid :	indépendantes et identiquement distribuées
NTSB :	National Transportation Safety Board
CAA :	Civil Aviation Authority
v.a :	variable aléatoire
U.S.A :	États Unis d'Amérique

Notations

y_F :	Point terminal supérieur
y :	Réalisation de la v.a Y
L_{MV} :	Log-vraisemblance conditionnelle aux données observées x
$F(.)$:	Fonction de répartition
$F_u(.)$:	Fonction de répartition des dépassements d'un seuil u
$\hat{F}(.)$:	Fonction de répartition estimée
$\overline{F}^{\leftarrow}$:	Inverse généralisé de $\overline{F}$
$F^{-1}(.)$:	Fonction de répartition inverse
$\bar{F}(.)$:	Fonction de survie
$S(.)$:	Fonction de survie
$h(.)$:	Fonction de risque
$H(.)$:	Fonction de risque cumulé
$R(.)$:	Fonction de fiabilité
$\xi(.)$:	Paramètre de forme GEV
$\sigma(.)$:	Paramètre d'échelle
$\mu(.)$:	Paramètre de localisation
$\Gamma(.)$:	Fonction Gamma
$\Lambda(.)$:	Fonction de répartition de Gumbel
$\psi(.)$:	Fonction de répartition de Weibull
$\Phi(.)$:	Fonction de répartition de Fréchet
$e(.)$:	Mean excess function
$E(.)$:	Espérance mathématique
M_s :	Moments Pondérés d'ordre s
$q(.)$:	Fonction quantile
$T(.)$:	Période de retour
N_u :	Nombre d'observations au dessus du seuil u
x_T :	Niveau de retour
φ :	Fonction de risque d'environnement
τ :	Réciproque de la fonction de risque
F_M :	Distribution du maximum d'un échantillon de loi F
F_W :	Distribution du minimum d'un échantillon de loi F
R_a :	Rugosité de surface

Introduction générale

La Théorie des Valeurs Extrêmes (TVE) est une branche des statistiques apparue entre 1920 et 1943 grâce à Fréchet (1927); Fisher et Tippet (1928) et Gnedenko (1943). Cette branche a connu un grand essor depuis les années cinquante bien que le théorème principal ait été énoncé par Fisher et Tippet dès 1928. Elle a pour but de modéliser et de décrire la survenue et l'intensité d'évènements dits extrêmes, c'est-à-dire qui présentent des variations de très grande amplitude (ayant une faible probabilité d'apparition). Lorsque le comportement de ces évènements est dû au hasard, on peut étudier leur loi. Ils sont dits extrêmes quand il s'agit de valeurs beaucoup plus grandes ou plus petites que celles observées habituellement.

Au cours des dernières décennies, nous avons pu observer dans le monde scientifique un intérêt renouvelé pour la modélisation et l'analyse statistique des évènements rares. L'étude de ces évènements revient à l'analyse des queues de distribution des fonctions ou de manière équivalente, à l'analyse de la plus grande ou de la plus petite observation d'un échantillon. En ce sens, nous pouvons considérer la théorie des valeurs extrêmes comme la contrepartie de la théorie statistique, qui est principalement basée sur l'étude des tendances centrales d'un échantillon plutôt que sur des observations extrêmes. La figure (0.0.1) illustre, d'une part, les centres d'intérêt de la statistique classique avec la mise en évidence des valeurs centrales et d'autre part, la statistique des valeurs extrêmes avec cette fois la mise en évidence des observations dans la queue de distribution.

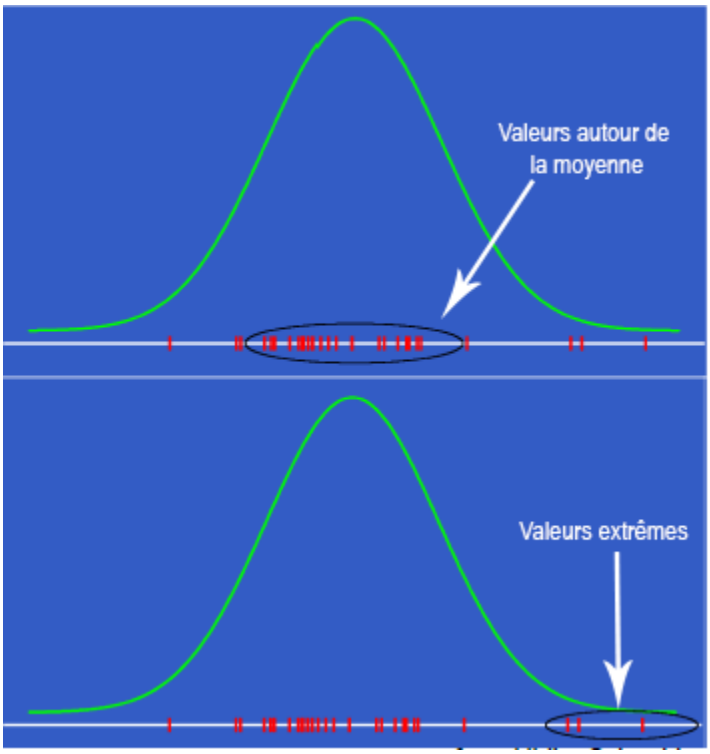


FIGURE 0.0.1 – Statistique classique et statistique des valeurs extrêmes

Cependant, la compréhension et l'étude des évènements extrêmes sont le plus souvent confrontées à un problème de manque d'information ou de données, notamment lors d'une analyse statistique des valeurs eloignées de la moyenne. De plus, les informations relatives aux évènements les plus fréquents occultent souvent celles des évènements extrêmes, ce qui peut conduire à des pertes d'information ou à des informations peu fiables. Par ailleurs, les approches statistiques classiques s'intéressent à toutes les valeurs observées sans distinction, fournissant ainsi des modèles appropriés pour la description de comportement central et non des informations spécifiques aux comportements extrêmes. Les outils probabilistes traditionnels développés ne sont plus adéquats à l'appréhension d'un univers où les valeurs ne sont pas du même ordre de grandeur et sont loin d'être concentrées autour de leur moyenne avec les variations de faible amplitude, pour plus de détails voir Coles (2001); Raggad (2009).

Un évènement extrême est un évènement qui a une très faible probabilité de se produire mais qui, lorsqu'il se produit, prend de très grandes ou petites valeurs et génère très souvent un grand impact. On notera la différence avec un évènement rare qui, par définition, est un évènement dont la probabilité d'occurence est faible. Le fait qu'un évènement soit rare n'implique pas qu'il soit extrême ; il est dépourvu de la notion de quantifiabilité (petite ou grandes valeur). A l'inverse, tout évènement extrême est rare au sens où il a une faible probabilité de se produire.

Du fait de son cadre probabiliste, la TVE permet de réaliser des extrapolations se rapportant aux comportements stochastiques et asymptotiques des extrêmes. Elle

permet aussi d'évaluer les évènements extrêmes et les conséquences pouvant en résulter, en cherchant à ajuster un modèle prédictif approprié à ces observations extrêmes. Ainsi, sous certaines conditions de régularité, dans la TVE, la loi du maximum ou du minimum d'un ensemble de variables aléatoires ne peut appartenir qu'à une seule famille de loi : Weibull lorsque le paramètre de forme est négatif, Gumbel lorsqu'il est nul et Fréchet lorsqu'il est positif.

Dans la TVE, on distingue principalement deux approches ; une première approche concernant la loi des valeurs extrêmes généralisées, en anglais (Generalized Extreme Value, GEV) et qui s'intéresse à la loi du maximum ou du minimum convenablement normalisée. Cette première approche repose sur le théorème de Fisher et Tippett. La seconde approche concerne la loi des excès, en anglais (Peaks-Over Threshold, POT). Elle consiste à utiliser les observations qui dépassent un certain seuil déterministe. Les différences entre ces observations et le seuil sont appelées excès. Cette approche est basée sur la convergence vers une loi de Pareto Généralisée, (en anglais GPD) assurée par le théorème de Pickands-Balkema-de Haan.

La première approche souvent appelée « approche des blocs de maxima » a été très critiquée dans la littérature, elle utilise le maximum de chaque bloc et un maximum d'un bloc peut être inférieur à une valeur qui n'est pas le maximum d'un autre bloc. Ainsi, il y a une perte d'information qui peut conduire à des résultats peu fiables. Pour pallier ces insuffisances, la seconde méthode dite aussi méthode des excès est assez utilisée bien que la sélection du seuil soit problématique. En fait, le seuil doit être assez grand pour que l'approximation par une loi GPD soit valide, mais pas trop élevé pour garder un nombre suffisant de dépassements pour une meilleure estimation.

Par ailleurs, la littérature concernant la Théorie des Valeurs Extrêmes est très fournie. Nous pouvons par exemple citer les ouvrages de Gumbel* (1958) qui constitue une référence classique dans la TVE. L'ouvrage de Galambos (1985) propose une description historique des principaux résultats relatifs à la TVE. Quant à Leadbetter *et al.* (1983); Resnick (1987), ils fournissent un traitement mathématique et probabiliste sur les notions de fonctions à variation lente et à variation regulière, tandis que Kotz et Nadarajah (2000); Coles (2001); Reiss et Thomas (2001); Castillo *et al.* (2005) font des rappels des résultats théoriques et proposent des applications pratiques de la TVE.

La TVE est couramment utilisée dans des domaines variés Kotz et Nadarajah (2000); Castillo *et al.* (2005). Il s'agit souvent de domaines où les statistiques sont utilisées comme outils prédictifs ou de gestion. En climatologie, on pourra par exemple se référer à Carter et Chalenor (1981), à (Walshaw et Anderson, 2000) pour la modélisation des champs éoliens, en océanographie à Tawn (1992); Dawson (2000); Harris (2001) et pour des données sportives à Robinson et Tawn (1995). Un ouvrage de référence aussi bien pour la théorie que pour ses applications financières est celui d'Embrechts *et al.* (1997). En hydrologie, domaine dans lequel la prévision des crues par exemple est particulièrement importante, voir Davison et Smith (1990); Katz (2002); Amar (2014). Dans les domaines des sciences humaines et sociales, la TVE peut également contribuer à la compréhension de nombreux phénomènes sociaux.

Par exemple, dans le domaine de la démographie, le débat initié par Gumbel en 1937, auquel Fréchet a pris une part active, concerne le concept de durée extrême de la vie humaine et sa mesure Thatcher (1999).

Depuis la décennie précédente, les domaines d'application de la TVE se sont élargis. Par exemple, on trouve de nouvelles applications en management avec Dahan et Mendelson (2001), dans les assurances et la gestion des risques où l'une des préoccupations est la prise en compte des grands sinistres, voir Raggad (2009); Deme (2013); Amar (2014). Citons pour le domaine de l'agro-alimentaire les travaux de Kawas et Moreira (2001) et Lavenda et Cipollone (2000) pour la thermodynamique des tremblements de terre. En médecine et biomédecine, on peut citer les travaux de Roberts (2000); Bermudez et Mendes (2012) et en ce qui concerne les changements climatiques, on peut se référer à Thompson *et al.* (2001); Christidis *et al.* (2011); Zwiers *et al.* (2011); Diallo (2017); Methni (2013). On retrouve aussi des applications en ingéniérie avec Makkonen (2008); Piera-Martinez (2008), en énergie renouvelable, voir Papatheou *et al.* (2017), dans le domaine de la recherche opérationnelle voir Romeijnders *et al.* (2012); Sha *et al.* (2017) et dans le domaine des sciences forestières pour la modélisation du diamètre des arbres, Gorgoso-Varela *et al.* (2014). La TVE a aussi été utilisée dans la modélisation du trafic routier et des accidents de circulation, voir Schmidt *et al.* (2014); Songchitruksa et Tarko (2006). Son usage apparaît également dans la modélisation des sytèmes électroniques de payement, Wang et Dey (2011) et l'usure des composants en ingéniérie par Benstock et Cegla (2016).

Dans le domaine de l'aéronautique et la navigation des satellites, on retrouve les travaux de Bureau (2012); Dey et Das (2016). La sensibilité de ces deux domaines d'activité est généralement mise en exergue lorsqu'interviennent des incidents graves tels que les crashs d'avions et les défaillances des réseaux de télécommunication. En outre, ces domaines exigent beaucoup d'investissements financiers, technologiques et humains et de ce fait la maîtrise, la prédiction et la gestion des risques relatifs à ces deux domaines sont importantes pour les compagnies d'assurance et autres structures associées. Nous proposons donc à partir de l'approche POT de la TVE une modélisation des victimes de l'Aviation Générale Américaine (AGA). Le modèle ainsi proposé se présente comme un outil d'aide à la décision pour les compagnies d'assurance et responsables de l'AGA.

Dans l'industrie manufacturière, la production en masse est l'une des conditions pour minimiser les coûts de production, cependant cette production de masse doit satisfaire aux exigences de qualité des commandes pour conserver la clientèle. Il arrive souvent qu'un incident grave non détecté entache toute la production en créant un défaut global de production. Cette situation peut s'avérer être l'impact de l'outil de coupe dont le bec est brisé sur la surface de rugosité de pièces usinées. Nous utilisons donc les statistiques d'ordres extrêmes pour modéliser la surface des meilleures pièces usinées, ainsi que les pièces défectueuses puis la fiabilité de l'outil de coupe.

Le terme de durée de vie a été initialement utilisé en médecine et en biostatistique, il désigne le temps écoulé jusqu'à la survenue d'un événement précis. L'événement

étudié communément appelé « décès » est le passage irréversible entre deux états généralement nommés « vivant et décès ». L'événement terminal n'est pas forcément la mort : il peut s'agir de l'apparition d'une maladie, par exemple le temps avant une rechute ou le temps entre le diagnostic et la guérison. Il faut noter que les concepts évoqués dans l'analyse de survie ont leur équivalent dans l'analyse de fiabilité, par exemple on parle de taux de décès pour les êtres vivants et de taux de défaillance pour les machines ou composants électroniques. De ce fait, dans ce travail, on désignera par système tout être vivant ou matériel, mais l'accent sera plutôt mis sur le cas du matériel. En outre, en s'intéressant à la performance d'un système, on est souvent amené à évaluer l'influence des facteurs d'environnement ou les conditions d'opération. Généralement, cette étude se fait en utilisant les modèles à risques proportionnels. C'est dans cette optique, que Cox (1972) a introduit son modèle semi-paramétrique pour prendre en compte les effets des variables d'environnement. Dans ce modèle, la fonction de risque est le produit de deux fonctions la première représentant la fonction de risque de base et la seconde les effets d'environnement.

Le modèle à risque proportionnel est dit semi-paramétrique lorsque la fonction de risque de base n'est pas spécifiée et paramétrique lorsqu'elle est connue. La plupart des fonctions de risque de base sont monotones et la plus utilisée est celle de Weibull. Cependant, l'étude des systèmes a montré que certains peuvent être soumis à deux phases d'évolution différentes. Une première phase où le système subit une dégradation croissante de sa performance et la suivante où le système retrouve progressivement toutes ses capacités. Ces deux types d'évolution peuvent être associées au concept de taux de défaillance ; pour la première phase, on parle de taux de défaillance croissant et la seconde de taux de défaillance décroissant. Pour cette raison, des modifications ont été apportées à la distribution de Weibull pour obtenir des fonctions de risque non monotones, pour plus de détails voir Park *et al.* (2016); Park et Park (2018). La prise en compte des conditions extrêmes d'opération et la surestimation des taux de défaillance s'avèrent problématiques pour la distribution de Weibull comme fonction de risque de base, voir Zio (2016). Pour pallier ces insuffisances, nous utilisons la TVE pour construire une fonction de risque croissante-décroissante.

De façon globale, le présent travail a pour objectifs d'illustrer et de mettre en application les méthodes de la TVE dans les études de fiabilité et de performance des systèmes. Les statistiques d'ordre extrême ont permis la modélisation de la surface de rugosité, de la fiabilité de l'outil de coupe et le nombre de pièces défectueuses par lots dans l'industrie manufacturière. De plus, l'approche POT a été mise en application dans la prédiction des victimes de crash dans l'Aviation Générale Américaine. Cependant cette méthode est basée sur la sélection du seuil adéquat qui est problématique dans la mesure où la plupart des méthodes de sélection de seuil sont subjectives. De ce fait, nous avons utilisé dans ce travail une approche basée sur les méthodes graphiques et une estimation intensive des paramètres pour déterminer l'intervalle du seuil optimal. Il faut aussi souligner le fait que ces modélisations ont permis de caractériser les lois des données de rugosité et d'avion afin de proposer des outils d'aide à la décision et de gestion.

Notre thèse est ainsi structurée en quatre chapitres. Le premier chapitre concerne la TVE sur ses fondements théoriques et les approches les plus utilisées. Le deuxième chapitre porte sur les concepts de base de l'analyse de survie, particulièrement l'analyse de fiabilité. Les deux premiers chapitres constituent de ce fait la partie théorique qui sert à introduire les concepts nécessaires pour contextualiser les questions de la présente thèse. Les chapitres 3 et 4 on fait l'objet de deux publications dans le journal « Measurement, 2018 (Quality control in machining using order statistics) », voir (Diamoutene *et al.*, 2018b) et le journal « Transportation Practice Research Part A, 2018 (Prediction of U.S. General Aviation fatalities from extreme value approach) », voir Diamoutene *et al.* (2018a) ; ils constituent en fait la partie empirique de la présente thèse.

Chapitre 1

Eléments de la Théorie des Valeurs Extrêmes

1.1 Introduction

La Théorie des Valeurs Extrêmes (TVE) a pour objectif l'étude du comportement asymptotique des grandes ou petites observations d'un échantillon de variables aléatoires indépendantes et identiquement distribuées (iid). L'approche classique en théorie de probabilités s'intéresse au comportement moyen et à la variabilité des phénomènes autour de la moyenne par le biais d'outils probabilistes comme par exemple la loi des grands nombres ou le théorème central limite. Le théorème fondamental de la Théorie des Valeurs Extrêmes, connu sous le nom de théorème de Fisher-Tippett, donne quant à lui les lois limites possibles du maximum de l'échantillon et permet ainsi d'avoir une certaine connaissance sur le comportement stochastique de la queue de distribution.

L'utilisation des lois des valeurs extrêmes repose sur des propriétés des statistiques d'ordre et sur des méthodes d'extrapolation. Plus précisément, elle repose principalement sur les distributions limites des extrêmes et leurs domaines d'attraction. En pratique, on souhaite estimer des petites probabilités ou des quantités dont la probabilité d'observation est très faible, c'est-à-dire proche de zéro. De plus, elle ne vise pas à modéliser ou à estimer la fonction de répartition inconnue F dans son ensemble, mais seulement ses queues de distribution qui sont utiles à la représentation des extrêmes. Les deux approches utilisées sont : la méthode des blocs de maxima et l'approche de dépassement de seuil ou Peaks Over Threshold (POT) qui modélise la loi des variables excédant un certain seuil fixé.
Dans la section (1.2), on s'intéressera à l'approche des blocs de maxima et aux différents résultats théoriques de la théorie des valeurs extrêmes univariées.
Avant d'entamer la section (1.2), nous présentons quelques exemples pratiques tirés de Coles (2001) qui illustrent différentes situations dans les modélisations utilisant la TVE.

Exemple 1.1.1 *La figure (1.1.1) illustre les maxima annuels du niveau de la mer*

au Port Pirie dans le sud de l'Australie durant la période 1923-1987.
A partir de ces données, nous pouvons faire des prédictions sur le niveau maximal de la mer susceptible de se reproduire dans cette région au cours des 100 ou 1000 ans à venir. Cependant, l'estimation du niveau de retour pour de longues périodes peut être confrontée à des incertitudes dues aux changements climatiques.

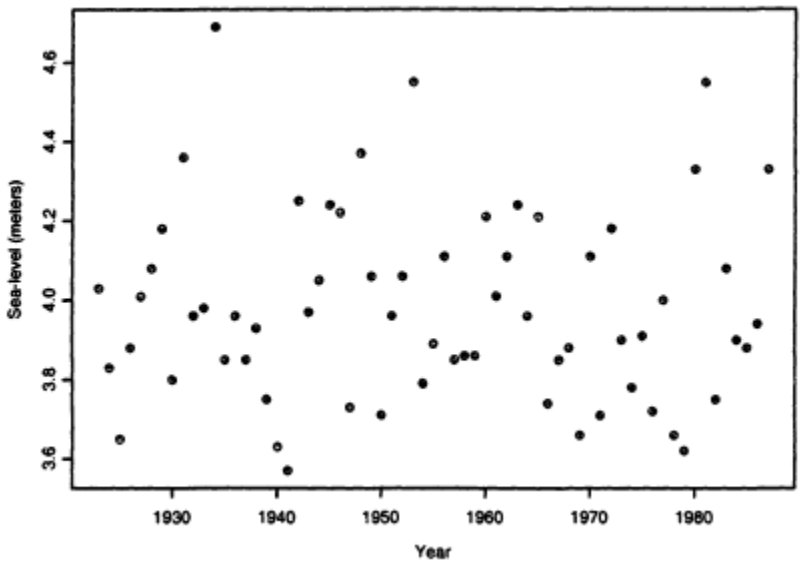


FIGURE 1.1.1 – Maxima annuels du niveau de la mer au Port Pirie, sud de l'Australie Coles (2001)

Exemple 1.1.2 *La figure (1.1.2) représente les températures minimales quotidiennes en Ohio. On remarque que la série met en évidence un phénomène cyclique annuel. De ce fait, il n'est pas raisonnable d'utiliser les modèles des valeurs extrêmes qui modélisent les processus à variation constante au cours du temps.*

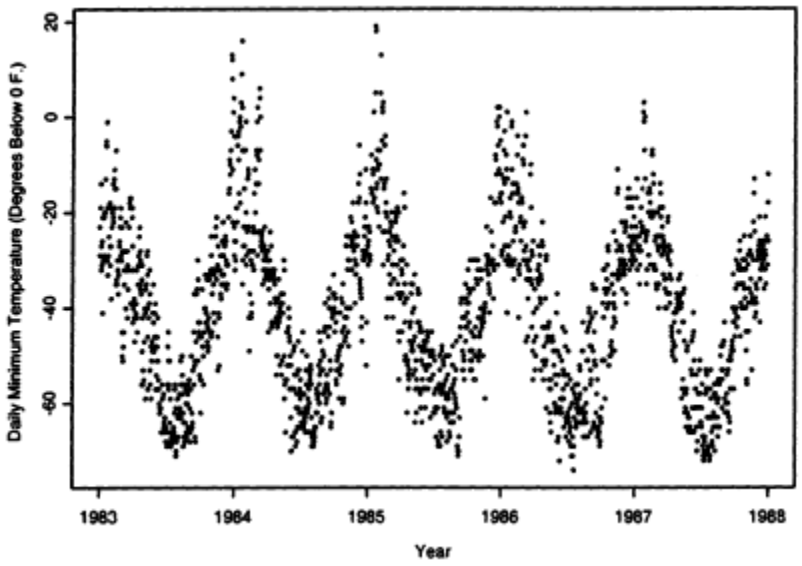


FIGURE 1.1.2 – Températures journalières minimales en Ohio, Coles (2001)

Exemple 1.1.3 *La figure (1.1.3) présente une série des précipitations journalières cumulées dans le sud-ouest de l'Angleterre, enregistrées durant la période allant de 1914 à 1962. Si le modèle statistique adéquat pour une telle série était connu, les données présentées pourraient être utilisées pour estimer les paramètres nécessaires. Cependant, si ce modèle n'est pas connu et en tenant compte du très grand nombre d'observations, on pourra utiliser un modèle asymptotique tel que l'approche POT en considérant comme événement extrême les précipitations au-dessus d'un certain seuil (par exemple 30mm) au lieu de celle des blocs de maxima.*

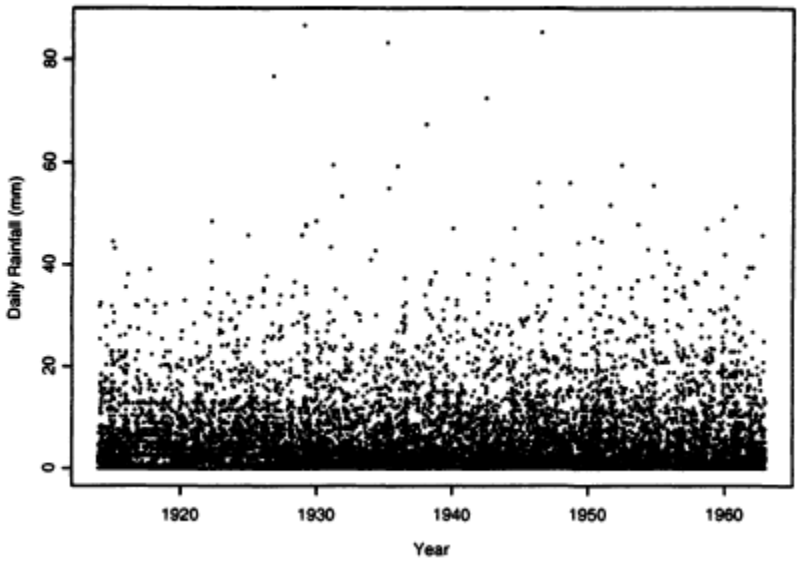


FIGURE 1.1.3 – Cumul des précipitations journalières (sud-ouest de l'Angleterre), Coles (2001)

Exemple 1.1.4 *Les hypothèses utilisées dans la modélisation des événements extrêmes de grande ou faible amplitude sont les mêmes. On retrouve cette situation dans le domaine de la fiabilité. Dans la conception, beaucoup de systèmes sont considérés comme un ensemble de plusieurs composantes, de telle sorte que l'ensemble du système tombe en panne, dans le cas d'un montage en série, si l'une des composantes est défaillante. On pourrait donc utiliser la TVE pour détecter le maillon le plus faible, sachant que la fiabilité du système est égale à celle de la composante la plus faible.*
La figure (1.1.4) illustre cette situation en présentant les valeurs de la résistance à la rupture de 63 fibres de verre d'une longueur de 1.5 cm, enregistrées dans des conditions expérimentales. Dans cette situation, nous pouvons considérer une fibre de verre comme un paquet de nombreuses fibres plus fines, de telle sorte que si l'une des petites fibres est défaillante, le paquet entier est défaillant.

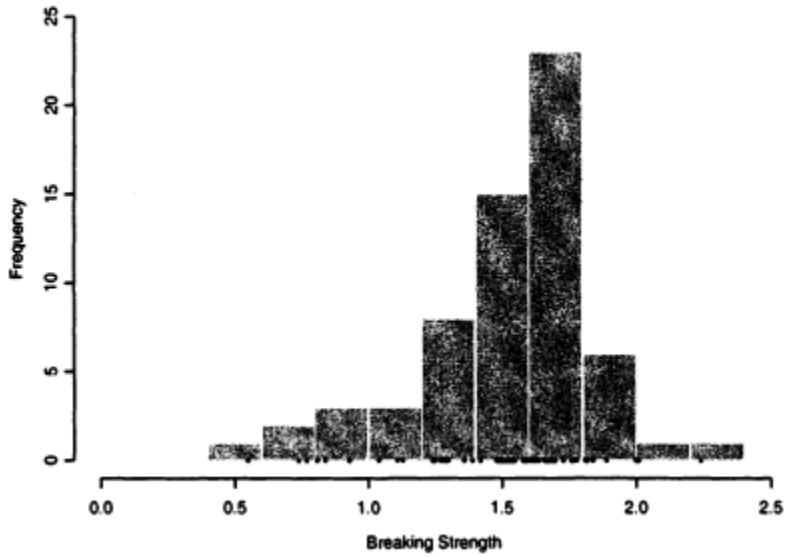


FIGURE 1.1.4 – Histogramme de résistance à la rupture de quelques fibres de verre, Coles (2001)

1.2 Approche des maxima ou des minima par blocs

Dans cette approche, on regroupe les observations par blocs de taille fixée, et on s'intéresse à la loi des maxima (minima) sur chaque bloc pour une loi des valeurs extrêmes généralisées (GEV). Il faut alors trouver un compromis entre la taille des blocs qui doivent être assez grands pour que l'approximation par la loi GEV soit réaliste et le nombre de blocs qui doit être assez grand pour obtenir une estimation précise des trois paramètres de la GEV. En pratique, le choix retenu est souvent de considérer les maxima (minima) annuels pour éviter les effets saisonniers, ce qui nécessite d'observer le phénomène sur de nombreuses années. Cependant, dans les domaines de l'industrie par exemple, les données ne sont souvent pas à l'échelle annuelle. Cette méthode présente l'inconvénient bien connu de ne pas prendre en compte toutes les informations présentes dans les données sur les évènements extrêmes puisqu'elle ne conserve qu'une valeur par bloc pour l'inférence statistique. La figure (1.2.1) illustre l'approche par une division en n blocs ; les bâtons en rouge indiquent les maxima des blocs tandis que ceux en jaune indiquent les minima.

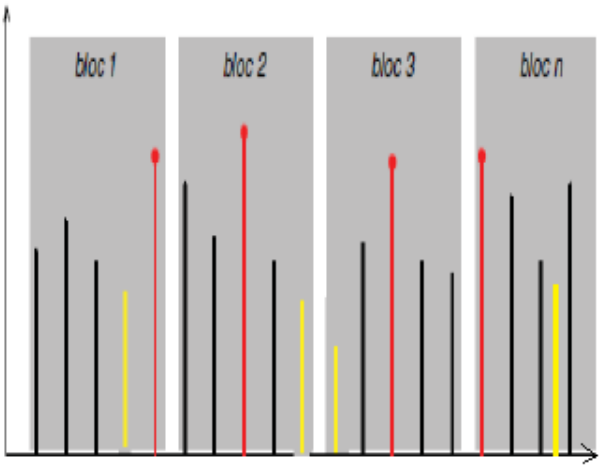


FIGURE 1.2.1 – Approche des blocs de maxima (minima)

1.2.1 Comportement asymptotique des extrêmes

Nous présentons ici une synthèse de la théorie des valeurs extrêmes univariées. Pour plus de détails et d'éventuelles démonstrations, on pourra se référer à Embrechts *et al.* (1997). Soit $(X_n)_{n\geq 1}$ une suite de copies indépendantes d'une variable aléatoire (v.a) X et de fonction de distribution $F(x) = P(X \leq x)$. Rangeons ces variables aléatoires par ordre croissant, on notera dans la suite l'échantillon ordonné :

$$X_{1:n} \leq ... \leq X_{n:n}$$

Une manière simple d'étudier le comportement des évènements extrêmes est de considérer les variables aléatoires :

$$M_n = \max\{X_1, ..., X_n\} \text{ et } W_n = \min\{X_1, ..., X_n\}.$$

Il faut noter que M_n et W_n représentent, respectivement, la plus grande et la plus petite valeur observées sur les n valeurs observées $X_1, ..., X_n$. Comme les variables aléatoires sont iid, on obtient pour toute réalisation x

$$\begin{aligned} P(M_n \leq x) &= P[\cap_{i=1}^{n}(X_i \leq x)] \\ &= \prod_{i=1}^{n} P(X_i \leq x) \\ &= \prod_{i=1}^{n} F(x) \\ &= [F(x)]^n. \end{aligned}$$

La difficulté provient du fait que l'on ne connaît pas, en général, la fonction de répartition F. C'est la raison pour laquelle on s'intéresse au comportement asymptotique de la $v.a$ M_n. Ainsi, en exhibant la famille de loi vers laquelle M_n convenablement normalisée va converger, on pourra remplacer F par cette dernière pour les grandes valeurs de n. De plus, les valeurs extrêmes se trouvent à droite et à la fin du support de la distribution et intuitivement le comportement asymptotique du maximum M_n caractérise la fin de la distribution Deme (2013). On notera par $x_F = \sup\{x \in \mathbb{R}, F(x) < 1\}$, le point terminal de F, c'est-à-dire la borne supérieure du support de F. Ce point terminal peut être fini ou infini, pour plus de détails voir Embrechts *et al.* (1997).

