

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ANALYSE DE LA SINISTRALITÉ D'UNE BASE DE DONNÉES À L'AIDE
DE LA THÉORIE DES VALEURS EXTRÊMES ET D'UN OUTIL DE
SIMULATION

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

YAPI ARNAUD CARMEL

JUIN 2021

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.04-2020). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Dans un premier lieu, je voudrais rendre toute la gloire à Dieu tout-puissant, le scientifique par excellence, pour sa grâce, ses bénédictions et la force qu'il n'a cessé de dispenser à mon égard, tout au long de mes études au sein de l'UQAM. Je voudrais par la suite dire un grand merci à mon directeur de recherche, Mathieu Pigeon, pour sa patience à mon égard, son professionnalisme et sa passion pour le travail bien accompli. Aussi, je suis reconnaissant envers les professeurs qui ont dispensé des cours d'une qualité supérieure, pendant ma période d'étude. De plus, toute ma profonde gratitude va à l'endroit d'Intact Corporation Financière, mon employeur, qui n'a cessé, par l'entremise des gestionnaires, de me soutenir dans tous mes projets de perfectionnement professionnel. Enfin, je dédie très spécialement ce travail à ma conjointe, Désiré, à mes enfants, Elicora et Jayden, à mon cher Père, et à la mémoire de ma défunte Mère, pour leur amour et leur soutien indéfectible dans mes projets de perfectionnement.

TABLE DES MATIÈRES

LISTE DES TABLEAUX	v
LISTE DES FIGURES	viii
RÉSUMÉ	xi
ABSTRACT	xii
INTRODUCTION	1
CHAPITRE I DÉFINITION ET LOIS DE PROBABILITÉS USUELLES	4
1.1 Distributions usuelles pour modéliser les montants des sinistres	4
1.1.1 Loi Gamma	5
1.1.2 Loi Lognormale	5
1.1.3 Loi Loggamma	5
1.1.4 Loi Weibull	6
1.1.5 Loi Pareto	6
1.1.6 Loi Pareto Généralisée	6
1.2 Notions sur la réassurance	7
1.2.1 Définition	7
1.2.2 Types et méthodes de réassurance	8
1.2.3 Tarification d'un traité de réassurance non-proportionnelle : le cas de l'excédent de sinistre par risque (XOL)	10
1.2.4 Expression de la prime pure de réassurance XOL pour une mo- délisation avec la loi GPD	11
CHAPITRE II THÉORIE DES VALEURS EXTRÊMES	14
2.1 La loi des maxima	14
2.2 Distribution des valeurs extrêmes généralisées (GEV)	15
2.2.1 Estimation de la loi du maximum	17
2.3 La loi des excès et la distribution de Pareto Généralisée	19

2.3.1	Estimation des paramètres	20
2.3.2	Propriété de stabilité	21
2.3.3	Theorème de Pickands-Balkema-de-Haan	22
2.4	Quantiles extrêmes	22
2.4.1	Nombre de dépassements	23
2.4.2	Le plus grand sinistre ou sinistre maximum possible	23
2.4.3	Notions de temps de retour, période de retour et niveau de retour	25
2.4.4	Valeur au risque pour une GPD	26
CHAPITRE III LES OUTILS GRAPHIQUES ET TESTS D'ADÉQUATION		28
3.1	Le Graphique Quantile-Quantile	28
3.1.1	Cas particulier de la distribution de Pareto Généralisée	29
3.1.2	Cas particulier de la distribution Exponentielle	29
3.2	La fonction d'excès moyen	31
3.2.1	Cas de la distribution exponentielle	31
3.2.2	Cas de la distribution Pareto Généralisée	32
3.3	Le test de Kolmogorov-Smirnov	34
3.3.1	Cas particulier dans le milieu de la réassurance	35
3.4	Le test d' Anderson-Darling	35
3.5	Le graphique de Gertensgarbe et Werner	36
3.6	Inférence paramétrique- Estimation par Maximum de vraisemblance .	38
3.6.1	Cas des données tronquées	39
CHAPITRE IV PRÉSENTATION DE L'OUTIL DE SIMULATION <i>LOSS SIMULATOR 2.0</i>		40
4.1	Le processus de simulation	42
4.1.1	L'ajustement des variables d'intérêts	42
4.1.2	Les distributions paramétriques	46
4.1.3	L'ajustement d'un modèle de copule	46

4.2	Les types de sinistres simulés	47
4.2.1	Les sinistres survenus ou ouverts	47
4.2.2	Les sinistres survenus mais non déclarés	50
4.2.3	Les sinistres futurs	51
4.2.4	Les sinistres ré-ouverts	52
	CHAPITRE V APPLICATIONS NUMÉRIQUES	53
5.1	Description des données	54
5.1.1	Statistiques descriptives des données initiales	54
5.1.2	Statistiques descriptives des données simulées	54
5.2	Ajustement des modèles de sévérité sur les données simulées	59
5.2.1	Modèle basé sur la théorie des valeurs extrêmes	59
5.2.2	Modèle basé sur les lois standards	66
5.3	Application de la théorie des valeurs extrêmes dans le milieu de la réassurance	67
5.3.1	Calcul des niveaux de retour pour différentes périodes de retour	67
5.3.2	Calcul de la prime de réassurance en excédent de sinistre	71
	CONCLUSION	75
	APPENDICE A COMPOSANTES DES PRIMES PURES DE RÉASSURANCE EN EXCÉDENT DE SINISTRE	76
	BIBLIOGRAPHIE	79

LISTE DES TABLEAUX

1.1	Répartition des primes et sinistres entre l'assureur et le réassureur	11
2.1	Estimation des paramètres de la distribution <i>GEV</i> par maximum de vraisemblance des pertes incendies danoises	18
3.1	Comportement de la fonction d'excès moyen	32
4.1	Exemple de facteur de développement année après année	49
4.2	Facteur de développement cumulatif selon la fonction lien. On définit par β les estimateurs par maximum de vraisemblance, d l'année de développement, l le montant du sinistre en cours, os le taux d'encouru et ε une variable aléatoire qui suit une loi normale centrée réduite.	50
5.1	Statistiques descriptives du montant des <i>pertes encourues</i> en responsabilité civile de la base initiale	55
5.2	Paramètres (Param.) de l'ajustement de la base de données initiale pour les <i>pertes encourues</i> en responsabilité civile	55
5.3	Statistiques descriptives des <i>pertes encourues</i> simulées en responsabilité civile selon le scénario 1.	56
5.4	Valeurs p de l'ajustement de la base de données initiale pour les <i>pertes encourues</i> en responsabilité civile par des copules.	58

5.5	Statistiques descriptive des <i>pertes encourues</i> simulées en responsabilité civile selon le scénario 2.	58
5.6	Paramètres de la loi Pareto Généralisée pour différents seuils pour les <i>pertes encourues</i> simulées selon le <i>scénario 1</i> . (<i>AD</i> pour le test d'Anderson-Darling)	65
5.7	Paramètres de la loi Pareto Généralisée pour différents seuils pour les <i>pertes encourues</i> simulées selon le <i>scénario 2</i> (<i>AD</i> pour le test d'Anderson-Darling)	66
5.8	Paramètres de l'ajustement des <i>pertes encourues</i> simulées, avec les lois standards	67
5.9	Niveau de retour selon les périodes de retour pour les <i>pertes encourues</i> simulées selon le <i>scénario 1</i>	69
5.10	Niveau de retour selon les périodes de retour pour les <i>pertes encourues</i> simulées selon le <i>scénario 2</i>	71
5.11	Prime pure de réassurance en excédent de sinistres pour les <i>pertes encourues</i> simulées selon le <i>scénario 1</i>	73
5.12	Prime pure de réassurance en excédent de sinistres pour les <i>pertes encourues</i> simulées selon le <i>scénario 2</i>	74
A.1	Fréquence annuelle pour le <i>scénario 1</i>	76
A.2	Probabilité de dépasser le seuil u pour le <i>scénario 1</i>	76
A.3	Sévérité pour le <i>scénario 1</i>	77
A.4	Fréquence annuelle pour le <i>scénario 2</i>	77

A.5	Probabilité de dépasser le seuil u pour le <i>scénario 2</i>	77
A.6	Sévérité pour le <i>scénario 2</i>	78

LISTE DES FIGURES

2.1	La méthodologie du block Maxima appliquée aux données des pertes incendies danoises.	19
2.2	Estimateur de Hill appliqué aux données des pertes incendies danoises	21
2.3	Graphe des niveaux de retour(EML) associés à des périodes de retour pour les pertes incendies danoises	27
3.1	« QQ-Plot » de la distribution exponentielle pour des simulations de Pareto (convexe), Exponentielle (linéaire) et Weibull (concave)	30
3.2	Graphe de la moyenne excès des simulations des distributions Exponentielle (pente constante), Pareto (pente croissante) et Weibull (pente décroissante)	34
3.3	Graphe de Gertensgarbe et Werner appliqué aux pertes incendies danoises.	37
4.1	Architecture de la simulation de l'outil <i>Loss Simulator 2.0</i> - extrait de CAS <i>Loss Simulator 2.0 Methodology</i>	41
4.2	Délai de rapport ou déclaration dans le cycle de développement d'un sinistre individuel	43
4.3	Délai de règlement dans le cycle de développement d'un sinistre individuel	43
4.4	Graphe des montants totaux des pertes encourues incendies danoises	44

4.5	Graphe du nombre total des sinistres cyclone par année des pertes encourues aux états unis	45
4.6	Types de sinistres dans le cycle de développement d'un sinistre individuel	47
5.1	Graphe de la fonction de densité des <i>pertes encourues</i> en responsabilité civile. Les données initiales sont en noire et les pertes simulées selon le scénario 1 sont en rouge.	57
5.2	Graphe de la fonction de densité des <i>pertes encourues</i> simulées en responsabilité civile. Les pertes simulées selon le scénario 1 sont en noire et les pertes simulées selon le scénario 2 sont en rouge.	59
5.3	Graphe de la fonction d'excès moyen empirique des <i>pertes encourues</i> simulées selon le <i>scénario 1</i>	60
5.4	Graphe du quantile-quantile exponentiel des <i>pertes encourues</i> simulées selon le <i>scénario 1</i>	61
5.5	Graphe de la fonction d'excès moyen empirique des <i>pertes encourues</i> simulées selon le <i>scénario 2</i>	62
5.6	Graphe du quantile-quantile exponentiel des <i>pertes encourues</i> simulées selon le <i>scénario 2</i>	62
5.7	Graphe de l'indice de queue $1/\xi$ des <i>pertes encourues</i> simulées selon le <i>scénario 1</i>	64
5.8	Graphe de l'indice de queue $1/\xi$ des <i>pertes encourues</i> simulées selon le <i>scénario 2</i>	65

- 5.9 Graphe du niveau de retour pour les *pertes encourues* simulées selon le *scénario 1*. Les données empiriques en noir, le modèle 1 en rouge, le modèle 2 en bleu et le modèle 3 en vert. 69
- 5.10 Graphe du niveau de retour pour les *pertes encourues* simulées selon le *scénario 2*. Les données empiriques en noir, le modèle 1 en rouge, le modèle 2 en bleu et le modèle 3 en vert. 70

RÉSUMÉ

Les compagnies d'assurance et de réassurance ont un intérêt particulier à comprendre la structure de leur sinistralité. En particulier, les sinistres extrêmes doivent être séparés des sinistres attritionnels par la détermination d'un seuil optimal. De plus, les compagnies sont intéressées à analyser les impacts que peuvent avoir des modifications dans les hypothèses liées aux distributions choisies sur la modélisation des sinistres graves. Dans ce travail, nous utilisons un outil de simulation nommé *Loss Simulator 2.0* et disponible librement via la *Casualty Actuarial Society* (CAS) pour obtenir plusieurs trajectoires futures pour une base de données. Cet outil nous permet, en particulier, de modifier différents paramètres du développement d'un sinistre, par exemple les distributions choisies ou la présence d'inflation, en plus de nous permettre d'ajouter de la dépendance entre certaines composantes. Ainsi, des modèles basés sur des distributions classiques et des modèles basés sur la théorie des valeurs extrêmes seront appliqués sur les bases de données simulées, puis comparés en analysant les niveaux de retour et les primes pures de réassurance en excédent de sinistres pour plusieurs tranches.

Mots clés : Distributions classiques, Sinistres attritionnels, Sinistres extrêmes, Théorie des valeurs extrêmes, Niveau de retour, *Loss simulator 2.0*, Prime pure de réassurance.

ABSTRACT

Insurance and reinsurance companies have a particular interest in understanding the structure of their claims experience. In particular, large claims must be separated from attritional claims by determining an optimal threshold. In addition, companies are interested in analyzing the impact that changes in the assumptions related to the chosen distributions can have on the modeling of severe claims. In this work, we use a simulation tool named *Loss simulator 2.0* and freely available via the *Casualty Actuarial Society* (CAS), to obtain several future trajectories for a database. This tool allows us, in particular, to modify different parameters of the development of a loss, for example the chosen distributions or the presence of inflation, in addition to allowing us to add dependence between certain components. Thus, models based on standard distributions and models based on extreme value theory will be applied on the simulated databases and then compared by analyzing the return levels and excess of loss reinsurance pure premiums for several layers.

Keywords : Standard distributions, Attritional claims, Large claims, Extreme value theory, Return level, *Loss simulator 2.0*, Reinsurance pure premium.

INTRODUCTION

L'étude de la sinistralité dans le milieu de l'assurance et de la réassurance présente plusieurs challenges auxquels les actuaires doivent faire face. En général, les sinistres sont classifiés en deux catégories : les sinistres attritionnels ou normaux d'une part et les sinistres extrêmes ou larges de l'autre part. Par définition, une sinistralité attritionnelle correspond aux sinistres récurrents, avec une importante fréquence, mais une sévérité faible. Tandis qu'une sinistralité extrême correspond aux sinistres rares, avec une sévérité plus importante. On distingue également deux types de développement de sinistres en assurance. Dans un premier temps, nous avons les sinistres à développements longs, pour lesquels les déclarations des sinistres auprès des assureurs arrivent souvent plusieurs années après la survenance des sinistres. Ainsi, la détermination des montants de sinistres dans ces branches peut découler de décisions à la suite d'un procès, ce qui fait que les estimations de ces montants sont plus longues et difficiles à anticiper et que le règlement final se fait longtemps après la date de déclaration. L'exemple le plus courant de ce type de sinistres est les sinistres de la branche « responsabilité civile ». Dans un deuxième temps, nous avons les sinistres à développements courts au cours desquels les déclarations des sinistres auprès des assureurs se font plus rapidement, les règlements finaux de ces sinistres se font rapidement une fois que les estimations sont faites. Comme exemple pour ce type de sinistre, on a les sinistres de la branche « dommages physiques pour l'automobile ou Dommages aux biens assurés ». De façon pratique les actuaires dans leurs analyses de tarification ou des réserves, traitent de manière distincte ces deux types de sinistralité. C'est ce que les auteurs (Denuit et Charpentier, 2005) ont appliqué en utilisant deux

distributions distinctes pour modéliser les sinistres normaux et les sinistres larges pour des pertes en automobile dans un objectif de tarification. L'une des difficultés qui est apparue dans cette démarche de séparer les sinistres en deux catégories, était de trouver le point optimal seuil qui déterminerait la frontière entre les deux catégories de sinistres. Cette problématique est beaucoup plus fréquente dans le milieu de la réassurance où l'on traite le plus souvent avec des sinistres extrêmes, et l'utilisation des théories sur les valeurs extrêmes. Pour répondre à cette problématique de seuil optimal, (Cebrián *et al.*, 2003) ont présenté les modèles statistiques propres à la théorie des valeurs extrêmes et leur mise en application sur des sinistres en assurance médicale de groupe dans le cadre d'un traité de réassurance excédent de sinistres. Ils ont montré que ces modèles nous aident pour déterminer un point seuil qui sert de frontière entre les sinistres normaux et les sinistres extrêmes pour une même distribution des sinistres. Toujours dans le même objectif de trouver le point seuil, Eric Gillard et Mathieu Ribatet dans (Charpentier, 2014), ont appliqué ces mêmes outils des valeurs extrêmes sur des sinistres Incendie et en y ajoutant les notions de « valeur au risque » et de « période de retour » qui sont couramment utilisées dans le milieu de la réassurance pour l'optimisation des programmes de réassurance et la prise de décision dans le choix des structures des traités de réassurance.

La question à laquelle nous essaierons de répondre dans ce mémoire est de montrer comment les compagnies d'assurance et de réassurance pourraient modéliser leur sinistralité extrême, tout en émettant diverses hypothèses sur le développement des sinistres (par exemple en modifiant les facteurs de développement des sinistres ou en introduisant la dépendance).

Pour répondre à la question, notre travail va se dérouler en cinq chapitres. Dans le premier chapitre, certaines lois usuelles pour modéliser les sinistres en assurance seront énoncées. Les modèles standards que nous allons définir sont issus de ces lois. Aussi, certaines définitions importantes en rapport avec la réassurance vont

être passées en revue également dans ce chapitre. Dans le deuxième chapitre, les outils statistiques et de probabilités qui sous-tendent la théorie des valeurs extrêmes seront énoncés. Quant au troisième chapitre, il sera question de présenter les outils graphiques très utiles dans la détermination d'un seuil optimal de valeurs extrêmes, ainsi que les tests d'adéquation. Les outils énoncés dans le deuxième et le troisième chapitre nous aideront à définir notre modèle basé sur la théorie des valeurs extrêmes. Le quatrième chapitre sera consacré à une procédure qui est utilisée pour générer par simulation des bases de données fictives, mais ayant des traits de caractères des catégories de sinistres à développements longs et/ou courts. Cet outil permettra d'émettre plusieurs hypothèses sur la distribution des sinistres d'une compagnie d'assurance ou de réassurance. Enfin, dans le cinquième chapitre, les résultats de l'application de plusieurs modèles basés sur la théorie des valeurs extrêmes d'une part, et sur les lois standards d'autre part, sur les bases de données simulées, sont présentés. L'impact de ces modèles sur certaines métriques essentielles dans le cadre d'un traité de réassurance en excédent des sinistres est présenté également dans ce dernier chapitre.

CHAPITRE I

DÉFINITION ET LOIS DE PROBABILITÉS USUELLES

Ce chapitre est consacré aux définitions et lois usuelles qui sont utilisées pour modéliser les montants des sinistres. Alors nous passerons en revue quelques distributions de probabilités paramétriques qui sont couramment utilisées pour modéliser les sinistres en général, puis nous présenterons les principaux critères et définitions en lien avec la réassurance.

1.1 Distributions usuelles pour modéliser les montants des sinistres

Nous présentons dans cette section certaines lois usuelles parmi les plus utilisées pour modéliser les montants des sinistres. Les propriétés telles que l'expression de la fonction de densité de probabilité, de l'espérance mathématique et de la variance seront présentées pour chaque distribution.

Soit X une variable aléatoire de fonction de répartition F et de fonction de densité f . On définit $E[X]$ et $\text{Var}[X]$ comme l'espérance mathématique et la variance de X .

1.1.1 Loi Gamma

La variable aléatoire X suit une distribution Gamma de paramètres α et β si sa fonction de densité s'écrit sous la forme :

$$f(x) = \frac{\beta^{-\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right)}{\Gamma(\alpha)} \mathbb{I}_{\mathbb{R}^+}(x),$$

avec $\alpha > 0$, $\beta > 0$, $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$ et $\mathbb{I}_{\mathbb{R}^+}(x)$ la fonction indicatrice des valeurs de x positifs. On a $E[X] = \alpha\beta$ et $\text{Var}[X] = \alpha\beta^2$.

1.1.2 Loi Lognormale

La variable aléatoire X suit une distribution Lognormale de paramètres μ et σ si sa fonction de densité s'écrit sous la forme :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \mathbb{I}_{\mathbb{R}^+}(x),$$

avec $\mu \in \mathbb{R}$ et $\sigma > 0$. De manière analogue, on dit que la distribution du logarithme de X suit une loi normale de paramètres μ et σ^2 . On a $E[X] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$ et $\text{Var}[X] = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$.

1.1.3 Loi Loggamma

La loi Loggamma est une distribution de e^X , où X suit une distribution Gamma de paramètres α et $1/\beta$. Sa fonction de densité de probabilité s'écrit sous la forme :

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{(\log x)^{\alpha-1}}{x^{\beta+1}} \mathbb{I}_{\mathbb{R}^+}(x),$$

avec $\alpha > 0$, $\beta > 0$ et $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$. On a $E[X] = \frac{1}{\left(1-\frac{1}{\beta}\right)^\alpha}$ et $\text{Var}[X] = \left(\frac{1}{\left(1-\frac{2}{\beta}\right)^\alpha} - \frac{1}{\left(1-\frac{1}{\beta}\right)^{2\alpha}}\right)$.

1.1.4 Loi Weibull

La variable aléatoire X suit une distribution Weibull de paramètres α et β si sa fonction de densité de probabilité s'écrit sous la forme :

$$f(x) = \alpha\beta^{-\alpha}x^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)^\alpha}\mathbb{I}_{\mathbb{R}^+}(x),$$

avec $\alpha > 0$, $\beta > 0$. On a $E[X] = \frac{\beta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)$ et $\text{Var}[X] = \frac{\beta^2}{\alpha}\left[2\Gamma\left(\frac{2}{\alpha}\right) - \frac{1}{\alpha}\left(\Gamma\left(\frac{1}{\alpha}\right)\right)^2\right]$.

1.1.5 Loi Pareto

La variable aléatoire X suit une distribution de Pareto de paramètres θ et a si sa fonction de densité de probabilité s'écrit sous la forme :

$$f(x) = \frac{\theta a^\theta}{x^{\theta+1}}(x),$$

avec $\theta > 0$ et $a > 0$. On a $E[X] = \frac{a\theta}{\theta-1}$ et $\text{Var}[X] = \frac{\theta a^2}{(\theta-2)(\theta-1)^2}$. La loi Pareto est un cas particulier de la loi Pareto Généralisée avec les paramètres $a = \beta/\xi$ et $\theta = 1/\xi$.

1.1.6 Loi Pareto Généralisée

La variable aléatoire X suit une distribution de Pareto Généralisée de paramètres ξ et β et si sa fonction de densité de probabilité s'écrit sous la forme :

$$f_{\xi,\beta}(x) = \begin{cases} \frac{1}{\beta}\left(1 + \frac{\xi x}{\beta}\right)^{-\frac{\xi+1}{\xi}} & \text{si } \xi \neq 0 \\ \frac{1}{\beta}\exp(-x/\beta) & \text{si } \xi = 0, \end{cases}$$

avec $\beta > 0$, $x \geq 0$ si $\xi \geq 0$; et $0 \leq x \leq -\beta/\xi$ si $\xi < 0$. On a $E[X] = \frac{\beta}{1-\xi}$ et $\text{Var}[X] = \frac{\beta^2}{(1-2\xi)(1-\xi)^2}$.

1.2 Notions sur la réassurance

Soit N , la variable aléatoire représentant le nombre de sinistres. Soit X_i , $i = 1, 2, \dots, N$, la variable aléatoire représentant le montant total pour le sinistre i , avec la condition que les X_i sont des variables aléatoires indépendantes et identiquement distribuées. On définit le montant total des sinistres par :

$$S = \begin{cases} \sum_{i=1}^N X_i, & N > 0 \\ 0, & N = 0. \end{cases}$$

Dans le cadre de la réassurance, le montant du sinistre i et le montant total des sinistres sont subdivisés en

$$X_i = R_i + L_i \quad \text{et} \quad S = R + L, \quad (1.1)$$

avec R une variable aléatoire représentant le montant total des sinistres retenus par l'assureur, R_i une variable aléatoire représentant le montant de sinistre i retenu par l'assureur, L une variable aléatoire représentant le montant total des sinistres payés par le réassureur et L_i une variable aléatoire représentant le montant de sinistre i payé par le réassureur.

1.2.1 Définition

La réassurance est définie dans (IIC, 2014) par « une entente contractuelle selon laquelle un réassureur accepte moyennant une contrepartie financière, d'indemniser une cédante (un assureur) d'une partie ou de la totalité des sinistres que ce dernier est tenu de payer aux termes des contrats d'assurance qu'il a émis. »

Le marché de la réassurance étant mondial, ceci facilite la répartition des risques acceptés par les réassureurs. Les compagnies d'assurance ont recours à la réassurance pour diverses raisons dont :

- la réduction de la probabilité d’avoir une perte qui pourrait mettre la compagnie d’assurance en faillite, comme les pertes dues à des catastrophes naturelles ;
- la stabilisation des résultats financiers en réduisant la volatilité ;
- l’augmentation de la capacité de souscription en offrant de la capacité financière supplémentaire aux compagnies d’assurance pour écrire des contrats avec des montants d’assurance plus élevés.

1.2.2 Types et méthodes de réassurance

1.2.2.1 Type de réassurance

On distingue deux types de réassurance : la réassurance par traité et la réassurance facultative. Un traité de réassurance correspond à une cession d’un groupe homogène de risque. Ainsi, un traité oblige l’assureur à céder toutes les affaires qui répondent aux conditions de l’entente, et le réassureur est obligé d’accepter. La réassurance facultative quant à elle, est une entente où l’assureur décide le moment de céder un seul risque ou police, et le réassureur a le libre choix d’accepter ou non ce risque.

1.2.2.2 Méthodes de réassurance

En ce qui concerne les méthodes de réassurance, nous avons la réassurance proportionnelle et la réassurance non-proportionnelle. En premier lieu, la réassurance est proportionnelle lorsque le réassureur accepte une portion de sinistres et reçoit la même portion de primes en contrepartie. Parmi les traités proportionnels, nous avons la quote-part et l’excédent de plein. Le traité quote-part est un traité dans lequel l’assureur cède le même pourcentage de primes et de sinistres sur tous les

risques qui font l'objet du traité. Ainsi, on a :

$$L_i = aX_i \quad \text{et} \quad L = aS,$$

avec $0 < a < 1$, appelé le coefficient de proportionnalité. Le traité en excédent de plein est un traité dans lequel la cession se fait en fonction de la portion du risque qui dépasse un niveau de capital ou plein de rétention conservé par l'assureur. Soit Q_i le montant de la limite de la police d'assurance qui a occasionné le sinistre i et K le montant du plein de rétention. Alors on a :

$$L_i = \left(1 - \frac{K}{Q_i}\right) X_i \mathbb{I}_{Q_i > K} \quad \text{et} \quad R_i = X_i \mathbb{I}_{Q_i \leq K} + K \left(\frac{X_i}{Q_i}\right) \mathbb{I}_{Q_i > K},$$

$$L = \begin{cases} \sum_{i=1}^N L_i, & N > 0 \\ 0, & N = 0 \end{cases} \quad \text{et} \quad R = \begin{cases} \sum_{i=1}^N R_i, & N > 0 \\ 0, & N = 0, \end{cases}$$

avec $\mathbb{I}_{Q_i > K}$ la fonction indicatrice qui indique si la limite de couverture de police d'assurance qui a occasionnée le sinistre i dépasse le plein de rétention. En second lieu, la réassurance est non-proportionnelle lorsque l'assureur fixe un montant de sinistre qu'il désire conserver appelé « rétention ou priorité », et que le réassureur accepte de couvrir tout ou partie du sinistre qui excède la rétention. Le réassureur peut décider d'inclure une limite de couverture de sorte à réduire son engagement en cas de sinistre. Ce type de réassurance peut avoir plusieurs tranches et chaque tranche est représentée par une rétention avec une limite. Parmi les traités non-proportionnels nous avons l'excédent de sinistre par risque, l'excédent de sinistre par évènement et l'excédent de perte annuelle. Pour la suite, nous définissons par r le montant de la rétention et l le montant de la limite de couverture. Un traité est en excédent de sinistre par risque lorsque la rétention et la limite sont définies par risque. Ainsi, sans l'introduction d'une limite de couverture par le réassureur, les engagements sont définis par :

$$L = \begin{cases} \sum_{i=1}^N (X_i - r)_+, & N > 0 \\ 0, & N = 0 \end{cases} \quad \text{et} \quad R = \begin{cases} \sum_{i=1}^N \min(X_i, r), & N > 0 \\ 0, & N = 0, \end{cases}$$

si on introduit une limite de couverture, on a :

$$L = \begin{cases} \sum_{i=1}^N \min\{(X_i - r)_+, l\}, & N > 0 \\ 0, & N = 0, \end{cases} \quad (1.2)$$

$$R = \begin{cases} \sum_{i=1}^N \min\{X_i, r\} \mathbb{I}_{x_i \leq r+l} + (X_i - (r+l)) \mathbb{I}_{x_i > r+l}, & N > 0 \\ 0, & N = 0, \end{cases} \quad (1.3)$$

avec $(X_i - r)_+ = \max\{X_i - r, 0\}$, $\mathbb{I}_{x_i \leq r+l}$ la fonction indicatrice qui indique si le montant de sinistre i est inférieur ou égal à la rétention plus la limite et $\mathbb{I}_{x_i > r+l}$ la fonction indicatrice qui indique si le montant de sinistre i est supérieur à la rétention plus la limite. Un traité est en excédent de sinistre par événement lorsque la rétention et la limite sont définies pour un événement qui génère plusieurs sinistres, alors que dans un traité en excédent de perte annuelle, la rétention et la limite sont définies pour protéger une accumulation de sinistres sur une période. Dans le cas d'un traité en excédent de perte annuelle, sans l'introduction d'une limite de couverture par le réassureur, on a :

$$L = \begin{cases} (\sum_{i=1}^N X_i - r)_+, & N > 0 \\ 0, & N = 0, \end{cases}$$

avec l'introduction d'une limite, on a :

$$L = \begin{cases} \min\{(\sum_{i=1}^N X_i - r)_+, l\}, & N > 0 \\ 0, & N = 0. \end{cases}$$

1.2.3 Tarification d'un traité de réassurance non-proportionnelle : le cas de l'excédent de sinistre par risque (XOL)

Pour rappel, dans un traité en excédent de sinistre par risque, le réassureur rembourse à l'assureur les montants de sinistres payés par celui-ci, lorsque le sinistre

dépasse un certain niveau appelé la priorité r , et dans la limite d'un montant l (l vs r).