Définition 1.2.1 *Soient F_1 et F_2 deux fonctions de répartition. On dit que F_1 et F_2 sont du même type si et seulement si il existe $a \in \mathbb{R}_+^*$ et $b \in \mathbb{R}$ tels que $F_1(ax+b) = F_2(x)$.*

Remarque 1.2.2 *Deux fonctions de répartition de même type sont donc égales modulo un paramètre d'échelle et de position.*

Théorème 1.2.3 *Fisher et Tippet (1928); Gnedenko (1943). Soit $(X_n)_{n\geq 1}$ une suite de n variables aléatoires iid et de même loi de probabilité F telle que $F(x) = P(X \leq x)$. S'il existe deux suites normalisantes réelles $(a_n > 0, b_n \in \mathbb{R}, n \geq 1)$ et une loi non dégénérée G telle que :*

$$\begin{aligned}\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) &= \lim_{n\to+\infty} F^n(a_n x + b_n) \\ &= G(x), \forall x \in \mathbb{R},\end{aligned}$$

alors G est du même type qu'une des trois types de loi suivantes :
Loi de Gumbel $(\xi = 0)$:

$$\Lambda_{\mu,\sigma}(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right), x \in \mathbb{R} \tag{1.2.1}$$

Loi de Fréchet $(\xi > 0)$:

$$\Phi_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^{\frac{-1}{\xi}}\right) & \text{si } x > \mu \\ 0 & \text{si } x \leq \mu \end{cases},$$

Loi de Weibull $(\xi < 0)$:

$$\Psi_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\left[-\left(-\frac{x-\mu}{\sigma}\right)^{\frac{1}{\xi}}\right] & \text{si } x < \mu \\ 1 & \text{si } x \geq \mu \end{cases}.$$

Remarque 1.2.4 *La loi de Weibull définie précédemment n'est pas la loi de Weibull standard qui est plutôt définie sur* $\mathbb{R}_+$.

Les trois lois de probabilité ci-dessus sont appelées lois des valeurs extrêmes. Pour la preuve, nous renvoyons le lecteur à Resnick (1987) et avec des développements dans Embrechts *et al.* (1997), p.152. Il faut aussi signaler que chacune des trois lois des valeurs extrêmes peut s'obtenir par une transformation fonctionnelle de l'autre. D'une manière analogue, on définit les lois des valeurs extrêmes associées au minimum.
Le théorème suivant établit l'unification des trois types de loi en une loi unique dite distribution généralisée des valeurs extrêmes pour le maximum.

Théorème 1.2.5 *Resnick (1987) Soit* $(X_n)_{n\geq 1}$ *une suite de variables aléatoires iid de fonction de répartition F. S'il existe deux suites normalisantes réelles* $(a_n)_{n\geq 1} > 0$ *et* $(b_n)_{n\geq 1} \in \mathbb{R}$ *et une loi* non *dégénérée* G_ξ telle que :

$$\begin{aligned}\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) &= \lim_{n\to\infty} F^n\left(a_n x + b_n\right) \\ &= G_\xi\left(x\right), \forall x \in \mathbb{R}\end{aligned}$$

alors G est donnée par :

$$G_{\mu,\sigma,\xi,}\left(x\right) = \exp\left(-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{\frac{-1}{\xi}}\right), \tag{1.2.2}$$

où x est tel que $1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0$, $-\infty < \mu < +\infty$, $\sigma > 0$ *et* $-\infty < \xi < +\infty$.
Les paramètres μ, σ *et* ξ *sont respectivement les paramètres de position, d'échelle et de forme.*

Une preuve détaillée de ce théorème est donnée dans l'ouvrage de Resnick (1987) et pour plus de détails voir Embrechts *et al.* (1997) et Galambos (1985).
Le comportement limite du maximum normalisé est ainsi décrit par la fonction de répartition G_ξ pour la plus grande partie des lois usuelles. Ainsi, G_ξ est appelée fonction de répartition de la loi des valeurs extrêmes, en anglais « Generalized Extreme Value distribution » notée GEV. Pour plus de détails voir Resnick (1987); Embrechts *et al.* (1997).
L'unification de la loi standard de la distribution généralisée des valeurs extrêmes en une seule fonction de répartition facilite l'étude du comportement du maximum ou du minimum. De plus, pour $\mu = 0$ et $\sigma = 1$, on obtient la forme standard des trois types de loi des valeurs extrêmes. Cette loi dépend du seul paramètre de forme ξ appelé indice des valeurs extrêmes, ainsi la relation (1.2.2) devient

$$G_{\xi}(x) = \exp\left(-(1+\xi x)^{\frac{-1}{\xi}}\right), \ \xi \neq 0, \text{ avec } (1+\xi x) > 0. \tag{1.2.3}$$

Le cas $\xi = 0$, dans la relation (1.2.3) peut être vu comme le cas limite lorsque $\xi \to 0$. On retrouve alors la loi de Gumbel ayant pour fonction de répartition :

$$G_0(x) = \exp(-\exp(-x)).$$

La figure (1.2.2) représente les fonctions de répartition et de densité pour la forme standard unifiée des lois des valeurs extrêmes.

1. Pour $\xi = 1$, loi de Fréchet, en bleu, où le support de la loi est $[-1, +\infty[$,
2. pour $\xi = -1$, loi de Weibull en rouge où le support de la loi est $]-\infty, 1]$,
3. pour $\xi = 0$, loi de Gumbel en noir où le support de la loi est $\mathbb{R}$.

Souvent on indique les limites des supports des lois sur les graphiques.

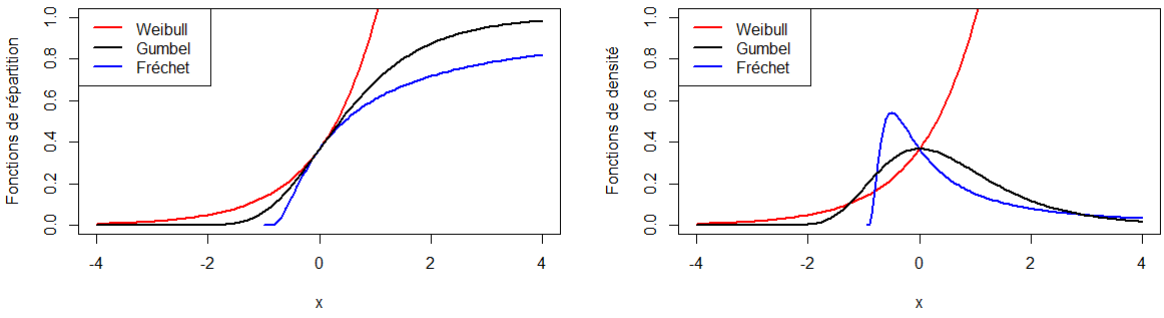


FIGURE 1.2.2 – À gauche fonctions de répartition, à droite fonctions de densité des lois des valeurs extrêmes

En outre, dans la TVE les informations les plus importantes se trouvent dans la queue de la distribution et sont caractérisées par le paramètre de forme ξ. En effet selon son signe, on distingue trois domaines d'attraction que l'on présentera dans la suite.

1.2.2 Domaines d'attraction des lois GEV

Il faut noter que le paramètre de forme ξ conditionne le type de la loi des valeurs extrêmes. Nous présentons dans ce qui suit les domaines d'attraction et les concepts théoriques qui leur sont associés dans les trois cas correspondant au signe du paramètre de forme ξ.

Définition 1.2.6 *On dit qu'une distribution F appartient au domaine d'attraction de H_ξ, et on note $F \in D(H_\xi)$ s'il existe des suites réelles $(a_n) > 0$ et $b_n \in \mathbb{R}$ telles que :*

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

1.2.2.1 Notion de fonction à variation régulière

La notion de fonction à variation régulière est très utilisée dans le contexte de la caractérisation des domaines d'attraction dans la théorie des valeurs extrêmes. Nous présentons ici quelques résultats principaux, pour plus de détails voir Bingham *et al.* (1987).

Définition 1.2.7 *Une fonction $f : \mathbb{R}_+ \to \mathbb{R}_+$ mesurable au sens de Lebesgue, est à variations régulières à l'infini si et seulement s'il existe un réel α tel que pour tout $x > 0$*

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\alpha$$

et on écrit $f \in \Re\nu_\alpha$, α est appelé indice (ou exposant) de la fonction à variation régulière f.

Remarque 1.2.8 *Pour $\alpha = 0$, on retrouve le cas de la fonction à variation lente dans la définition suivante.*

Définition 1.2.9 *Une fonction L est dite à variation lente si $L(t) > 0$ pour t assez grand et si pour tout $x > 0$, on a :*

$$\lim_{t\to\infty} \frac{L(tx)}{L(t)} = 1$$

Théorème 1.2.10 *Resnick (1987)* Représentation de Karamata
Toute fonction à variation lente L à l'infini s'écrit sous la forme

$$L(x) = c(x) \exp\left(\int_1^x \triangle(t)\, t^{-1} dt\right), \tag{1.2.4}$$

où $c(.) > 0$ et $\triangle(.)$ sont deux fonctions mesurables telles que

$$\lim_{x\to+\infty} c(x) = c_0 \in]0; +\infty[\text{ et } \lim_{x\to+\infty} \triangle(x) = 0.$$

Si la fonction $c(.)$ est une constante, alors on dit que L est normalisée. La relation (1.2.4) implique que si L est *normalisée alors L est dérivable, de dérivée L' avec pour tout $x > 0$, $L'(x) = \dfrac{\triangle(x) L(x)}{x}$. En particulier, on a $\lim\limits_{x\to\infty} \dfrac{xL'(x)}{L(x)} = 0$.*

1.2.2.2 Domaine d'attraction de la loi de Gumbel

La loi présente dans la queue une décroissance de type exponentiel, ce qui permet de caractériser dans ce cas, le domaine d'attraction de Gumbel noté $D(\Lambda)$. La caractérisation des fonctions de répartition du domaine d'attraction de Gumbel est plus complexe car il n'y a pas de lien direct entre la queue de la loi et les fonctions à variation lente, pour plus de détails voir Delmas et Jourdain (2006).
VonMises (1936) a donné une caractérisation simple pour le domaine d'attraction de Gumbel, formulée par le biais du théorème suivant :

Théorème 1.2.11 *VonMises (1936) S'il existe une fonction mesurable R, appelée fonction auxiliaire telle que :*

$$\lim_{x \to x_F} \frac{1 - F(t + xR(t))}{1 - F(t)} = exp(-x)$$

où $x_F = \sup\{x \in \mathbb{R},, F(x) < 1\}$, est le point terminal de F, alors $F \in D(\Lambda)$.

1.2.2.3 Domaine d'attraction de la loi de Weibull

Les lois de ce domaine sont bornées à droite et, par conséquent, le point terminal x_F est fini ; le domaine d'attraction de Weibull est noté $D(\Psi_\xi)$. Une caractérisation d'appartenance à ce domaine d'attraction est donnée par le théorème suivant, pour la preuve voir Gnedenko (1943) :

Théorème 1.2.12 *Gnedenko (1943) Une fonction de répartition F appartient au $D(\Psi_\xi)$ si et seulement si $x_F < +\infty$ et*

$$\overline{F}\left(x_F - \frac{1}{x}\right) = x^{\frac{-1}{\xi}} L(x)$$

avec : $\overline{F}$ est la fonction de survie donnée par $\overline{F}(x) = 1 - F(x)$, L est une fonction à variation lente.

1.2.2.4 Domaine d'attraction de la loi de Fréchet

Les lois appartenant à ce domaine d'attraction sont caractérisées par une queue à décroissance lente (polynomiale) à l'infini, et un point terminal $x_F = +\infty$. Elles sont dites aussi lois à queues lourdes. Une caractérisation de ce domaine d'attraction noté $D(\Phi_\xi)$ est donnée par le théorème suivant, pour la preuve voir Gnedenko (1943) :

Théorème 1.2.13 *VonMises (1936) Une fonction de répartition F appartient au $D(\Phi_\xi)$ si et seulement si sa fonction de survie est donnée par :*

$$\overline{F}(x) = x^{\frac{-1}{\xi}} L(x)$$

où L est une fonction à variation lente.

Remarque 1.2.14 *Les variables aléatoires dans les trois domaines d'attraction de Gumbel, de Fréchet et de Weibull sont liées par la relation suivante, pour plus de détails voir Embrechts* et al. *(1997) :*

$$X \in D(\Phi_\xi) \Longleftrightarrow \log\left(X^{\frac{1}{\xi}}\right) \in D(\Lambda) \Longleftrightarrow -X^{-1} \in D(\Psi_\xi)$$

Le tableau suivant présente quelques lois qui appartiennent aux domaines d'attraction de Gumbel, Fréchet et *Weibull.*

Domaines d'attraction	Gumbel $\xi = 0$	Fréchet $\xi > 0$	Weibull $\xi < 0$
Lois	Normale	Cauchy	Uniforme
	Exponentielle	Pareto Généralisée	reverse Burr
	Log normale	Student	Beta
	Weibull	Log-gamma	
	Gumbel		

TABLE 1.2.1 – Domaines d'attraction et quelques lois associées

Exemple 1.2.15 *Considérons une variable aléatoire X suivant une loi exponentielle de paramètre* 1. *Sa fonction de répartition est donnée par :*

$$F(x) = \begin{cases} 1 - \exp(-x) & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}.$$

Le support de la loi étant $\mathbb{R}_+$ on a $x_F = +\infty$ d'où $X_{n:n} \to +\infty$. Effectuons la normalisation suivante :

$$\begin{aligned} P\left(\frac{X_{n:n} - \log(n)}{1} \leq x\right) &= P(X_{n:n} \leq x + \log(n)) \\ &= (F(x + \log(n)))^n \\ &= (1 - \exp(-x - \log(n)))^n \\ &= \left(1 - \frac{\exp(-x)}{n}\right)^n \\ &\to \exp(-\exp(-x)),\ n \to +\infty. \end{aligned}$$

Il en résulte donc, d'après la relation (1.2.1), que $X_{n:n}$ converge vers une loi de Gumbel après normalisation.

1.2.3 Méthodes d'estimation des paramètres des lois GEV

Les lois GEV se distinguent fondamentalement selon le paramètre ξ. De ce fait, on distinguera trois cas dans l'estimation des paramètres. Le premier cas correspond à la loi de Gumbel pour $\xi = 0$, le deuxième correspond à celle de Weibull pour $\xi < 0$ et le troisième correspond à celle de Fréchet pour $\xi > 0$.
Dans la littérature, il existe plusieurs méthodes d'estimation des paramètres des lois GEV. On peut citer par exemple, les méthodes d'estimation empirique Gumbel et Mustafi (1967), la méthode du maximum de vraisemblance Smith (1987), Prescott et Walden (1980), la méthode des moments Christopeit (1994), la méthode des moments de probabilité pondérés Greenwood *et al.* (1979). La comparaison de ces méthodes d'estimation s'effectue généralement dans un cadre empirique, pour plus de détails voir Benkhaled (2007).
Par ailleurs, d'autres approches principalement non paramétriques, ont été proposées pour l'estimation de l'indice de queue. A titre d'exemple, on peut citer l'estimateur de Pickands, Pickands (1975), l'estimateur de Hill, Hill (1975) pour le modèle de Fréchet et l'estimateur de Dekkers, Dekkers *et al.* (1989).

1.2.3.1 Méthodes d'estimation pour les lois du domaine d'attraction de Gumbel

a) Méthode du maximum de vraisemblance

Cette méthode consiste à chercher les paramètres $\varpi = (\mu, \sigma)$ qui maximisent la fonction de vraisemblance d'un échantillon de taille n est

$$l(x_1, ..., x_n, \mu, \sigma) = \prod_{i=1}^{n} f\left(x_1, ..., x_n, \mu, \sigma\right).$$

Par suite, sa log-vraisemblance est :

$$L_{MV}\left(\varpi\right) = \log l\left(\varpi\right)$$

en résolvant le système suivant représentant les conditions nécessaires d'optimalité :

$$\begin{cases} \dfrac{\partial L_{MV}\left(\varpi\right)}{\partial \mu} & = 0 \\ \dfrac{\partial L_{MV}\left(\varpi\right)}{\partial \sigma} & = 0 \end{cases},$$

on obtient une estimation des paramètres $\varpi = (\mu, \sigma)$. Cependant, la solution n'est pas explicite et il faut passer par une résolution numérique en utilisant par exemple la méthode de *Newton-Raphson.*

b) Méthode des moments pour la loi Gumbel

La méthode consiste à associer respectivement les deux premiers moments m_1 et m_2, respectivement à la moyenne $(\overline{x_n})$ et à la variance empirique S_X^2.

Dans le cas d'une variable aléatoire issue d'une loi de Gumbel, la méthode des moments nous conduit au système suivant :

$$\begin{cases} E(X) & = \mu + \gamma\sigma \\ V(X) & = \dfrac{1}{6}\pi^2\sigma \end{cases},$$

où $\gamma \approx 0.57721$ représente la constante d'Euler. La résolution de ce système d'équations nous permet d'obtenir les valeurs estimées $\widehat{\mu}$ et $\widehat{\sigma}$:

$$\begin{cases} \widehat{\sigma} & = \dfrac{\sqrt{6}S_X}{\pi} \\ \widehat{\mu} & = \overline{x_n} - \gamma\widehat{\sigma} \end{cases}$$

c) Méthode des moments pour la loi Weibull à deux paramètres

La méthode consiste à associer respectivement les deux premiers moments théoriques m_1 et m_2 aux deux premiers moments observés à savoir la moyenne $(\overline{x_n})$ et la variance empirique σ^2.
La méthode des moments est une autre technique utilisée dans l'estimation des paramètres. Nous présentons ici la procédure pour la loi de Weibull à deux paramètres (k, λ), définie par :

$$\begin{cases} f_{k,\lambda}(y) & = \dfrac{k}{\lambda}\left(\dfrac{y}{\lambda}\right)^{k-1} \exp^{-\left(\frac{y}{\lambda}\right)^k} \\ F_{k,\lambda}(y) & = 1 - \exp^{-\left(\frac{y}{\lambda}\right)^k} \end{cases}, \tag{1.2.5}$$

où f et F sont respectivement les fonctions de densité et de répartition.
Soit $y_1, ..., y_n$ un ensemble de données pour lesquelles nous allons chercher les deux premiers moments empiriques à partir de la relation (1.2.6) définie comme suit :

$$\widehat{M_m} = \frac{1}{m}\sum_{i=1}^{m} y_i^m, \tag{1.2.6}$$

où $\widehat{M_m}$ est une estimation de M_m. Dans la distribution de Weibull, le moment d'ordre m peut être obtenu à travers la fonction Gamma $\Gamma(y)$. De plus, la moyenne des observations peut aussi être exprimée en fonction des paramètres k et λ. Pour la distribution de Weibull donnée dans la relation (1.2.5), le moment d'ordre m, est défini par :

$$\mu_m = \left(\frac{1}{\lambda}\right)^{\frac{m}{k}} \Gamma\left(1 + \frac{m}{k}\right), \tag{1.2.7}$$

où la fonction Γ est définie par $\Gamma(s) = \int_0^{+\infty} q^{s-1}e^{-q}dq$, $(s > 0)$.
A partir de la relation (1.2.7), nous pouvons trouver les moments d'ordre 1 et 2 comme suit :

$$\mu_1 = \left(\frac{1}{\lambda}\right)^{\frac{1}{k}} \Gamma\left(1 + \frac{1}{k}\right) = M_1, \tag{1.2.8}$$

$$\mu_2^2 = \mu_1^2 + \sigma^2 = \left(\frac{1}{\lambda}\right)^{\frac{2}{k}} \left[\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right] = M_2, \tag{1.2.9}$$

où σ^2 est la variance, M_1 et M_2 sont respectivement la moyenne arithmétique et le moment centré d'ordre 2. Lorsqu'on divise M_2 par le carré de M_1, on obtient l'expression suivante :

$$\frac{M_2}{M_1^2} = \frac{\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)}{\Gamma^2\left(1+\frac{1}{k}\right)}. \tag{1.2.10}$$

En prenant la racine carrée de la relation (1.2.10), on obtient le coefficient de variation noté C_v, défini comme suit :

$$C_v = \frac{\sqrt{\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)}}{\Gamma\left(1+\frac{1}{k}\right)}.$$

Ainsi, les valeurs de k et λ peuvent être obtenues par le système d'équations suivant :

$$\begin{cases} \sigma &= \lambda\left[\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right]^{\frac{1}{2}} \\ C_v &= \sqrt{\dfrac{\Gamma\left(1+\frac{2}{k}\right)}{\Gamma^2\left(1+\frac{1}{k}\right)} - 1}. \end{cases} . \tag{1.2.11}$$

Après certaines transformations, voir Justus *et al.* (1977); Abul *et al.* (2014), nous avons :

$$k = \left(\frac{0.9874}{C_v}\right)^{1.0983}, \lambda = \frac{\overline{y}}{\Gamma\left(1+\frac{1}{k}\right)}. \tag{1.2.12}$$

1.2.3.2 Méthodes d'estimation pour les lois des domaines d'attraction de Fréchet et Weibull

Il existe plusieurs méthodes d'estimation des paramètres pour les lois de Fréchet et de Weibull, telles la méthode du maximum de vraisemblance, la méthode des moments, la méthode des moments de probabilité pondérés, etc. Mais dans cette partie, nous présentons celle du maximum de vraisemblance commune aux deux domaines d'attraction.

Méthode du maximum de vraisemblance

La méthode consiste à chercher les paramètres $\theta = (\mu, \sigma, \xi)$ qui maximisent la fonction de log-vraisemblance. La maximisation conduit à résoudre le système suivant :

$$\begin{cases} \dfrac{\partial L_{MV}(\theta)}{\partial \mu} & = 0 \\ \dfrac{\partial L_{MV}(\theta)}{\partial \sigma} & = 0 \\ \dfrac{\partial L_{MV}(\theta)}{\partial \xi} & = 0 \end{cases}.$$

Les problèmes relatifs à l'estimation par la méthode du maximum de vraisemblance ont été étudiés par Smith (1985). Les résultats obtenus sont :

1. Si $\xi > -0.5$, alors les estimations du maximum de vraisemblance ne possèdent pas de propriété asymptotique telles que la convergence vers la vraie valeur du paramètre inconnu, l'invariance par rapport à une transformation paramétrique et l'efficacité asymptotique.
2. Si $-1 < \xi < -0.5$, alors les estimateurs du maximum de vraisemblance ne possèdent pas de propriétés asymptotiques standards.
3. Si $\xi < -1$, alors l'obtention des estimateurs du maximum de vraisemblance n'est pas garantie.

1.3 Approche des dépassements de seuil : loi GPD

L'approche par dépassements de seuil, en anglais « Peaks-Over Threshold » 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.
Considérons un échantillon de variables aléatoires i.i.d $Y_1, ..., Y_n$. Soit u un seuil fixé tel que $u < y_F$ et 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 définis par $Z_j = Y_{i_j} - u$, pour $j = 1, ..., N_u$ voir la figure (1.3.1).
La figure (1.3.1) est composée de deux graphiques sur celui de gauche on peut remarquer le seuil u en trait horizontal et les observations au-dessus du seuil, par exemple Y_2 et Y_{10}. Quant à celui de droite, il indique le seuil u et les excès au-dessus de ce seuil.

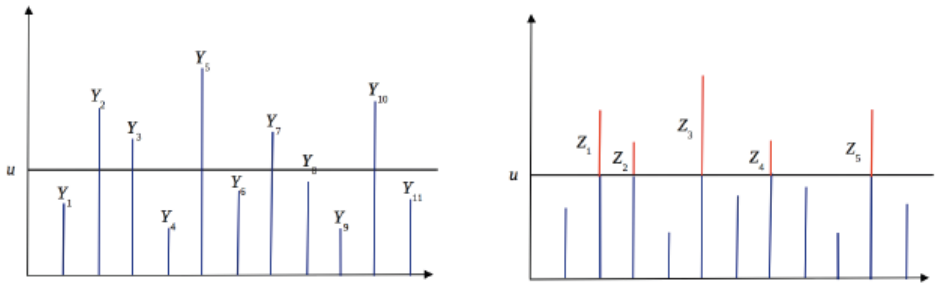


FIGURE 1.3.1 – Excès au-dessus du seuil

1.3.1 Modélisation des excès

L'approche des blocs de maxima conduit souvent à des pertes d'information lors de la sélection des maxima ou des minima par blocs. La prise en compte de cette insuffisance a conduit à une nouvelle approche dite des dépassements de seuil Peaks Over Threshold (POT). De plus, certains blocs peuvent contenir plusieurs valeurs extrêmes provenant de la distribution initiale, alors que d'autres peuvent ne pas en contenir, pour plus de details voir Katz (2002). Contrairement à l'approche des maxima par blocs, la méthode POT consiste à utiliser toutes les observations, appelées excès, qui dépassent un certain seuil suffisamment élevé. L'objectif est d'analyser leur comportement asymptotique.
La méthode POT a été proposée par Pickands (1975) et reprise par Haan et Rootzén (1993). Les premières applications étaient dans le domaine de l'hydrologie. Concernant les aspects théoriques de la méthode, ils ont été abondamment développés par plusieurs auteurs. On peut citer Todorovic et Zelenhasic (1970); Todorovic et Rousselle (1971); Smith (1987); Davison et Smith (1990); Reiss et Thomas (2001).
L'idée d'utiliser un nombre croissant de statistiques d'ordre de l'échantillon a ensuite été plus largement développée dans le cadre de l'approche POT, via l'approximation de la loi des excès au-delà d'un seuil par des Generalized Pareto Distribution (GPD). Plus précisément, soit $u < y_F$ et F_u la fonction de répartition des excès définie par

$$F_u(y) = Pr(Z \leq y \mid Y > u) = Pr(Y - u \leq y \mid Y > u) = \frac{F(u+y) - F(u)}{\overline{F}(u)},$$

où $\overline{F} = 1 - F$ est la fonction de survie. On a alors le résultat suivant dû à Balkema et Haan (1974) et Pickands (1975).

Théorème 1.3.1 *Balkema et Haan (1974); Pickands (1975) Si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes, (Fréchet, Gumbel ou Weibull), alors il existe une fonction σ_u strictement positive et un $\xi \in \mathbb{R}$ tel que*

$$\lim_{u \to y_F} \sup_{y \in]0, y_F - u[} \mid F_u(y) - G_{\xi,\sigma_u}(y) \mid = 0;$$

où $y_F = \sup\{y \in \mathbb{R}, F(y) < 1\}$ *est le point terminal de* F *et* G_{ξ,σ_u} *est la fonction de répartition de la loi de Pareto Généralisée définie par :*

$$G_{\xi,\sigma_u}(y) = \begin{cases} 1 - \left(1 + \dfrac{\xi}{\sigma_u} y\right)^{\frac{-1}{\xi}}, & \text{si } \xi \neq 0 \\ 1 - \exp\left(-\dfrac{y}{\sigma_u}\right), & \text{si } \xi = 0 \end{cases}. \tag{1.3.1}$$

Remarque 1.3.2 *Selon le signe de* ξ*, nous avons les cas suivants :*
$\xi > 0$ *: distribution de type Pareto à queue lourde,*
$\xi < 0$ *: distribution de type Beta bornée au dessus de* $u - \dfrac{\sigma_u}{\xi}$,
$\xi = 0$ *: distribution de type exponentielle à queue légère.*

La figure (1.3.2) représente les fonctions de répartition et de densité des lois de Pareto pour un paramètre d'échelle fixé à 1. Ainsi, en faisant varier le paramètre de forme entre -1 et 1, on obtient différentes lois GPD comme suit, $\xi = -1$ pour la loi de Pareto II en rouge, $\xi = 0$ pour la loi exponentielle en noir et $\xi = 1$ pour la loi de Pareto en bleu. En remplaçant par exemple dans la relation (1.3.1) $\xi = 0$ et $\sigma_u = 1$; on obtient respectivement les fonctions de répartition et de densité suivantes :

$$G_{0,1}(y) = 1 - \exp(-y) \text{ et } g_{0,1}(y) = \exp(-y).$$

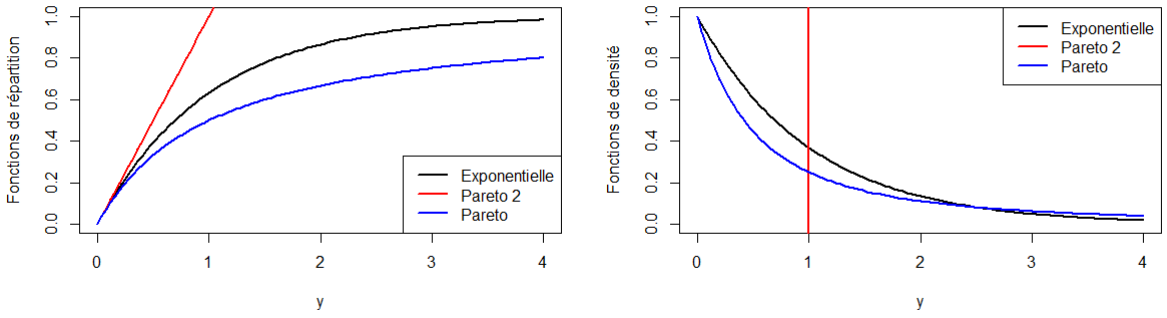


FIGURE 1.3.2 – À gauche : fonctions de répartition et à droite : fonctions de densité des lois de Pareto Généralisée

1.3.2 Méthodes pour la détermination du seuil

La détermination du seuil est l'étape la plus délicate dans l'implémentation de l'approche POT, étant donné que la qualité du modèle en dépend. La convergence des

excès vers une GPD passe par la détermination d'un seuil adéquat (pas très bas et pas très haut). Par la suite, il faut estimer les paramètres et les niveaux de retour. Le choix du seuil doit être un compromis de sorte que le seuil déterminé soit suffisamment grand pour pouvoir utiliser les résultats asymptotiques, mais pas trop élevé afin d'obtenir des estimations précises. Cependant, le choix d'un seuil faible peut conduire à des incertitudes sur le nombre d'observations extrêmes et par conséquent produire des estimations biaisées et une mauvaise approximation de la loi asymptotique. Dans cette optique, plusieurs méthodes de détermination de seuils sont proposées dans la littérature. On distingue principalement deux approches, l'approche graphique et celle dite numérique. La plupart de ces méthodes sont subjectives et il est nécessaire de quantifier les incertitudes dues à ces méthodes.

1.3.2.1 Méthodes graphiques

a- Fonction moyenne des excès

La fonction moyenne des excès, obtenue par une méthode graphique, peut être illustrée de deux façons grâce au logiciel R :

1. par la fonction « mean excess function »
2. par la fonction « mean residual life plot ».

Elle est définie comme suit :

Définition 1.3.3 *On appelle fonction moyenne des excès, mean excess function ou Mean residual life plot au-dessus d'un seuil u, la fonction $e(u)$définie par :*

$$e(u) = E[Y - u \mid Y > u]$$

avec $0 \leq u \leq y_F$.

Dans le logiciel R, elle est donnée par la fonction « Meplot ou Mrlplot » ; ces deux fonctions jouent le même rôle mais présentent la linéarité sous deux formes différentes. La moyenne des excès peut être utilisée pour guider le choix du seuil adéquat u^*.

Proposition 1.3.4 *Si* $[Z = Y - u^* \mid Y > u^*] \backsim G_{\xi,\sigma_{u^*}}$, $e(u)$ *est linéaire en* u *pour* $u > u^*$, *Coles (2001)*

Démonstration Si $Z \backsim G_{\xi,\sigma_{u^*}}$ alors $E(Z) = \dfrac{\sigma_{u^*}}{1-\xi} = C$.

Donc pour $u > u^*$, $E[Y - u \mid Y > u] = \dfrac{\sigma_{u^*} + \xi u}{1-\xi} = C + \dfrac{\xi}{1-\xi}u$, où C=$E(Z)$.
Ainsi, si au-dessus d'un seuil u^*, la GPD est une approximation de la loi des excès $Z = Y - u^*$ avec $\xi < 1$, alors on a :

$$E[Z - u^* \mid Z > u^*] \approx \frac{\sigma_{u^*}}{1-\xi}.$$

Comme l'approximation par une GPD reste valable pour tout seuil $u > u^*$, alors on a :

$$E\left[Y - u \mid Y > u\right] \approx \frac{\sigma_u}{1-\xi},$$

avec $\sigma_u = \sigma_{u^*} + \xi\left(u - u^*\right)$. Ainsi, pour tout $u > u^*$:

$$E\left[Y - u \mid Y > u\right] \approx \frac{\sigma_{u^*} + \xi\left(u - u^*\right)}{1-\xi}.$$

La fonction $e\left(.\right)$ est une fonction linéaire de u. Par conséquent, l'identification du seuil u^* consiste en la recherche de la linéarité sur le graphe défini par $(u,\, e\left(u\right))$ et appelé « Mean Residual Life Plot ou Mean excess plot ». En pratique, la fonction $e\left(.\right)$ est approchée sur la base de l'approximation empirique $\widehat{e_n\left(.\right)}$:

$$\widehat{e_n\left(u\right)} = \frac{\sum_{i=1}^{n}\left(Y_i - u\right)^+}{\sum_{i=1}^{n}\left(I_{Y_i > u}\right)} = \frac{1}{N_u}\sum_{i=1}^{n}\left(Y_i - u\right)^+, \tag{1.3.2}$$

c'est-à-dire la somme des excès au-dessus du seuil u divisée par le nombre N_u de données qui excèdent le seuil u.

b- Méthodes des graphes des paramètres d'échelle et de forme (Parameter stability plot), fonction tcplot

Encore appelée « stable scale and shape parameters », cette méthode permet de déterminer un seuil requis en ajustant les données à une distribution de GPD en utilisant un seuil différent. La stabilité des paramètres (forme et échelle) peut alors être contrôlée et localisée.
Cette technique est implémentée dans le logiciel R avec des packages spécifiques. Ces paquets disposent d'outils objectifs pour guider le choix du seuil adéquat en examinant tout simplement la stabilité des paramètres de forme et d'échelle ξ et σ. Ce graphe établit un lien direct entre les valeurs des paramètres estimés (ξ et σ) et les seuils potentiels u^*. Les paramètres estimés au-dessus des seuils sont ceux pour lesquels le modèle GPD devient valable.

Exemple 1.3.5 *Les figures (1.3.3) et (1.3.4) illustrent le « MRLplot » et le « parameter stability plot » pour les données de pluies anglaises, Coles (2001) a montré que le meilleur choix du seuil est approximativement* $u = 30mm$ *et qu'un seuil de* $20mm$ *peut être trop bas pour la validité de l'hypothèse des valeurs extrêmes.*

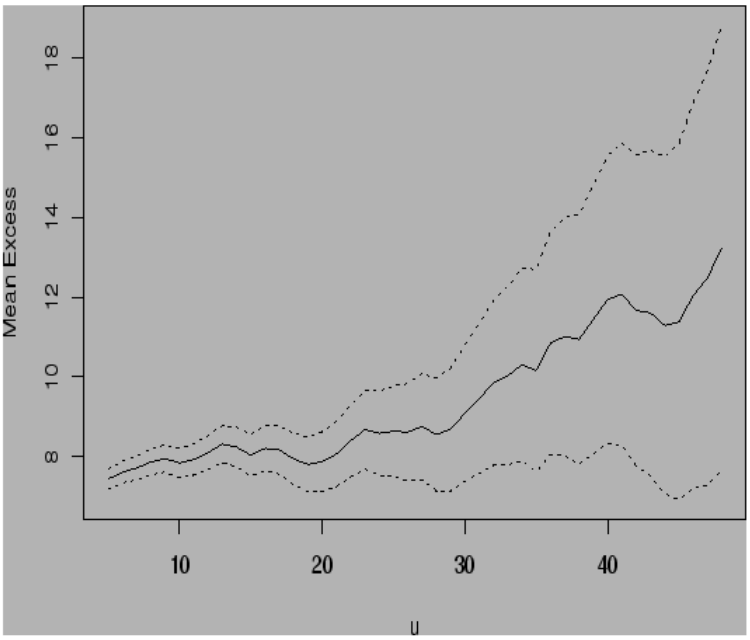


FIGURE 1.3.3 – Graphe de la fonction moyenne des excès (MRlplot) pour des données de pluie anglaises, (Coles, 2001)

Sur la figure (1.3.3), les pointillés encadrant la courbe de la fonction moyenne des excès indiquent les intervalles de confiance associés au seuil.

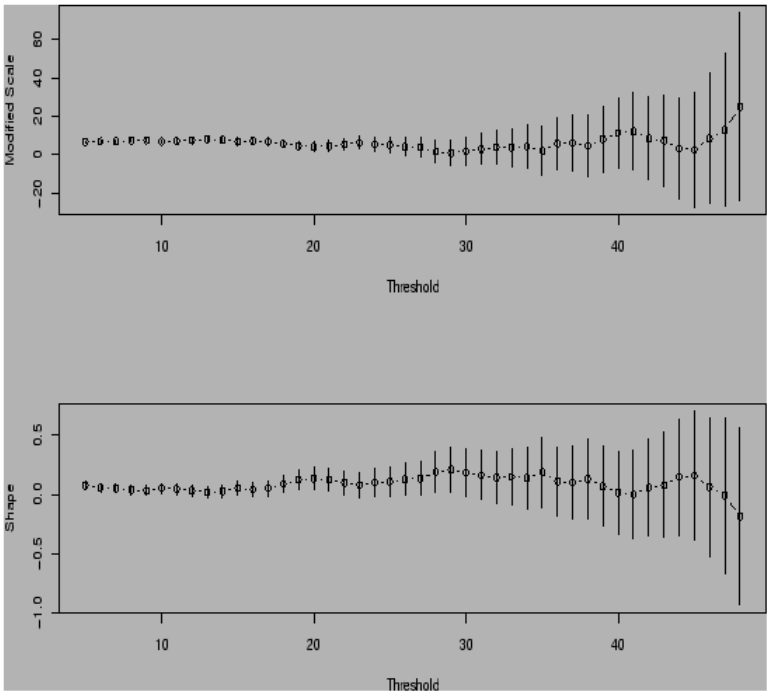


FIGURE 1.3.4 – Graphe de la stabilité des paramètres (tcplot)

Sur la figure (1.3.3), on constate une linéarité et une stabilité entre 23 et 32 mm. De ce fait, on peut affirmer que le seuil est compris entre 23 et 32 mm. En outre, sur la figure (1.3.4), on remarque une stabilité sur l'intervalle $[25, 32]$; ainsi, le seuil se situer autour de 30 mm.

1.3.2.2 Une approche de compromis

Dans cette approche de compromis, nous combinons les méthodes graphiques et une estimation intensive des paramètres de forme et échelle par différentes méthodes, ainsi que les erreurs standards sur les paramètres estimés pour les seuils potentiels obtenus. Les erreurs standards sur les paramètres estimés pour différents seuils permettent de guider le choix du seuil et de la méthode d'estimation qui sont les principales sources d'incertitudes dans l'approche POT.

1.3.3 Estimation des paramètres du modèle GPD

Dans cette sous-section, nous nous intéressons aux différentes méthodes d'estimation du paramètre ξ ou σ qui interviennent dans la distribution asymptotique des valeurs extrêmes. On distingue principalement deux méthodes, à savoir les méthodes paramétriques et non-paramétriques. Parmi les méthodes paramétriques, on peut citer la méthode du maximum de vraisemblance, la méthode des moments, la méthode des moments de probabilités pondérées Hosking et Wallis (1987), les méthodes de régression Beirlant et Goegebeur (2003). Concernant les méthodes non paramétriques, on peut noter l'estimateur de Pickands (1975), l'estimateur de Hill (1975), l'estimateur des moments Dekkers *et al.* (1989). Dans le cadre de ce travail, on s'intéressera aux méthodes du maximum de vraisemblance, des moments et des moments de probabilités pondérées.

1.3.3.1 Méthode du maximum de vraisemblance

Considérons un échantillon $Y_1, ..., Y_{k_n}$ iid de loi GPD G_{ξ,σ_u}. La fonction de log-vraisemblance est obtenue à partir de la loi GPD, ce qui nous donne :

$$\log\left(L\left(\xi, \sigma_u\right)\right) = -k_n \log\left(\sigma_u\right) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k_n} \log\left(1 + \frac{\xi}{\sigma_u} Y_i\right), \tag{1.3.3}$$

lorsque $1 + \dfrac{\xi}{\sigma_u} Y_i > 0$ pour $i = 1, ..., k_n$ sinon $\left(L_{MV}\left(\xi, \sigma_u\right)\right) = -\infty$.
La log-vraisemblance définie par la relation (1.3.3) est maximisée par des méthodes numériques telles que l'algorithme de Newton-Raphson. Les estimateurs du maximum de vraisemblance sont asymptotiquement gaussiens et efficaces pour $\xi > -\dfrac{1}{2}$, pour plus de détails voir Smith (1987). De plus, Davison et Smith (1990) ont montré que l'estimateur du maximum de vraisemblance a souvent des problèmes de convergence et d'efficacité pour des échantillons de petite taille ($n < 500$).

1.3.3.2 Méthode des moments

La méthode des moments fut introduite par Hosking et Wallis (1987) pour estimer les paramètres de la loi GPD. La condition d'existence de l'espérance et de la variance d'une variable aléatoire Y de loi GPD $G_{\xi,\sigma}$ est $\xi < \frac{1}{2}$.
Pour une GPD, G_{ξ,σ_u} (voir équation (1.3.1)), l'estimation par la méthode des moments (voir Hosking et Wallis 1987) est basée sur l'hypothèse que

$$E\left[\left(1+\frac{\xi}{\sigma_u}Y\right)^r\right] = \frac{\sigma_u}{1-r\xi}, \text{ si } 1-r\xi > 0$$

Dans ce cas, on a :

$$\begin{cases} E(Y) = & \dfrac{\sigma_u}{1-\xi} \\ V(Y) = & \dfrac{\sigma_u^2}{(1-\xi)^2(1-2\xi)} \end{cases}.$$

On peut donc exprimer les paramètres de la GPD en fonction en fonction de ξ et σ en fonction de $E(Y)$ et $V(Y)$, on a donc :

$$\begin{cases} \xi & = \dfrac{1}{2}\left(1-\dfrac{E(Y)^2}{V(Y)}\right) \\ \sigma & = \dfrac{E(Y)}{2}\left(1+\dfrac{E(Y)^2}{V(Y)}\right) \end{cases};$$

où $\overline{Y}$ et S_Y^2 sont respectivement les estimateurs empiriques des moments d'ordre 1 et 2 de l'échantillon. Ainsi, en remplaçant $E(Y)$ et $V(Y)$ par leurs estimateurs empiriques respectifs, il vient que :

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

et on obtient les estimateurs de $\widehat{\xi}$ et $\widehat{\sigma_u}$ définis respectivement par

$$\begin{cases} \widehat{\xi} & = \dfrac{1}{2}\left(1-\dfrac{\overline{Y}^2}{S_Y^{*2}}\right) \\ \widehat{\sigma_u} & = \dfrac{1}{2}\overline{Y}\left(\dfrac{\overline{Y}^2}{S_Y^{*2}}+1\right) \end{cases}. \tag{1.3.4}$$

1.3.3.3 Méthode des moments de probabilité pondérés

A partir de la théorie des moments pondérés, Hosking *et al.* (2003) ont proposé une estimation. Cette approche consiste à utiliser les deux moments pondérés M_0 et M_1 avec $M_s = E\left[X\left(1-G_{\xi,\sigma}(Y)\right)^s\right]$. Nous avons

$$M_s = \frac{\sigma_u}{(s+1)(s-\xi+1)}, \text{ si } s > \xi - 1$$

Un estimateur de ce moment est donné par

$$\widehat{M_s} = \frac{1}{k_n}\sum_{i=1}^{k_n}\left(1-\frac{i}{k_n+1}\right)^s Y_{i,k_n},$$

où $Y_{1,k_n},...,Y_{k_n,k_n}$ sont les statistiques ordonnées associées à $Y_1,...,Y_{k_n}$.
Cette dernière équation conduit pour $s=0$ et $s=1$ aux estimateurs suivants :

$$\begin{cases} \widehat{\xi} & = 2-\dfrac{\widehat{M_0}}{\widehat{M_0}-2\widehat{M_1}} \\ \widehat{\sigma_u} & = 2\dfrac{\widehat{M_0}\widehat{M_1}}{\widehat{M_0}-2\widehat{M_1}} \end{cases}. \tag{1.3.5}$$

Les estimateurs des moments pondérés sont asymptotiquement gaussiens pour $-1 < \xi < \frac{1}{2}$, voir Hosking *et al.* (2003). Ils ont aussi étudié leurs performances avec des simulations. Une extension du domaine de validité à $-1 < \xi < \frac{3}{2}$ a été introduite par Diebolt *et al.* (2004).

1.3.4 Concepts de quantiles extrêmes

1.3.4.1 Quelques éléments théoriques sur les quantiles extrêmes

Soit F la fonction de répartition associée à une loi X.

Définition 1.3.6 *Le quantile d'ordre $1-\alpha$ de la fonction de répartition F est défini par :*

$$q(\alpha) = \overline{F}^{\leftarrow}(\alpha) = \inf\left\{y : \overline{F}(y) \leq \alpha\right\},\ \alpha \in]0,1[,$$

où $\overline{F}^{\leftarrow}$ est l'inverse généralisée de $\overline{F}$.
Rappelons que l'inverse généralisée d'une fonction coïncide avec l'inverse classique lorsque celle-ci existe.

Définition 1.3.7 *Dans le cas d'une variable aléatoire continue Y, $P(Y=y)=0$. Cependant il y a une probabilité $\overline{F}(y) = P(Y \geq y)$ que la variable aléatoire soit supérieure ou égale à y. Ainsi, on a alors la fonction appelée période de retour donnée par :*

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

Exemple 1.3.8 *Considérons un dé à six faces parfaitement équilibré. Cet exemple est un cas classique en théorie des probabilités d'une variable aléatoire discrète suivant une loi uniforme sur l'ensemble $\{1,2,3,4,5,6\}$. La probabilité d'obtenir l'une des six faces est $\frac{1}{6}$. Prenons par exemple la face relative au chiffre 4, on peut s'attendre à*

obtenir cette face au bout d'un certain nombre de lancers ou plus précisément on peut s'attendre à l'obtenir en moyenne tous les $T = 6$ lancers. T est appelée période de retour.

Définition 1.3.9 *Le niveau de retour y_N est le niveau que l'on peut s'attendre à atteindre ou dépasser, en moyenne une fois toutes les N années.*

Cette fonction représente le nombre d'observations tel que, en moyenne, il y ait une observation égale ou supérieure à y. Il en résulte que la période de retour augmente lorsque y augmente. On peut donc définir la fonction niveau de retour comme l'inverse de la période de retour :

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

Exemple 1.3.10 *Supposons que la figure (1.3.5) représente la distribution de survie d'une variable aléatoire continue Y définie sur $[0, 10]$. Sur la figure (1.3.5), on peut lire par exemple pour $\alpha = 0.1$, $q(0.1) = 8$. Dans le contexte du niveau d'un cours d'eau, si le nombre d'observations est annuel et le niveau du cours d'eau en mètres, alors on aura : $T(8) = \dfrac{1}{\overline{F}(8)} = 10\,ans$ et $q\left(\dfrac{1}{10}\right) = 8m$. On peut donc s'attendre à ce que le niveau du cours d'eau atteigne ou dépasse $8m$ dans dix ans.*

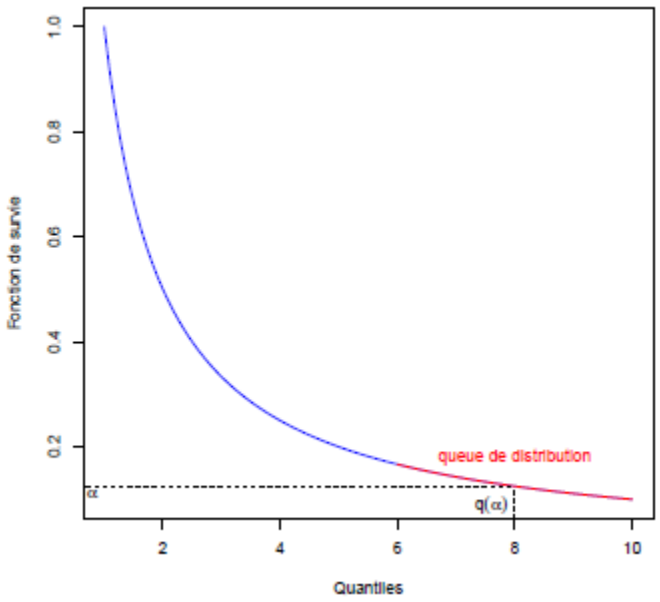


FIGURE 1.3.5 – Quantiles extrêmes et queue de distribution

La figure (1.3.5) montre la relation entre fonction de survie et quantile extrême. La queue de la distribution est donnée sur le graphique en rouge. On remarque donc

que pour une valeur de la fonction de survie inférieure ou égale à 0.1, le quantile (niveau de retour) d'ordre 0.1 est supérieur ou égal à 8.

1.3.4.2 Estimation de quantiles extrêmes pour la loi GPD

Nous rappelons que la loi des excès au delà d'un seuil u suffisamment élevé est donnée par la relation suivante :

$$F_u(y) = Pr(X - u \leq y \mid X > u) = \frac{F(u+y) - F(u)}{1 - F(u)} = \frac{F(u+y) - F(u)}{\overline{F}(u)},$$

où $\overline{F} = 1 - F$ est la fonction de survie.
De ce qui précède, on peut écrire :

$$F_u(y)\,\overline{F}(u) = F(u+y) - F(u) = -1 + F(u+y) + 1 - F(u),$$

par suite, on a

$$F_u(y)\,\overline{F}(u) = \overline{F}(u) - \overline{F}(u+y),$$

donc, on aura :

$$\overline{F}(u+y) = \overline{F}(u)\,\overline{F}_u(y). \tag{1.3.6}$$

Un résultat limite pour $F_u(y)$ fut proposé par Balkema et Haan (1974) et Pickands (1975). De ce fait, on a l'approximation suivante, pour u assez grand $F_u(y) \approx G_{\xi,\sigma}(y)$, $u \to \infty$, $y \geq 0$. Ce dernier résultat implique aussi l'expression suivante

$$\overline{F_u}(y) \approx \overline{G_{\xi,\sigma}}(y), \; u \to \infty, \;\; y \geq 0. \tag{1.3.7}$$

Un estimateur naturel pour $\overline{F}_u$ est la fonction de survie empirique

$$\widehat{\overline{F}}(u) = \widehat{\overline{F}}_n(u) = \frac{1}{n}\sum_{i=1}^{n} I_{\{Y_i > u\}} = \frac{N_u}{n} \tag{1.3.8}$$

où $I_{\{Y_i>u\}}$ est une fonction indicatrice ayant l'expression suivante

$$I_{\{Y_i>u\}} = \begin{cases} 1 & \text{si } Y_i > u, \; \forall i = 1, ..., n. \\ 0 & \text{sinon} \end{cases}$$

Par ailleurs, à partir de l'approximation (1.3.7), on peut proposer un estimateur pour $\overline{F_u}(y)$, soit

$$\widehat{\overline{F}}_u(y) = \widehat{\overline{G_{\xi,\sigma}}}(y) = \left(1 + \frac{\widehat{\xi}}{\widehat{\sigma_u}} y\right)^{\frac{-1}{\widehat{\xi}}}. \tag{1.3.9}$$

En utilisant les relations (1.3.6), (1.3.8) et (1.3.9), on obtient

$$\widehat{\overline{F}}(u+y) = \widehat{\overline{F}}(u)\,\widehat{\overline{F}}_u(y) = \frac{N_u}{n}\left(1+\frac{\widehat{\xi}}{\widehat{\sigma_u}}y\right)^{\frac{-1}{\widehat{\xi}}}.$$

Le niveau de retour y_N dépassé en moyenne toutes les N années est alors obtenu comme la solution de l'équation (1.3.10) :

$$\frac{N_u}{n}\left(1+\frac{\widehat{\xi}}{\widehat{\sigma_u}}(y-u)\right)^{\frac{-1}{\widehat{\xi}}} = \frac{1}{m}, \text{ avec } y > u \tag{1.3.10}$$

Selon les valeurs de ξ en résolvant l'équation (1.3.10), on a :

$$\widehat{y_N} = \begin{cases} u + \dfrac{\widehat{\xi}}{\widehat{\sigma_u}}\left[\left(m\dfrac{N_u}{n}\right)^{\widehat{\xi}} - 1\right], & \text{si } \xi \neq 0 \\ u + \widehat{\sigma_u}\log\left(m\dfrac{N_u}{n}\right), & \text{si } \xi = 0 \end{cases}. \tag{1.3.11}$$

Remarque 1.3.11 Il est souvent adéquat d'estimer le niveau de retour en termes d'échelle annuelle. S'il existe n_y observations par an sur une période de N années avec $m = N * n_y$ mesures, la relation (1.3.11) devient

$$\widehat{y_N} = \begin{cases} u + \dfrac{\widehat{\xi}}{\widehat{\sigma_u}}\left[\left(Nn_y\dfrac{N_u}{n}\right)^{\widehat{\xi}} - 1\right], & \text{si } \xi \neq 0 \\ u + \widehat{\sigma_u}\log\left(Nn_y\dfrac{N_u}{n}\right), & \text{si } \xi = 0 \end{cases}. \tag{1.3.12}$$

.

1.4 Adéquation des modèles des valeurs extrêmes

1.4.1 Introduction

Dans la pratique, il est souvent nécessaire d'analyser la validité d'un modèle retenu pour des prédictions. Il s'agit donc de s'assurer que les modèles sélectionnés décrivent bien les données étudiées et produisent de bonnes estimations afin de donner plus de précision aux prédictions et estimations. De ce fait, nous disposons de certains outils de diagnostic tels que les graphiques Probabilité-Probabilité (P-P plot) et Quantile-Quantile (Q-Q plot) ; il y a aussi les tests usuels d'adéquation comme ceux de Kolmogorov Smirnov (KS), Anderson Darling (AD), Wald, rapport de vraisemblance et autres pour juger de la pertinence des modèles.

1.4.2 Probabilité et Quantile Plot

Les (P-P plot) et (Q-Q plot) sont très utilisés pour tester l'adéquation des modèles des valeurs extrêmes. Ces outils graphiques permettent de comparer la fonction de répartition empirique pour un (P-P plot) ou les quantiles extrêmes (Q-Q plot) d'un échantillon de données à ceux d'un échantillon provenant d'une loi théorique (GPD ou GEV) par exemple. Lorsque la loi observée est la même que la loi théorique, les points sur ces graphiques sont confondus avec la première bissectrice dans le plan.
Supposons donc que les valeurs observées $y_{(1)}, \ldots, y_{(n)}$ sont ordonnées dans l'ordre croissant, une (P-P plot) est l'ensemble des points

$$\left\{ \left(\hat{F}\left(y_{(i)}\right), \frac{i}{n+1} \right) : i = 1, \ldots, n \right\},$$

où $\hat{F}$ est la distribution empirique estimée et $i/(n+1)$ est la valeur prise par la fonction de distribution empirique dans l'intervalle $\left[y_{(i)}, y_{(i+1)}\right[$.
Concernant le (Q-Q plot), comme son nom l'indique, il compare les quantiles extrêmes. Il est constitué de l'ensemble des couples

$$\left\{ \left(\hat{F}^{-1}\left(\frac{i}{n+1}\right), x_{(i)} \right) : i = 1, \ldots, n \right\}.$$

Pour plus de détails, voir Coles (2001).

1.4.3 Tests d'adéquation de Kolmogorov Smirnov

Soient $x_1, ..., x_n$, n réalisations d'une variable aléatoire X et de fonction de répartition F. On se demande s'il est raisonnable de supposer que X suit la loi caractérisée par F et on pose les deux hypothèses de test :

1. H_0 :X suit la loi de F,
2. H_1 :X suit une autre loi.

Ainsi, on utilise la fonction de répartition empirique F_n de F définie par $F_n(x) = \dfrac{Card\left(\{i \mid x_i \leq x\}\right)}{n}$. La statistique de test utilisée est définie par :

$$D_n = \sup_{x \in \mathbb{R}} \left| F_n(x) - F(x) \right|.$$

On compare la valeur obtenue D_n à une valeur critique $D_\alpha(n)$ fournie par les tables de Kolmogorov-Smirnov. Si $D_n > D_\alpha(n)$, on rejette l'hypothèse H_0 avec un risque α de se tromper.

1.4.4 Quantification de l'incertitude par la méthode du bootstrap dans la théorie des valeurs extrêmes

On distingue principalement deux approches : l'approche paramétrique qui est une méthode dans laquelle des hypothèses de base sont émises sur la distribution de

l'échantillon de base et l'approche non paramétrique. Dans cette dernière approche, il peut y avoir des hypothèses sur la distribution de l'échantillon de base ; cependant, elle ne nécessite pas que la loi parent appartienne à une famille paramétrique.
La technique du bootstrap a été introduite par Efron (1979). C'est la méthode de réplication des échantillons la mieux fondée théoriquement. Elle consiste à créer, à partir d'un échantillon de base, un grand nombre d'échantillons par tirage aléatoire avec remise. Sur chaque échantillon, les statistiques auxquelles on s'intéresse sont calculées, ce qui permet d'approcher leur dispersion. De ce fait, on peut estimer la variance ou la loi des paramètres caractéristiques de la distribution de l'échantillon et construire des intervalles de confiance, lorsque la distribution des paramètres est analytiquement complexe. Cette méthode semble donc adéquate pour la quantification des incertitudes dans la théorie des valeurs extrêmes.
Dans ce travail, on s'intéressera uniquement à l'approche non paramétrique pour quantifier les incertitudes sur les estimations.

Définition 1.4.1 *Le bootstrap est une technique de reéchantillonnage permettant de simuler la distribution d'un estimateur quelconque pour en apprécier le biais, la variance, l'erreur quadratique moyenne ou encore pour en estimer un intervalle de confiance, même si la loi théorique est inconnue.*

1.4.4.1 Approche non paramétrique du bootstrap

Cette approche est généralement utilisée dans les situations où l'on ne peut pas faire l'hypothèse que la distribution des observations de certains paramètres appartient à une famille connue. Ici, notre approche consiste à analyser les incertitudes sur les estimations des quantiles extrêmes (niveaux de retour), des paramètres de forme et d'échelle à travers l'approche non paramétrique. Cette analyse se base sur la comparaison des résultats observés et simulés.

1.4.4.2 Bootstrap pour l'estimation d'une erreur standard

Dans l'estimation d'une erreur standard par bootstrap, les statistiques d'intérêt peuvent être une moyenne, une médiane, un coefficient de corrélation, un coefficient de variation, une erreur moyenne quadratique, etc.
La figure (1.4.1) illustre la procédure de bootstrap d'estimation d'une erreur standard pour une moyenne.