Dans la Table 1.1, sont réparties les primes et les sinistres entre l'assureur et le réassureur, avec P le montant total des primes récoltées par l'assureur sur toutes les polices sujettes au traité. La prime cédée au réassureur est un taux de prime t qui est appliqué sur l'ensemble des primes sujettes au traité P . Le montant total des sinistres est réparti comme décrit dans l'équation 1.1. De façon pratique la

Table 1.1: Répartition des primes et sinistres entre l'assureur et le réassureur

Ref	Risque total	Risque conservé	Risque cédé
Primes	P	$(1 - t)P$	tP
Sinistres	$S = \sum_{i=1}^n X_i$	$\sum_{i=1}^n R_i$	$\sum_{i=1}^n L_i$

portion de prime cédée au réassureur est exprimée sous la forme d'un taux sur la limite de couverture appelée « rate on line » :

$$\text{Rate on line} = 100 \times \frac{\text{prime cédée en réassurance}}{\text{limite de la tranche du traité}}. \quad (1.4)$$

1.2.4 Expression de la prime pure de réassurance XOL pour une modélisation avec la loi GPD

Soit X_i , le montant du sinistre i modélisée par une loi de Pareto Généralisée. Le calcul de la prime pure correspond à calculer l'espérance mathématique du coût annuel des sinistres qui sont rapportés à l'assureur. Ainsi, deux composantes essentielles déterminent ce calcul, la fréquence et la sévérité :

$$E[S] = E[N] E[X_i]. \quad (1.5)$$

En appliquant l'équation 1.5, Eric Gillard et Mathieu Ribatet dans (Charpentier, 2014) définissent les composantes de la prime pure de réassurance. Dans un

premier temps, la fréquence annuelle est généralement modélisée par une distribution de Poisson en utilisant un modèle linéaire généralisé avec une fonction de lien logarithme. Ainsi, si N suit une distribution de Poisson de paramètre λ , la fréquence annuelle s'écrit sous la forme :

$$E[N] = \exp(Z\beta), \quad (1.6)$$

avec Z la variable explicative qui est l'année dans notre situation, et β l'estimateur par maximum de vraisemblance.

Étant donné que l'utilisation de la distribution de Pareto Généralisée nécessite la détermination d'un seuil optimal u des valeurs extrêmes, la fréquence sera ainsi ajustée avec la probabilité pour des sinistres individuels de dépasser ce seuil u :

$$P(X > u) = \frac{\text{Nombre de sinistres dont le montant dépasse } u}{\text{Nombre total des sinistres}}. \quad (1.7)$$

Dans un deuxième temps, la sévérité est définie avec une loi de Pareto Généralisée. Alors pour tout seuil optimal u , on assume que tous les sinistres individuels au-delà de ce seuil suivent une distribution de Pareto Généralisée de fonction de répartition G et de fonction de densité de probabilité L de paramètres ξ et β . Ainsi le montant espéré du montant payé par les réassureurs pour une rétention de $r > u$ et une limite de l s'écrit sous la forme :

$$E[X|X > u] = \int_r^{r+l} L(x)dG(x) + (r+l)(1 - G(r+l)). \quad (1.8)$$

Finalement en combinant les équations 1.6, 1.7 et 1.8, la prime pure que doit payer une compagnie d'assurance pour céder aux réassureurs tous les sinistres dont les montants dépassent r dans la limite de l (l xs r) s'écrit sous la forme :

$$\text{Prime pure} = E[N] P(X > u) E[X|X > u]. \quad (1.9)$$

Dans certains traités, le réassureur donne la possibilité à l'assureur de remettre en vigueur la limite de couverture d'une tranche dans le cas où celle-ci est totalement épuisée par des sinistres. La remise en vigueur peut être moyennant une prime ou gratuite. Pour obtenir plus de détails, (Mata, 2000) développe les principes de la tarification avec une remise en vigueur de la limite de la tranche d'un traité en excédent de sinistre.

CHAPITRE II

THÉORIE DES VALEURS EXTRÊMES

Dans ce chapitre, nous allons décrire les principaux éléments de base de la théorie des valeurs extrêmes. Dans un premier temps, nous présenterons la loi des maxima, qui avec certaines hypothèses de convergence va conduire aux lois des valeurs extrêmes généralisées ou « Generalized Extreme Values (GEV) ». Ensuite, nous présenterons la loi des excès ou « Peaks over threshold (POT) » qui aussi avec des hypothèses de convergences va s'apparenter aux lois des valeurs extrêmes généralisées. Enfin, les propriétés des quantiles extrêmes qui sont très utilisées dans le milieu de la réassurance sont abordées.

2.1 La loi des maxima

Soit X , une variable aléatoire dont la fonction de répartition est F et soit $\{X_1, \dots, X_n\}$ un échantillon aléatoire indépendant et identiquement distribué de cette variable. On définit $X_{n:n}$ comme le maximum de $\{X_1, \dots, X_n\}$. Ainsi, dans la loi des maxima, on s'intéresse à la fonction de répartition de $X_{n:n}$ dont l'expression est la suivante :

$$Pr[X_{n:n} \leq x] = (F(x))^n. \quad (2.1)$$

Si nous notons x_F la borne supérieure du support de X , on aura

$$x_F = \sup\{x \in \mathbb{R} | F(x) < 1\} \leq \infty.$$

En appliquant la limite pour $n \rightarrow \infty$ sur l'équation 2.1, on observe :

$$\lim_{n \rightarrow \infty} Pr[X_{n:n} < x] = \begin{cases} 0, & \text{pour tout } x < x_F, \\ 1, & \text{pour tout } x \geq x_F. \end{cases} \quad (2.2)$$

Le résultat de l'équation 2.2 n'est pas utile pour étudier le comportement asymptotique des lois des maxima. Pour être capable d'obtenir une approximation de la limite du maximum, on applique le théorème de la limite centrale en introduisant des constantes de normalisation $a_n \in \mathbb{R}$ et $b_n > 0$ telles que :

$$Pr\left[\frac{X_{n:n} - a_n}{b_n} < x\right] = (F(a_n + b_n x))^n \rightarrow G(x), \quad n \rightarrow \infty$$

où G est une fonction de répartition non dégénérée.

Ainsi, il est possible de démontrer que les lois limites pour le maximum sont des lois stables. Par définition, une loi de probabilité de fonction de répartition G est dite max-stable si, pour tout $n \geq 2$, il existe des constantes $a_n \in \mathbb{R}$ et $b_n > 0$ telles que

$$(G(a_n + b_n x))^n = G(x), \quad \text{pour tout } x \in \mathbb{R}.$$

2.2 Distribution des valeurs extrêmes généralisées (GEV)

Après avoir déterminé la convergence des lois des maxima et leur stabilité, nous allons présenter les lois limites possibles pour les maxima.

Théorème 2.2.1 (Fisher et Tippett, 1928). *Pour une suite de variables aléatoires indépendantes et de même fonction de répartition F , s'il existe des constantes*

de normalisation $a_n \in \mathbb{R}$ et $b_n > 0$, alors

$$\lim_{n \rightarrow \infty} Pr \left[\frac{X_{n:n} - a_n}{b_n} < x \right] = G(x). \quad (2.3)$$

Alors G est appelée la fonction de répartition de la loi des valeurs extrêmes généralisées (GEV), qui est représentée par (Jenkinson, 1955) et (Mises, 1936) sous la forme suivante :

$$G_{\mu, \sigma, \xi}(x) = \begin{cases} \exp\left(-\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right), & \text{si } \xi \neq 0 \text{ et } \mu + \frac{\xi x}{\sigma} > 0, \\ \exp\left(-\exp\left(-\left(\frac{x-\mu}{\sigma}\right)\right)\right), & \text{si } \xi = 0. \end{cases} \quad (2.4)$$

Le paramètre ξ est appelé l'indice de queue et il permet de classer la loi de GEV en trois domaines selon l'expression des fonctions de répartition et la valeur de ξ . Le premier domaine est le max-domaine d'attraction de G_+ qui correspond aux lois de Fréchet :

$$G_+(x) = \exp\left(-\left(1 + \xi x\right)^{-1/\xi}\right), \quad \text{pour } x \geq -1/\xi \text{ et } \xi > 0.$$

Il s'agit des lois à queues épaisses et dites « de Pareto ». On peut citer la loi de Cauchy, la loi de Pareto, la loi Log-gamma et la loi Student.

Le deuxième domaine est le max-domaine d'attraction de Λ qui correspondent aux lois de Gumbel :

$$\Lambda(x) = \exp\left(-\exp(x)\right), \quad \text{pour } x \in \mathbb{R} \text{ et } \xi = 0.$$

Il s'agit des lois à queues fines et dites « exponentielle ». On peut citer la loi exponentielle, la loi gamma et la loi normale.

Le troisième domaine est le max-domaine d'attraction de G_- qui correspondent aux lois de Weibull :

$$G_-(x) = \exp\left(-\left(1 + \xi x\right)^{-1/\xi}\right), \quad \text{pour } x \leq -1/\xi \text{ et } \xi < 0.$$

Il s'agit des lois bornées à droite. On peut citer la loi uniforme et la loi bêta.

2.2.1 Estimation de la loi du maximum

Nous voulons trouver une estimation des paramètres de la loi du maximum pour la variable aléatoire X_i , $i = 1, 2, \dots, n$. La difficulté dans cet exercice est que le maximum des n variables aléatoires est une seule observation. Dans cette situation, nous ne pourrions pas appliquer les méthodes d'estimation habituelles. Devant ce constat, une idée proposée par (Gumbel, 1958) nous permet d'éviter cette contrainte, par la subdivision de l'échantillon des variables aléatoires en plusieurs blocs. Cette méthode, plus connue sous le nom de « block Maxima », permet de déterminer les maximums de chaque bloc, puis en utilisant cet échantillon de maxima, l'on serait en mesure d'appliquer les méthodes usuelles d'estimation de paramètres telles que le maximum de vraisemblance. Ainsi les n variables aléatoires X_1, \dots, X_n sont subdivisées en m blocs de tailles $k = n/m$. On désigne par Y_i le maximum que nous observons pour le i^e bloc :

$$Y_i = \max\{X_{(i-1)k+1}, \dots, X_{ik}\}.$$

Il faut noter que la propriété de la max-stabilité est respectée car pour tout n , $k \in \mathbb{N}$, on a :

$$\max\{Y_1, \dots, Y_m\} = \max\{\max(X_1, \dots, X_k), \dots, \max(X_{(m-1)k+1}, \dots, X_{mk})\}.$$

On définit la log-vraisemblance de l'échantillon Y_1, \dots, Y_m de la manière suivante, si Y_i suit une distribution $GEV(\xi, \mu, \sigma)$:

$$\ell = -m \log(\sigma) - (1 - \xi^{-1}) \sum_{i=1}^m \log\left(1 + \xi \left(\frac{Y_i - \mu}{\sigma}\right)\right) - \sum_{i=1}^m \left(1 + \xi \left(\frac{Y_i - \mu}{\sigma}\right)\right)^{-1/\xi},$$

si $\xi \neq 0$

$$\ell = -m \log(\sigma) - \sum_{i=1}^m \exp\left(1 + \xi \left(\frac{Y_i - \mu}{\sigma}\right)\right) - \sum_{i=1}^m \left(1 + \xi \left(\frac{Y_i - \mu}{\sigma}\right)\right), \quad \text{si } \xi = 0.$$

Il est important de signaler que cette méthode peut entraîner des pertes de données par le simple fait que nous utilisons seulement une partie des données (l'ensemble des maxima) au lieu d'utiliser toutes les données de l'échantillon pour trouver les estimateurs. La Figure 2.1 nous montre graphiquement la méthodologie du « block Maxima » appliquée aux données des pertes incendies danoises. L'échantillon est subdivisé en plusieurs blocs annuels et chaque maximum annuel est représenté en rouge. Les estimations des paramètres de la distribution *GEV* par maximum de vraisemblance, sur les maxima des pertes incendies danoises, sont résumées dans la Table 2.1.

Table 2.1: Estimation des paramètres de la distribution *GEV* par maximum de vraisemblance des pertes incendies danoises

	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
Estimation	38	29	0,64
Écart-type	11	11	0,41

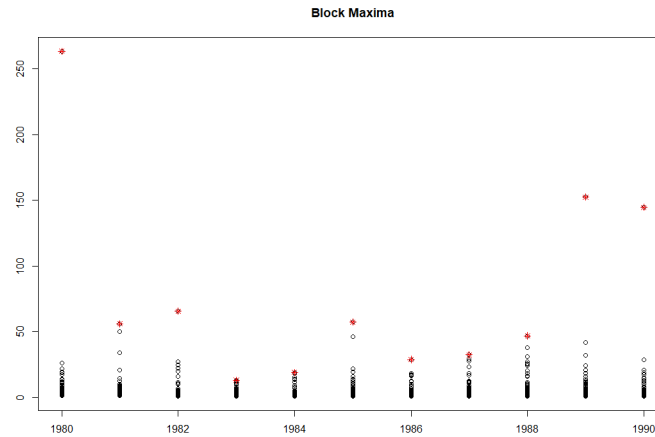


Figure 2.1: La méthodologie du block Maxima appliquée aux données des pertes incendies danoises.