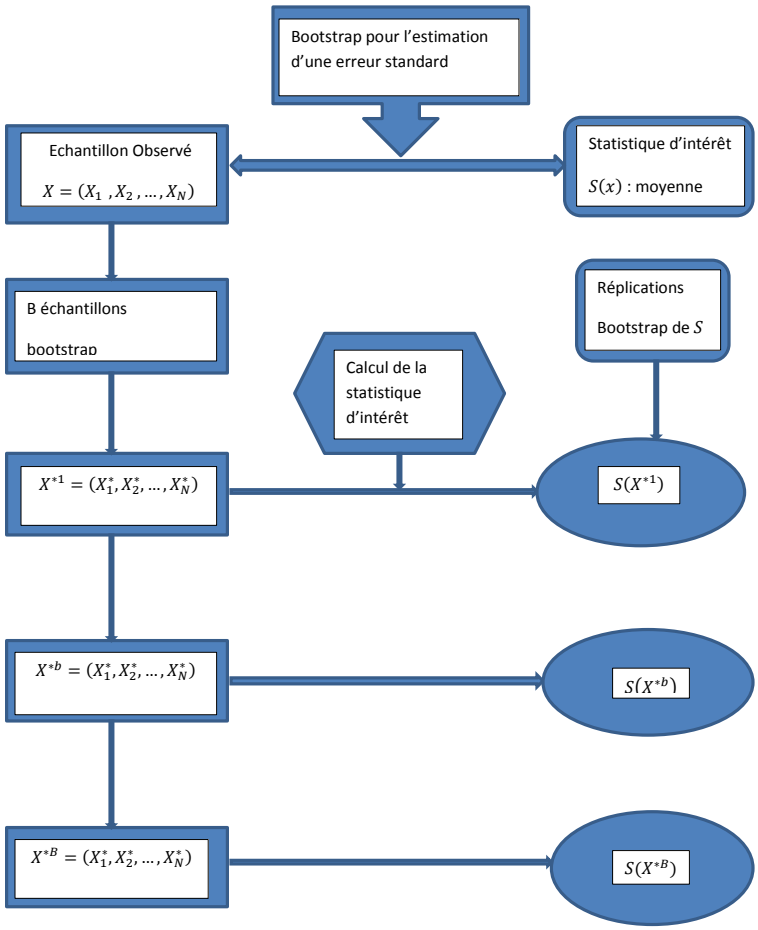


FIGURE 1.4.1 – Bootstrap pour l'estimation d'une erreur standard

Remarque 1.4.2 *L'estimation du bootstrap de l'erreur standard correspond à l'écart-type des réplications bootstrap ; ceci est obtenu par l'équation suivante :*

$$\sqrt{\frac{\sum_{b=1}^{B}\left[S\left(X^{*b}\right)-S^{*}\right]^{2}}{B-1}}$$

avec $S^* = \dfrac{\sum_{b=1}^{B} S\left(X^{*b}\right)}{B}$ *et* S *peut être une moyenne, un coefficient de variation, une erreur quadratique moyenne, etc.*

1.4.4.3 Quantification des incertitudes sur les estimations dans l'approche POT

La quantification des incertitudes porte sur les niveaux de retour x_T estimés et les intervalles de confiance respectifs associés ; ainsi que les paramètres de forme ξ et d'échelle σ de la GPD.

Définition 1.4.3 *L'erreur quadratique moyenne en anglais (root mean square error, rmse), est une mesure de la différence entre les valeurs prédites et observées. Elle est définie par :*

$$rmse = \sqrt{\frac{\sum_{i=1}^{n} \left(\widehat{y_i} - y_i\right)^2}{n}}$$

où $\widehat{y_i}$ *est la* $i-$*ème valeur estimée,* y_i *celle observée et* n *la taille de l'échantillon.*

L'algorithme du bootstrap peut être résumé pour l'approche POT comme suit :

1. Les données sont considérées comme un vecteur z_{obs} de n observations indépendantes rangées dans l'ordre croissant.
2. Un échantillon de n observations pris au hasard avec répétition d'éléments du vecteur z_{obs} pour obtenir les données de bootstrap est noté Z^*. Mais dans notre approche, les données sont constituées d'un certain nombre de valeurs au-dessus d'un seuil sélectionné. Ainsi, un échantillon aléatoire de taille le nombre d'excès est selectionné avec remise à partir des données initiales pour obtenir la base de données du bootstrap.
3. Une méthode d'estimation des paramètres de forme ξ et d'échelle σ de la GPD est utilisée afin de déterminer la statistique d'intérêt pour différentes périodes de retour.
4. Déterminer la statistique d'intérêt, $\theta^* = \theta\left(Z\right)$.
5. Répéter les étapes 3 et 4 un grand nombre de fois (B fois) pour obtenir une estimation de la distribution du bootstrap.
6. Calculer le ($rmse$) pour les paramètres de la GPD et la statistique d'intérêt.
7. Déterminer l'intervalle de confiance pour la statistique d'intérêt.

1.5 Conclusion

Dans ce chapitre, nous avons présenté les deux approches de la théorie des valeurs extrêmes et les éléments théoriques y afférant. Il ressort de cette présentation et de la littérature concernant ces deux approches que l'approche POT est la plus

utilisée mais très sensible au choix du seuil pour obtenir une bonne approximation des excès au-dessus du seuil considéré. En outre, dans les domaines de l'industrie ou de l'aéronautique où les données ne sont pas toujours saisonnières ou annuelles, les divisions en blocs s'avèrent problématiques. Ainsi, l'approche POT semble être la plus adaptée pour de nombreuses situations telles que l'analyse des mesures de surface de rugosité, la prédiction des pièces de rechange et la gestion des risques. Dans l'optique de la quantification des incertitudes dues à l'approche POT, nous avons abordé l'approche du bootstrap non-paramétrique dans la TVE ; plus précisément avec l'approche POT.

Le chapitre 2 porte sur l'analyse de survie ou de fiabilité des systèmes dans le cadre des modèles à risques proportionnels. Il établit le lien entre la théorie des valeurs extrêmes et les modèles de fiabilité.

Chapitre 2

Analyse de survie et santé des systèmes dans un cadre extrême

2.1 Introduction

Le terme analyse de fiabilité tire sa source de l'analyse de survie. L'analyse de survie concerne les êtres vivants, par exemple le suivi d'un patient après un traitement spécifique. Quant à l'analyse de fiabilité, elle s'intéresse aux systèmes matériels tels que les outils électriques, mécaniques ou tout autre système industriel.
L'analyse de survie est un domaine important des biostatistiques qui consiste à étudier le fonctionnement, les durées de vie ou l'évolution des systèmes. La théorie de la fiabilité sert à étudier l'aptitude de systèmes à fonctionner correctement durant une période donnée. Plus précisément, la fonction de fiabilité est la probabilité pour qu'un matériel ou un système fonctionne sans défaillance pendant une durée de temps. Ce matériel ou un système peut être technique (une machine, un bâtiment, un outil industriel, etc), biologique (plantes, êtres vivants, etc).

2.2 Concepts de base de l'analyse de fiabilité

La fiabilité a pris son développement depuis la dernière guerre mondiale. Elle est devenue une science à part entière avec des applications dans de nombreux domaines. Elle a pour fondements mathématiques la statistique et le calcul de probabilités qui sont nécessaires à la compréhension et à l'analyse des données de fiabilité.
La détermination de la fiabilité d'un système électronique, mécanique ou autre nécessite tout d'abord de connaître la loi de fiabilité (ou la loi de défaillance) de chacun des composants intervenant dans le système.
Certaines terminologies sont couramment utilisées pour indiquer la fin d'une durée de vie telles que par exemple :

* une défaillance par l'ingéniérie de la fiabilité,
* décès dans le domaine des sciences actuarielles et biostatistiques,
* une époque pour les processus ponctuels.

De plus, certains concepts sont associés à des objets, comme par exemple :
* un système ou un composant dans l'ingéniérie de la fiabilité,
* un individu pour les sciences actuarielles,
* un organisme en biostatistiques.

2.2.1 Fonctions décrivant l'évolution d'un système

On distingue principalement cinq fonctions représentant l'évolution d'un système, chacune de ces fonctions peut avoir une dénomination particulière selon le domaine d'application :

1. le complémentaire de la fonction de répartition notée S ou R. On l'appelle aussi fonction de survie ou fonction de fiabilité,
2. la fonction de répartition notée F, elle est aussi connue sous le nom de fonction de défaillance en fiabilité des systèmes,
3. la fonction de densité de probabilité notée f, elle est aussi connue sous le nom de densité de défaillance,
4. la fonction de risque notée h, elle est aussi appelée taux de défaillance instantané ou risque instantané,
5. la fonction de risque cumulée notée H, elle est aussi connue sous le nom de fonction de risque intégrée.

2.2.1.1 Fonction de survie S ou de fiabilité R

Soit X une variable aléatoire modélisant l'évolution d'un système. La fonction de survie S est, pour t fixé, la probabilité de survivre jusqu'à l'instant t, c'est-à-dire

$$S(t) = P(X > t), \ t \geq 0.$$

Remarque 2.2.1 *Dans un contexte multivarié* $S(t) \neq 1 - F(t)$ *car* $\overline{F}(t) = P(X > t) \neq 1 - F(t)$.

2.2.1.2 Fonction de répartition F

La fonction de répartition ou « *cumulative distribution en anglais* » représente, pour t fixé, la probabilité de mourir avant l'instant t, c'est-à -dire

$$F(t) = P(X \leq t) = 1 - S(t),$$

et donc $S(t) = 1 - F(t) = \overline{F}(t)$.

2.2.1.3 Fonction de densité de probabilité f

C'est la fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$,

$$F(t) = \int_0^t f(u)\, du.$$

Si la fonction de répartition admet une dérivée au point t, alors

$$f(t) = \lim_{h \to 0} \frac{P(t \leq X < t + h)}{h} = F'(t) = -S'(t).$$

Pour t fixé, la fonction de densité représente la probabilité que l'évènement d'intérêt se produise dans un intervalle de temps rapporté à la largeur de cet intervalle.

2.2.1.4 Fonction de risque instantané

Le risque instantané ou encore taux d'incidence, pour t fixé, caractérise la probabilité de mourir ou d'avoir une défaillance dans un intervalle de temps après t, sous l'hypothèse d'avoir survécu jusqu'au temps t rapporté à la largeur de cet intervalle :

$$h(t) = \lim_{x \to 0} \frac{P(t \leq X < t + x \mid X \geq t)}{x} = \frac{f(t)}{S(t)}. \tag{2.2.1}$$

2.2.1.5 Fonction de risque cumulé

La fonction ou taux de risque cumulé est l'intégrale du risque instantané h défini par :

$$H(t) = \int_0^t h(u)\, du = -\log[S(t)]. \tag{2.2.2}$$

On peut déduire de la relation (2.2.2), une expression de la fonction de survie en fonction du risque cumulé :

$$S(t) = \exp(-H(t)) = \exp\left(-\int_0^t h(u)\, du\right).$$

On en déduit que

$$f(t) = h(t) \exp\left(-\int_0^t h(u)\, du\right).$$

2.2.2 Fonction de fiabilité d'un système

Le terme fiabilité est un néologisme introduit il y a plus de cinquante ans pour traduire le terme anglo-saxon *reliability*. La Commission Électronique Internationale donne à la fiabilité la définition suivante :

Définition 2.2.2 *La fiabilité notée R (Reliability) caractérise l'aptitude d'un système ou d'un matériel à accomplir une fonction requise dans des conditions données pendant un intervalle de temps donné.*

Considérons un système matériel dont on s'intéresse à la fiabilité. Soit T la variable aléatoire associée au temps de fonctionnement du système. On peut distinguer deux évènements A et B définis comme suit :

* A : le système est en bon état de fonctionnement à l'instant t.

* B : le système est défaillant à l'instant $t + \triangle t$.

On a alors : $P(A) = P(T > t)$et $P(B) = P(T \leq t + \triangle t)$. Par conséquent,

$$\begin{aligned} P(A \cap B) &= P(t < T \leq t + \triangle t) \\ &= F(t + \triangle t) - F(t) \\ &= [1 - R(t + \triangle t)] - [1 - R(t)] \\ \text{soit } P(A \cap B) &= R(t) - R(t + \triangle t). \end{aligned}$$

On en déduit donc que :

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{R(t) - R(t + \triangle t)}{R(t)}.$$

Remarque 2.2.3 *On appelle fonction de défaillance la fonction F définie par :*

$$F(t) = P(T \leq t)\ , t \geq 0,$$

où T représente la durée de vie pour une v.a.

Le nombre $F(t)$ représente la probabilité que le système ait une défaillance avant l'instant t.

La figure (2.2.1) illustre la fonction de fiabilité et de défaillance.

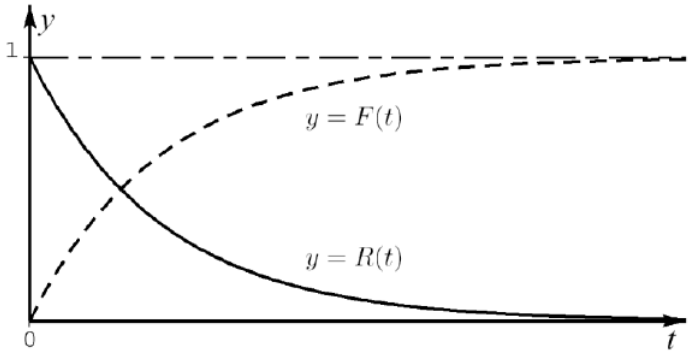


FIGURE 2.2.1 – Fonction de fiabilité et de défaillance

2.2.3 Taux de défaillance instantané ou fonction de risque instantané

Définition 2.2.4 *Soit un système ayant pour fonction de fiabilité R, le taux de défaillance instantané à l'instant t, est noté $h(t)$, et défini par :*

$$h(t) = \lim_{\triangle t \to 0} \left[\frac{1}{\triangle t} \times \frac{R(t) - R(t + \triangle t)}{R(t)} \right] \; t \geq 0.$$

De plus, le terme $h(t) \triangle t$, mesure la probabilité qu'une défaillance du système se produise dans l'intervalle de temps $[t, t + \triangle t]$ sachant que le système a bien fonctionné jusqu'à l'instant t. On a aussi la fonction taux de défaillance cumulé, noté $H(t)$ et défini par :

$$H(t) = \int_0^t h(x)\, dx.$$

Ainsi on peut aussi écrire,

$$\begin{aligned} h(t) &= -\frac{dR(t)}{dt} \times \frac{1}{R(t)} \\ &= \frac{dF(t)}{dt} \times \frac{1}{R(t)} \\ h(t) &= \frac{f(t)}{1 - F(t)}. \end{aligned}$$

Remarque 2.2.5 *On peut exprimer les fonctions de fiabilité et de défaillance en fonction du risque instantané ou du risque cumulé comme suit*

$$R(t) = \exp\left(- \int_0^t h(x)\, dx \right) \text{ et } F(t) = 1 - \exp\left(- \int_0^t h(x)\, dx \right).$$

Proposition 2.2.6 *Si h est une fonction de risque, alors elle vérifie les propriétés suivantes :*

1. $h(t) \geq 0$, *pour tout* $t \geq 0$,
2. $\int_0^{+\infty} h(t)\, dt = +\infty$.

La fonction de fiabilité ou de survie examine le risque de défaillance d'un organisme, d'un système matériel, etc pouvant se produire sur une période donnée. Pour suivre la durée de vie d'un système à travers la distribution associée à l'évolution du système, la fonction de risque est utilisée.
En fait, la fonction de risque est plus informative sur la défaillance des systèmes que les autres distributions caractérisant la durée des systèmes. C'est dans cette optique que Cox et Oakes (1984) donnent les raisons pour lesquelles l'utilisation de la fonction de risque peut être une bonne idée :

1. Elle peut mettre en évidence le fait de considérer le risque immédiat lié à un système d'être en vie à un âge t,
2. la comparaison de groupes d'individus est souvent faite par la fonction de risque,
3. les modèles de risque sont souvent adéquats lorsqu'il y a des données censurées ou plusieurs types de défaillance,
4. la comparaison avec une distribution exponentielle est particulièrement simple en termes de fonction de risque.

La fonction de risque peut prendre plusieurs dénominations selon le domaine d'intérêt, ainsi

* dans les sciences de l'ingénierie, on parle de taux de défaillance,

* en actuariat, elle est connue sous le nom de force de mortalité,

* dans les sciences de la vie et les sciences sociales, on parle d'âge ou taux de décès,

* en économie, elle est connue sous le nom de ratio de Mills, voir Mills (1926),

* concernant les processus ponctuels et la Théorie des Valeurs Extrêmes on parle plutôt de taux de fonction ou intensité de fonction.

2.3 Quelques lois usuelles de fiabilité

Dans cette section, nous rappelons quelques lois de probabilité couramment utilisées dans l'étude de la fiabilité des systèmes.

2.3.1 Loi normale

La loi normale est une loi de probabilité utilisée dans la modélisation de nombreux phénomènes. C'est une loi continue et symétrique dépendant de deux paramètres qui représentent bien la fin de vie des systèmes ou dispositifs subissant un phénomène de vieillissement tels que l'usure, la fatigue ou la corrosion. Dans les études de fiabilité, la variable T étant positive ou nulle, on s'intéresse à l'intervalle $[0, +\infty[$.

2.3.2 Loi log-Normale

Une variable aléatoire continue positive T est distribuée suivant une loi Log-Normale, si son logarithme népérien est distribué suivant une loi Normale. Cette loi est souvent représentative de la distribution de la mortalité ou de la durée de réparation des matériels. La distribution Log-Normale est un modèle fréquemment utilisé en fiabilité, car elle concerne des variables aléatoires positives et le paramètre de forme lui permet des représentations variées : en particulier, elle s'applique lorsque les observations faites sont les conséquences d'un effet multiplicatif de différentes causes indépendantes et aléatoires.

2.3.3 Loi exponentielle

C'est une loi qui ne dépend que d'un seul paramètre (le taux de défaillance λ). Elle s'applique d'une manière générale aux matériels qui subissent des défaillances brutales ou à des systèmes. La loi exponentielle est associée au processus Poissonnien qui est un processus qui génère des évènements dont les temps d'occurrence sont *iid*.

2.3.4 Loi de Weibull

C'est une loi très souple, représentative d'une grande variété de phénomènes aléatoires, qui est souvent utilisée dans le domaine de fiabilité des systèmes. La loi de Weibull est une loi représentative des trois phases de la durée de vie d'un système : jeunesse, maturité et vieillesse.
Cette distribution est souvent utilisée dans le domaine de l'analyse de la durée de vie grâce à sa flexibilité. Elle convient pour modéliser des structures comportant une grande quantité de petits défauts. C'est une distribution utilisée pour évaluer les valeurs extrêmes des pannes. Elle modélise aussi de nombreuses situations d'usure de matériels.

2.3.5 Distribution des valeurs extrêmes

La Théorie des Valeurs Extrêmes est de plus en plus utilisée dans le domaine des sciences modernes telles que l'ingénierie et l'assurance, pour plus de détails voir Embrechts *et al.* (1997), Reiss et Thomas (1997) et Castillo *et al.* (2005). Cette théorie permet d'estimer la distribution de durée de vie d'un système à travers la distribution asymptotique des valeurs extrêmes. Mahdavi et Mahdavi (2011) utilisent la distribution des valeurs extrêmes de type 1 ou de Gumbel pour estimer la fiabilité des systèmes en série ou en parallèle.

Remarque 2.3.1 *Le tableau (2.3.1) présente les fonctions de densité des lois usuelles de fiabilité répertoriées dans le tableau (1.2.1) comme appartenant toutes au domaine d'attraction de Gumbel.*

Lois usuelles	Fonctions de densité
loi normale $T \sim N(\mu, \sigma^2)$	$f_T(t;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]$, $t \geq 0$, μ est l'espérance et σ l'écart-type
loi log-normale $T \sim \log - N(\mu, \sigma^2)$	$f_T(t;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}t} \exp\left[-\frac{(\ln t-\mu)^2}{2\sigma^2}\right]$, $t \geq 0$, où μ et σ sont les paramètres de la loi normale
loi exponentielle $T \sim Exp(\lambda)$	$f_T(t;\lambda) = \lambda \exp(-\lambda t)$, $t \geq 0$
loi de Weibull $T \sim W(k, \lambda)$	$f_T(t;k,\lambda) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{t}{\lambda}\right)^k\right]$, $t > 0$, $k > 0$ est le paramètre de forme et $\lambda > 0$ est celui d'échelle
loi de Gumbel $T \sim G(\mu, \sigma)$	$f_T(t;\mu,\sigma) = \frac{1}{\sigma} \exp\left[-\frac{t-\mu}{\sigma} - \exp\left(-\frac{t-\mu}{\sigma}\right)\right]$, le paramètre de forme est 0, μ le paramètre de position et σ celui d'échelle

TABLE 2.3.1 – Fonctions de densité de quelques lois usuelles de fiabilité

2.4 Modèles à risques proportionnels pour la fiabilité des systèmes

Le modèle à risques proportionnels a été introduit par Cox en 1972 dans le but d'estimer les effets des covariables influençant les temps de défaillance d'un système. Ce modèle a tout d'abord été utilisé intensivement dans la biomédecine et récemment de plus en plus dans la fiabilité en ingéniérie. Kumar et Klefsjo (1994) proposèrent une synthèse des travaux sur l'utilisation des modèles à risques proportionnels. On distingue principalement deux familles de modèles à risques proportionnels, les modèles semi-paramétriques et les modèles paramétriques.

2.4.1 Modèles à risques proportionnels

2.4.1.1 Présentation du modèle à risques proportionnels de Cox

Afin de prendre en compte l'effet des variables d'environnement ou covariables dans les analyse de survie, le modèle à risques proportionnels a été développé par Cox en 1972. Ce modèle a tout d'abord été utilisé dans le champ médical, ensuite pour la détermination de la fiabilité des systèmes électroniques et par la suite pour les systèmes de manière générale.

Dans le modèle à risques proportionnels *(Proportional Hazard Model),* la fonction de risque h dépend du temps mais également de covariables $Z = (z_1, z_2, ..., z_n)$ relatives

à l'environnement, on a donc $h = h(t, Z)$. Plus précisément on fait l'hypothèse que la fonction de risques proportionnels est égale au produit d'une fonction de risque de base $h_0(t)$ qui ne dépend pas des covariables et d'une fonction de risque $\varphi(Z)$, dépendant uniquement des covariables Z. Cependant, dans le cas le plus général, les covariables peuvent aussi dépendre du temps. Dans un formalisme mathématique, la forme générale des modèles paramétriques à risques proportionnels est :

$$h(t, Z) = h_0(t)\,\varphi(Z) \tag{2.4.1}$$

avec :
* $h(t, Z)$ la fonction de risque du système, fonction du temps t et de l'état Z des facteurs d'influence,
* $h_0(t)$ distribution représentant la fonction de risque de base du système, fonction uniquement du temps. Les plus couramment utilisées comme fonctions de risque de base sont les distributions de Weibull, exponentielle, etc. Il faut noter que ces distributions sont toutes monotones.
* $\varphi(Z)$ une fonction appellée fonction d'influence ou risque d'environnement et qui dépend uniquement de l'état des facteurs d'influence ou variables d'environnement,
* $\varphi[Z(t)]$ lorsque les covariables dépendent du temps.

Remarque 2.4.1 *Lorsque la fonction $h_0(t)$ n'est pas spécifiée, on parle alors de modèles semi-paramétriques à risques proportionnels.*

Le modèle de Cox se situe dans un cadre semi-paramétrique car la fonction de risque de base n'est pas spécifiée. De plus, dans ce modèle, le risque lié aux covariables ou facteurs d'influence est sous la forme exponentielle.
Le modèle de Cox est largement utilisé en analyse de données de survie ou de fiabilité. Il est employé lorsque l'objectif est d'évaluer l'effet de covariables ou variables d'environnement sur la durée de vie ou le fonctionnement d'un système. Le modèle de Cox a été proposé pour l'étude de la fiabilité sous vieillissement accéléré, pour plus de détails voir Lanternier *et al.* (2008) et son utilisation est bien répandue dans les applications biomédicales, on peut citer Ohno-Machado (2001); Austin (2012); Choi *et al.* (2011).

2.4.1.2 Facteurs d'influence ou variables d'environnement

La fonction d'influence ou d'environnement peut être dépendante ou indépendante du temps. Plusieurs facteurs et des plus variés pouvant entrer en jeu, Z sera un vecteur dont les composantes $\{z_k\}_{k=1,\ldots,p}$ peuvent être discrètes ou continues. Ainsi, la fonction d'influence sera sous la forme

$$\varphi(Z) = \exp(B^t Z) = \exp(\sum_{k=1}^{p} b_k z_k) \tag{2.4.2}$$

où

* p est le nombre de facteurs d'influence pris en compte dans le modèle,
* $B = (b_1, ..., b_p)^t$, le vecteur des coefficients du modèle,
* $Z = (z_1, ..., z_p)^t$, le vecteur des états des facteurs d'influence ou d'environnement.

Les facteurs d'influence sont des éléments internes ou externes qui ont pour effet d'agir sur la fiabilité du système. L'influence peut être positive en améliorant la performance ou négative en réduisant la performance.

2.4.1.3 Modèles paramétriques classiques à risques proportionnels

Lorsque la fonction de risque de base est spécifiée, on dit que le modèle à risque proportionnel est paramétrique. Les distributions les plus couramment utilisées comme risque de base sont les distributions de Weibull, exponentielle, log-normale, etc.
Afin de prendre en compte l'environnement dans les analyses de survie, le modèle à risques proportionnels ou *Proportional Hazard Model* a été développé par Cox. Ce modèle a tout d'abord été utilisé dans le champ médical et ensuite pour la détermination de la fiabilité de systèmes.

Exemple 2.4.2 *Le modèle à risques proportionnels de Weibull : dans ce modèle, la fonction de risque de base est de type Weibull. Ainsi pour une distribution de Weibull à deux paramètres, la fonction de risque de base est*

$$h_0(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1},$$

où k est le paramètre d'échelle et λ le paramètre de forme.
** Si $k > 1$, alors la fonction de risque de base traduit une situation de dégradation,*
** Tandis que si $k < 1$, alors la fonction de risque de base traduit une situation d'amélioration. Dans ce cas, le modèle à risques proportionnels est de la forme suivante :*

$$h(t, Z) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \varphi(Z). \tag{2.4.3}$$

Lorsque la fonction d'influence est sous forme exponentielle, en utilisant les relations (2.4.2) et (2.4.3) nous obtenons

$$h(t, Z) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp(\sum_{k=1}^{p} b_k z_k).$$

2.4.2 Modèles GEV à risques proportionnels

L'approche que nous proposons cherche à modéliser à partir de la TVE, le taux de défaillance ou la performance d'un système sur une période donnée. Nous utilisons particulièrement la loi standard de Gumbel dans cette modélisation, afin de prendre

en compte des conditions d'opération ou des variables d'environnement extrêmes qui ne sont pas bien prises en compte par les fonctions de risque de base classiques telles que les fonctions de risque de Weibull, exponentielle, etc. Nous nous intéressons tout particulièrement au cas où le système ne fonctionne pas correctement, puis fonctionne normalement sur une période. Ceci est caractérisé par une fonction de risque non monotone (croissante puis décroissante).

2.4.2.1 Modélisation de la fonction de risque par la Théorie des Valeurs Extrêmes généralisées

Supposons que T suit une distribution de Weibull, alors $\log T$ suit une distribution de Gumbel pour le minimum des extrêmes, pour plus de détails voir Kim et Ibrahim (2000). Pour obtenir une fonction flexible, nous considérons que $\log T$ suit une loi GEV. En utilisant la relation (1.2.2) et le fait que $\log T \sim GEV(0,1,\xi)$, nous obtenons la densité de probabilité suivante :

$$f_\xi(t) = \begin{cases} \dfrac{\exp\left[-(1+\xi\log t)^{-\frac{1}{\xi}}\right]}{t(1+\xi\log t)^{\frac{1}{\xi}+1}} & \text{si } \xi \neq 0\,;\, t \neq \exp\left(\frac{-1}{\xi}\right)\,;\, t>0 \\ \frac{1}{t^2}\exp\left(\frac{-1}{t}\right) & \text{si } \xi = 0\,;\, t>0 \end{cases}. \tag{2.4.4}$$

Dans ce cas la fonction de survie correspondante $S_\xi = P(T>t)$ est donnée par :

$$S_\xi(t) = \begin{cases} 1-\exp\left[-(1+\xi\log t)^{-\frac{1}{\xi}}\right] & \text{si } \xi \neq 0;\, t>0 \\ 1-\exp\left(-\frac{1}{t}\right) & \text{si } \xi = 0;\, t>0 \end{cases}. \tag{2.4.5}$$

2.4.2.2 Fonction de risque croissante-décroissante

En pratique la plupart des fonctions de risque sont monotones et celles non monotones sont rares dans la litterature. Les plus connues sont la fonction de risque pour la distribution Log-normale et la distribution inverse de Gauss. Cependant leurs utilisations sont limitées car elles nécessitent beaucoup de conditions. Les conditions suffisantes et nécessaires pour la caractérisation des fonctions de risque croissante-décroissante ont été données par Glaser (1980). Nous supposons que la fonction de densité f est continue et deux fois différentiable sur $]0,+\infty[$. Le reste des conditions sont établies à partir de la fonction inverse de la fonction de risque notée τ et définie par

$$\tau(t) = \frac{1}{h(t)} = \frac{S(t)}{f(t)}; \tag{2.4.6}$$

et sa dérivée première est donnée par :

$$\tau'(t) = \tau(t)\,\psi(t) - 1 \tag{2.4.7}$$

où ψ est définie par :

$$\psi(t) = -\frac{f'(t)}{f(t)},$$

$$\psi'(t) = [\psi(t)]^2 - \frac{f''(t)}{f(t)},$$

ψ est continue et dérivable sur $]0;+\infty[$.
Pour des raisons pratiques, nous utilisons une autre expression pour la relation (2.4.7), à savoir

$$\tau'(t) = \int_t^{+\infty} \left[\frac{f(y)}{f(t)}\right] [\psi(t) - \psi(y)]\, dy \tag{2.4.8}$$

et

$$\tau''(t) = \tau'(t)\,\psi(t) + \psi'(t)\,\tau(t); \tag{2.4.9}$$

pour plus de détails voir Horst (2014) et Glaser (1980).

Proposition 2.4.3 *Si h est une fonction de risque croissante-décroissante, alors elle vérifie les propriétés suivantes :*

1. $h(t) \geq 0$, *pour tout* $t \geq 0$;
2. $\int_0^{+\infty} h(t)\, dt = +\infty$;
3. *Supposons qu'il existe* $t_0 > 0$ *tel que :* $\psi'(t) > 0, \forall t \in]0, t_0[$, $\psi'(t_0) = 0$, *et* $\psi'(t) < 0, \forall t \in]t_0, +\infty[$. *S'il existe* $x_0 > 0$ *tel que :* $\tau'(x_0) = 0$, *alors h est une fonction de risque croissante-décroissante.*

DÉMONSTRATION Pour les propriétés 1 et 2, voir la preuve de la proposition 2.2.5.