2.3 La loi des excès et la distribution de Pareto Généralisée

Après l'étude de la loi des maximums et des distributions des valeurs extrêmes généralisées (GEV), nous allons nous intéresser à une autre méthode couramment utilisée dans la théorie des valeurs extrêmes, surtout dans le domaine de la réassurance. Cette méthode s'intéresse à la distribution d'une variable aléatoire X sachant que $X > u$ où u est une valeur suffisamment grande. Aussi, cette méthode met en lumière la distribution de Pareto Généralisée qui a des propriétés asymptotiques similaires aux distributions GEV. Soit X une variable aléatoire avec fonction de répartition H . Cette variable aléatoire suit une distribution de Pareto Généralisée de paramètre ξ si :

$$H(x) = \begin{cases} 1 - (1 + \xi x)^{-1/\xi}, & \text{si } \xi \neq 0 \\ 1 - \exp(-x), & \text{si } \xi = 0, \end{cases} \quad (2.5)$$

pour $x \geq 0$, si $\xi \geq 0$; et pour $0 \leq x \leq -1/\xi$, si $\xi < 0$. Cette distribution peut s'écrire aussi avec trois paramètres en remplaçant x par $(x - \mu)/\beta$. Enfin, la version

la plus courante que nous utiliserons pour la suite est celle qui s'écrit avec deux paramètres. On définit par $GPD(\xi, \beta)$ la fonction de répartition de la loi Pareto Généralisée avec deux paramètres :

$$GPD_{\xi, \beta}(x) = \begin{cases} 1 - \left(1 + \xi\left(\frac{x}{\beta}\right)\right)^{-1/\xi}, & \text{si } \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\beta}\right), & \text{si } \xi = 0, \end{cases}$$

pour $x \geq 0$, si $\xi \geq 0$; et pour $0 \leq x \leq -\beta/\xi$, si $\xi < 0$.

2.3.1 Estimation des paramètres

L'indice de queue ξ et le paramètre β peuvent être estimés par maximum de vraisemblance. Pour un seuil u suffisamment grand et pour un nombre k d'observations x_1, \dots, x_k excédant u , la log-vraisemblance s'écrit :

$$\ell(\xi, \beta) = -k \log \beta - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^k \log\left(1 + \xi\left(\frac{x_i}{\beta}\right)\right),$$

avec $\xi \neq 0$ et $\left(1 + \xi\left(\frac{x_i}{\beta}\right)\right) > 0$, pour tout $i = 1, \dots, k$.

L'estimation de l'indice de queue ξ introduite dans (Hill, 1975) est couramment utilisée en utilisant les statistiques d'ordre. L'idée est de proposer d'écrire l'estimateur de l'indice de queue pour les k plus larges observations d'un échantillon sous la forme suivante :

$$\hat{\xi} = \frac{\frac{1}{k} \sum_{i=1}^k \log(x_{n-i+1:n}) - \log(x_{n-k+1:n})}{\frac{1}{k} \sum_{i=1}^k \log\left(\frac{i}{n+1}\right) - \log\left(\frac{k}{n+1}\right)}.$$

Pour k suffisamment grand, le dénominateur tend vers 1 et l'expression de l'estimateur devient :

$$\hat{\xi}_k^{Hill} = \frac{1}{k} \sum_{i=1}^k \log(x_{n-i+1:n}) - \log(x_{n-k+1:n}).$$

La Figure 2.2 est une représentation graphique de l'estimateur de Hill selon les statistiques d'ordre, appliqué aux pertes incendies danoises. Le tracé en rouge représente l'écart-type autour de l'indice de queue.

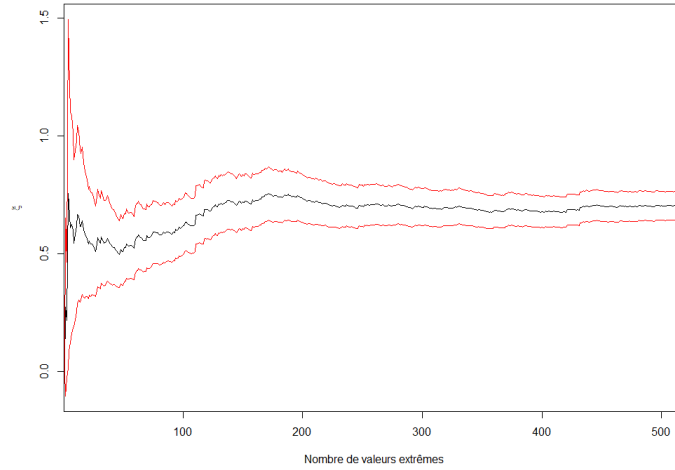


Figure 2.2: Estimateur de Hill appliqué aux données des pertes incendies danoises

2.3.2 Propriété de stabilité

La loi de Pareto Généralisée (GPD) est stable pour les montants excédant un certain niveau suffisamment grand.

Propriété 2.3.1. *Si $X \sim GPD$ avec pour fonction de répartition $GPD_{\xi,\beta}$, alors $(X - u|X > u)$ a aussi une de fonction de répartition qui s'écrit $GPD_{\xi,\beta+\xi u}$, pour tout seuil $u > 0$.*

Démonstration.

$$\begin{aligned}
 Pr[X - u > x|X > u] &= \frac{\overline{GPD}_{\xi,\beta}(x+u)}{\overline{GPD}_{\xi,\beta}(u)} \\
 &= \left(1 + \xi \left(\frac{x}{\beta + \xi u}\right)^{-1/\xi}\right) \\
 &= \overline{GPD}_{\xi,\beta+\xi u}(x), \quad \text{si } \xi \neq 0.
 \end{aligned}$$

□

2.3.3 Théorème de Pickands-Balkema-de-Haan

Ce théorème nous permet de mettre en parallèle la loi de Pareto Généralisée (GPD) et la loi des valeurs extrêmes généralisées (GEV) du maximum, en assimilant la loi limite des excès à la loi GPD.

Théorème 2.3.2 ((Pickands III *et al.*, 1975), (Balkema et De Haan, 1974)).
Considérons G la fonction de répartition d'une variable aléatoire X qui appartient à l'un des domaines d'attraction G_+, G_- ou Λ . Il existe alors une fonction $\beta(u)$ positive qui vérifie la relation suivante :

$$\lim_{n \rightarrow x_F} \sup_{0 < x < x_F} \{|Pr[X - u \leq x | X > u] - H_{\xi, \beta(u)}(x)|\} = 0, \quad (2.6)$$

avec $H_{\xi, \beta}(x)$ qui est la fonction de répartition de la loi GPD et $x_F = \sup\{x \in \mathbb{R} | G(x) < 1\} \leq \infty$.

À partir des résultats de ce théorème, on peut comprendre qu'il existe une relation entre la loi GPD et la loi GEV. Ainsi (Pickands III *et al.*, 1975) a pu montrer que pour n'importe quelle loi, l'approximation GPD selon l'équation 2.6 n'est vérifiée que s'il existe des constantes de normalisation et une loi non dégénérée comme vérifié dans l'équation 2.3. Ainsi l'indice de queue ξ de la loi GEV dans l'équation 2.4 est le même que le paramètre ξ de la loi GPD dans l'équation 2.5, pour des montants excédant un seuil u suffisamment grand.

2.4 Quantiles extrêmes

Dans cette section, nous allons présenter des métriques couramment utilisées dans le milieu de la réassurance en rapport avec la théorie des quantiles des valeurs extrêmes. Lors de l'application numérique, nous allons montrer comment leur interprétation est utilisée dans la tarification d'un traité de réassurance.

2.4.1 Nombre de dépassements

Le nombre d'observations excédant le seuil u suffisamment grand est approximé par une loi de Poisson. Cette notion nous permet de déterminer la fréquence conditionnelle utilisée pour la tarification des traités en excédent de sinistres.

Propriété 2.4.1. *Soit N_u le nombre d'observations excédant le seuil u dans un échantillon de taille n et X une variable aléatoire de fonction de répartition F . Si la suite des seuils (u_n) vérifie :*

$$\lim_{n \rightarrow \infty} n\bar{F}(u_n) = \tau,$$

alors, $N_u \xrightarrow{\mathcal{L}} \text{Poisson}(\tau)$, lorsque $n \rightarrow \infty$.

Démonstration. Soit :

$$N_u = \sum_{i=1}^n \mathbb{I}[X_i > U_n] \sim \text{Bin}(n, \bar{F}(u_n)),$$

$$E[N_u] = n\bar{F}(u_n),$$

$$\lim_{n \rightarrow \infty} E[N_u] = \lim_{n \rightarrow \infty} n\bar{F}(u_n) = \tau,$$

avec $\mathbb{I}[X_i > U_n]$ la fonction indicatrice qui indique les fois où la valeur de X est supérieur à U_n . □

Ce qui confirme que la variable aléatoire $N \xrightarrow{\mathcal{L}} \text{Poisson}(\tau)$, lorsque $n \rightarrow \infty$. De façon analogue, le nombre de sinistres au delà d'un seuil u suffisamment grand suit une loi $\text{Poisson}(\lambda_u)$, avec $\lambda_u = n\overline{\text{GPD}}_{\xi, \beta}(u)$. Ainsi cela vient justifier la formule 1.6 utilisée pour la fréquence.

2.4.2 Le plus grand sinistre ou sinistre maximum possible

Soit $\{X_1, \dots, X_n\}$, une suite de variables aléatoires indépendantes et de même fonction de répartition $\text{GPD}_{\xi, \beta}$. Le plus grand sinistre ou sinistre maximum pos-

sible (PML) qui touchera un portefeuille d'assurés au cours de l'année prochaine dans les conditions les plus défavorables, correspond au maximum de l'échantillon aléatoire qui touchera ce portefeuille. Cette notion est utilisée pour déterminer le pire scénario de pertes qui puissent affecter une compagnie d'assurance ou une entreprise en générale.

Proposition 2.4.2. *Soit $\{X_1, \dots, X_n\}$, une suite de variables aléatoires indépendantes et de même fonction de répartition $GPD_{\xi, \beta}$. Si une variable aléatoire $N \sim \text{Poisson}(\lambda)$, qui est indépendante de X_i , alors $M_N = \max\{X_1, \dots, X_N\} \sim GEV(\mu, \sigma, \xi)$ avec les paramètres qui sont les suivants : $\mu = \beta\xi^{-1}(\lambda^\xi - 1)$ et $\sigma = \beta\lambda^\xi$.*

Ainsi, le sinistre maximum possible avec une probabilité de ϵ selon une loi GPD s'écrit sous la forme :

$$PML_\epsilon = u + \left(\left(\frac{-\lambda}{\ln(1-\epsilon)} \right)^\xi - 1 \right) \frac{\beta}{\xi}. \quad (2.7)$$

Démonstration. Pour $\xi \neq 0$,

$$\begin{aligned} \Pr[M_N \leq x] &= \Pr[\max\{X_1, \dots, X_n\} \leq x] \\ &= \sum_{n=0}^{\infty} \Pr[\max\{X_1, \dots, X_n\} \leq x] \Pr[N = n] \\ &= \sum_{n=0}^{\infty} (\Pr[X_1 \leq x])^n \Pr[N = n] \\ &= \sum_{n=0}^{\infty} \frac{(\lambda GPD_{\xi, \beta}(x))^n}{n!} \exp(-\lambda) \\ &= \exp\left(-\lambda \left(1 + \frac{\xi(x-u)}{\beta}\right)^{-1/\xi}\right) \\ &= \exp\left(-\left(1 + \xi \frac{(x-u) - \xi^{-1}\beta(\lambda^\xi - 1)}{\beta\lambda^\xi}\right)^{-1/\xi}\right). \end{aligned}$$

Ainsi pour une probabilité de ϵ tel que $\Pr[M_N \leq PML_\epsilon] = 1 - \epsilon$,

$$1 - \epsilon = \exp\left(-\left(1 + \xi \frac{(PML_\epsilon - u) - \xi^{-1}\beta(\lambda^\xi - 1)}{\beta\lambda^\xi}\right)^{-1/\xi}\right)$$

$$(PML_\epsilon - u) = \left(\left(\frac{-\lambda}{\ln(1 - \epsilon)}\right)^\xi - 1\right) \frac{\beta}{\xi}$$

$$PML_\epsilon = u + \left(\left(\frac{-\lambda}{\ln(1 - \epsilon)}\right)^\xi - 1\right) \frac{\beta}{\xi}.$$

Pour $\xi = 0$

$$\Pr[M_N \leq x] = \exp\left(-\exp\left(-\frac{1}{\beta}(x - \beta \ln \lambda)\right)\right).$$

□

2.4.3 Notions de temps de retour, période de retour et niveau de retour

Ce concept est beaucoup plus utilisé dans le milieu de la réassurance pour déterminer entre autres les rétentions, les limites, le nombre de remise en vigueur d'une structure de réassurance, aussi pour approximer le sinistre maximum probable (*EML*) qui est le sinistre maximum qui peut arriver dans les conditions normales.

Définition 2.4.3. Soit (Y_1, \dots, Y_n) , les montants des maxima annuels indépendants et identiquement distribués et u un seuil suffisamment grand. Le temps de retour est la variable aléatoire associée au premier dépassement du seuil u :

$$N(u) = \inf\{i \geq 1 | Y_i \geq u\}.$$

La période de retour T quant à elle, se définit comme le temps moyen de retour $E[N(u)]$.

On note que $N(u)$ suit une loi Géométrique :

$$\begin{aligned}\Pr[N(u) = k + 1] &= \Pr[Y_1 < u, \dots, Y_k < u, Y_{k+1} \geq u] \\ &= \Pr[Y_1 < u] \Pr[Y_2 < u] \dots \Pr[Y_k < u] \Pr[Y_{k+1} \geq u] \\ &= (1 - p)^k p,\end{aligned}$$

avec $p = \Pr[Y \geq u]$. Alors la période de retour T est :

$$T = E[N(u)] = \frac{1}{p} = \frac{1}{\Pr[Y \geq u]}. \quad (2.8)$$

Par exemple le seuil u associé à un tremblement de terre de 500 ans comme période de retour (soit une probabilité de réalisation de 0,2%) vérifie $\Pr[Y \geq u] = 1/500$ ou $u = VaR[Y, 1/500]$. On définit aussi le niveau de retour ou sinistre maximum probable (EML) comme la VaR associée à la période de retour tel que :

$$T = E\left[N\left(VaR\left[Y; \frac{1}{T}\right]\right)\right]. \quad (2.9)$$

2.4.4 Valeur au risque pour une GPD

Soit $N_u = \sum_{i=1}^n \mathbb{I}[X_i > u]$ le nombre d'observations qui excèdent le seuil u , pour un échantillon de variables aléatoires X_1, \dots, X_n et de même fonction de répartition F . Pour $x > u$ on écrit :

$$\begin{aligned}\bar{F}(x) &= \Pr[X \geq u] \Pr[X > x | X > u] \\ &= \bar{F}(u) \Pr[X > x | X > u] \\ &= F(u) \bar{F}_u(x - u),\end{aligned}$$

avec $F_u(x) = \Pr[X - u \leq x | X > u] \sim GPD(\xi, \beta)$.

Un estimateur naturel de $\bar{F}(x)$ est l'estimateur empirique :

$$\hat{F}(x) = 1 - \frac{N_u}{n} \left(1 + \hat{\xi} \frac{x - u}{\hat{\beta}}\right)^{-1/\hat{\xi}},$$

pour tout $x > u$ et u suffisamment grand.

Un estimateur de $VaR[X; p]$ est défini par :

$$\hat{x}_p = u + \frac{\hat{\beta}}{\hat{\xi}} \left(\left(\frac{n}{N_u} (1-p) \right)^{-\hat{\xi}} - 1 \right).$$

En faisant référence à (2.9), nous définissons le niveau de retour EML associé à une période de retour de m années par :

$$\Pr[X > u] \left(1 + \xi \left(\frac{EML - u}{\beta} \right) \right)^{1-\xi} = \frac{1}{m}.$$

Si $(X|X > u) \sim GPD(\xi, \beta)$ alors

$$EML = \begin{cases} u + \frac{\beta}{\xi} ((m \Pr[X > u])^\xi - 1), & \text{si } \xi \neq 0 \\ u + \log(m \Pr[X > u]), & \text{si } \xi = 0. \end{cases} \quad (2.10)$$

La Figure 2.3 nous présente les différents niveaux de retour (EML), selon les périodes de retour appliquées aux pertes incendies danoises.

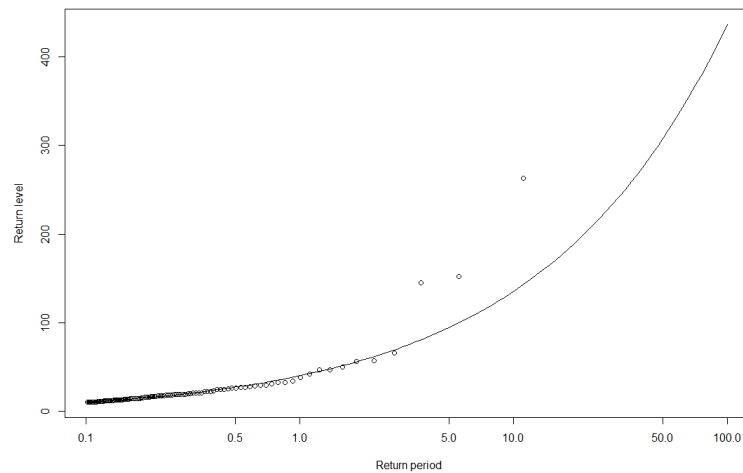


Figure 2.3: Graphe des niveaux de retour(EML) associés à des périodes de retour pour les pertes incendies danoises

CHAPITRE III

LES OUTILS GRAPHIQUES ET TESTS D'ADÉQUATION

Dans cette section, certains outils graphiques et les tests les plus couramment utilisés dans la modélisation des valeurs extrêmes sont énumérés. Nous présenterons des cas particuliers avec la loi Pareto Généralisée et/ou certaines lois usuelles.

On considère un échantillon X_1, \dots, X_n de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition F . On définit la statistique d'ordre associée à cet échantillon par :

$$\min(X_1, \dots, X_n) = X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n} = \max(X_1, \dots, X_n)$$

3.1 Le Graphique Quantile-Quantile

Ce graphique qui est appelé couramment « QQ-Plot », est un excellent outil visuel pour tester comment une famille de distribution de probabilité s'adapte à un jeu de données. Pour une distribution F donnée, il s'agit de comparer les quantiles théoriques ($F^{-1}(p)$) et les quantiles empiriques ($\widehat{F}^{-1}(p)$) pour plusieurs valeurs de p dans l'intervalle $[0, 1]$. Ainsi, une distribution est considérée adéquate à une base de données si l'observation du couple des points ($F^{-1}(p), \widehat{F}^{-1}(p)$) est alignée avec la droite qui relie les points quantile 25 % et 75 %. En pratique, l'idée est de représenter le couple de points ($F^{-1}(\frac{i}{n+1}), x_{i:n}$) pour $i = 1, \dots, n$.

3.1.1 Cas particulier de la distribution de Pareto Généralisée

La représentation graphique se fait avec une estimation du quantile qui est associé à l'observation de valeur $\frac{i}{n+1}$ et qui s'écrit sous la forme :

$$Q_p = \begin{cases} u + \frac{\beta}{\xi} \left(\frac{(1-p)^{-\xi}}{P(X>u)} - 1 \right) & \text{si } \xi \neq 0 \\ u + \beta \ln \left(\frac{1-p}{P(X>u)} \right) & \text{si } \xi = 0. \end{cases} \quad (3.1)$$

Le « QQ-Plot » pour la distribution de Pareto Généralisée se fait sur des observations qui dépassent un certain seuil optimal u , c'est pourquoi on admet que $P(X > u) = 1$ dans l'équation 3.1 et que les couples des points pour le graphe sont les suivants : $(\widehat{Q}_{(\frac{i}{n+1})}, x_{1:n})$, pour $i = 1, \dots, n$.

3.1.2 Cas particulier de la distribution Exponentielle

Le « QQ-Plot » de la distribution exponentielle est couramment utilisé pour déterminer si une distribution de probabilité possède une queue épaisse ou fine. La fonction quantile de cette loi est donnée par :

$$F^{-1}(p) = -\lambda \ln(1 - p),$$

avec $p \in]0, 1[$ et la fonction de répartition pour tout $x \geq 0$ s'écrit sous la forme :

$$F(x) = 1 - \exp(-\lambda x).$$

Les couples des points du « QQ-Plot » exponentielle en pratique sont les suivants : $\left(\frac{-\ln(1-\frac{i}{n+1})}{\lambda}, x_{i:n} \right)$, pour $i = 1, \dots, n$.

Si on observe une forme convexe du graphe lorsqu'il est comparé avec la droite de pente $1/\lambda$, alors vraisemblablement la distribution de la variable aléatoire sous-jacente a une queue plus épaisse, car les quantiles empiriques croissent plus vite que les quantiles théoriques exponentielle. Dans le cas contraire, c'est-à-dire si

le graphe a une forme concave, alors vraisemblablement la distribution de la variable aléatoire sous-jacente a une queue plus fine. Enfin, si les points du graphe sont presque alignés à la droite de pente $1/\lambda$, alors vraisemblablement la distribution de la variable aléatoire sous-jacente est exponentielle. Dans la Figure 3.1, nous présentons des graphes de « QQ-Plot » exponentielle de trois échantillons de simulation. Le premier graphe représente un échantillon de simulations de Pareto, qui est une distribution à queue épaisse avec une forme convexe. Quant au deuxième graphe, il représente un échantillon de simulation exponentielle avec une forme linéaire. Le dernier graphe, enfin, représente un échantillon de Weibull, qui est une distribution à queue fine avec une forme concave.

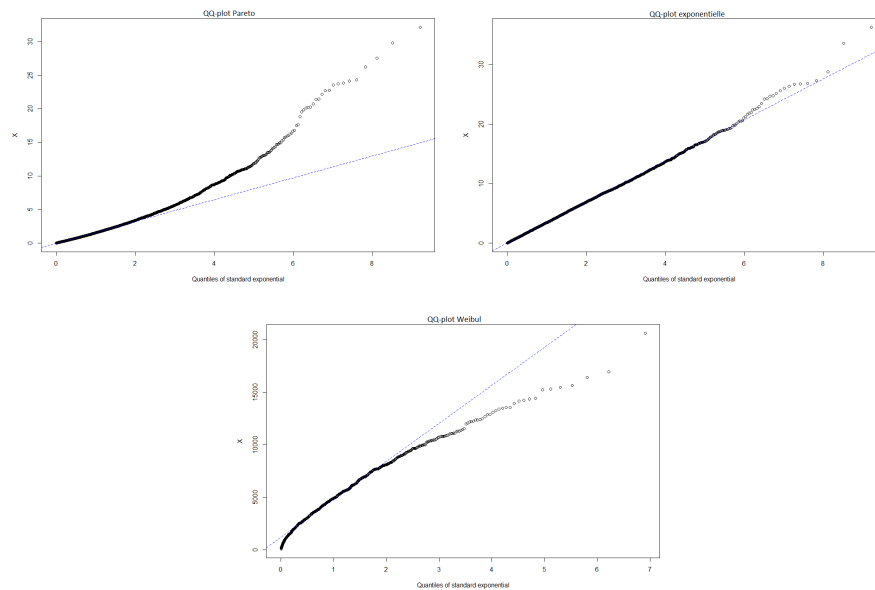


Figure 3.1: « QQ-Plot » de la distribution exponentielle pour des simulations de Pareto (convexe), Exponentielle (linéaire) et Weibull (concave)

3.2 La fonction d'excès moyen

L'analyse de la fonction d'excès moyen porte sur le comportement de la distribution conditionnelle pour les événements qui dépassent un certain seuil u positif ($X - u | X > u$). L'idée est de représenter graphiquement l'excédent moyen pour un échantillon de données en fonction d'un seuil u (par exemple des données de sinistres) :

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

En pratique l'excédent moyen conditionnel est estimé par sa version empirique :

$$e_{k:n} = \frac{1}{k} \sum_{i=1}^n X_{n-i+1:n} - X_{n-k:n}.$$

Ainsi, le graphe de la fonction d'excès moyen est représenté par le couple de points $(X_{k:n}, e_{k:n})$. Cette fonction est aussi très utile dans la détermination d'un seuil acceptable u_0 qui dans un premier temps doit être relativement grand pour respecter les critères de convergence en loi, et dans un second temps doit avoir un nombre considérable d'observations supérieures à ce seuil optimal u_0 pour être en mesure de trouver des estimateurs adéquats. Aussi pour tout seuil $u > u_0$, le graphe de la fonction d'excès moyen doit être relativement linéaire. Enfin, en observant le comportement de cette fonction, il est possible d'avoir une idée du comportement de la queue de distribution de la variable aléatoire à l'étude, comme résumé dans la Table 3.1 et observé dans la Figure 3.2.

3.2.1 Cas de la distribution exponentielle

Pour une variable aléatoire X qui suit une loi exponentielle de paramètres λ , on a $F(X) = 1 - \exp(-\lambda x)$. La fonction de répartition de la distribution conditionnelle

Table 3.1: Comportement de la fonction d'excès moyen

Graphe de $e(u)$	Distribution de la Variable aléatoire X
Pente nulle	Exponentielle
Pente croissante	Distribution à queue plus épaisse que Exponentielle
Pente décroissante	Distribution à queue plus fine que Exponentielle

$(X - u|X > u)$ s'écrit sous la forme :

$$\begin{aligned}
 F_u(x) &= P(X - u \leq x | X > u) \\
 &= \frac{F(u+x) - F(u)}{1 - F(u)} \\
 &= \frac{(1 - \exp(-\lambda(x+u))) - (1 - \exp(-\lambda u))}{1 - (1 - \exp(-\lambda u))} \\
 &= 1 - \exp(-\lambda x) \\
 &= F(x)
 \end{aligned}$$

Puisque $F_u(x) = F(x)$, alors

$$e(u) = E[X - u | X > u] = E[X] = \lambda.$$

Ainsi, la fonction d'excès moyen s'écrit de la sorte :

$$e(u) = \lambda, \quad \text{pour tout } u > 0.$$

3.2.2 Cas de la distribution Pareto Généralisée

Pour une variable aléatoire $(X - u|X > u)$ qui suit une loi de Pareto Généralisée de paramètres β et ξ , la fonction d'excès moyen s'écrit sous la forme :

$$E[X - \tilde{u} | X > \tilde{u}] = \frac{\beta + \xi \tilde{u}}{1 - \xi}, \quad \text{avec : } \begin{cases} \tilde{u} \geq u \\ \xi < 1. \end{cases}$$

En effet, selon le théorème 2.3.2, on déduit que $F_u(x)$ est équivalent à $H_{\xi, \beta(u)}(x)$ avec H qui est la fonction de répartition de la loi *GPD* de paramètres ξ et $\beta(u) = \beta + \xi u$. Puisque $F_u(x) = H_{\xi, \beta(u)}(x)$, alors

$$e(u) = E[X - u | X > u] = E[X] = \frac{\beta(u)}{1 - \xi} = \frac{\beta + \xi u}{1 - \xi}.$$

Cette fonction est une fonction linéaire de u . On spécifie que $\xi < 1$ pour que $E[X - \tilde{u} | X > \tilde{u}]$ nous donne une valeur finie.

Dans la tarification des traités de réassurance en excédent de sinistres, une analogie est faite entre la méthode du « burning cost » et la fonction d'excès moyen. Ainsi pour un traité en excédent de sinistre avec une rétention r , pour tout k tel que $r = X_{n-k:n}$, la version empirique de la prime pure s'écrit de la sorte :

$$\hat{\Pi}(r) = \frac{k}{n} e_{k,n}.$$

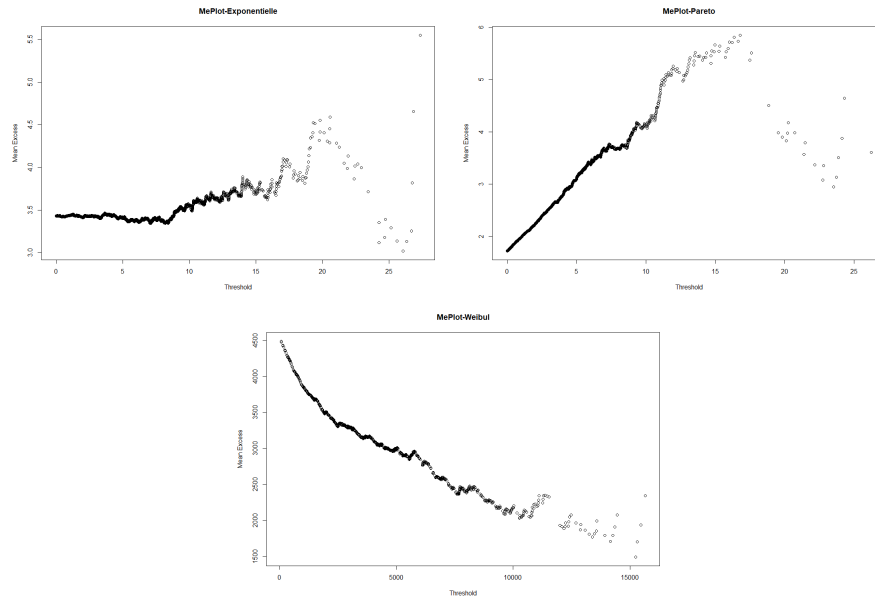


Figure 3.2: Graphe de la moyenne excès des simulations des distributions Exponentielle (pente constante), Pareto (pente croissante) et Weibul (pente décroissante)

3.3 Le test de Kolmogorov-Smirnov

On définit par G la fonction de répartition de la loi de distribution connue. Les hypothèses qui sont formulées pour le test sont :

$$H_0 : F = G, \quad (3.2)$$

$$H_1 : F \neq G. \quad (3.3)$$

Le test de Kolmogorov-Smirnov permet de tester de façon non-paramétrique si deux échantillons ont une distribution commune (ou vraisemblablement comparable), ou si un échantillon est semblable à une loi de probabilité dont la distribution est bien connue. Dans ce cas, les hypothèses sont plutôt formulées sous la forme d'une mesure de l'écart maximum entre la fonction de répartition empirique

et la fonction de répartition théorique d'une distribution connue.