3) Pour $t_0 > 0$, soit $x_0 \in]t_0, +\infty[$, supposons que $\tau''(x_0) < 0$ avec $\tau'(x_0) = 0$. En utilisant la relation (2.4.9), il s'en suit que $\tau''(x_0) = \tau(x_0)\,\psi'(x_0)$. D'où,

$$\tau''(x_0) < 0 \Longleftrightarrow \psi'(x_0) < 0 \Longleftrightarrow x_0 < t_0.$$

Supposons qu'il existe $t_0 > 0$ tel que $\psi'(t) > 0$, pour $t \in]0, t_0[$, $\psi'(t_0) = 0$ et $\psi'(t) < 0$ pour $t > t_0$ et $x_0 > t_0$, en utilisant la relation (2.4.8), on obtient $\tau'(t) < 0$ pour tout $t \geq 0$. Puisque $x_0 > t_0$, alors $\tau'(x_0) < 0$, ce qui est absurde. par conséquent $x_0 < t_0$ et $\tau''(x_0) < 0$. On peut donc affirmer qu'il existe une unique racine nommée x_0 pour l'équation $\tau'(x_0) = 0$ et τ atteint son maximum en ce point.

Lemme 2.4.4 *Supposons que h est une fonction de risque et qu'il existe* $t_0 > 0$ *tel que :* $\psi'(t) > 0, \forall t \in]0, t_0[$, $\psi'(t_0) = 0$, *et* $\psi'(t) < 0, \forall t \in]t_0, +\infty[$. *Si* $\lim_{t \to 0^+} f(t) = 0$, *alors h est une fonction de risque croissante-décroissante.*

Définition 2.4.5 *Une fonction de durée de vie $F(t)$ avec $t \in [0, +\infty[$ est une fonction de risque croissante-décroissante s'il existe $t_0 > 0$ tel que $H(t) = -\ln[1 - F(t)]$. Cette fonction est convexe sur $[0, t_0[$ puis concave sur $[t_0, +\infty[$.*

2.4.2.3 Modélisation d'une nouvelle fonction de risque de base

Par définition, la fonction de risque est telle que

$$h_{\xi}(t) = \frac{f_{\xi}(t)}{S_{\xi}(t)}.$$

En se plaçant dans le cas où $\xi = 0$, et en utilisant les relations (2.4.4) et (2.4.5) nous obtenons l'expression :

$$h_0(t) = \frac{1}{t^2\left(\exp\left(\frac{1}{t}\right) - 1\right)},$$

représentée sur la figure (2.4.1).

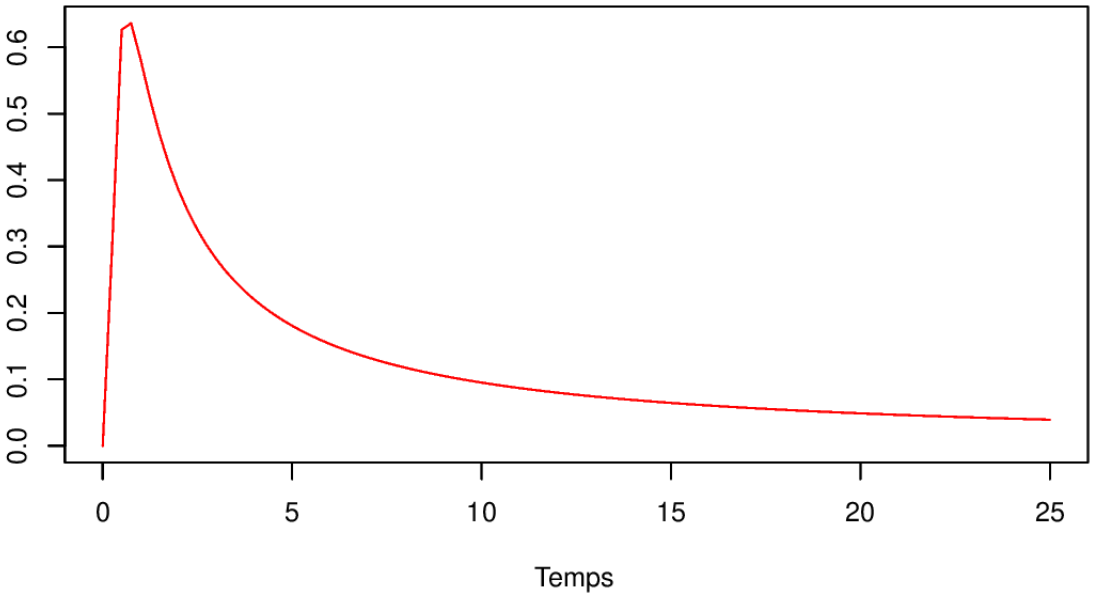


FIGURE 2.4.1 – Fonction de risque de base

Proposition 2.4.6 *La fonction* $t \longmapsto \dfrac{1}{t^2\left(\exp\left(\frac{1}{t}\right) - 1\right)}$ *est une fonction de risque croissante-décroissante.*

DÉMONSTRATION Pour $\xi = 0$, on a $f_0(t) = \dfrac{1}{t^2}\exp\left(\dfrac{-1}{t}\right)$, $S_0(t) = 1 - \exp\left(\dfrac{-1}{t}\right)$, $h_0(t) = \dfrac{f_0(t)}{S_0(t)}$ *et* $H_0(t) = -\ln\left(1 - \exp\left(\dfrac{-1}{t}\right)\right)$. Ainsi, il vient que : $h_0(t) = \dfrac{1}{t^2\left(\exp\left(\frac{1}{t}\right) - 1\right)}$.

1) Les fonctions f_0 et S_0 étant positives, alors $h_0(t) = \dfrac{f_0(t)}{S_0(t)} \geq 0$.

2) $\int_0^{+\infty} h_0(t)\,dt = [H(t)]_0^{+\infty} = [-\ln S_0(t)]_0^{+\infty} = \ln S(0) - \ln S_0(+\infty) = +\infty$.
3) La fonction $\psi(t) = -\dfrac{f_0^{'}(t)}{f_0(t)} = \dfrac{2t-1}{t^2}$ est continue et dérivable sur $]0,+\infty[$ et $\psi^{'}(t) = \dfrac{2(-t+1)}{t^3}$. De plus $\psi^{'}(1) = 0$, $\psi^{'}(t) > 0, \forall t \in]0,1[$, $\psi^{'}(t) < 0$, $\forall t \in]1,+\infty[$; d'où $t_0 = 1$ et les hypothèses sont vérifiées. En outre $\tau(t) = \dfrac{S_0(t)}{f_0(t)} = \dfrac{1-\exp(-\frac{1}{t})}{\frac{\exp(-\frac{1}{t})}{t^2}} = t^2\left(\exp(-\frac{1}{t}) - 1\right)$. En utilisant la relation (2.4.7), on obtient

$$\begin{aligned}\tau^{'}(t) &= \tau(t)\,\psi(t) - 1 \\ &= (2t-1)\exp(\frac{1}{t}) - 2t;\end{aligned}$$

et avec le logiciel Geogebra on a $x_0 = 0.63$ et $\tau^{'}(0.63) = 0$, d'où la fonction h_0 est une fonction de risque croissante-décroissante. Ainsi le modèle à risque proportionnel obtenu est

$$h(t,Z) = \frac{1}{t^2\left(\exp\left(\dfrac{1}{t}\right) - 1\right)}\exp(\sum_{k=1}^{p} b_k z_k). \tag{2.4.10}$$

Dans la sous-sous-section (2.4.3) suivante, nous présentons une application potentielle du modèle dans le domaine de l'industrie et nous faisons aussi une comparaison entre le modèle proposé et le modèle flexible de Weibull proposé par Park et Park (2018).

2.4.3 Application du modèle en usinage

L'exemple proposé relève du domaine de l'usinage à enlèvement de matière. Il s'agit d'évaluer les performances d'un outil de coupe en utilisant comme indicateur de surveillance la rugosité des surfaces usinées. Pour certains types d'usinage et aussi en tenant compte du couple outil-matière, la durée de vie de l'outil peut être difficilement évaluable, ceci à cause de phénomènes aléatoires caractérisés par des micro-cassures sur l'outil. Ces phénomènes, rares, sur une durée totale d'usinage provoquent des valeurs extrêmes de l'indicateur de surveillance et l'on mesure ainsi tout l'intérêt d'utiliser la TVE pour construire notre fonction de risque.

2.4.3.1 Modélisation de la rugosité

La rugosité des surfaces des pièces usinées dépend de plusieurs facteurs. Certains facteurs sont systématiques tels les conditions géométriques de l'outil de coupe associées aux conditions cinématiques du processus de génération. Elles se modélisent par un terme constant noté rugosité de génération R_g. D'autres facteurs sont associés à l'usure de l'outil et aux erreurs aléatoires d'usinage, modélisées par des fonctions déterministes ou aléatoires. Ces deux derniers facteurs se rajoutent généralement à R_g.

Alam *et al.* (1997) ont montré que la valeur de la rugosité de la surface usinée en fonction du temps d'usinage $R(t)$ peut s'exprimer, de façon déterministe, comme une combinaison linéaire de R_g et d'un terme ct^m, où les paramètres c et m dépendent des conditions de coupe utilisées lors de l'usinage.

$$R(t) = R_g + ct^m \tag{2.4.11}$$

Les deux paramètres c et m sont généralement obtenus par des techniques d'identification.
Les erreurs aléatoires d'usinage se caractérisent par une dispersion importante sur les résultats expérimentaux des mesures de $R(t)$.
Le processus d'usinage a été développé dans le cadre des travaux menés par Noureddine *et al.* (2012) et porte sur un couple outil-matière composé d'un carbure métallique de nuance P20 pour ce qui est de l'outil, et d'un acier à traitement thermique de nuance C38, de diamètre 100mm et de longueur 150mm, pour ce qui est de la pièce.
La rugosité maximale imposée par le dessin de définition est : $R_a = R_M = 4\mu$m.
Le critère R_a (rugosité moyenne arithmétique) est celui utilisé par Axinte et Dewes (2002), étant donné qu'il révèle les détails avec une bonne précision.
Les mesures ont été effectuées grâce à un rugosimètre électronique à palpeur inductif (Hommel Werke), avec une longueur de base de 2mm et « un cut-off de 0,75mm ».
Ces mesures présentant une dispersion notable, il a été nécessaire de les réaliser sur un minimum de 5 séries de 10 pièces chacune.
L'ensemble des données relatives aux 5 courbes ont permis d'identifier les valeurs respectives de c et m. Nous retiendrons le paramètre $m = 0.1017$ comme paramètre de risque pour la fonction de risque proposée dans la sous-sous-section (2.4.3.2).

2.4.3.2 Étude de la performance d'un outil de coupe

Dans le processus d'usinage, il peut arriver que le bec de l'outil de coupe subisse une micro-cassure, entraînant ainsi une dégradation de la rugosité de la surface des pièces usinées, (augmentation de la valeur de rugosité) mais qui peut être immédiatement atténuée par un auto-affutage de l'outil, pour plus de détails voir Inoue *et al.* (2003); Jiang (2014); Karpuschewski *et al.* (2014), qui finalement rehausse la qualité de surface, (diminution de la valeur de rugosité). Des études expérimentales menées par Noureddine *et al.* (2012) ont mis en évidence le phénomène d'auto-affutage avec la conséquence importante de l'existence de valeurs maximales dans les différents lots de pièces usinées. Par la suite, nous supposons donc que la fonction de risque de base associée à l'outil de coupe est la fonction h_0 définie dans la sous-sous-section (2.4.2.3). Les conditions d'opération de l'outil de coupe représentent les variables d'environnement dans la relation (2.4.10).

a) Le modèle à risque proportionnel GEV

En remplaçant dans la relation (2.4.10) la valeur de l'impact des conditions d'opérations, on obtient

$$h\left(t, 0.1017\right) = \frac{0.1017}{t^2\left(\exp\left(\frac{1}{t}\right) - 1\right)}. \tag{2.4.12}$$

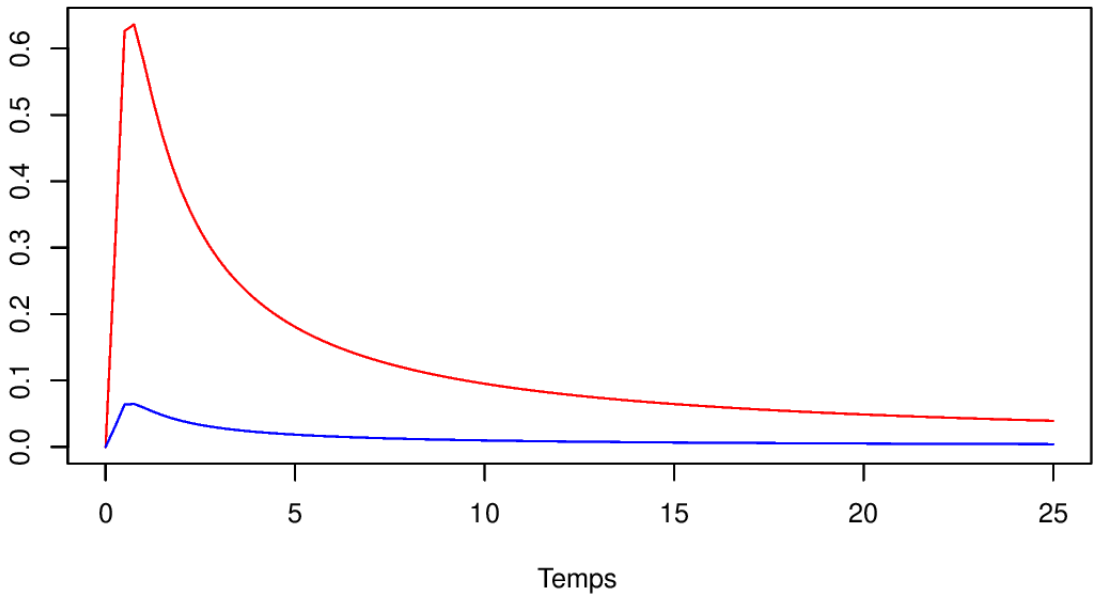


FIGURE 2.4.2 – Comparaison entre la fonction de risque de base et le modèle à risque proportionnel GEV

La figure (2.4.2) illustre la fonction de risque de base en rouge et celle à risques proportionnels GEV en bleu. On remarque que la courbe de la fonction à risques proportionnels GEV en bleu est en dessous de celle de base en rouge ; ceci montre l'impact des conditions d'opération sur la performance de l'outil de coupe puisque le taux de défaillance a considérablement diminué d'échelle.

b) Exemple de modèle à risques proportionnels croissant-décroissant : modèle de Park

Pour la prise en compte de la non monotonicité dans la modélisation de certains phénomènes en utilisant la distribution de Weibull, **Park et Park (2018)** ont proposé une fonction de risque, qu'ils nomment fonction de risque flexible de Weibull. La forme générale de cette fonction de risque est la suivante :

$$h(t; \mu, \alpha, \beta) = \left(\log\left(t+1\right)\right)^{\beta-1} \exp\left(\mu + \alpha t\right)\left(\alpha \log\left(t+1\right) + \frac{\beta}{t+1}\right), \tag{2.4.13}$$

où α, μ et β sont les paramètres du modèle.
Cette fonction de risque peut être croissante puis décroissante ou décroissante puis croissante selon les valeurs des paramètres. Ces paramètres du modèle de Park sont ajustés de façon à ce que ce modèle approche le plus possible le modèle à risques proportionnels GEV proposé.
Par exemple, pour $\alpha = 0$, $\mu = 0$ et $\beta = 1.5$, on obtient la fonction de risque de base croissante puis décroissante définie comme suit :

$$h_0\left(t; 0, 0, 1.5\right) = \frac{1.5}{t+1}\left(\log\left(t+1\right)\right)^{0.5}.$$

Ainsi, en utilisant la valeur relative aux conditions d'opérations ($m = 0.1017$) on obtient finalement

$$h\left(t; 0, 0, 1.5, 0.1017\right) = \frac{0.15255}{t+1}\left(\log\left(t+1\right)\right)^{0.5}.$$

Il faut noter que dans ce modèle, on peut obtenir plusieurs fonctions de risque de base croissante-décroissante en choisissant convenablement les valeurs des paramètres du modèle.
La figure (2.4.3) montre les représentations graphiques des fonctions à risques proportionnels GEV et Park. Sur cette figure, on remarque globalement que les courbes de la fonction à risques proportionnels de Park en vert, bleu et jaune sont globalement au-dessus de celle de la fonction à risque GEV en rouge. Ceci se traduit concrètement par une surestimation du taux de défaillance par rapport au modèle à risque proportionnel GEV. Il faut aussi noter que la surestimation du taux de défaillance par la distribution de Weibull dans les conditions d'opération extrêmes en général a été évoquée par Zio (2016). En outre, le modèle représenté par la courbe en jaune semble mieux ajuster le modèle à risque GEV proposé.

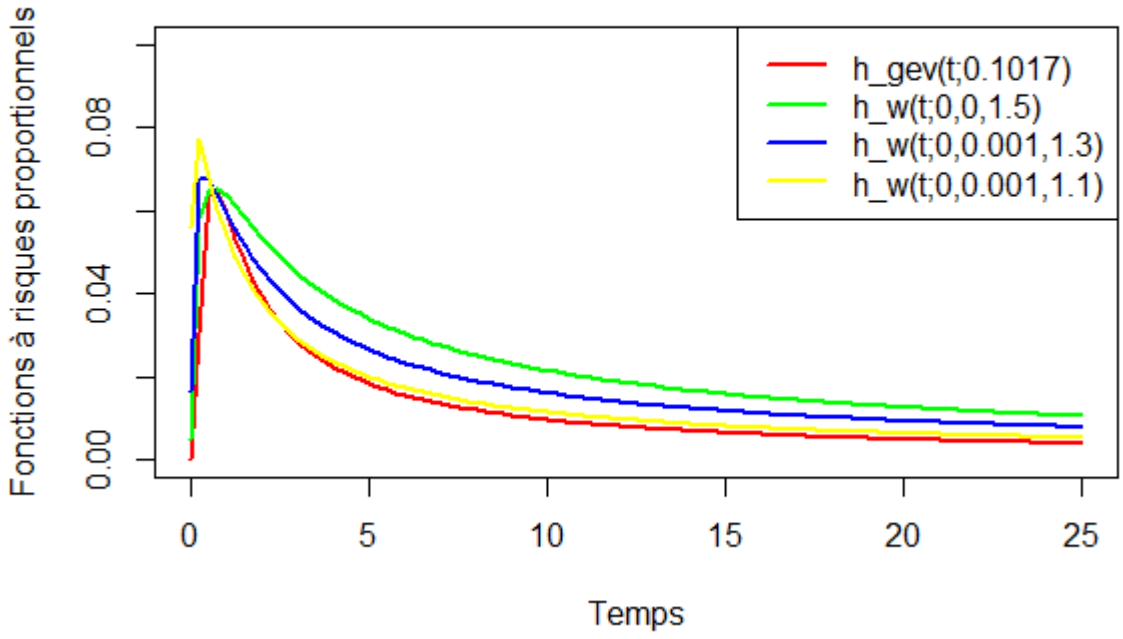


FIGURE 2.4.3 – Comparaison entre la fonction de risque proportionnel GEV et la fonction flexible de Weibull

Conclusion

Dans ce chapitre, nous avons rappelé les principaux concepts de l'analyse de survie, particulièrement ceux relatifs à l'analyse de fiabilité des systèmes et les lois usuelles de fiabilité. Nous nous sommes aussi intéressés au modèle à risque proportionnel et particulièrement au modèle paramétrique à risque proportionnel dans la prise en compte de la non monotonicité et des facteurs d'environnement ou d'opération dits extrêmes. C'est dans cette optique que nous avons construit, à partir de la TVE, une fonction de risque croissante-décroissante tenant compte des conditions extrêmes d'opération. Un exemple d'application du modèle dans le domaine de l'industrie a été proposé et un ajustement par une approche graphique du modèle proposé a aussi été fait à partir de la fonction de risque proposée par Park et Park (2018). De cet ajustement, il ressort que le modèle flexible de Weibull fait en général une certaine surestimation du taux de défaillance par rapport au modèle proposé. En perspective, une généralisation du modèle permettra de réaliser une étude comparative théorique qui pourra être illustrée par un cas d'étude.
Dans le chapitre 3 suivant, nous nous intéressons essentiellement à la statistique d'ordre extrême dans le contrôle de qualité en industrie.

Chapitre 3

Modélisation par les statistiques d'ordre pour le contrôle de qualité en usinage

3.1 Statistiques d'ordres et contrôle de qualité dans l'industrie

Les statistiques d'ordre sont utilisées dans des domaines variés tels que le génie civil avec Longuet-Higgins (1952) qui utilise la distribution de Rayleigh et les statistiques d'ordre pour modéliser la hauteur des vagues et estimer la rigidité des digues de protection ; Bernard et Ahmed (1958) pour la modélisation des données de santé. Quant au domaine de l'industrie, on peut citer Surajit (2005); Castillo *et al.* (2005) qui étudient la durée de vie des ampoules et les tests de défaillance par les statistiques d'ordre, Richard *et al.* (2007) travaillent en ingénierie structurale en utilisant les statistiques d'ordre extrêmes, tandis que Arak (2003); Sita et Satya (2014) appliquent les statistiques d'ordre dans l'étude de la fiabilité des logiciels et la compression de données. Par ailleurs, les statistiques d'ordre décrivant les variables aléatoires dans l'ordre de la magnitude sont très utilisées dans les méthodes statistiques et les inférences. Ces domaines des statistiques s'intéressent au rang ainsi qu'à l'intensité des observations en combinant les techniques conventionnelles de la statistique. En outre, la statistique d'ordre considère seulement le rang relatif quelles que soient les observations originales mesurées selon une échelle ordinaire ou non. Dans la pratique, les deux distributions les plus importantes des statistiques d'ordre sont les distributions du minimum et du maximum car elles sont les valeurs critiques utilisées en ingénierie, physique, médecine, etc, voir Arak (2003); Richard *et al.* (2007). Pour notre part, nous nous intéressons au contrôle de qualité dans l'industrie manufacturière par la construction d'un processus fonctionnel basé sur les statistiques d'ordre pour la détection des fautes tolérables sur les pièces usinées.

3.1.1 Statistiques d'ordre

Soit $Y_1, ..., Y_m$ un échantillon aléatoire de taille m, de fonction de densité commune $f_Y(y)$ et de répartition commune $F_Y(y)$. Rangeons $Y_1, ..., Y_m$ dans l'ordre croissant de magnitude et soit $Y_{1:m} < Y_{2:m}, ..., < Y_{m:m}$ les valeurs ordonnées. La première, la deuxième,..., la m-ème statistique d'ordre notées respectivement par $Y_{1:m}, Y_{2:m}, ..., Y_{m:m}$ sont définies par :

$$\begin{aligned} Y_{1:m} &= \min\{Y_1, ..., Y_m\} \\ Y_{2:m} &= \text{deuxième plus petite } Y_i \\ &\vdots \\ Y_{m:m} &= \max\{Y_1, ..., Y_m\}. \end{aligned}$$

Le résultat suivant caractérise les fonctions de probabilité d'une statistique d'ordre quelconque. Pour la démonstration voir Castillo *et al.* (2005), pp.164-165.

Théorème 3.1.1 *Supposons que $Y_1, ..., Y_m$ sont des variables aléatoires iid de fonction de répartition commune $F_Y(y)$. Alors la fonction de répartition de la $r-$ième statistique d'ordre $Y_{r:m}$ définie par $Y_{1:m} < Y_{2:m}, ..., Y_{r:m}, ... < Y_{m:m}$ est donnée par :*

$$F_{Y_{r:m}}(y) = \sum_{j=r}^{m} C_m^j [F_Y(y)]^j [1 - F_Y(y)]^{m-j}, \tag{3.1.1}$$

avec $1 \leq r \leq m$ et sa fonction de densité associée est définie par

$$f_{Y_{r:m}}(y) = \frac{m!}{(r-1)!(m-r)!} (F_Y(y))^{r-1} [1 - F_Y(y)]^{m-r} f(y). \tag{3.1.2}$$

3.1.2 Statistiques d'ordre extrêmes

Soit Y une variable aléatoire de fonction de densité $f_Y(y)$ et de répartition $F_Y(y)$. Soit $(Y_1, ..., Y_m)$ un échantillon aléatoire iid de fonction de répartition $F_Y(y)$. La densité jointe et la fonction de répartition sont respectivement déterminées par

$$f_Y(y_1, ..., y_m) = \prod_{i=1}^{n} f_Y(y_i) \text{ et } F_Y(y_1, ..., y_m) = \prod_{i=1}^{n} F_Y(y_i). \tag{3.1.3}$$

Les statistiques d'ordres sont très importantes en pratique et particulièrement le minimum $Y_{1:m}$ et le maximum $Y_{m:m}$. Elles sont des valeurs critiques utilisées en ingénierie, physique, médecine, etc, pour plus de détails voir Castillo *et al.* (2005). On désignera par W_n et M_n les statistiques d'ordre associées respectivement au minimum et au maximum d'un échantillon. Les statistiques d'ordre W_m et M_m sont appelées statistiques d'ordre extrêmes.

La fonction de répartition F_{M_m} de la statistique d'ordre du maximum M_m, est telle que pour toute réalisation y,

$$\begin{aligned} F_{M_m}(y) &= P(M_m \leq y) \\ &= P(\max(Y_1, ..., Y_m) \leq y) \\ &= P[\cap(Y_i \leq y)] \\ \text{Ainsi, } F_{M_m}(y) &= \prod_{i=1}^{m} P(Y_i \leq y) \\ &= \prod_{i=1}^{m} F_Y(y) \\ \text{Soit} \quad F_{M_m}(y) &= [F_Y(y)]^m . \end{aligned} \tag{3.1.4}$$

Ainsi, on obtient la fonction de densité correspondante en dérivant la relation (3.1.4) et on a

$$f_{M_m}(y) = m f_Y(y) [F_Y(y)]^{m-1} .$$

La fonction de répartition de la statistique d'ordre du minimum, W_m est obtenue de manière similaire comme suit :

$$\begin{aligned} F_{W_m}(y) &= P(W_m \leq y) \\ &= P(\min(Y_1, ..., Y_m) \leq y) \\ &= 1 - P(min(Y_1, ..., Y_m) > y) \\ &= 1 - P[\cap_{i=1}^{m}(Y_i > y)] \\ \text{Ainsi,} \quad F_{W_m}(y) &= 1 - \prod_{i=1}^{m} P(Y_i > y) \\ &= 1 - \prod_{i=1}^{m} (1 - F_Y(y)) \\ \text{Soit} \quad F_{W_m}(y) &= 1 - [1 - F_Y(y)]^m . \end{aligned} \tag{3.1.5}$$

De même, la fonction de densité correspondante est obtenue en dérivant la relation (3.1.5) et on a

$$f_{W_m}(y) = m f_Y(y) [1 - F_Y(y)]^{m-1} .$$

On peut extraire des sous-sections (3.1.1) et (3.1.2) le corollaire suivant :

Corollaire 3.1.2 *Les distributions F_{M_m} et F_{W_m} des statistiques d'ordre extrême M_m et W_m sont données respectivement par :*

** La distribution du maximum*

$$F_{M_m}(y) = [F(y)]^m \text{ et } f_M(y) = m f(y) [F(y)]^{m-1} . \tag{3.1.6}$$

** La distribution du minimum*

$$F_{W_m}(y) = 1 - [1 - F(y)]^m \text{ et } f_W(y) = m f(y) [1 - F(y)]^{m-1} . \tag{3.1.7}$$

avec $y \in \mathbb{R}$.

Exemple 3.1.3 *Supposons que l'échantillon $Y_1, ..., Y_m$ suit une loi standard de Gumbel de paramètres $\mu = 0$ et $\beta = 1$. Ainsi, on obtient*

$$F_Y(y) = \exp(-\exp(-y)),\ y \in \mathbb{R}. \tag{3.1.8}$$

En dérivant F_Y, on obtient

$$f_Y(y) = \exp(-y)\exp(-\exp(-y))\ y \in \mathbb{R}. \tag{3.1.9}$$

** La distribution du maximum en utilisant la relation (3.1.6) est définie par*

$$F_{M_m}(y) = [\exp(-\exp(-y))]^m \text{ et } f_M(y) = m\exp(-y)[\exp(-\exp(-y))]^m$$

** La distribution du minimum en utilisant la relation (3.1.7) est, quant à elle, définie par*

$$F_{W_m}(y) = 1 - [1 - \exp(-\exp(-y))]^m$$

$$f_{W_m}(y) = m\exp(-y)\exp(-\exp(-y))[1 - \exp(-\exp(-y))]^{m-1}$$

Pour des échantillons de taille $m = 10$, $m = 100$ et $m = 1000$, nous présentons graphiquement les distributions du minimum et du maximum. Concernant la figure (3.1.1), les couleurs sont associées aux tailles comme suit verte pour la taille 10, rouge pour la taille 100 et bleue pour la taille 1000.

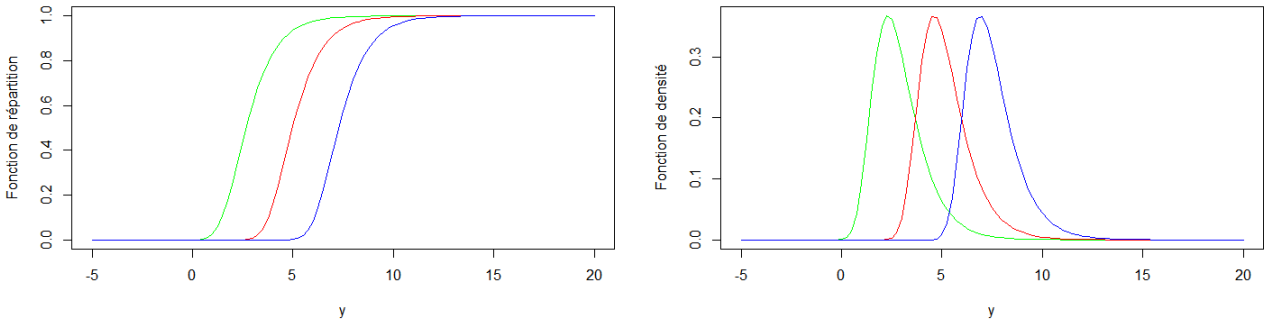


FIGURE 3.1.1 – Fonction de répartition à gauche et de densité à droite du maximum pour un échantillon de taille 10, 100 et 1000 pour une loi de Gumbel standard

La figure (3.1.1) montre la fonction de répartition et la fonction de densité de la statistique d'ordre du maximum, M_m. Nous observons aussi qu'il y a un changement de position et d'échelle lorsque la taille de l'échantillon augmente.