La statistique pour le test de Kolmogorov-Smirnov tient compte de la distribution empirique de l'échantillon F :

$$D_n = \sup_x |F(x) - G(x)|.$$

Autrement,

$$D_n = \max\left\{\max\left(\frac{i}{n} - G(x_i), 0\right); \max\left(G(x_i) - \frac{i-1}{n}, 0\right)\right\},$$

avec x_i la i^e observation de l'échantillon, et n le nombre total des observations.

Si $D_n \geq K_\alpha$, nous sommes dans la région critique et le test est rejeté; avec K_α qui est le quantile à $(1 - \alpha)$ de la Table de Kolmogorov-Smirnov.

3.3.1 Cas particulier dans le milieu de la réassurance

En pratique, les données reçues par les réassureurs en excédent de sinistre de la part des cédantes sont souvent des données tronquées. Dans ce contexte, une modification de la fonction de répartition théorique doit se faire de la sorte :

$$G'(x) = \frac{G(x) - G(r)}{G(l) - G(r)}, \quad \text{avec : } \begin{cases} r : \text{la rétention} \\ l : \text{la limite de la tranche.} \end{cases}$$

3.4 Le test d' Anderson-Darling

Dans le même principe que le test de Kolmogorov-Smirnov, l'idée est de calculer une distance entre la fonction de répartition empirique et la fonction de répartition théorique puis de comparer l'écart dans la Table d'Anderson-darling.

La statistique pour le test d'Anderson-Darling tient compte de la distribution empirique de l'échantillon F :

$$A_n^2 = n \int_{-\infty}^{+\infty} \frac{(F(x) - G(x))^2}{G(x)(1 - G(x))} dG(x).$$

Autrement,

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^n [(2i-1) \ln Z_i + (2n+1-2i) \ln(1-Z_i)],$$

avec $Z_i = G(x_i)$, x_i la i^e observation de l'échantillon, et n le nombre total des observations.

Si $A_n^2 \geq C_\alpha$, nous sommes dans la région critique et l'hypothèse H_0 3.2 est rejetée ; avec C_α qui est le quantile à $(1-\alpha)$ de la Table d'Anderson-Darling.

3.5 Le graphique de Gertensgarbe et Werner

Il s'agit d'une procédure qui est très bien connue dans la détection d'un niveau de seuil optimal u , à partir duquel les observations $x > u$ sont considérées comme des valeurs extrêmes, de sorte à appliquer la théorie des valeurs extrêmes sur ces données en dépassement. Ainsi (Gerstengarbe et Werner, 1989) ont développé cette procédure en se basant sur l'observation du comportement de la différence entre les observations ordonnées d'une série. En pratique, ils admettent un principe essentiel selon lequel en comparant la série des différences des observations extrêmes, on trouvera une différence de comportement avec la série des différences des observations non-extrêmes. De ce fait, la version séquentielle du test de Mann-Kendall est utilisée pour être capable d'identifier le changement dans la série des observations. On définit par $\Delta_i = x_{i:n} - x_{i-1:n}$ pour $i = 2, 3, \dots, n$ la série des différences des observations ordonnées. Alors le point optimal comme niveau du seuil pour les valeurs extrêmes sera obtenu en observant le comportement des séries $\{\Delta_i, i = 2, 3, \dots, n\}$.

Pour appliquer les test de Mann-Kendall, on définit les séries normalisées des U_i par :

$$U_i = \frac{U_i^* - \frac{i(i-1)}{4}}{\sqrt{\frac{i(i-1)(i+5)}{72}}},$$

avec $U_i^* = \sum_{k=1}^i n_k$ et $n_k =$ nombre de valeurs dans $\Delta_2, \dots, \Delta_k$ qui sont inférieures à Δ_k . La série de points U_i correspond à la série croissante.

Une deuxième série nommée U_p est calculée selon la même procédure, mais les différences des séries commencent par $\Delta_n, \dots, \Delta_2$.

Lorsque nous traçons les deux séries de points U_i et U_p , il est déduit que le point d'intersection de ces deux séries sera probablement le point seuil d'entrée dans la zone extrême. La Figure 3.3 nous illustre l'application de cette procédure sur les pertes incendies danoises, où le point d'intersection des séries U_i et U_p nous suggère un seuil de 9,4 .

Il faut souligner que cette procédure peut ne pas toujours fonctionner selon les données analysées. En effet, l'application de la version séquentielle du test de Mann-Kendall pour vérifier l'hypothèse nulle d'absence de changement dans les données observées présente une limitation à moins que les données observées soient uniformément distribuées. En plus, par définition, les séries des différences Δ_i ne sont pas uniformément distribuées, ce qui rejette l'hypothèse nulle du test de Mann-Kendall.

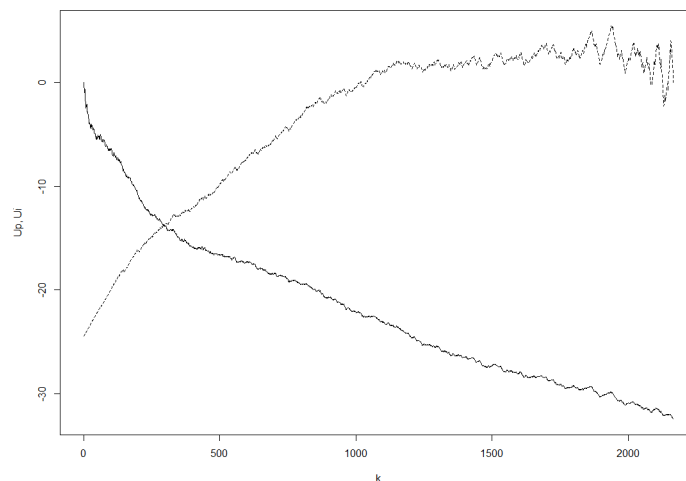


Figure 3.3: Graphe de Gertensgarbe et Werner appliqué aux pertes incendies danoises.

3.6 Inférence paramétrique- Estimation par Maximum de vraisemblance

Dans cette section, nous présentons brièvement une méthode d'inférence paramétrique, car tout au long de ce mémoire, nous serons appelés à estimer les paramètres pour une distribution paramétrique donnée.

Soit x_i $\{i = 1, \dots, n\}$ les réalisations de X_i . On assume que F se définit par $F(x, \theta)$ avec $\theta \in \Theta \subset \mathbb{R}^d$.

Il existe plusieurs méthodes qui permettent de trouver un estimateur $\hat{\theta}$ de θ , telles que la méthode des moments, la méthode des quantiles, ou la méthode du maximum de vraisemblance. Cette dernière sera celle que nous allons privilégier par la suite. L'estimation par le maximum de vraisemblance a pour but de maximiser la fonction de vraisemblance par rapport à θ . Ce qui est particulier dans cette méthode est que son estimateur présente les caractéristiques suivantes : il est défini de façon unique, son expression est explicite dans la plupart des cas, il est asymptotiquement sans biais avec une limite finie et est convergent.

Soit f la fonction de densité de probabilité de la variable aléatoire X continue, on définit la fonction de vraisemblance par :

$$\mathcal{L}(\theta, x_1, \dots, x_n) = \prod_{i=1}^n f_x(x_i; \theta).$$

De façon pratique, il est plus aisé de maximiser la log-vraisemblance par rapport à θ :

$$\ell(\theta, x_1, \dots, x_n) = \sum_{i=1}^n \log(f_x(x_i; \theta)).$$

3.6.1 Cas des données tronquées

Soit Y_i , pour $i = 1, \dots, n$, les montants tronqués par rapport à X_i , avec l la limite et r la rétention. La fonction de vraisemblance s'écrit :

$$\mathcal{L}(\theta, y_1, \dots, y_n) = \prod_{i=1}^n f_y(y_i; \theta) = \prod_{i=1}^n \frac{f_x(x_i)}{F_X(l) - F_X(r)}.$$

La log-vraisemblance s'écrit alors :

$$\ell(\theta, y_1, \dots, y_n) = \sum_{i=1}^n \log f_y(y_i; \theta) = \sum_{i=1}^n \log f_x(x_i) - n \ln[F_X(l) - F_X(r)].$$

CHAPITRE IV

PRÉSENTATION DE L'OUTIL DE SIMULATION *LOSS SIMULATOR 2.0*

L'outil *Loss Simulator 2.0* a été développé par des actuaires et mis à la disposition de la *Casualty Actuarial Society* (CAS) dans l'optique de simuler des données futures d'une base de données existante selon plusieurs hypothèses. Plusieurs types de données de sinistres peuvent être modélisées : les sinistres ouverts, les sinistres fermés, les sinistres ré-ouverts, les sinistres survenus, mais non déclarés, les frais de dépenses. De plus, il est possible d'ajouter de la dépendance entre certaines composantes. Cet outil est utilisé dans ce mémoire pour générer des montants de sinistres simulés selon la distribution la plus appropriée à une base de données existante, pour différentes branches d'assurance, afin de mener une analyse sur la théorie des valeurs extrêmes. La Figure 4.1 présente de façon graphique l'architecture de l'outil. Nous n'allons pas dans le détail de l'architecture dans ce travail, mais allons plutôt présenter les principaux éléments qui rentrent en ligne de compte dans la simulation.

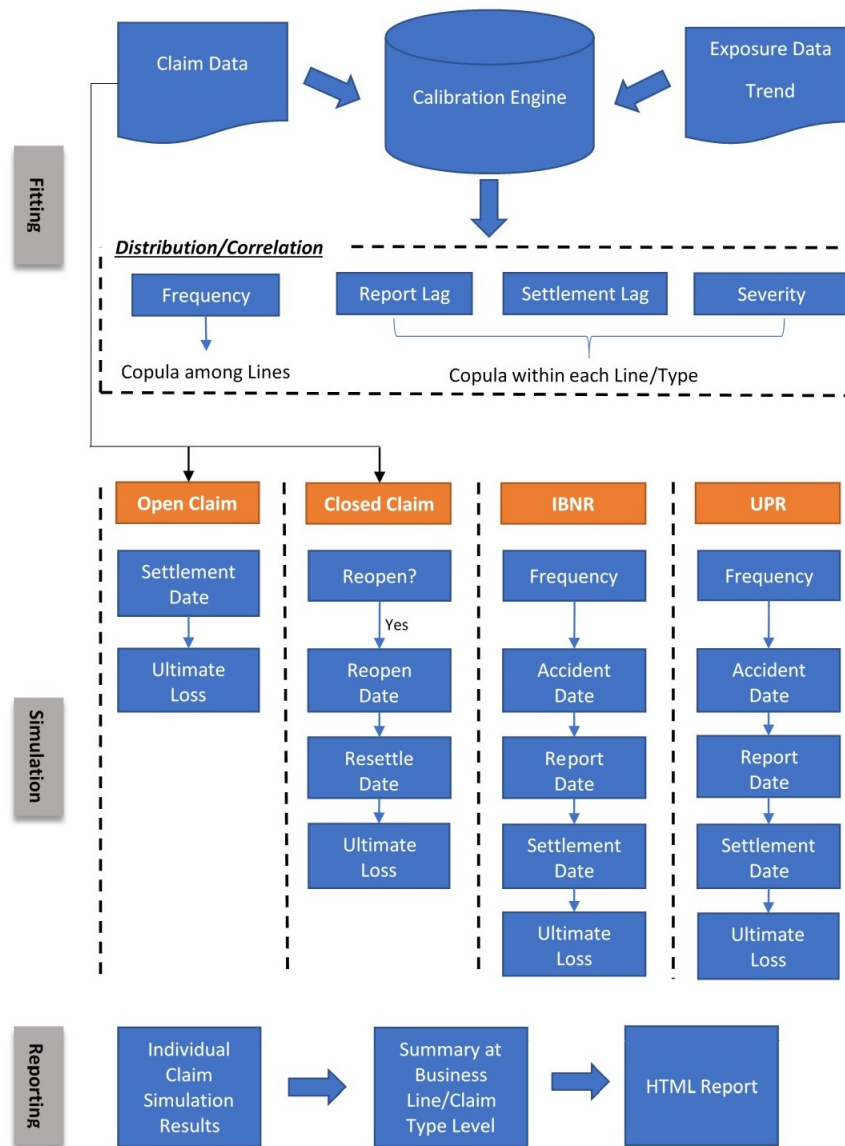


Figure 4.1: Architecture de la simulation de l'outil *Loss Simulator 2.0* - extrait de CAS *Loss Simulator 2.0 Methodology*

4.1 Le processus de simulation

Avant de commencer les étapes de la simulation, nous avons besoin d'une base de données contenant des informations du service sinistre et qui doivent contenir au minimum les informations suivantes : le numéro de sinistre, la branche d'affaires, le type de sinistre, le statut du sinistre, le montant du sinistre, la date de la perte et la date de déclaration. D'autres informations peuvent être obtenues pour rendre la simulation plus complète et réaliste, mais ces informations minimums sont requises pour que l'outil puisse produire un résultat. Dans un premier temps, nous définissons les variables qui vont servir de base pour la simulation. Pour chaque variable d'intérêts, un modèle de distribution paramétrique sera ajusté. Pour certaines variables, il sera possible d'ajuster un modèle avec les copules.

4.1.1 L'ajustement des variables d'intérêts

À partir d'une base de données existante, quatre modèles sont ajustés : un pour le délai de déclaration d'un sinistre, un pour le délai de règlement d'un sinistre, un pour la sévérité et un pour la fréquence. En plus de ces quatre modèles, une franchise et une limite peuvent être directement appliquées sur les données.

4.1.1.1 Le délai de déclaration d'un sinistre

La date à laquelle l'assureur est avisé de l'existence d'un sinistre est souvent différente de la date à laquelle le sinistre est survenu comme illustré à la Figure 4.2 qui décrit le cycle de développement d'un sinistre individuel :

$$\text{délai de déclaration} = \text{date de déclaration} - \text{date de survenance.}$$

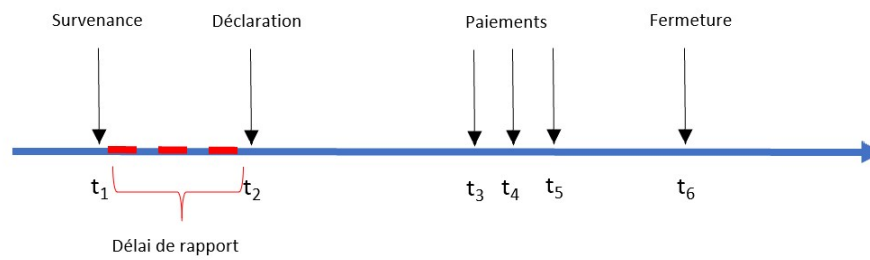


Figure 4.2: Délai de rapport ou déclaration dans le cycle de développement d'un sinistre individuel

4.1.1.2 Le délai de règlement d'un sinistre

De plus, il existe un délai entre le moment où le sinistre est rapporté et l'instant où les paiements sont effectués jusqu'à la fermeture du dossier comme illustré à la Figure 4.3 :

délai de règlement = date de fermeture – date de déclaration.