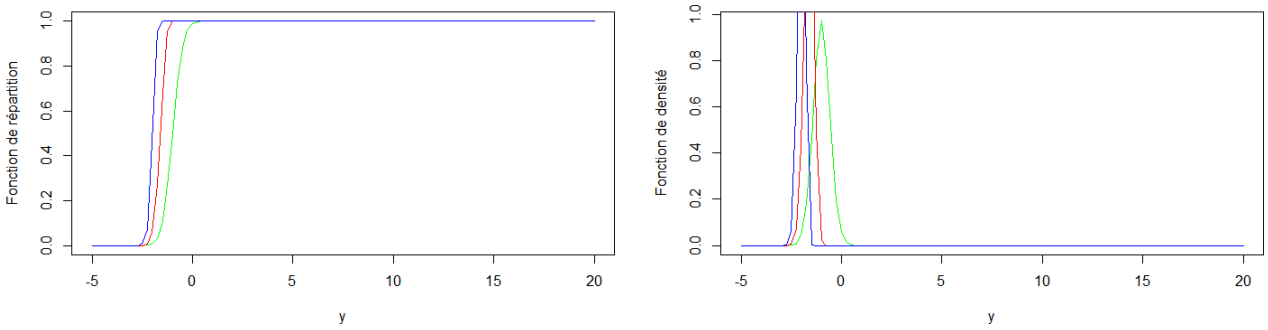


FIGURE 3.1.2 – Fonction de répartition à gauche et de densité à droite du minimum pour un échantillon de taille 10 (verte), 100 (rouge) et 100 (bleue) pour une loi de Gumbel standard.

De même, la figure (3.1.2) montre la fonction de répartition et de densité de la statistique d'ordre du manimum W_m. On remarque aussi qu'il y a changement de position et d'échelle lorsque la taille de l'échantillon augmente ; mais ce changement est dans le sens opposé de la figure (3.1.1).

Remarque 3.1.4 *Les statistiques d'ordre du minimum et du maximum sont des cas spécifiques de la théorie des valeurs extrêmes et les plus importants dans l'approche des statistiques d'ordre.*

3.1.3 Contrôle statistique de qualité dans l'industrie manufacturière

Le contrôle statistique de la qualité des produits dans un système de production industrielle est très important dans la mesure où il permet de s'assurer que les critères fixés au préalable sont atteints ou non. Cette étape utilise beaucoup de méthodes statistiques qui sont souvent connues sous le nom de statistique industrielle [1] ; elle peut se situer dans les trois niveaux suivants du processus de production : avant la production de masse pour vérifier la fonctionnalité du système, pendant la production pour faire l'évaluation à mi-parcours de la production et la dernière phase concerne les ultimes vérifications avant la livraison des produits. Généralement dans l'industrie, on distingue deux types de contrôle statistique :

1. le contrôle de fabrication et le contrôle en cours de fabrication ;
2. le contrôle de réception.

1. La statistique industrielle est une discipline scientifique qui développe des méthodes quantitatives et des règles pour la conduite d'études et la prise de décision dans le processus manufacturier

3.1.3.1 Contrôle de fabrication et contrôle en cours de fabrication

La qualité d'un lot de pièces dépend essentiellement du nombre de pièces défectueuses que le lot peut contenir. En ce qui concerne la qualité d'une pièce usinée, elle peut être étudiée des deux manières suivantes :

* le contrôle qualitatif : la pièce est passée au calibre et l'on vérifie si elle correspond ou non aux tolérances imposées par le cahier des charges ;

* le contrôle quantitatif : la pièce est mesurée, et l'on vérifie si la valeur obtenue est comprise dans l'intervalle de tolérance.

3.1.3.2 Le contrôle de réception

Ce contrôle est en quelque sorte une contre-expertise du client ; il permet de s'assurer que le cahier de charges a été pris en compte par le fabricant. Ce contrôle a aussi pour objectif de porter un jugement sur la proportion de pièces défectueuses dans un lot.

.

3.2 Contrôle de qualité en usinage

3.2.1 Introduction

Dans l'industrie manufacturière, la surface finale des pièces usinées n'est pas toujours homogène. Nous nous intéressons d'abord aux critères de surface liés à la rugosité, pour plus de détails voir Ilhan et Mehmet (2011). Dans les processus d'usinage de pièces, la surface des pièces dépend de plusieurs facteurs. Certains de ces facteurs sont systématiques tels que les conditions géométriques de l'outil de coupe associées aux conditions cinématiques du processus de génération. D'autres facteurs sont associés à l'usure de l'outil de coupe et aux défaillances aléatoires voir Noureddine *et al.* (2012); Zhang *et al.* (2015). L'évolution du comportement de l'outil de coupe est aussi aléatoire. L'apparition d'une faute d'usinage peut se situer au début du processus d'usinage et du coup peut causer d'énormes pertes économiques. Une telle situation peut être identifiée rapidement et il faut envisager les possibles réparations pour retrouver le processus normal de production ou décider d'arrêter la production. Ici, comme c'est souvent le cas dans l'industrie manufacturière, nous considérons un traitement par lots possédant un certain nombre de pièces à usiner avec un dépassement tolérable de la rugosité de surface des pièces usinées. Les valeurs de rugosité, définies dans la sous-sous-section (2.4.3.1), sont ordonnées afin d'appliquer les concepts des statistiques d'ordre.
C'est dans ce but qu'une approche basée sur la statistique d'ordre est particulièrement adéquate pour améliorer les processus d'usinage.
Généralement, la distribution la plus utilisée dans les modèles de fiabilité est vue comme une paramétrisation de la famille de Weibull. Les applications des statistiques d'ordre dans la modélisation de la surface de rugosité sont rares dans la litérature. Le

contrôle de qualité de la surface des pièces usinées dans le processus manufacturier est crucial. Ce contrôle se situe à plusieurs niveaux, particulièrement à la fin du processus.
La section est organisée comme suit : après l'introduction, la sous-sous-section (3.2.2) concerne la présentation des données, suivie de la partie consacrée à la méthodologie adoptée et la dernière partie porte sur les résultats obtenus.

3.2.2 Présentation des données

Les données proviennent de l'étude expérimentale effectuée par Noureddine *et al.* (2012). Cette étude porte sur l'impact de l'outil de coupe sur la rugosité de surface des pièces usinées. La rugosité de surface est donnée par $R_a(t)$ et dépend des conditions géométriques de l'outil de coupe associées aux conditions cinématiques du processus manufacturier, pour plus de détails voir Noureddine *et al.* (2012). Les paramètres du modèle sont estimés à partir de cinq séries de dix pièces chacune ; sur chaque lot, dix mesures sont effectuées. Ici, nous considérons seulement le signal dynamique de la surface de rugosité. L'étude expérimentale a aussi montré qu'à partir des cinq séries de dix pièces chacune, le minimum de la rugosité de surface était de $1.79\mu m$ et que toute mesure de rugosité supérieure ou égale à $4\mu m$ est une indication pour détecter une pièce défectueuse.

3.2.3 Méthodologie

Soit Y une variable aléatoire provenant d'une population donnée. Les relations (3.1.1) et (3.1.2) définies dans la section (3.1). Supposons que $y, y_2, ..., y_m$ sont les valeurs prises par la variable aléatoire Y et m la taille de la population ; les observations sont rangées dans l'ordre croissant $(y_{1:m}, y_{2:m}, ..., y_{m:m})$ avec $y_{1:m} \leq y_{2:m} \leq ... \leq y_{m:m}$.
Selon les applications, la magnitude d'une ou plusieurs statistiques d'ordres peuvent indiquer l'état, la performance ou l'intégrité d'un système et les propriétés statistiques de $y_{r:m}$ peuvent donner des indications sur la conception et les activités de maintenance ou de remplacement. L'évènement $(Y_{r:m} \leq y)$ est caractérisé par :

$$\begin{cases} f_{Y_{r:m}}(y) & = \dfrac{m!}{(r-1)!\,(m-j)!} F^{j-1}(y)\left[1-F(y)\right]^{m-r} f(y) \\ F_{Y_{r:m}}(y) & = \sum_{j=r}^{m} C_m^j F^j(y)\left[1-F(y)\right]^{m-j} \end{cases}; \tag{3.2.1}$$

avec $F_{Y_{r:m}}(y) + R_{Y_{r:m}}(y) = 1$, où R est la fonction de fiabilité et est définie par :

$$R_{Y_{r:m}}(y) = \sum_{j=1}^{r-1} C_m^j F^j(y)\left[1-F(y)\right]^{m-j}. \tag{3.2.2}$$

En particulier, pour les cas extrêmes minimum $(Y_{1:m} \leq y)$ et maximum$(Y_{m:m} \leq y)$, on a respectivement les équations suivantes :

$$\begin{cases} f_{Y_{1:m}}(y) & = m\left[1-F(y)\right]^{m-1} f(y) \\ F_{Y_{1:m}}(y) & = 1-\left[1-F(y)\right]^{m} \end{cases}. \tag{3.2.3}$$

$$\begin{cases} f_{Y_{m:m}}(y) & = mF(y)^{m-1} f(y) \\ F_{Y_{m:m}}(y) & = F^m(y) \end{cases}. \tag{3.2.4}$$

Dans le contexte de l'hypothèse de l'iid, les séries des m expériences suivent une loi binomiale avec une probabilité de réussite de $p = F(y)$. La probabilité $F_{Y_{r:m}}(y)$ associée à l'évènement $Y_{r:m} \leq y$ est facilement calculée comme la probabilité d'avoir au moins r des Y inférieures ou égales à y.
Soit $Y_{1:m} = \min(y_1, y_2, ..., y_m)$ est la plus petite statistique d'ordre et correspond à l'échantillon du minimum, $Y_{m:m} = \max(y_1, y_2, ..., y_m)$ est la plus grande statistique d'ordre et correspond à l'échantillon du maximum. Comme nous l'avions tantôt souligné, ces statistiques d'ordre sont des cas spécifiques de la théorie des valeurs extrêmes lorsque m est grand et les plus importants dans l'approche des statistiques d'ordre.
Le but de ce chapitre est de proposer un outil d'aide à la décision permettant de définir des critères de contrôle de qualité dans un processus manufacturier. Pour ce faire, deux critères sont mis en évidence : la rugosité mesurée pour chaque pièce et le nombre de fautes tolérables (mesures hors-limites) par lot de dix pièces. Pour atteindre cet objectif, différentes étapes ont été menées comme suit :

* Estimation des paramètres des distributions parents par la méthode des moments ;

* Expression de la densité et de la fonction de répartition pour les statistiques d'ordre dans chaque lot de pièces ;

* Utilisation de la fonction de répartition associée à la statistique d'ordre pour déterminer le nombre de pièces défectueuses identifiées par lot ;

* Estimation des distributions du minimum et du maximum des mesures de rugosité avec les statistiques d'ordre extrême ;

* Résultats attendus : nombre de pièces défectueuses et les meilleures valeurs de rugosité.

Dans le reste du chapitre, y désignera la valeur de la rugosité de surface.

3.2.4 Résultats

Le domaine d'attraction de Gumbel est celui qui contient le plus de lois utilisées en fiabilité. De plus, la loi de Weibull standard est beaucoup utilisée dans la modélisation de phénomènes liés à l'usure. Pour cette modélisation, nous utilisons donc la loi de Weibull à deux paramètres représentée par la relation (1.2.5).

3.2.4.1 Estimation des paramètres

Pour l'estimation des paramètres, on utilise la méthode des moments définie dans la sous-section (1.2.3). Dans cette méthode illustrée dans la sous-section (1.2.3), nous avons besoin de la valeur moyenne, l'écart-type estimé, le coefficient de variation, les estimations des paramètres de forme et d'échelle par lot. Le tableau (3.2.1) présente

les valeurs obtenues pour chaque lot. Par exemple, pour le premier lot et en utilisant la relation (1.2.5) et le tableau (3.2.1) , on obtient :

$$\begin{cases} f_{3.769,3.526}(y) & = \dfrac{3.769}{3.526}\left(\dfrac{y}{3.526}\right)^{2.769} e^{-\left(\frac{y}{3.526}\right)^{3.769}} \\ F_{3.769,3.526}(y) & = 1 - e^{-\left(\frac{y}{3.526}\right)^{3.769}} \end{cases}.$$

Lots N^0	1	2	3	4	5
Moyenne $\bar{y}$	3.138	3.079	3.151	3.110	3.110
Ecart-type estiméσ	0.925	0.923	0.938	0.905	0.898
Coefficient de Variation	0.295	0.300	0.298	0.291	0.289
Paramètre de forme k	3.769	3.700	3.728	3.826	3.855
Paramètre d'échelle λ	3.526	3.499	3.540	3.494	3.494

TABLE 3.2.1 – Paramètres de Weibull pour les cinq lots

La figure (3.2.1) donne les fonctions de densité et de répartition pour chaque lot.

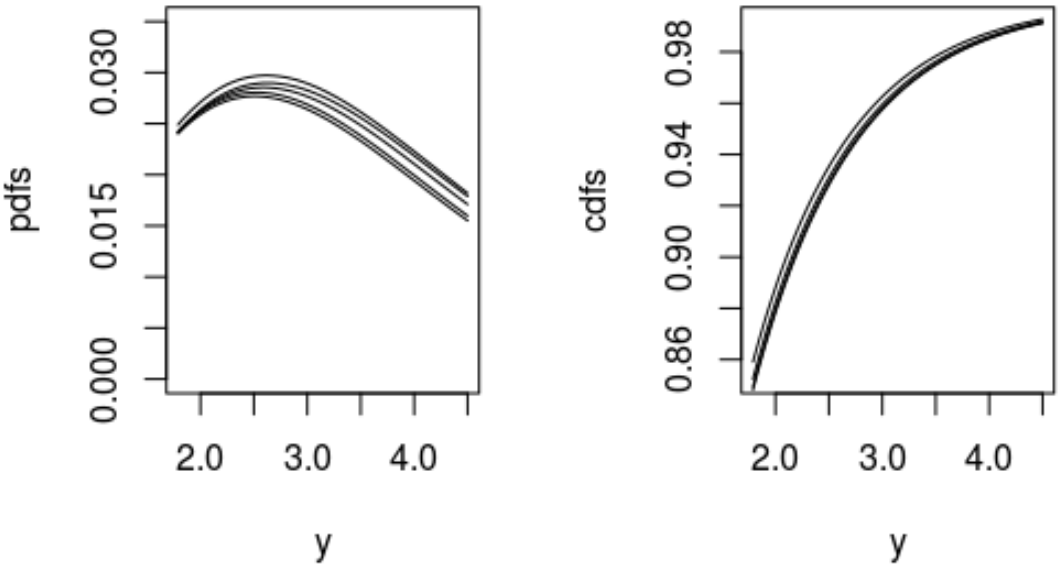


FIGURE 3.2.1 – Fonction de densité à gauche et de répartition à droite de Weibull pour les rugosités de surface des pièces usinées

3.2.4.2 Distribution des observations ordonnées pour les cinq lots

Les dix observations ont été ordonnées pour chaque lot. Les fonctions de densité et de répartition pour la r-ième statistique d'ordre de l'échantillon de taille 10 pour le premier lot. En utilisant les résultats du tableau 3.2.1, les relations (3.1.1) et (3.1.2), on obtient :

$$\begin{cases} f_{Y_{r:10}}(y) & = \dfrac{10!}{(r-1)!\,(10-j)!}\left(1-e^{-\left(\frac{y}{3.526}\right)^{3.769}}\right)^{j-1}.\left(e^{-\left(\frac{y}{3.526}\right)^{3.769}}\right)^{10-r} \\ & \quad \times \left(\frac{3.769}{3.526}\right)\left(\frac{y}{3.526}\right)^{2.769} e^{-\left(\frac{y}{3.526}\right)^{3.769}} \\ F_{Y_{r:10}}(y) & = \sum_{j=r}^{10} C_{10}^{j}\left(1-e^{-\left(\frac{y}{3.526}\right)^{3.769}}\right)^{j}.\left(e^{-\left(\frac{y}{3.526}\right)^{3.769}}\right)^{10-r} \end{cases}$$

De manière analogue, on obtient pour les autres lots de mesure les fonctions de densité et de répartition pour la r-ième statistique d'ordre.
A partir des résultats décrits dans Noureddine *et al.* (2012), nous supposons qu'une pièce est défectueuse lorsque $y > 4\mu$m. Il est aussi montré que l'origine de la faute est souvent due à une micro-cassure au niveau du bec de l'outil de coupe. Cette faute peut survenir à tous les niveaux du processus manufacturier. Si la faute arrive au début de l'opération, les effets du bris du bec peuvent diminuer avec le phénomène de l'auto-affutage. Cette situation est bien différente de l'usure classique de l'outil de coupe où l'évolution du comportement de l'outil de coupe peut être étudiée. Ici, nous modélisons la fiabilité de l'outil de coupe qui est directement liée aux conditions de coupe intégrant le bris de l'outil.
La probabilité d'avoir la dixième mesure inférieure ou égale à 4μm est $F_{Y_{10:10}}(4)$ et la fiabilité correspondante associée à l'outil de coupe est $R_{Y_{10:10}}(4)$; cette fiabilité indique qu'il y a entre 11 et 13% de chance d'avoir au moins une mauvaise pièce. Le tableau (3.2.2) présente les valeurs pour chaque lot.

Lots N^0	1	2	3	4	5
$F_{Y_{10:10}}(4)$	0.87	0.86	0.86	0.88	0.89
$R_{Y_{10:10}}(4)$	0.13	0.14	0.14	0.12	0.11

TABLE 3.2.2 – Probabilité d'avoir la dixième mesure inférieure ou égale à 4μm et fiabilité correspondante associée à l'outil de coupe

Considérons le cas de $F_{Y_{9:10}}(4)$ qui indique la probabilité d'avoir au moins huit bonnes pièces. Le tableau 3.2.3 montre que $F_{Y_{9:10}}(4) \simeq 1$, ceci signifie qu'il est quasiment certain d'avoir au moins huit bonnes pièces par lot. La fiabilité correspondante associée à l'outil de coupe est $R_{Y_{9:10}}(4)$; elle indique qu'il y a 1% de chance d'avoir plus de deux pièces défectueuses dans chaque lot. En outre, la magnitude de la mesure ($y = 4\mu$m) est assez indicative dans le comportement de l'outil de coupe sur la rugosité de surface des pièces usinées.

Lots N^0	1	2	3	4	5
$F_{Y_{9:10}}(4)$	0.99	0.99	0.99	0.99	0.99
$R_{Y_{9:10}}(4)$	0.01	0.01	0.01	0.01	0.01

TABLE 3.2.3 – Probabilité d'obtention d'au moins huit mesures inférieures ou égales à 4μm et fiabilité correspondante associée à l'outil de coupe

3.2.4.3 Distribution de minima et maxima

Dans cette modélisation, la distribution de la première statistique d'ordre est celle du minimum et correspond donc à la meilleure qualité obtenue pour les pièces. La fonction de densité du minimum est obtenue à partir de la relation (3.2.3) en posant $r = 1$; par exemple pour le premier lot on a :

$$f_{Y_{1:10}}(y) = 10.f_{3.769,3.526}(y).\left[1 - F_{3.769,3.526}(y)\right]^9.$$

La fonction de densité $f_{Y_{1:10}}(y)$ pour les cinq lots est donnée sur la figure (3.2.2). Les courbes dans les différentes couleurs représentent la distribution du minimum pour chaque lot.

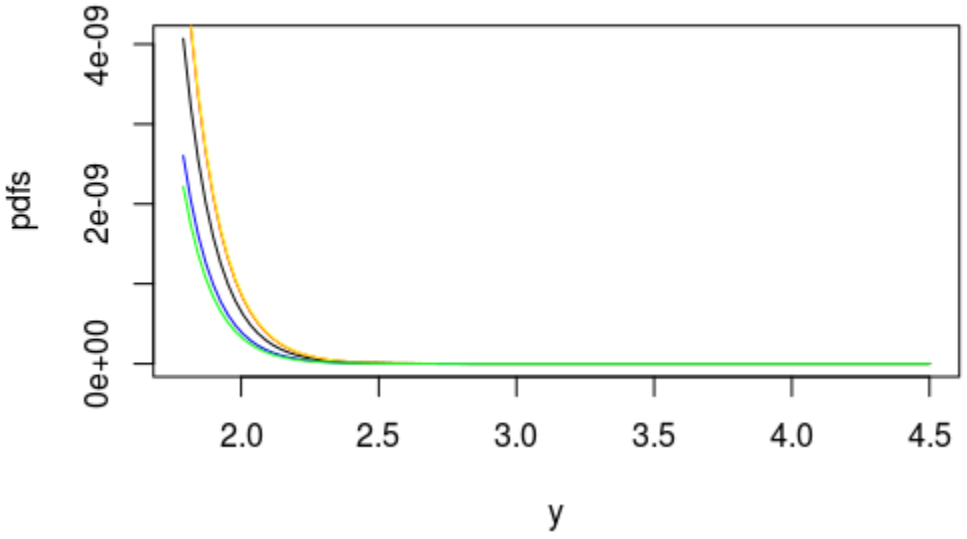


FIGURE 3.2.2 – Distribution des minimums pour les cinq séries de mesure

On constate que sur la figure (3.2.2), les valeurs élevées de $f_{Y_{1:10}}(y)$ sont obtenues pour les petites valeurs de rugosité de surface et les meilleures valeurs de rugosité de surface sont inférieures ou égales à $2.25\,\mu m$. Ainsi, on peut imposer des critères de qualité dans la ligne de production.

Dans cette même logique, on peut étudier la fonction de densité des maxima en utilisant la relation (3.2.4) et en posant $r = m$, on obtient :

$$f_{y_{10:10}}(y) = 10.f_{3.769,3.526}(y).[F(y)]^9,$$

La figure (3.2.3) confirme le fait que les valeurs élevées de $f_{Y_{10:10}}(y)$ sont au-dessus de $4\,\mu m$. C'est une indication de pièces défectueuses dans la ligne de production.

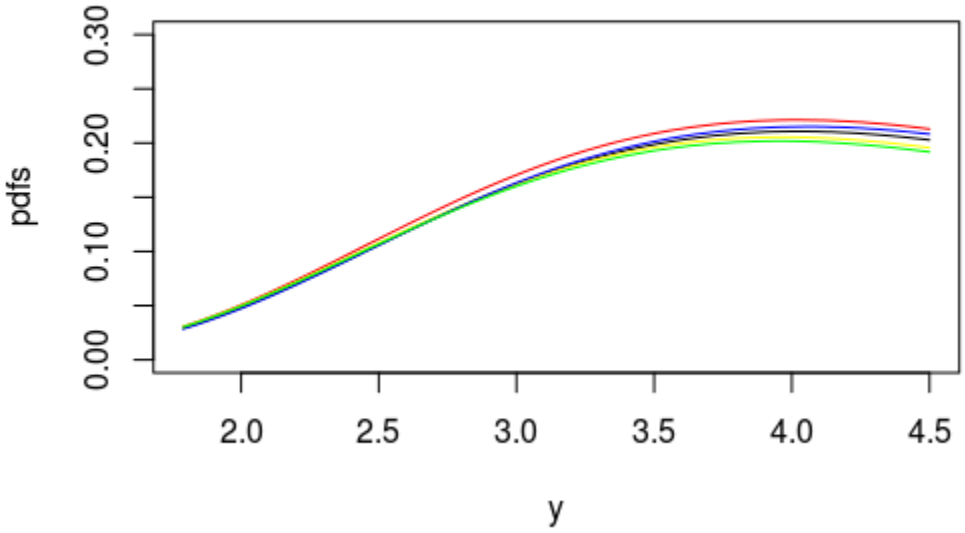


FIGURE 3.2.3 – Distribution des maximums pour les cinq séries de mesure

3.2.4.4 Diagramme des différentes étapes de la procédure

Nous proposons un diagramme qui illustre les étapes pratiques de notre approche ; la méthodologie proposée est composée des étapes suivantes :

1. Analyse des données : elle concerne l'analyse des données brutes provenant des composantes industrielles ;
2. Estimation des paramètres par la méthode des moments ;
3. Calculs : cette étape porte sur l'utilisation de la fonction de répartition des statistiques d'ordre comme outil d'aide à la décision dans la détermination du nombre de pièces défectueuses par lot.

En somme, nous avons les meilleures valeurs de rugosité de surface, le nombre de pièces défectueuses et les mesures de rugosité correspondantes dans chaque lot.

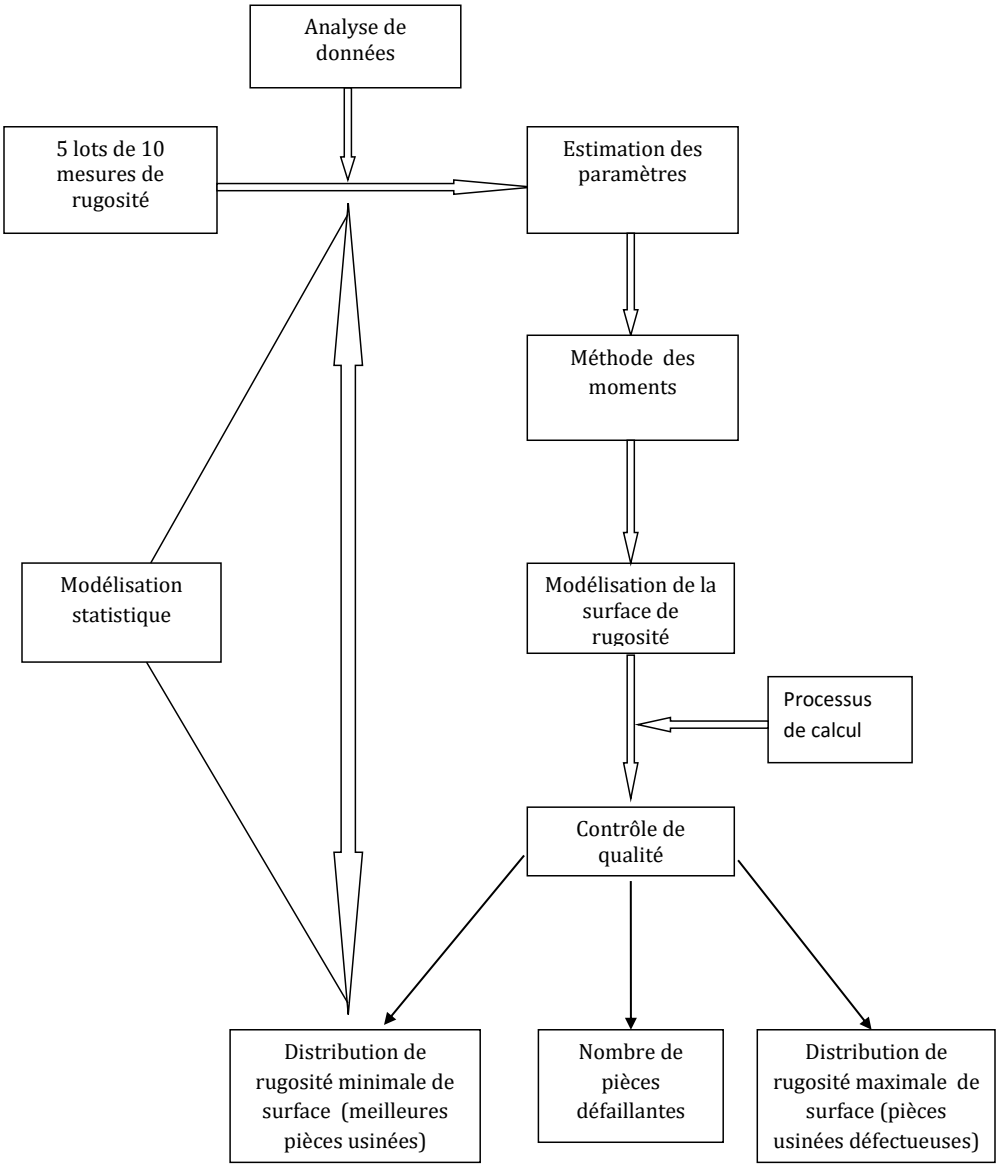


FIGURE 3.2.4 – Diagramme des différentes étapes

3.3 Conclusion

Dans le domaine de la production industrielle, les exigences du cahier de charges incluent les critères de qualité attendus de la production en ligne. Pour satisfaire les exigences de qualité, il est indispensable de faire des tests avant la production en masse. Pour atteindre cet objectif, nous avons considéré une situation expérimentale dans un contexte de rugosité avec des conditions spécifiques de l'outil de coupe. Nous avons donc procédé à l'usinage de cinq lots de dix pièces chacun. Les deux critères de qualité retenus sont : la valeur de rugosité minimale et le nombre de pièces défectueuses par lot. Une pièce est dite mauvaise lorsque sa rugosité est supérieure à $4\,\mu m$. Il a aussi été montré que les données de rugosité utilisées suivaient une distribution de Weibull avec un paramètre de forme $k > 1$, ce qui traduit une situation de dégradation. Les statistiques d'ordre extrêmes dans la rugosité des surfaces sont appliquées pour mettre en évidence les deux critères de qualité considérés. En utilisant les calculs associés aux statistiques d'ordre, nous avons montré deux résultats importants à savoir le nombre de pièces défectueuses tolérables par lot (deux pièces) et la distribution du minimum donne le meilleur intervalle pour la rugosité des surfaces $[1.79, 2.25]$. Il en résulte que la fiabilité de l'outil de coupe n'est pas satisfaisante autour de $4\,\mu m$.

Chapitre 4

Modélisation des extrêmes pour la gestion des risques d'accidents dans l'aéronautique

4.1 Gestion des risques d'accidents dans l'aviation

Aujourd'hui, avec les avancées de la technologie, les accidents d'avion ont considérablement baissé partout dans le monde. Cependant, il peut arriver qu'un seul accident cause à lui seul autant de victimes que tous les autres accidents sur une échelle annuelle ; on parle alors d'accident extrême. Ainsi, la gestion des risques d'accidents peut se situer à deux niveaux principaux. Le premier concerne l'état de l'avion avant le vol et les conditions de vol, le second concerne la gestion d'un accident à savoir les questions d'assurances relatives aux victimes, les motifs du crash, etc. Dans ce travail, notre approche de gestion de risque concerne plutôt le second niveau qui est essentiel pour les compagnies de transport dans les problèmes de capacité d'anticipation et d'assurance.