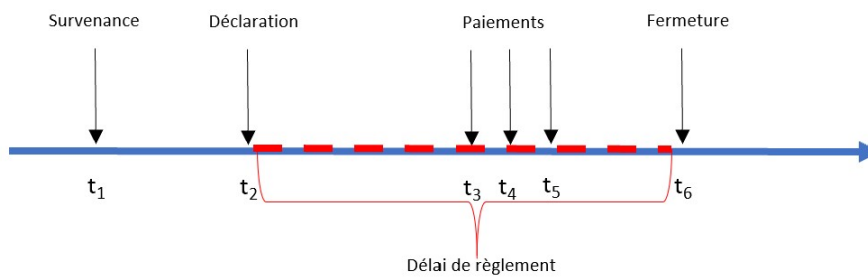


Figure 4.3: Délai de règlement dans le cycle de développement d'un sinistre individuel

4.1.1.3 La sévérité

La sévérité nous permet d'évaluer le coût d'un sinistre, comme illustré à la Figure 4.4 qui montre les montants totaux des pertes encourues pour des sinistres incendies, au Danemark, entre 1980 et 1990. Pour la simulation, les montants utilisés doivent être réduits de tout indice de sévérité. Par définition, l'indice de sévérité est un coefficient d'augmentation appliqué aux sinistres déjà payés pour tenir compte de l'inflation, des changements survenus sur les exigences imposées par la loi et de l'évolution du coût de la vie. Par exemple un sinistre payé en 2010 pour un montant de 10 000, sera évalué en 2021 par un montant 110 000, en considérant un indice de sévérité de 10% de 2015 à 2021.

$$\text{sévérité} = \frac{\text{montant total des pertes}}{\text{indice de sévérité}}.$$

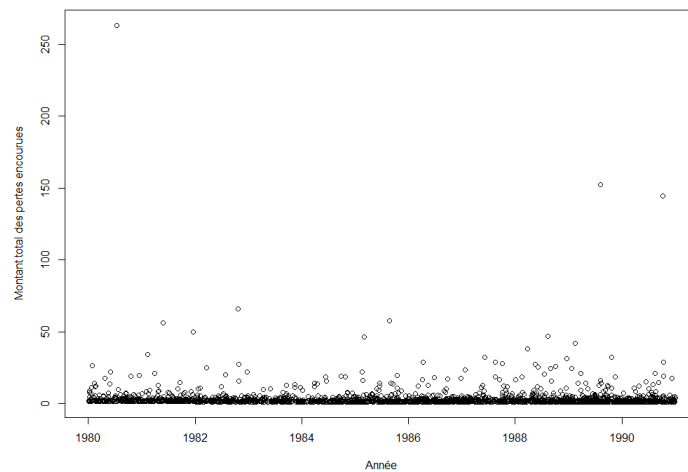


Figure 4.4: Graphe des montants totaux des pertes encourues incendies danoises

4.1.1.4 La fréquence

Pour calculer la fréquence, on détermine le nombre de sinistres survenus selon la date de survenance comme illustré dans la Figure 4.5 qui montre le nombre de sinistres cyclones survenus aux États-Unis de 1990 à 2005. Un ajustement est fait pour tenir compte du fait qu'il peut avoir des sinistres survenus, mais non déclarés, par l'application d'un facteur qui tient compte du délai de déclaration d'un sinistre :

$$\text{nombre de sinistres} = \frac{\text{nombre de sinistres déclarés}}{F[\text{délai de déclaration} < T]},$$

avec $T = \text{date d'évaluation} - \text{date du milieu du mois}$ et F la fonction de répartition du délai de déclaration.

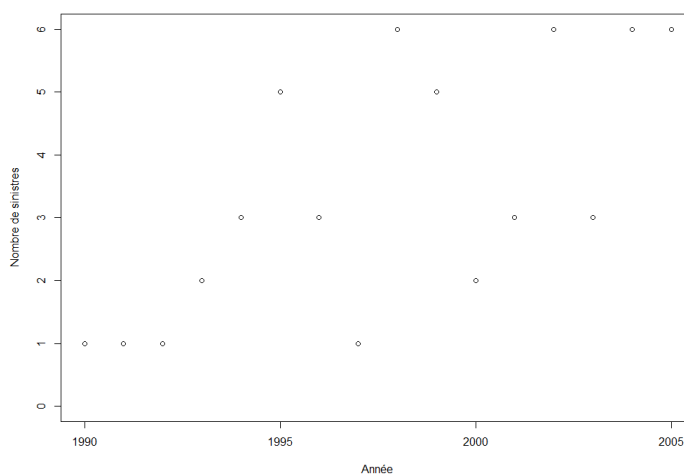


Figure 4.5: Graphe du nombre total des sinistres cyclone par année des pertes encourues aux états unis

L'impact de l'évolution du portefeuille année après année est exclu en divisant par l'indice de l'exposition. Par définition, l'indice d'exposition est un facteur appliqué à l'exposition pour tenir compte du changement des affaires que la compagnie

souscrit au fil du temps. Par exemple, une branche d'affaires qui a son exposition qui dépend du chiffres d'affaires, qui valait 100 000 en 2015, sera évalué à 110 000 en 2021 avec un facteur de 1,1.

$$\text{nombre de sinistre ajusté} = \frac{\text{nombre de sinistre}}{\text{indice d'exposition}}.$$

4.1.2 Les distributions paramétriques

Plusieurs distributions paramétriques peuvent être utilisées pour modéliser les variables suivantes : le délai de rapport, la fréquence, le délai de règlement et la sévérité. Parmi les distributions pouvant être utilisées pour une variable aléatoire discrète, on retrouve : la Poisson, la Binomiale négative, et la Géométrique. Parmi les distributions pouvant être utilisées pour une variable aléatoire continue, on retrouve : l'Exponentielle, la Normale, la Lognormale, la Weibull, la Pareto, la Gamma et l'Uniforme. Pour la franchise et la limite, l'outil utilise directement les données qui sont dans la base de données initiale.

4.1.3 L'ajustement d'un modèle de copule

Une étude de la corrélation entre différentes variables est réalisée. Il est possible d'étudier la corrélation à l'intérieur de chaque ligne d'affaires entre les variables aléatoires représentant la sévérité, le délai de rapport et le délai de règlement. De même, la corrélation entre les lignes d'affaires peut être étudiée entre les variables aléatoires représentant les fréquences. Plusieurs familles de copules peuvent être utilisées dans l'outil : la copule Gaussienne, la copule t, la copule de Clayton, la copule de Franck, la copule de Gumbel et la copule de Joe.

4.2 Les types de sinistres simulés

L'outil nous permet de générer quatre types de sinistres comme illustrés sur la Figure 4.6. Premièrement, les sinistres survenus, ou ouverts, sont les sinistres survenus et pour lesquels une déclaration a été faite à l'assureur, mais le paiement n'a pas été complété. Deuxièmement, les sinistres survenus, mais non déclarés sont ceux qui sont survenus, mais qui n'ont pas fait l'objet d'une déclaration à l'assureur. Troisièmement, les sinistres futurs sont ceux qui ne sont pas encore survenus. Enfin, les sinistres ré-ouverts sont les sinistres qui ont été fermés par le service des sinistres de l'assureur, mais pour lesquels de nouvelles informations nécessitent que ces sinistres soient ré-ouverts.

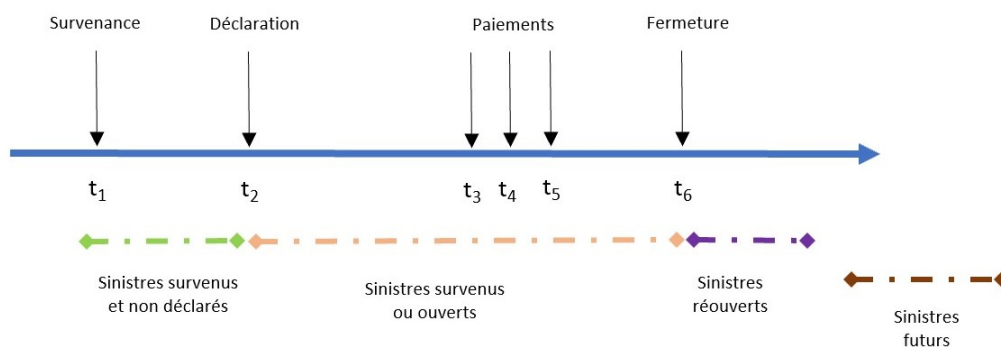


Figure 4.6: Types de sinistres dans le cycle de développement d'un sinistre individuel

4.2.1 Les sinistres survenus ou ouverts

Pour les sinistres ouverts, la date de survenance, la date de déclaration et le montant encouru de la perte sont connus. L'outil permettra alors de simuler les trois variables manquantes que sont la date de règlement, le montant de la perte à l'ultime et les frais liés à la perte.

4.2.1.1 Le délai de règlement

Une fonction de répartition est utilisée pour simuler le délai de règlement en supposant que celui-ci est supérieur au temps qui s'est écoulé entre la date d'évaluation et la date de déclaration. On définit par F la fonction de distribution tronquée du délai de règlement :

$$F(x < X) = \frac{F_{SL}(x < X) - F_{SL}(x < T)}{1 - F_{SL}(x < T)}, \quad (4.1)$$

avec T = la date d'évaluation - la date de déclaration, et F_{SL} est la fonction de répartition du délai de règlement.

4.2.1.2 La perte ultime

Quatre méthodes sont disponibles dans l'outil pour déterminer les pertes à l'ultime. Dans un premier temps, on peut utiliser un facteur de développement année après année. L'outil exige que nous indiquions les valeurs de la moyenne du facteur de développement ainsi que l'écart-type qui va contrôler la partie aléatoire de la simulation. Par exemple, la Table 4.1 présente des hypothèses de facteur de développement d'un sinistre quelconque. Ainsi, nous allons déterminer le montant à l'ultime pour un sinistre ouvert d'un montant de 10 000 et déclaré dans un délai de 1 jour. Si on assume que la date d'évaluation correspond à la deuxième année de développement, et que le sinistre se fermera à la quatrième année, la perte ultime est simulée selon la formule suivante :

$$\text{Perte ultime} = 10000 \times (1,15 + 0,04 \times \varepsilon_1) \times (1,1 + 0,09 \times \varepsilon_2),$$

avec ε_1 et ε_2 des nombres aléatoires qui suivent une loi normale centrée réduite. Ensuite, l'outil nous permet d'utiliser un facteur de développement à partir d'une fonction lien. Cette fonction lien peut être la fonction identité, la fonction exponentielle, la fonction logarithme ou la fonction inverse comme présenté dans la

Table 4.1: Exemple de facteur de développement année après année

Année de développement	Facteur de développement	
	Moyenne	Écart-type
1	1,2	0,06
2	1,15	0,04
3	1,1	0,09
4	1,05	0,1
5	1	0

Table 4.2. Par exemple, nous allons déterminer la perte ultime pour un sinistre avec un montant encouru de 10 000, qui est à sa deuxième année de développement et un taux d'encouru de 0,1. Si nous avons un facteur de développement cumulatif qui est $\exp(0,1 + 0,001 \times d + 0,005 \times l + 0,002 \times os + \varepsilon)$, alors la perte ultime est simulée selon la formule suivante :

$$\text{Perte ultime} = 10000 \times \exp(0,1 + 0,001 \times 2 + 0,005 \times 10000 + 0,002 \times 0,1 + \varepsilon),$$

avec ε qui suit une loi normale centrée réduite.

Enfin, l'outil nous permet d'utiliser directement la distribution de sévérité ou une distribution de sévérité conditionnelle. Dans ce cas, la perte ultime sera le montant maximum entre la perte encoure actuelle et la perte encourue simulée avec la distribution de sévérité après l'application de la franchise et de la limite. La perte ultime est par la suite ajustée avec un indice de sévérité.

4.2.1.3 Les frais

Les frais qui sont simulés sont les frais liés directement au sinistre, par exemple l'embauche d'un expert en sinistre pour évaluer une perte. Dans le cas où la perte

Table 4.2: Facteur de développement cumulatif selon la fonction lien. On définit par β les estimateurs par maximum de vraisemblance, d l'année de développement, l le montant du sinistre en cours, os le taux d'encouru et ε une variable aléatoire qui suit une loi normale centrée réduite.

Fonction Lien	Facteur de développement cumulatif
Identité	$\beta_0 + \beta_1 d + \beta_2 l + \beta_3 os + \dots + \varepsilon$
Exponentielle	$\log(\beta_0 + \beta_1 d + \beta_2 l + \beta_3 os + \dots + \varepsilon)$
Logarithme	$\exp(\beta_0 + \beta_1 d + \beta_2 l + \beta_3 os + \dots + \varepsilon)$
Inverse	$(\beta_0 + \beta_1 d + \beta_2 l + \beta_3 os + \dots + \varepsilon)^{-1}$

ultime est simulée à partir d'un facteur de développement, les mêmes facteurs peuvent être utilisés pour simuler les frais. Par contre, si la perte ultime est simulée à partir d'une distribution de sévérité, les frais seront simulés soit à partir d'une fonction de régression comme dans la Table 4.2, soit à l'aide de facteurs année après année comme dans la Table 4.1. Ainsi, on aura :

$$\text{Frais simulés} = \text{frais actuels} \times \text{facteur de développement simulé.}$$

4.2.2 Les sinistres survenus mais non déclarés

Le nombre de sinistres survenus, mais non déclarés est simulé en fonction de la distribution de la fréquence et de la distribution du délai de déclaration selon l'équation :

$$\text{Nombre de sinistres non déclarés} = f \times p \times \text{Indice d'exposition,}$$

avec f la fréquence mensuelle simulée et p la probabilité de déclarer un sinistre après la date d'évaluation.