4.2 Modélisation dans l'aviation par la méthode POT et application aux données de l'Aviation Générale Américaine

4.2.1 Introduction

L'Aviation Générale (AG) est un terme relatif aux opérations de l'aviation civile autres que les vols réguliers et non réguliers voir pour plus de détails Crane (1997) et National Transportation Statistics NTS (2016). Elle est la principale composante de l'aviation civile dans beaucoup de pays développés tels que les États-Unis d'Amérique (U.S.A), l'Australie, la Nouvelle-Zélande, etc. En pratique, les opérations de l'AG concernent les parachutes, les jets privés, certains petits avions, voir National

Transportation Safety Board NTSB (2011) et Civil Aviation Authority CAA (2013). La majorité du trafic aérien mondial se situe dans cette catégorie et la plupart des aéroports du monde sont destinés à l'aviation générale, voir Crane (1997). Le secteur de l'AG regroupe aussi toutes les activités de l'aviation qui ne sont pas prises en compte par l'aviation commerciale au service des passagers et marchandises ou les opérations militaires. Cependant, pour certaines raisons de sécurité et de conditions d'exploitation aérienne, certains vols commerciaux peuvent se retrouver dans l'Aviation Générale.

Entre le premier Janvier 2015 et le 17 Mai 2016, on a comptabilisé dans l'Aviation Générale Américaine (AGA) un total de 1546 accidents qui ont causé 466 décès et 384 blessés selon les données de la NTSB. En 2011, l'AGA était responsable de 95% des accidents d'avions et 94% des accidents fatals. De plus 92% des accidents de l'aviation civile américaine sont imputables à l'AGA, voir NTSB (2011) ; de ce fait, elle est une composante importante de l'aviation civile américaine, voir Shetty et Hansman (2012).

Par ailleurs, l'AGA couvre aussi un large éventail d'activités commerciales et non commerciales, y compris les aéroclubs, les cours de pilotage, l'aviation d'agriculture, les aéronefs de faible tonnage et la maintenance. Elle est notamment très importante en Amérique du Nord, avec plus de 6300 aéroports utilisés aux (U.S.A), voir NTSB (2011) et (NTS, 2016).

Entre 1990 et 2014, il y a eu aux (U.S.A) un total de 8402 accidents fatals qui ont causé au total 14984 décès selon les données de la NTSB. Parmi ces 8402 accidents fatals, environ 26 accidents peuvent être considérés comme extrêmes et chacun d'eux était responsable d'au moins 39 décès. De plus, dans l'approche des valeurs extrêmes, un évènement extrême est un évènement qui a une faible probabilité, mais implique de sérieuses conséquences. Dans notre contexte, un accident extrême est un accident qui aura causé au moins 39 décès. Selon NTS (2016) et Shetty et Hansman (2012), il y a une tendance globale à la baisse du nombre de décès annuels. Cependant, ceci n'est pas le cas pour les accidents extrêmes, car un tel accident peut causer plus de victimes que le nombre total de décès à l'échelle annuelle.

La prédiction et la gestion des accidents extrêmes sont très importantes pour les assurances et les industries aéronautiques telles que Boeing et Airbus, voir Boeing (2016) et Airbus (2016). De ce fait, en utilisant les données historiques, l'estimation du niveau de retour de décès dus aux accidents extrêmes peut être un outil d'aide à la décision aux organismes de gestion de l'AG afin d'améliorer les futures opérations, les maintenances, les conditions de vols pour éviter dans la mesure du possible de tels accidents.

Dans la littérature, certains auteurs ont travaillé sur les accidents d'avion ; Sobieralski (2012) propose des travaux sur le coût des accidents dans l'AGA, les travaux sur la prédiction des accidents dus aux heures de vols des pilotes ont été proposés par Knetch (2015). Quant à Fultz et Ashley (2016), ils étudient l'impact des conditions météorologiques extrêmes dans l'aviation générale américaine, Farah et Azevedo (1952) travaillent sur l'analyse de la sécurité dans la navigation aérienne en utilisant la théorie des valeurs extrêmes. Sha *et al.* (2017) utilisent aussi la théorie des valeurs

extrêmes pour la prévision des pièces de rechange dans l'aéronautique en se basant sur les données historiques des stocks. Dey et Das (2016) proposent une approche pour la prédiction des décès dus aux blessures extrêmes à la suite d'accidents d'avion. Mais dans la littérature, il est rare de voir des travaux concernant la modélisation des décès dus aux accidents extrêmes dans l'aviation générale américaine en utilisant la théorie des valeurs extrêmes ; notamment l'approche du « Peaks-Over-Threshold, (POT) » en anglais et les excès au-dessus du seuil en français.
Nous pouvons donc modéliser le niveau de retour du nombre de décès dus aux accidents extrêmes dans l'aviation générale américaine à partir de la théorie des valeurs extrêmes. On distingue principalement deux approches dans la modélisation des extrêmes, pour plus de détails voir les sections (1.2) et (1.3). L'approche des blocs de maxima est inutile lorsqu'on a une seule donnée par bloc ; la deuxième grande valeur d'un bloc peut être plus grande que la plus grande valeur d'un autre bloc et ceci n'est pas en général comptabilisé. L'approche POT permet de pallier ces insuffisances en utilisant la distribution généralisée de Pareto (GPD). En effet, les progrès de la théorie des valeurs extrêmes ont mis en évidence l'émergence de la distribution généralisée de Pareto comme la famille naturelle pour la modélisation des excédents au-dessus d'un seuil fixé. L'approche POT a montré son importance et ses succès dans un grand nombre de problèmes d'analyses statistiques relatifs à de nombreux domaines, voir Schmidt *et al.* (2014) et Castillo *et al.* (2005). Cependant, en dépit des avancées théoriques et de beaucoup de situations d'applicabilité, la modélisation de cette distribution en pratique n'est pas triviale.
Dans ce chapitre, nous nous intéressons aux accidents fatals dans l'AGA entre 1990 et 2014 et au nombre de victimes en résultant. Une prédiction du nombre possible de victimes pour un accident extrême dans les futures opérations est faite à travers l'approche POT en utilisant la GPD. Nous utilisons trois méthodes d'estimation des paramètres et choisissons par la suite la méthode dont les erreurs standards sur les paramètres estimés sont les plus faibles. Dans cette approche, les deux principaux facteurs qui affectent l'exactitude des estimations sont le choix du seuil adéquat d'une part et la méthode d'estimation d'autre part. Une discussion sur les différentes méthodes d'estimation a été faite par Bermudez et Kotz (2010) ; quant à Scarrot et Macdonald (2012), ils ont proposé une synthèse approfondie des méthodes de recherche du seuil. Ici, nous utilisons une approche graphique et une méthode d'estimation intensive des paramètres de la GPD, pour obtenir un intervalle raisonnable de seuil ; nous quantifions aussi les incertitudes dans l'inférence non-paramétrique, pour plus de détails voir Carpenter et Bithel (2000) et Yongcheng (2008).
Le reste de la section (4.2) est structuré comme suit : après l'introduction, nous présentons les données dans la sous-section (4.2.2), la sous-section (4.2.3) présente la méthodologie utilisée et la sous-section (4.2.4) concerne les résultats obtenus.

4.2.2 Présentation des données

Les données utilisées dans cette modélisation proviennent de l'Agence de Sureté de l'Aviation Générale Américaine qui est une division de la NTSB. Ces données

contiennent des informations d'accidents d'avions relatives aux USA et aux eaux internationales dans leur voisinage. Les données sont constituées de 8402 accidents fatals de l'aviation générale entre 1990 et 2014. Ces accidents ont causé au total 14984 décès. Parmi ces accidents, 26 peuvent être considérés comme extrêmes. De façon générale, on constate une diminution du nombre de victimes annuellement comme l'illustre la figure (4.2.1).

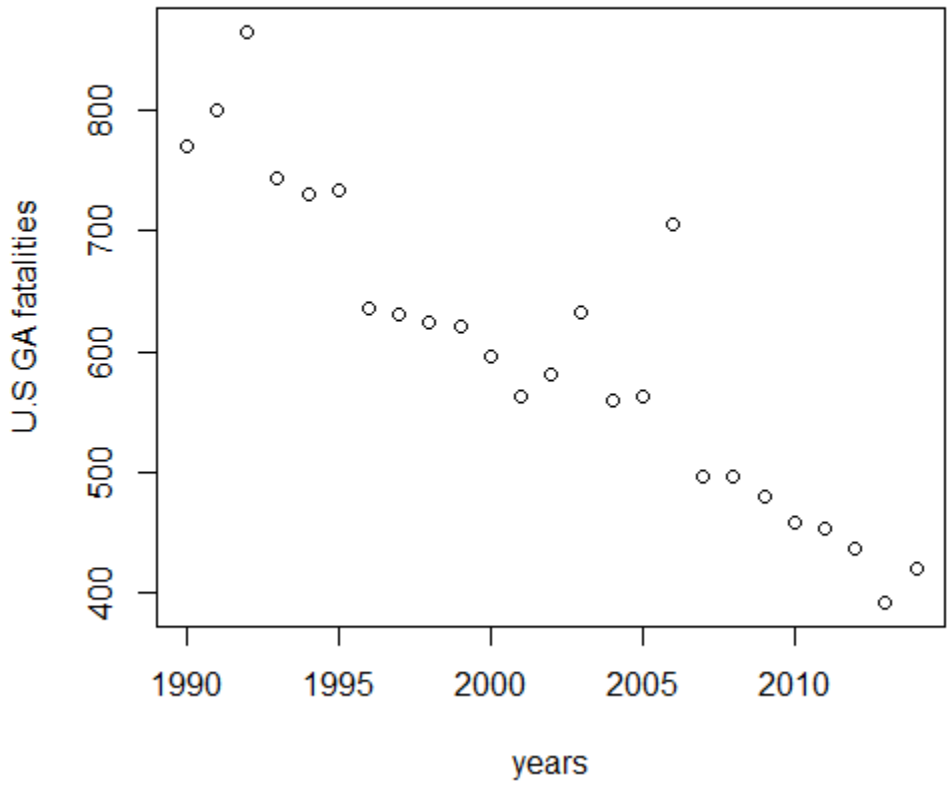


FIGURE 4.2.1 – Victimes annuelles

La figure 4.2.2 montre une concentration de valeurs entre 1 et 120 et quelques valeurs isolées ; ceci confime la présence d'accidents extrêmes dans les données et on remarque aussi que la plupart des accidents ont fait moins de 39 décès.

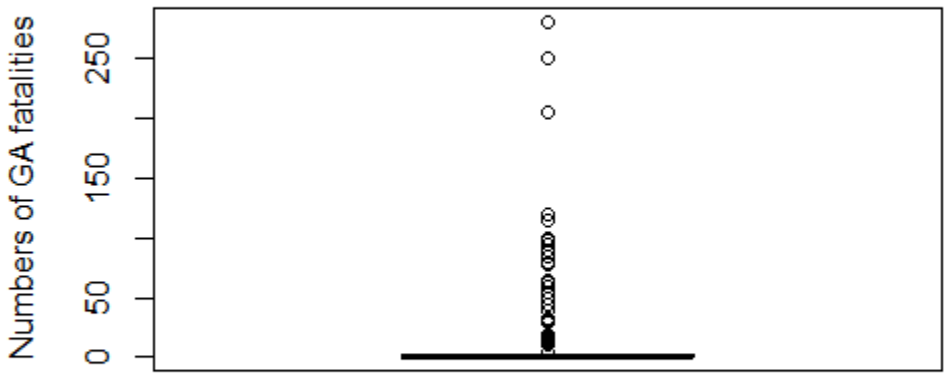


FIGURE 4.2.2 – Concentration des accidents de l'Aviation Générale Américaine

Les figures (4.2.1) et (4.2.2) montrent la nuance qu'il y a entre la tendance à la déclinaison pour les victimes annuelles et le nombre de victimes pouvant résulter d'un accident extrême.

4.2.3 Méthodologie

L'objectif de cette modélisation est de développer un modèle prédictif qui pourra fournir les estimations de la loi du nombre de victimes possibles pour un accident extrême dans les futures opérations de l'Aviation Générale Américaine.
Dans ce travail, une approche graphique et une autre, dite méthode d'estimation intensive des paramètres de la GPD, sont utilisées, pour obtenir un intervalle raisonnable du seuil. Nous quantifions les incertitudes dans les inférences en utilisant l'approche du bootstrap non paramétrique, voir la sous-section (1.4.4) et pour plus de détails voir Carpenter et Bithel (2000) et Yongcheng (2008).

4.2.4 Résultats

4.2.4.1 Détermination graphique du seuil

Pour plus de détails sur le « Mean Residual Life Plot », voir la sous-sous-section (1.3.2.1). L'intervalle de seuil raisonnable est celui où il débute une tendance linéaire sur le graphique. Les deux graphiques « Meplot et Mrlplot », donnés en figure (4.2.3) jouent le même rôle mais présentent la linéarité sous deux formes différentes

et permettent de guider le choix de l'intervalle de seuil optimal. De ce fait, La figure (4.2.3) montre une certaine stabilité et linéarité dans l'intervalle $[40, 70]$. Ainsi, le seuil raisonnable doit se situer dans cet intervalle.

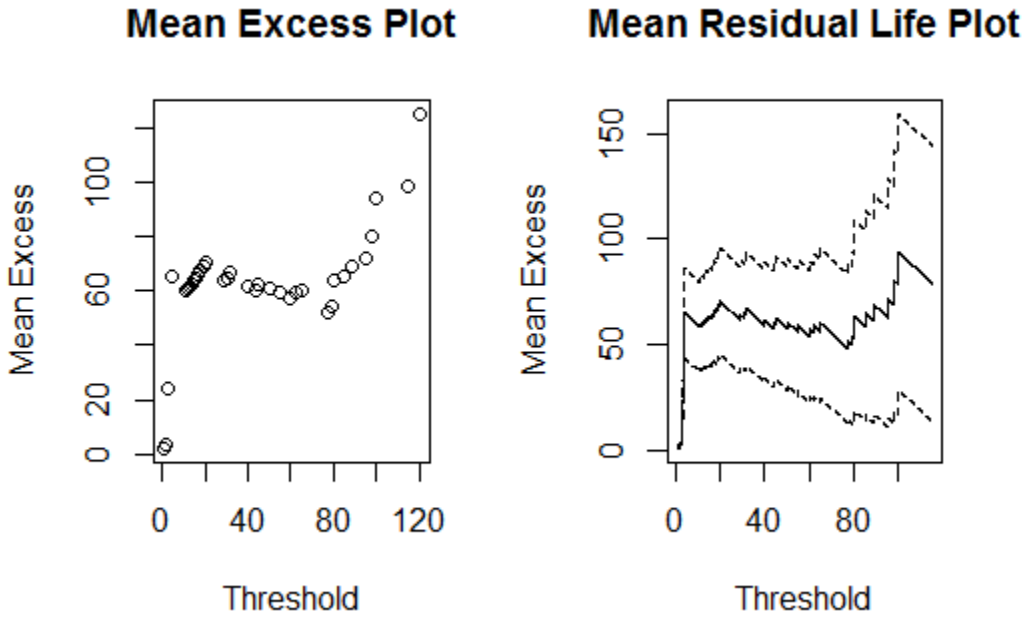


FIGURE 4.2.3 – Fonction moyenne des excès

4.2.4.2 Graphe de stabilité des paramètres de forme et d'échelle

La figure (4.2.4) permet de guider le choix de l'intervalle adéquat en donnant des estimations du paramètre d'échelle à gauche et celui de forme à droite ; sur cette figure on remarque que la stabilité est plus forte sur l'intervalle $[40, 50]$. En utilisant les figures (4.2.3) et (4.2.4), on peut donc réduire l'intervalle du seuil à l'intervalle $[40, 50]$. Ainsi, la partie linéaire du graphique permet de guider le choix de l'intervalle adéquat pour la stabilité des paramètres de la GPD.

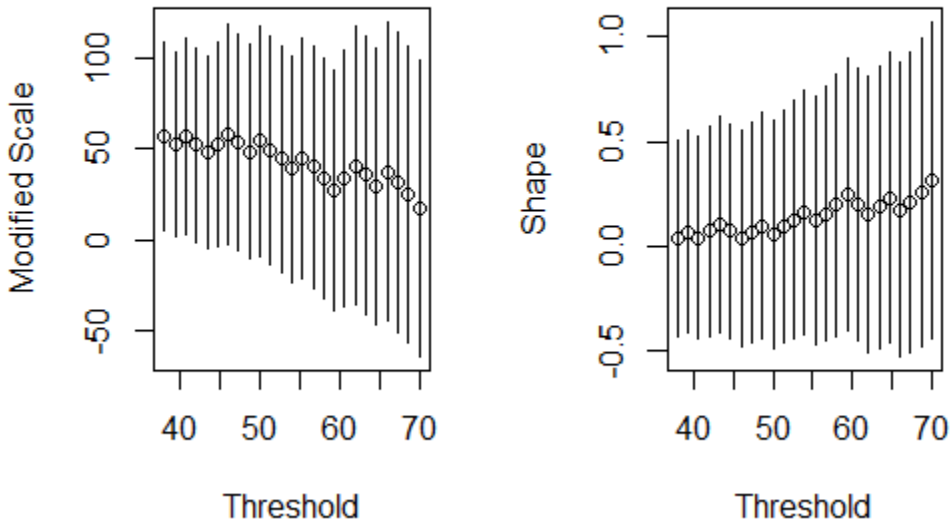


FIGURE 4.2.4 – Graphe de stabilité des paramètres de forme et d'échelle

4.2.4.3 Estimation des paramètres de la GPD pour différents seuils

Dans cette partie, en utilisant le logiciel R et les méthodes du maximum de vraisemblance, des moments et des probabilités de moments pondérés et la tranche de 40-50, nous déterminons des estimations des paramètres de forme ξ, d'échelle σ et les erreurs standards associées. Pour plus de détails voir la sous-section (1.3.3).

a) Méthode du maximum de vraisemblance pour différents seuils

u	$\widehat{\xi}$	$\widehat{\sigma}$	$se\left(\widehat{\xi}\right)$	$se\left(\widehat{\sigma}\right)$
40	0.0255	60.2172	0.2449	19.8103
41	0.0465	57.9678	0.2492	19.2504
42	0.0678	55.7567	0.2537	18.7157
43	0.0903	53.5635	0.2592	18.2129
44	0.0534	57.4497	0.2588	19.7149
45	0.0113	62.1133	0.2600	21.6100
46	0.0328	59.7771	0.2634	20.9387
47	0.0533	57.5980	0.2670	20.3540
48	0.0763	55.2884	0.2714	19.7162
49	0.1035	52.7837	0.2772	19.0017
50	0.0566	57.5925	0.2780	21.0430

TABLE 4.2.1 – Estimation des paramètres par la méthode du maximum de vraisemblance

L'analyse du tableau (4.2.1) montre une certaine instabilité dans l'estimation du paramètre de forme. Dans les tranches 40-43 et 45-49, le paramètre de forme augmente et il en est de même pour les erreurs standards, pour les seuils 45 et 50, nous avons des valeurs isolées. On remarque aussi que la méthode du maximum de vraisemblance est sensible au choix du seuil.

b) Méthode des moments pour différents seuils

u	$\widehat{\xi}$	$\widehat{\sigma}$	$se\left(\widehat{\xi}\right)$	$se\left(\widehat{\sigma}\right)$
40	0.0458	58.7971	0.2325	18.7472
41	0.0604	56.9565	0.2392	18.4507
42	0.0748	55.1594	0.2473	18.2158
43	0.0890	53.4050	0.2568	18.0524
44	0.0659	56.5108	0.2481	18.8886
45	0.0378	60.2645	0.2410	20.0530
46	0.0524	58.3999	0.2475	19.7094
47	0.0668	56.5788	0.2551	19.4256
48	0.0810	54.8007	0.2641	19.2103
49	0.0949	53.0647	0.2748	19.0751
50	0.0684	56.6743	0.2630	20.0320

TABLE 4.2.2 – Estimation des paramètres par la méthode des moments

Dans la méthode des moments, le paramètre de forme est positif. Ceci montre que le support de la loi n'est pas borné. L'analyse des résultats traduit presque la même situation que dans la méthode du maximum de vraisemblance, mais les erreurs associées sont moins élevées.

c) Méthode des probabilités de moments pondérés pour différents seuils

u	$\widehat{\xi}$	$\widehat{\sigma}$	$se\left(\widehat{\xi}\right)$	$se\left(\widehat{\sigma}\right)$
40	0.0486	58.6268	0.2509	19.5998
41	0.0802	55.7556	0.2513	18.6905
42	0.1119	52.9478	0.2528	17.8111
43	0.1436	50.2032	0.2554	16.9630
44	0.0924	54.9078	0.2579	18.8842
45	0.0279	60.8787	0.2640	21.3670
46	0.0594	57.8661	0.2638	20.3908
47	0.0909	55.1166	0.2646	19.4454
48	0.1224	52.3300	0.2665	18.5319
49	0.1539	49.6064	0.2698	17.6514
50	0.0940	55.1094	0.2720	19.9820

TABLE 4.2.3 – Estimation des paramètres pour la méthode des probabilités de moments pondérés

Le tableau (4.2.3) donne les valeurs du paramètre de forme et les erreurs standards associées. On remarque des variations semblables mais avec des erreurs standards associées plus grandes que dans les deux autres méthodes. On constate aussi que toutes ces méthodes sont sensibles au choix du seuil. Cependant, l'analyse globale des estimations dans les trois méthodes montre que les méthodes du maximum de vraisemblance et des moments sont moins sensibles aux variations du seuil que celle des probabilités de moments pondérés. Pour les estimations du niveau de retour, nous allons utiliser les méthodes du maximum de vraisemblance et des moments.

4.2.4.4 Estimation du niveau de retour

Dans cette sous-section, nous utilisons les estimations des paramètres obtenues dans les tableaux (4.2.1) et (4.2.2) pour les seuils 40, 45 et 50, afin de déterminer le niveau de retour des victimes sur des périodes de 5, 10 et 25 ans. En remplaçant dans la relation (1.3.12) les estimations des paramètres et les périodes de retour correspondantes, nous obtenons le tableau (4.2.4)

5 ans			10 ans			25 ans		
μ	vict.(MV)	vict.(MM)	μ	vict.(MV)	vict.(MM)	μ	vict.(MV)	vict.(MM)
40	133	132	40	177	177	40	236	238
45	140	139	45	184	183	45	243	245
50	141	141	50	186	185	50	247	248

TABLE 4.2.4 – Estimation possible de quantiles pouvant être attendus dans 5, 10 et 25 ans

Le tableau 4.2.4 résume le niveau de retour ou quantile extrême pour différentes périodes de retour avec les deux méthodes d'estimation retenues. L'analyse des résultats obtenus dans le tableau 4.2.4 montre que pour chaque période de retour, les niveaux de retour estimés sont presque les mêmes pour les deux méthodes d'estimation retenues et pour différents seuils. Cependant, du seuil 45 au seuil 50, nous remarquons qu'il y a un léger changement au niveau du nombre de décès ; ceci signifie que l'intervalle optimal pour le seuil est l'intervalle $[45, 50]$. Par exemple, pour la méthode des moments, une période de retour $T = 5$ ans et un seuil $u = 45$, le niveau de retour est de 139 décès ; ceci signifie que tous les 5 ans, on peut s'attendre au moins à 139 victimes à la suite d'un accident extrême. Une autre interprétation avec l'année de référence 2014 est que pour chaque période de retour de 5 ans après 2014, il est possible d'avoir 139 victimes à la suite d'un accident extrême. Pour ces prédictions, nous avons supposé que les conditions de vol, la technologie, le volume du trafic permettent de maintenir le modèle utilisé. Cependant, vu les avancées de la technologie dans le domaine de l'aviation, les conditions d'opérations sont amenées à s'améliorer avec le temps. Ainsi, la prise en compte de ces changements dans les modèles mathématiques pourrait améliorer la prédiction.

4.2.4.5 Quantification des incertitudes

Dans le domaine des assurances et la gestion des risques, la gestion et la prédiction d'évènements à risque extrême tels que les victimes d'accidents d'avion est très importante. Cependant, la plupart des modèles mathématiques ne peuvent pas représenter parfaitement la réalité des phénomènes. Les adéquations du modèle et l'inférence des paramètres sont très souvent influencées par des incertitudes. D'où la prise en compte de ces incertitudes s'avèrent nécessaires lorsque nous faisons des extrapolations sur des données observées.
Dans la littérature, plusieurs méthodes ont été proposées pour l'estimation des différentes sources d'incertitudes dans la statistique des valeurs extrêmes, pour plus de détails voir, Wehner (2010), Scarrot et Macdonald (2012), Zhang *et al.* (2015) et Wendy et Noriszura (2012). Dans la suite du travail, nous nous intéressons aux incertitudes relatives à l'adéquation du modèle et des paramètres dues à la sélection du seuil.

a) Incertitudes sur l'adéquation du modèle

Nous utilisons le (P-P plot) et (Q-Q plot) pour tester l'adéquation du modèle GPD, pour plus de détails voir la sous-section (1.4.2). Comme nous l'avions indiqué dans les sections précédentes, les méthodes des moments et maximum de vraisemblance ont été utilisées au détriment des probabilités de moments pondérés car les erreurs standards sur le paramètre de forme sont plus faibles pour les deux premières méthodes que dans la troisième citée. Pour ce faire, nous utilisons l'approche graphique pour quantifier les incertitudes sur l'adéquation du modèle avec les deux méthodes et les seuils 40, 45 et 50.

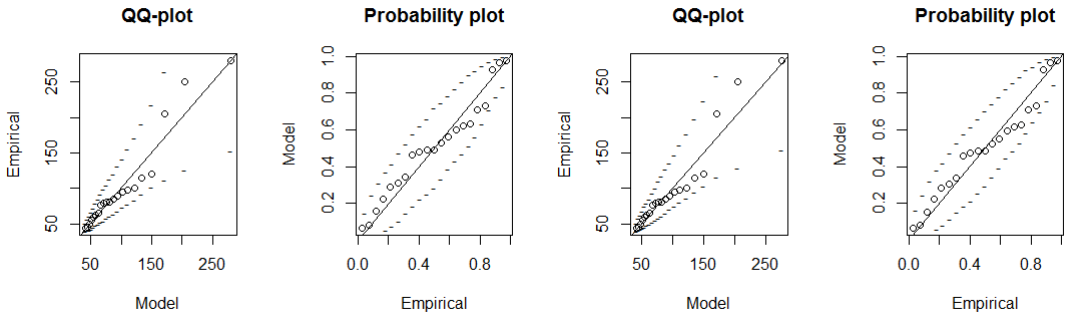


FIGURE 4.2.5 – QQ-plot et PP-plot par MV (droite) et MM (gauche) avec le seuil $u = 40$

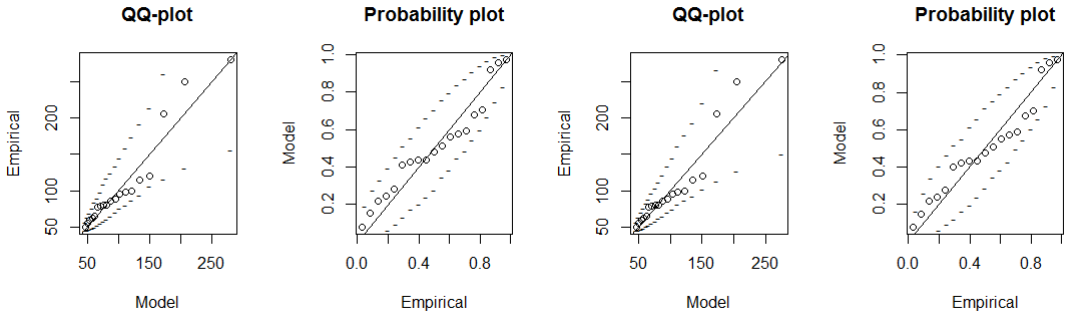


FIGURE 4.2.6 – QQ-plot et PP-plot par MV (droite) et MM (gauche) avec le seuil $u = 45$

Les figures (4.2.5) et (4.2.6) illustrent le QQ-plot et le PP-plot pour les seuils 40 et 45 avec les deux méthodes retenues. Dans la méthode (MV), pour les seuils 40 et 45, malgré une grande similitude globale, nous constatons néanmoins une légère déviation du QQ-plot par rapport à la première bissectrice ; tandis que dans la méthode (MM), le QQ-plot suit exactement la première bissectrice, il en est de même pour le PP-plot dans les deux cas.

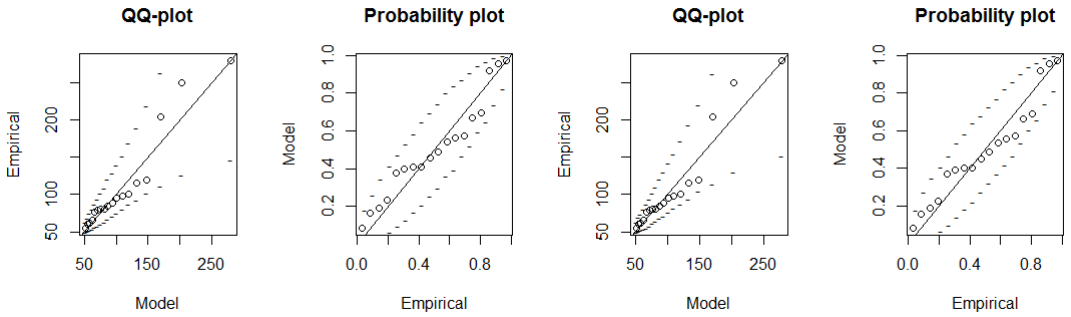


FIGURE 4.2.7 – QQ-plot et PP-plot par MV (droite) et MM (gauche) avec le seuil $u = 50$

La figure (4.2.7) montre que le QQ-plot et le PP-plot suivent la première bissectrice dans chacune des deux méthodes.
La synthèse de l'analyse graphique des (4.2.5), (4.2.6) et (4.2.7) montre que le QQ-plot et le PP-plot pour la méthode des moments suivent exactement la première bissectrice ; ceci signifie que la méthode des moments est plus adéquate dans cette modélisation. Par la suite, on utilisera la méthode des moments pour les estimations.

b) Incertitudes sur les méthodes d'estimation des paramètres

La principale source d'incertitude dans la modélisation des évènements extrêmes tels que les décès dus aux accidents d'avions avec la GPD est le choix du seuil. Dans ce travail, nous utilisons une méthode graphique et une estimation intensive des paramètres avec trois méthodes d'estimation pour sélectionner l'intervalle du seuil et ainsi obtenir le seuil recherché. On distingue plusieurs outils de mesure de l'incertitude sur les paramètres de la GPD tels que le coefficient de variation (CV), l'erreur standard (se) et les intervalles de confiance des paramètres estimés. Dans notre approche, pour chaque seuil sélectionné, l'erreur standard des paramètres de forme et d'échelle est calculée. Les résultats obtenus suggèrent que les méthodes des moments et du maximum de vraisemblance sont plus adéquats pour cette modélisation de GPD que celle des probabilités de moments pondérés.
Dans la sous-section précédente, le tableau (4.2.4) récapitule les prédictions pour différentes périodes de retour et les deux méthodes retenues. Ici, dans l'objectif de quantifier les incertitudes sur les estimations des paramètres et des quantiles extrêmes, nous nous intéressons aux erreurs standards, aux coefficients de variation, aux intervalles de confiance. Cependant l'intervalle de confiance à 95% pour le niveau de retour x_T dépend de la normalité asymptotique de $\widehat{x_T}$. Cette normalité asymptotique n'est pas toujours vérifiée en pratique pour des données à faible observation, pour plus de détails voir Carpenter et Bithel (2000) et Naess et Clausen (2001).
Nous voulons construire un intervalle de confiance sans prendre en compte la normalité asymptotique ; de ce fait, nous utilisons l'approche non paramétrique du bootstrap qui ne fait aucune hypothèse sur le modèle ou la distribution. Notre base de

données est considérée comme un vecteur z_{obs} de n observations indépendantes, voir la sous-section 1.4.4 pour l'algorithme du bootstrap.
Dans notre simulation, $B = 1000$, les statistiques d'intérêt sont : le niveau de retour x_T, le paramètre de forme ξ et d'échelle σ. La procédure est repétée 1000 fois ; pour chaque période de retour (T), nous avons 1000 valeurs pour le niveau de retour x_T. En utilisant cette base de mille valeurs, le niveau de retour de décès moyen, l'erreur standard, le coefficient de variation, le minimum et le maximum de victimes possibles sont calculés pour les seuils 40, 45 et 50.
Les résultats obtenus sont dans les tableaux (4.2.5), (4.2.6) et (4.2.7).

Seuil 40					
T	vict.moy	S.E	C.V	vict.min	vict.max
5	131	21.32474	16.23269	81	207
10	174	29.47477	16.92119	92	273
25	227	44.64727	19.65789	100	343

TABLE 4.2.5 – Récapitulatif des statistiques par bootstrap pour le seuil 40

Seuil 45					
T	vict.moy	S.E	C.V	vict.min	vict.max
5	137	21.32474	16.23269	91	224
10	179	33.2863	18.65247	94	289
25	231	45.7403	19.76416	94	362

TABLE 4.2.6 – Récapitulatif des statistiques par bootstrap pour le seuil 45

Seuil 50					
T	vict.moy	S.E	C.V	vict.min	vict.max
5	137	22.11979	16.114	86	247
10	179	32.6103	18.25288	94	278
25	232	47.86963	20.63656	93	359

Table 4.2.7 – Récapitulatif des statistiques par bootstrap pour le seuil 50

L'analyse des résultats des tableaux (4.2.5), (4.2.6) et (4.2.7) montre que le nombre de décès moyen, l'erreur standard, le coefficient de variation, le minimum et le maximum de décès augmentent avec la période de retour. Ceci signifie que l'incertitude augmente avec la période de retour. Nous remarquons aussi qu'il n'y a significativement pas une grande différence entre les résultats pour les seuils 45 et 50 ; ceci est une indication sur le fait que le modèle donne presque les mêmes résultats pour chaque valeur de seuil comprise entre 45 et 50.

Selon l'adéquation du modèle et les résultats de l'approche du bootstrap non paramétrique, nous pouvons dire que le seuil optimal est autour de 45.

GPD	$rmse(\sigma)$	$rmse(\xi)$	$rmse(x_5)$	$rmse(x_{10})$	$rmse(x_{25})$
$GPD\,(40, 58.7971, 0.0458)$	18.4212	0.2740	20.6004	32.4851	46.3645
$GPD\,(45, 60.26450, 0.03779)$	20.1151	0.3053	24.3064	33.9565	47.7765
$GPD\,(50, 56.67434, 0.06837)$	19.6865	0.2386	26.2439	34.1284	49.2163

TABLE 4.2.8 – Erreurs quadratiques moyennes pour les niveaux de retour, les paramètres de forme et d'échelle

Les résultats du tableau (4.2.8) montrent que l'erreur quadratique moyenne ($rmse$) pour les paramètres de forme et d'échelle est presque la même pour les seuils 40, 45 et 50. Il en est de même pour les niveaux de retour estimés. Ce qui confirme les résultats des tableaux 4.2.5, 4.2.6 et 4.2.7. D'où nous pouvons donc constater qu'il n'y a pas de problème d'adéquation du modèle du à la sélection du seuil autour de 45.

4.2.4.6 Comparaison entre quantile extrême observé et simulé

Dans la procédure du bootstrap, la base de données initiales est constituée d'un certain nombre d'excédents au-dessus des seuils 40, 45 et 50. Un échantillon de taille égale au nombre d'excédents est sélectionné avec remise à partir des données initiales (observations au-dessus du seuil) pour obtenir les données de bootstrap.
En analysant les tableaux (4.2.5), (4.2.6), (4.2.7) et (4.2.9), nous constatons que le nombre moyen de victimes possibles par l'approche du bootstrap utilisé et le nombre de victimes auquel on peut s'attendre à partir des données initiales sont bien dans des intervalles de confiance à 95% pour différentes périodes de retour et les seuils 40, 45 et 50. Il faut aussi noter que le tableau (4.2.9) présente les estimations obtenues à partir des données initiales ; tandis que les tableaux (4.2.5), (4.2.6), (4.2.7) présentent les estimations obtenues à partir du bootstrap.

		Seuils		
		$\mu = 40$	$\mu = 45$	$\mu = 50$
T- ans	5	132[97, 178]	139 [99, 189]	141 [102, 189]
	10	177 [111, 239]	183 [109, 243]	185[111, 250]
	25	238[118, 311]	245 [123, 322]	248 [118, 326]

TABLE 4.2.9 – Estimation des niveaux de retour attendus pour les différents seuils avec un intervalle de confiance à 95%

4.3 Conclusion

Les principales sources d'incertitude dans la modélisation avec la GPD sont dues à la détermination du seuil optimal et la méthode d'estimation. Beaucoup d'auteurs ont travaillé dans ce sens, Bermudez et Kotz (2010) font une étude extensive de plusieurs méthodes d'estimation des paramètres de la GPD. Dans leurs travaux, ils arguaient que le succès de GPD sur une base de données dépendait substantiellement du processus d'estimation des paramètres. Quant à Castillo et Daoudi (2009), ils ont proposé une approche permettant la détection d'anomalies dans l'estimation de la GPD pour de petits échantillons ; Far et Wahab (2016), proposent une approche appelée méthode hybride permettant de spécifier la valeur de seuil adéquat pour le modèle POT ; tandis que Zhang *et al.* (2015) se sont intéressés à la méthode POT en travaillant sur la quantification des incertitudes statistiques, etc. Cependant il n'y a pas d'approche universelle pour obtenir un seuil exact ; dans toutes les méthodes, il y a une certaine subjectivité.
L'estimation des risques d'accidents extrêmes de l'Aviation Générale Américaine est très importante car la majeure partie des accidents sont dans cette catégorie d'aviation. D'où la question de la prédiction du nombre éventuel de victimes dues à un accident extrême est un réel challenge ; mais la parfaite modélisation des victimes de l'Aviation Générale Américaine n'est pas du tout triviale. Nous proposons à partir de la théorie des valeurs extrêmes, une approche basée sur la GPD, une prédiction du nombre de décès pouvant résulter d'un accident extrême dans l'Aviation Générale Américaine dans des futures opérations. Cette approche peut s'appliquer à toute donnée rétrospective pour d'autres périodes ; dans de tels cas, on déterminera le nouveau seuil et les paramètres de la GPD. Le modèle permet à partir des données historiques d'avoir approximativement le nombre de décès auquel on peut s'attendre à des périodes précises dites périodes de retour en années.
Dans le but de quantifier les incertitudes dues aux méthodes d'estimation, nous utilisons l'approche non paramétrique du bootstrap ; pour vérifier l'adéquation du modèle GPD, nous utilisons une approche graphique. Les résultats obtenus montrent que l'approche non paramétrique donne de précises estimations des incertitudes associées aux paramètres estimés. Il faut aussi rappeler que les niveaux de retour prédits sont faits sur l'hypothèse que la technologie, le volume du trafic sont les mêmes qu'au moment de la modélisation. Cette question s'avère très complexe dans le domaine de l'aviation au vu des avancées technologiques croissantes et les conditions de vol ; ceci peut donc être vu comme une autre source d'incertitude dans les prédictions faites. Le futur challenge de ce travail pourrait être la prise en compte de cette incertitude en utilisant une approche visant à intégrer les progrès technologiques et scientifiques dans les modèles de prédiction.

Conclusion Générale

Dans cette thèse, nous avons abordé différents aspects de la TVE univariées sous l'hypothèse de l'indépendance des observations dans le cadre de l'analyse de fiabilité des systèmes. De nos jours, la performance et la fiabilité des systèmes sont des éléments très importants dans les systèmes de productions de masse ou les activités logistiques telles que l'aviation. Notre objectif est de proposer à partir de la TVE, des outils d'aide à la décision et de prédiction dans le domaine de l'industrie et de l'aviation.

Après avoir dans un premier chapitre rappelé les deux principales approches et les méthodes d'estimation des paramètres dans la TVE, nous nous sommes intéressés à l'analyse de fiabilité en rappelant les principales notions de base. Notre attention s'est beaucoup plus portée sur la fonction de risque, plus précisément la fonction de risque de base des modèles paramétriques à risques proportionnels et les statistiques d'ordre extrêmes. La prise en compte de la non monotonicité et des facteurs d'environnement ou d'opérations dits extrêmes est problématique ; des modifications ont été apportées à la distribution de Weibull afin de satisfaire la condition de la non monotonicité, cependant on constate une surestimation du taux de défaillance et la non prise en compte des conditions extrêmes d'opérations. C'est dans cette optique que nous avons construit à partir de la TVE une fonction de risque croissante-décroissante tenant compte des conditions extrêmes d'opérations et proposé un exemple d'application dans le domaine de l'industrie.

Dans le troisième chapitre, nous nous sommes intéressés aux statistiques d'ordre extrêmes, notamment à leurs applications dans le domaine du machinisme, plus précisément dans celui de la rugosité. Les statistiques d'ordre extrêmes dans la rugosité des surfaces sont appliquées pour mettre en évidence les critères de qualité. En utilisant les calculs associés aux statistiques d'ordre, nous avons montré deux résultats fondamentaux à savoir le nombre de pièces défectueuses par lots et la distribution du minimum donnant le meilleur intervalle pour la rugosité des surfaces.

Dans le dernier chapitre, nous nous sommes intéressés à l'approche POT, notamment dans une modélisation en aéronautique. Dans cette approche, la détermination du seuil et la méthode d'estimation sont les principales sources d'incertitudes ; beaucoup de méthodes ont été proposées pour la détermination du seuil optimal mais il n'existe pas de méthode universelle. Nous proposons une approche mixte consistant à utiliser une méthode graphique combinée avec une estimation intensive des paramètres afin de déterminer le seuil optimal. Cette approche a été appliquée à des

données de l'AGA ; ainsi nous avons obtenu un modèle de prédiction des victimes dues aux accidents extrêmes. Nous avons donc fait des estimations des niveaux de retour du nombre d'éventuelles victimes ; cependant, ces estimations sont faites avec des données antérieures qui dépendent de la technologie, du volume du trafic, du type d'avion et des conditions de vol de l'époque. En outre, dans le domaine de l'aviation, les systèmes s'améliorent avec les avancées technologiques. Ces avancées permettent globalement de baisser le nombre de victimes dues aux accidents extrêmes. C'est ainsi qu'une autre source d'incertitude due aux avancées technologiques vient s'ajouter aux deux premières mentionnées.

Nous envisageons dans un travail futur de combiner l'approche POT à d'autres méthodes permettant la prise en compte des avancées technologiques dans les prédictions. En termes de perspectives, nous pensons aussi généraliser la fonction de risque croissante-décroissante proposée en introduisant des paramètres de forme et d'échelle pour plus de flexibilité.

Bibliographie

Abul, K., Mohammad, G. et Talal, Y. (2014). Statistical diagnosis of the best weibull methods for wind power assessment for agricultural applications. *Energies*, 7:3056–3085.

Airbus (2016). *Commercial Aviation Accidents,A Statistical analysis, 1958-2016*. Airbus, Toulouse.

Alam, A. M., Chieragatti, R., Akamatsu, M. et Caumel, Y. (1997). Effect of machining procedures on the surface state of a low carbon steel. *Scripta Materialia*, (37):1559–1567.

Amar, A. (2014). *Apport de la théorie des valeurs extrêmes à la modélisation et la gestion des risques boursiers, financiers et hydro-météorologiques*. Thèse de doctorat, Université Mohamed V-AGDAL, Faculté des sciences de Rabat, Maroc.

Arak, M. (2003). Order statistics from a logistic distribution and applications to survival and reliability analysis. *IEEE transactions on reliability*, 52(2):200–206.

Austin, P. (2012). Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31:3946–3958.

Axinte, D. A. et Dewes, R. C. (2002). Surface integrity of hot work tool steel after high speed milling- experimental data and empirical models. *Materials Processing Technology*, 127(3):325–335.

Balkema, A. et Haan, L. D. (1974). Residual life time at a great age. *The Annals of Probability*, 2(5):20–21.

Beirlant, J. et Goegebeur, Y. (2003). Regression with response distribution of pareto-type. *Computational Statistics and Data Analysis*, 42(4):595–619.

Benkhaled, A. (2007). Statistical distributions of annual maximum rainfalls depths in the area of cheliff. comparison of techniques and results. *Courrier du Savoir*, 8:83–91.

Benstock, D. et Cegla, F. (2016). Sample selection for extreme value analysis of inspection data collected from corroded surfaces. *Corrosion Science*, 103:206–214.

Bermudez, P. D. Z. et Mendes, Z. (2012). Extreme value theory in medical sciences : Modeling total high cholesterol levels. *Journal of Statistical theory and Practice*, 6(3):468–491.

Bermudez, Z. et Kotz, S. (2010). Parameter estimation of the generalized pareto distribution- part ii. *Journal of Statistical Planning and Inference*, 140:1374–1388.

Bernard, G. et Ahmed, S. (1958). Applications of order statistics to health data. *A.J.P.H*, 48(10):1388–1394.

Bingham, N., Goldie, C. et Teugels, J. (1987). *Regular Variation,Volume 27.* Encyclopedia of Mathematics and its Applications, Cambridge University Press.

Boeing (2016). *Statistical Summary of Commercial Jet Airplane Accidents Worldwide Operations,1959-2013.* Aviation Safety-Boeing Commercial Airplanes, Washington.

Bureau, J. (2012). *Définition et analyse statistique d'une mesure d'intégrité pour données GPS/EGNOS.* Thèse de doctorat, Université Toulouse 3 Paul Sabatier, France.

CAA (2013). *Global fatal accident review.* CAA, Norwich.

Carpenter, J. et Bithel, J. (2000). Bootstrap confidence intervals when, which, what ? a pratical guide for medical statisticians. *Statistics in medicine*, 19:1141–1164.

Carter, D. et Chalenor, P. (1981). Estimating return values of environmental parameters. *Quaterly journal of the Royal Meteorological Society*, 107:259–266.

Castillo, E., Hadi, A., Balakrishnan, N. et Sarabia, M. (2005). *Extreme Value and Related Models with Applications in Engineering and the Sciences.* Wiley, WILEY-INTERSCIENCE.

Castillo, J. et Daoudi, J. (2009). Estimation of generalized pareto distribution. *Statistics and Probability Letters*, 79(5):684–693.

Choi, I., Wells, B., Yu, C. et Kattan, M. (2011). Modeling medical prognosis :survival analysis techniques. *Journal of Biomedical Informatics*, 44:595–606.

Christidis, N., Stott, P. et Brown, S. (2011). The role of human activity in the recent warming of extremely warm daytime temperature. *Journal of climate*, 24:1922–1930.

Christopeit, N. (1994). Estimating parameters of an extreme value law by the method of moments. *Journal of Statistical Planning and Inference*, 41:173–186.

Coles, S. (2001). *An introduction to statistical modeling of extreme values.* Springer-Verlag, New York.

Cox, D. (1972). Regression models and life-tables. *Journal of Royal Statistic Society, series B*, pages 187–220.

Cox, D. et Oakes, D. (1984). *Analysis of Survival Data*. Biometrika, London-New York.

Crane, D. (1997). *Dictionary of Aeronautical terms*. Aviation Supplies- Academics, Baltimore.

Dahan, E. et Mendelson, H. (2001). An extreme value analysis of concept testing. *Management Science*, 47:102–116.

Davison, A. et Smith, R. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B*, 52(3):393–442.

Dawson, T. (2000). Maximum wave crests in heavy seas. *journal of Offshore Mechanics and Artic Engineering-Transactions of the AMSE*, 122:222–224.

Dekkers, A., Einmahl, J. et Haan, L. D. (1989). A moment estimator for the index of an extreme value distribution. *Annals of Statistcs*, 17(4):1833–1855.

Delmas, J. et Jourdain, B. (2006). *Modèles aléatoires, Applications aux sciences de l'ingénieur*. Springer, Berlin.

Deme, E. H. (2013). *Quelques contributions à la théorie univariée des valeurs extrêmes et estimation des mesures de risque actuariel pour des pertes $\tilde{A}$ à queues lourdes*. Thèse de doctorat, Université Gaston Berger de Saint Louis, Sénégal.

Dey, A. et Das, K. (2016). Quantifying the risk of extreme aviation accidents. *Physica A*, 463(1):345–355.

Diallo, M. (2017). *Méthodes d'analyse géostatistiques avec des applications à la prédiction en environnement et aux changements climatiques*. Thèse de doctorat, Université Professeur Joseph Ki-Zerbo de Ougadougou, Burkina Faso.

Diamoutene, A., Kamsu-Foguem, B., Noureddine, F. et Barro, D. (2018a). Prediction of u.s. general aviation fatalities from extreme value approach. *Transportation Research Part A*, 109:65–75.

Diamoutene, A., Noureddine, F., Kamsu-Foguem, B. et Barro, D. (2018b). Quality control in machining using order statistics. *Measurement*, 116:596–601.

Diebolt, J., Guillou, A. et Rached, I. (2004). A new look at probability-weighted moments estimators. *Comptes rendus de l'académie des sciences*, 338(8):629–634.

Efron, B. (1979). Bootstrap methods : Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Embrechts, P., Kluppelberg, C. et Mikosch, T. (1997). *Modelling extremal events*. Springer Verlag, Berlin.

Far, S. et Wahab, A. (2016). Evaluation of peaks over threshold method. *Ocean Science*.

Farah, H. et Azevedo, C. (1952). Safety analysis of passing maneuvers using extreme value theory. *IATSS Research*, 41:12–21.

Fisher, R. et Tippet, L. (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.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathématiques*, 6:93–116.

Fultz, A. et Ashley, W. (2016). Fatal weather-related general aviation accidents in the united states. *Physical Geography*.

Galambos, J. (1985). *The asymptotic theory of Extreme Order Statistics*. Wiley, New-York.

Glaser, R. (1980). Bathub and related failure rate characterizations. *Journal of American Statistical Association*, 75:667–672.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3):423–453.

Gorgoso-Varela, J. J., Parlange, M. et Naveau, P. (2014). Use of gumbel and weibull functions to model extreme values of diameter distributions in forest stands. *Annals of forest Science,Springer Verlag/EDP Sciences*, 71(7):741–750.

Greenwood, J., Landweher, J., Matalas, N. et Wallis, J. (1979). Probability weighted moments : Definition and relation to parameters of several laws expressible in inverse form. *Water Resources Research*, 15:1049–1054.

Gumbel*, E. (1958). *Statistics of Extremes*. University Press, New-York.

Gumbel, E. et Mustafi, C. (1967). Some analytical properties of bivariate extremal distributions. *American Statistical Association journal*, 62:569–589.

Haan, L. D. et Rootzén, H. (1993). On the estimation of high quantiles. *Journal of Planning and Infrence*, 35(1):1–13.

Harris, R. (2001). The accuracy of design values predicted from extreme values analysis. *journal of Wind Engineering and Industrial Aerodynamics*, 89:153–164.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistcs*, 5:1163–1174.

Horst, R. (2014). *The hazard rate, theory and inference*. Justus-Liebig-University, Germany.

Hosking, J. et Wallis, J. (1987). Parameter and quantile estimation for the generalized pareto distribution. *TECHNOMETRICS*, 29(3):339–349.

Hosking, J., Wallis, J. et Wood, E. (2003). Estimation of the generalized extreme value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261.

Ilhan, A. et Mehmet, C. (2011). Modeling and prediction of surface roughness in turning operations using artificial neural network and multiple regression method. *Expert Systems with Applications*, 38(5):5826–5832.

Inoue, T., Tsuchida, Y., Suzuki, T. et Aoki, W. (2003). Alumina ceramics cutting by self-sharpening edge technique of diamond. *Journal of Materials Processing Technology*, 143-144:657–661.

Jiang, W. (2014). Bio-inspired self-sharpening cutting tool surface for finish hard turning of steel. *CIRP Annals-Manufacturing Technology*, 63:517–520.

Justus, C., Hargraves, W., Amir, M. et Denise, G. (1977). Methods for estimating wind speed frequency distributions. *Journal of applied meteorology*, 17:350–533.

Karpuschewski, B., Goldau, H. et Stolze, R. (2014). Process force and technology model for designing and controling finishing operations with rotating grinding tools. *CIRP Annals-Manufacturing Technology*, 63:337–340.

Katz, R. (2002). Techniques for esmating uncertainty in climate change. *Scenarios and impact studies, Climate research*, 20:167–185.

Kawas, M. et Moreira, R. (2001). Characterization of product quality attributes of tortilla chips during the frying process. *Journal of Food Engineering*, 47:97–107.

Kim, S. W. et Ibrahim, J. (2000). On bayesian inference for proportional hazards models using noninformative priors. *Lifetime Data Analysis*, 6(4):331–341.

Knetch, W. (2015). *Civil Aerospace Medical Institute*. FAA, Washington.

Kotz, S. et Nadarajah, S. (2000). *Extremes Value Distributions. Theory and Applications*. Imperial College Press, London.

Kumar, D. et Klefsjo, B. (1994). Proportional hazards model : a review. *Reliability Engineering and System Safety*, 44:177–188.

Lanternier, B., Lyonnet, P. et Toscano, R. (2008). Modèle à hasard proporttionnel pour la fiabilité des composants mécaniques. *Mécanique and Industries*, 9:397–405.

Lavenda, B. et Cipollone, R. (2000). Extreme value statistics and thermodynamics of earthquakes : aftershock sequences. *Annali di Geofisica*, 43:967–982.

Leadbetter, M., Lindgren, G. et Rootzén, H. (1983). *Extremes and related properties of random sequences and series*. Springer Verlag, New York.

Longuet-Higgins, M. (1952). On the statistical distribution of the heights of sea waves. *Journal of marine research*, 9:245–266.

Mahdavi, M. et Mahdavi, M. (2011). Statistical reliability based modeling of series and parallel operating systems using extreme value theory. *International Journal of Industrial and Manufacturing Engineering*, 5(7):1359–1363.

Makkonen, L. (2008). Problems extreme value analysis. *Structural Safety*, 30:405–419.

Methni, J. E. (2013). *Contributions à l'estimation de quantiles extrêmes. Applications à des données environnementales*. Thèse de doctorat, Université de Grenoble, France.

Mills, J. (1926). Tables of the ratio area to bounding ordinate for any portion of normal curve. *Biometrika*, 18:395–400.

Naess, A. et Clausen, P. (2001). Combination of the peaks-over-threshold and bootstrapping methods for extreme value prediction. *Structural Safety*, 23:315–330.

Noureddine, R., Noureddine, F. et Benamar, A. (2012). Surface roughness measurement for model-based fault detection in turning process. *International review of mechanical engineering*, 6(7):1411–1417.

NTS (2016). *National Transportation Statistics*. U.S. Departement of Transportation, Washington.

NTSB (2011). *Review of U.S. civil aviation accidents calendar year*. National Transportation Safety Board, Washington.

Ohno-Machado, L. (2001). Modeling medical prognosis :survival analysis techniques. *Journal of Biomedical Informatics*, 34:428–439.

Papatheou, E., Dervilis, N., Maguire, A., Campos, C., Antoniadou, I. et Worden, K. (2017). Performance monitoring of a turbine using extreme function theory. *Renewable Energy*, 113:490–1502.

Park, S. et Park, J. (2018). A general class of flexible weibull distributions. *Communications for Statistical Applications and Methods*, 47(4):767–778.

Park, S., Park, J. et Choi, Y. (2016). A new flexible weibull distribution. *Communications for Statistical Applications and Methods*, 23(5):399–409.

Pickands, J. (1975). Statistical inference using extreme order order statistics. *Annals of Statistics*, 3(1):119–131.

Piera-Martinez, M. (2008). *Modélisation des comportements extrêmes en ingénierie*. Thèse de doctorat, Université Paris Sud-Paris XI,France.

Prescott, P. et Walden, A. (1980). Maximum likelihood estimation of the parameters of the generalized extreme value distribution. *Biometrika*, 67(3):723–724.

Raggad, B. (2009). Fondements de la thérie des valeurs extrêmes, ses principales applications et son apport à la gestion des risques du marché pétrolier. *Mathematics and Social Sciences*, pages 29–63.

Reiss, R. et Thomas, M. (1997). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and other fields*. Birkh auser, Verlag,Basel.

Reiss, R. et Thomas, M. (2001). *Statistical Analysis of Extreme Values*. Birkhauser, Verlag,Boston,2nd edition.

Resnick, S. (1987). *Extreme Values, Regular variation and Point Processes*. Springer Verlag, New-York.

Richard, A., Krajcinovic, D. et Mastilovic, S. (2007). Statistical damage mechanics and extreme value theory. *International journal of damage mechanics*, 16(1):55–76.

Roberts, S. (2000). Extreme value statistics for novelty detection in biomedical data processing. *IEEE proceeding-Science Measurement and Technology*, 147:363–367.

Robinson, M. et Tawn, J. (1995). Statistics for exceptional record. *Applied Statistics*, 44:499–511.

Romeijnders, W., Teunter, R. et Jaarveld, W. (2012). A two step method for forecasting spare parts demand usimg information on component repairs. *European Journal of Operational Research*, 220(2):386–393.

Scarrot, C. et Macdonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT Statistcal Journal*, 10(1):33–60.

Schmidt, F., Zhou, X. Y. et Toulemonde, F. (2014). A peaks-over-threshold analysis of extreme traffic load effects on bridge. *Archives-ouvertes, hal-00949698*.

Sha, Z., Rommert, D., Rex, V. W. W. et Alex, J. (2017). An improved method for forecasting spare parts demand using extreme value theory. *European Journal of Operational Research*, 261(1):161–181.

Shetty, K. et Hansman, R. (2012). *Current and historical trends in General Aviation in the United States*. MIT ICAT, Cambridge.

Sita, K. et Satya, R. (2014). Pareto type 2 software reliability growth model : an order statistics approach. *International journal of computer sciences trend and technology*, 2(4):49–54.

Smith, R. (1985). Maximum likelihood estimation in a class of non regular cases. *Biometrika*, 72:67–90.

Smith, R. (1987). Estimating tails of probability. *The Annals of Statistics*, 3:1174–1207.

Sobieralski, J. (2012). The cost of general aviation accidents in the u.s. transportation reserach part a : Policy and practice. *REVSTAT Statistcal Journal*, 47:19–27.

Songchitruksa, P. et Tarko, A. (2006). The extreme value theory approach to safety estimation. *Accident Analysis and Prevention*, 38:811–822.

Surajit, P. (2005). Order statistics for some common hazard rate functions with an application. *International journal of quality & reliability management*, 22(2):201–210.

Tawn, J. (1992). Estimating probabilities of extreme sea levels. *Applied Statistics*, 41:77–93.

Thatcher, R. (1999). The long term-pattern of adult mortality and highest attained age. *Journal of Royal Statistic Society series A*, 162(1):5–43.

Thompson, M., Reynolds, J., Cox, L., Guttorp, P. et Sampson, P. (2001). A review of statistical methods for the meteorological adjustement of the tropospheric ozone. *Atmospheric Environment*, 35:617–630.

Todorovic, P. et Rousselle, J. (1971). Some problems on flood analysis. *Water Resources*, 7:1144–1150.

Todorovic, P. et Zelenhasic, E. (1970). A stochastig model for flood analysis. *Water Resources*, 6:1641–1648.

VonMises, R. (1936). La distribution de la plus grande de n valeur. *Revue de Mathématiques, Union Interbalcanique*, 1:141–160.

Walshaw, D. et Anderson, C. (2000). A model for extreme wind gusts. *Applied Statistics*, 49:499–508.

Wang, B. et Dey, D. (2011). Generalized extreme value regression for binary response data : an application to b2b electronic payments system adoption. *The annals of Applied Staistics*, 4(4):2000–2023.

Wehner, W. (2010). Sources of uncertainty in the extreme value statistics of climate data. *Extremes*, 19:205–217.

Wendy, S. et Noriszura, I. (2012). Analysis of t-return level for partial duration rainfall series. *Sains Malaysiana*, 41:1389–1401.

Yongcheng, Q. (2008). Bootstrap and empirical likelihood methods in extremes. *Extremes*, 11:81–97.

Zhang, Y., Cao, Y. et Dai, J. (2015). Quantification of statistical uncertainties in performing the peaks over threshold method. *Journal of Marine Science and Technology*, 23(5):717–726.

Zio, E. (2016). Challenges in the vulnerability and risk analysis of critical infrastructures. *Reliability Engineering and System Safety*, 152:137–150.

Zwiers, F., Zhang, X. et Feng, Y. (2011). Antropogenic influence on long return period daily temperature extremes at regional scales. *Journal of climate*, 24:881–892.