Par exemple, on souhaite évaluer en date du 31 décembre 2020 le nombre de

sinistres survenus, mais non déclarés basé sur les hypothèses que la fréquence mensuelle suit une distribution Poisson de moyenne 100 ($\lambda = 100$), que le délai de déclaration suit une distribution Exponentielle de 150 jours de moyenne ($\lambda = 0,01$) et que l'indice de sévérité est de 1. Si le nombre de sinistres simulé avec la distribution Poisson est 117 en octobre 2020, ainsi le nombre de sinistres survenus, mais non déclarés sera de 55 ($117 \times \exp(-0,01 \times 76) \times 1$), où 76 jours correspond au nombre de jours entre la date d'évaluation et le milieu du mois d'octobre. Une fois que le nombre de sinistres est déterminé, les étapes suivantes sont réalisées :

- la date de survenance pour chaque période d'accident est simulée avec une distribution uniforme ;
- le délai de rapport est simulé selon une fonction de répartition F :

$$F(x < X) = \frac{F_{RL}(x < X) - F_{RL}(x < T)}{1 - F_{RL}(x < T)},$$

avec T = la date d'évaluation - la date de survenance, et F_{RL} est la fonction de distribution du délai de déclaration ;

- le délai de règlement est simulé selon l'Équation 4.1,

$$\text{date de règlement} = \text{date de déclaration} + \text{délai de règlement};$$

- la perte ultime est simulée selon la méthode basée sur la distribution de sévérité en tenant compte de la limite et la franchise ; et
- les frais sont simulés selon les mêmes hypothèses que pour les frais des sinistres ouverts.

4.2.3 Les sinistres futurs

La simulation des sinistres futurs se fait en utilisant le même processus que celui des sinistres survenus, mais non déclarés, avec la seule exception que le nombre de sinistres ne dépend pas du délai de rapport.

4.2.4 Les sinistres ré-ouverts

Chaque sinistre fermé peut être simulé avec une possibilité d'être ré-ouvert ou non. Pour ce faire, on utilise soit une régression qui peut être linéaire, log-linéaire ou exponentielle, soit une table avec des probabilités de réouverture. Ainsi, lorsqu'un sinistre est ré-ouvert, la simulation de ce sinistre se fait selon les étapes suivantes :

- le délai de réouverture est simulé par une fonction de répartition F :

$$F(x < X) = \frac{F_{RO}(x < X) - F_{RO}(x < T)}{1 - F_{RO}(x < T)},$$

avec T = la date d'évaluation - la date de règlement précédent, et F_{RO} est la fonction de répartition du délai de réouverture (la différence entre la date de réouverture et la date de règlement précédent) ;

- le délai de règlement après la réouverture selon les mêmes hypothèses que l'Équation 4.1

date de repaiement = date de réouverture + délai de règlement;

- la perte ultime simulée selon les mêmes méthodes que pour les sinistres ouverts ; et
- les frais simulés avec les mêmes méthodes que pour les frais des sinistres ouverts.

CHAPITRE V

APPLICATIONS NUMÉRIQUES

Dans cette partie, nous allons présenter les résultats de plusieurs analyses qui ont pour but de comparer les impacts de différentes hypothèses sur des modèles de réassurance, basés d'une part sur la loi Pareto Généralisée (« Peaks Over Threshold »), puis sur des lois standards sélectionnées arbitrairement d'autre part. Pour ce faire nous utilisons des bases de données sinistres provenant des données simulées de sinistres futurs pour la ligne d'affaires Responsabilité civile d'une compagnie d'assurance fictive à partir de l'outil *Loss Simulator 2.0*. Ainsi, deux bases de données de sinistres futures allant des années fictives A1 à A12 sont générées selon les scénarios de simulations suivants :

- scénario 1 : à partir de la base de données initiale en responsabilité civile, 12 années de *pertes encourues* futures de la ligne d'affaires responsabilité civile sont générées, avec des facteurs de développement long, sans dépendance entre les variables, avec une franchise de 0 et une limite de 10 millions ;
- scénario 2 : à partir de la base de données initiale en responsabilité civile, 12 années de *pertes encourues* futures de la ligne d'affaires responsabilité civile sont générées, avec des facteurs de développement long, avec une dépendance de queue plus accentuée à l'aide d'une copule, avec une franchise de 0 et une limite de 10 millions.

Le modèle basé sur la loi Pareto Généralisée sera comparé avec d'autres modèles

en analysant les quantiles extrêmes et la tarification de traités de réassurance en excédent de sinistre.

5.1 Description des données

Dans cette section, nous présenterons les statistiques descriptives des données initiales et des données simulées. Les données initiales qui serviront de base pour notre simulation proviennent de la base de données nommée *claimdata* disponible sur le site internet de la CAS. La ligne d'affaires Responsabilité civile dans la base *claimdata* contient 1 805 sinistres de 2012 à 2016, avec 15 variables. La variable d'intérêt est *pertes encourues* qui représentent les sinistres payés plus les réserves à l'ultime. Ainsi, avec l'outil *Loss Simulator 2.0*, nous allons simuler les sinistres futurs pour cette base de données selon les *scénarios 1 et 2*.

5.1.1 Statistiques descriptives des données initiales

Dans la Table 5.1, nous présentons un résumé des données à travers les principales statistiques descriptives pour la variable *pertes encourues*. Nous constatons que 75 % des sinistres ont une valeur inférieure à 2 avec un maximum de 11 119, ce qui confirme à priori l'existence de valeurs extrêmes dans la base de données initiale en responsabilité civile.

5.1.2 Statistiques descriptives des données simulées

Dans cette section, nous allons présenter brièvement les données qui ont été générées avec l'outil *Loss Simulator 2.0* en se basant sur les données initiales. Les paramètres de fréquence et de sévérité ont été sélectionnés après une analyse d'ajustement des différentes distributions avec les lois les plus utilisées. Ainsi,

Table 5.1: Statistiques descriptives du montant des *pertes encourues* en responsabilité civile de la base initiale

Moyenne	271	Minimum	0,8
Quantile 25%	1,4	Maximum	11 119
Quantile 50%	1,7	Étendue	11 119
Quantile 75%	2	Écart-Type	710
Quantile 90%	1 280	Nombre de données	1 828
Coef. asymétrie	4,6	Coef. variation	2,6

dans la Table 5.2, les lois les plus appropriées sont classées selon les meilleurs résultats en ordre croissant de la valeur du critère *AIC*.

Table 5.2: Paramètres (Param.) de l'ajustement de la base de données initiale pour les *pertes encourues* en responsabilité civile

Type	Loi	Param. 1	Param. 2	Score (<i>AIC</i>)
Fréquence mensuelle	Poisson	$\lambda : 25$		406,3
	Bin. négative	$\mu : 25$	$\beta : 0,49$	424,4
Sévérité	Lognormale	$\mu : 1,60$	$\sigma : 2,56$	12 843,02
	Inv. Gauss	$\mu : 271,9$	$\lambda : 11,99$	14 509,96
	Weibull	$\theta : 21,64$	$\beta : 0,31$	15 468,44
	Gamma	$\alpha : 1 461$	$\beta : 0,18$	16 276,19

5.1.2.1 Les données simulées selon les critères du scénario 1

Nous avons généré 35 510 sinistres sur 12 ans allant de *A1* à *A12* en responsabilité civile (*scénario 1*) comme présenté dans la Table 5.3. Les paramètres de la fré-

quence suivent une loi Poisson de moyenne mensuelle λ de 25 et les paramètres de sévérité suivent une loi Lognormale de moyenne μ de 1,60 et un écart-type σ de 2,56. Nous constatons que 75 % des sinistres ont une valeur inférieure à 62 avec un maximum de 456 652, ce qui confirme à priori l'existence de valeurs extrêmes dans la base de données simulées en responsabilité civile. Aussi la représentation graphique de la fonction de densité des données initiales et des données simulées selon le *scénario 1*, à la Figure 5.1, confirme la présence de valeurs extrêmes dans les données simulées.

Table 5.3: Statistiques descriptives des *pertes encourues* simulées en responsabilité civile selon le scénario 1.

Moyenne	240	Minimum	1
Quantile 25%	4,0	Maximum	456 652
Quantile 50%	14,4	Étendue	456 651
Quantile 75%	62,3	Ecart Type	3 806
Quantile 90%	256	Nombre de données	35 510
Coef. asymétrie	79,9	Coef. variation	15,9

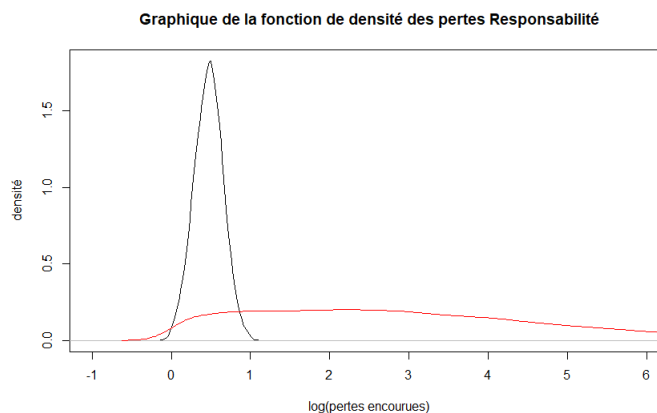


Figure 5.1: Graphe de la fonction de densité des *pertes encourues* en responsabilité civile. Les données initiales sont en noire et les pertes simulées selon le scénario 1 sont en rouge.

5.1.2.2 Les données simulées selon les critères du scénario 2

Pour le *scénario 2*, nous utiliserons les mêmes paramètres (fréquence et sévérité) que le *scénario 1*, puis nous allons introduire de la dépendance entre les distributions marginales du délai de déclaration, du délai de règlement et des montants des pertes encourues (*scénario 2*). Étant donné que la ligne d'affaire responsabilité civile est considérée comme une branche à développement des sinistres longs, nous verrons si l'introduction de la dépendance a un impact sur les résultats obtenus en appliquant les outils d'analyse des valeurs extrêmes. Alors la Table 5.4 présente les résultats des tests d'hypothèses de trois différentes copules qui sont classées par ordre décroissant, selon les valeurs p . Le choix des copules testées tient compte des copules qui ont un coefficient de dépendance de queue supérieur non nulle, parmi les copules disponibles dans l'outil *Loss Simulator 2.0*. Ainsi, la dépendance a été introduite à l'aide d'une copule de Gumbel à trois dimensions avec un paramètre α de 1,04.

Ainsi, nous avons généré 35 991 sinistres sur 12 ans allant de $A1$ à $A12$ en respon-

Table 5.4: Valeurs p de l'ajustement de la base de données initiale pour les *pertes encourues* en responsabilité civile par des copules.

Copule	valeur p
Gumbel	0,26
Student	0,11
Normal	0,05

sabilité civile (*scénario 2*) comme présenté dans la Table 5.5. Nous constatons que 75 % des sinistres ont une valeur inférieure à 68 avec un maximum de 589 243, ce qui confirme à priori l'existence de valeurs extrêmes dans la base de données simulées selon le scénario 2. Aussi la représentation graphique de la fonction de densité des données simulées selon le *scénario 1* et des données simulées selon le *scénario 2*, à la Figure 5.2, nous confirme que l'introduction de la dépendance a permis de générer une distribution avec une queue plus allongée à droite.

Table 5.5: Statistiques descriptive des *pertes encourues* simulées en responsabilité civile selon le scénario 2.

Moyenne	741	Minimum	1
Quantile 25%	4,3	Maximum	589 243
Quantile 50%	15,3	Étendue	589 242
Quantile 75%	68,7	Ecart Type	14 734
Quantile 90%	286	Nombre de données	35 991
Coef. asymétrie	29,9	Coef. variation	19,9

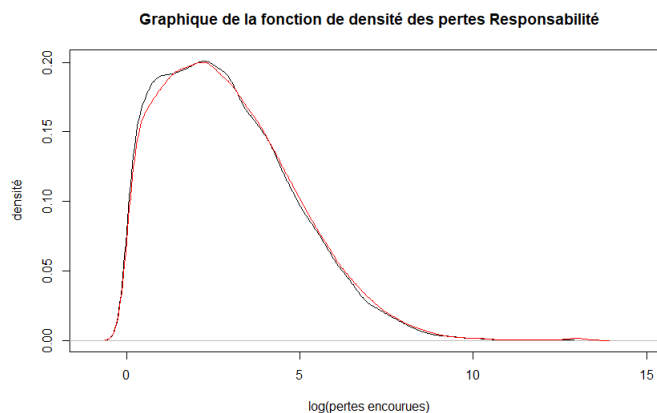


Figure 5.2: Graphe de la fonction de densité des *pertes encourues* simulées en responsabilité civile. Les pertes simulées selon le scénario 1 sont en noire et les pertes simulées selon le scénario 2 sont en rouge.

5.2 Ajustement des modèles de sévérité sur les données simulées

Dans cette partie, nous allons présenter les différents modèles de sévérité qui seront ajustés et qui seront utilisés pour notre analyse comparative. Ainsi, nous aurons un modèle basé sur la théorie des valeurs extrêmes et un autre basé sur les lois standards.

5.2.1 Modèle basé sur la théorie des valeurs extrêmes

Dans cette section, nous allons montrer les résultats de l'application de la loi des excès (« Peaks Over Threshold ») sur les deux scénarios de données simulées. Ainsi, les résultats des analyses de la queue de distribution et de la détermination d'un seuil optimal de valeurs extrêmes pour chaque scénario sont présentés.

5.2.1.1 Analyse de la queue de distribution

Nous allons déterminer si la distribution de la variable *pertes encourues* a une queue épaisse ou fine pour chaque scénario. La conclusion de cette analyse préliminaire permettra de savoir si nous pouvons utiliser la théorie des valeurs extrêmes sur ces bases de données. Au niveau de la Figure 5.3, nous avons le tracé du graphique de la fonction d'excès moyen empirique qui présente une tendance à la hausse, ce qui indique que la distribution simulée selon le *scénario 1* semble avoir une queue épaisse. Aussi, nous avons observé le graphique du quantile-quantile exponentiel appliqué sur la distribution des *pertes encourues* simulées. La Figure 5.4 nous montre clairement que la forme du graphe est convexe comparée à la droite de pente $1/\lambda$. Nous pouvons conclure que la distribution *pertes encourues* simulées selon le *scénario 1* indique une queue épaisse.

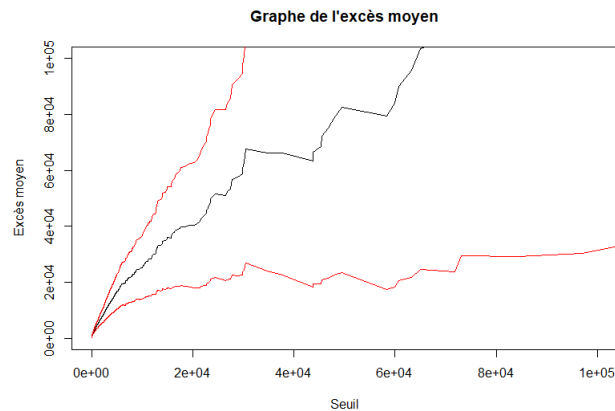


Figure 5.3: Graphe de la fonction d'excès moyen empirique des *pertes encourues* simulées selon le *scénario 1*

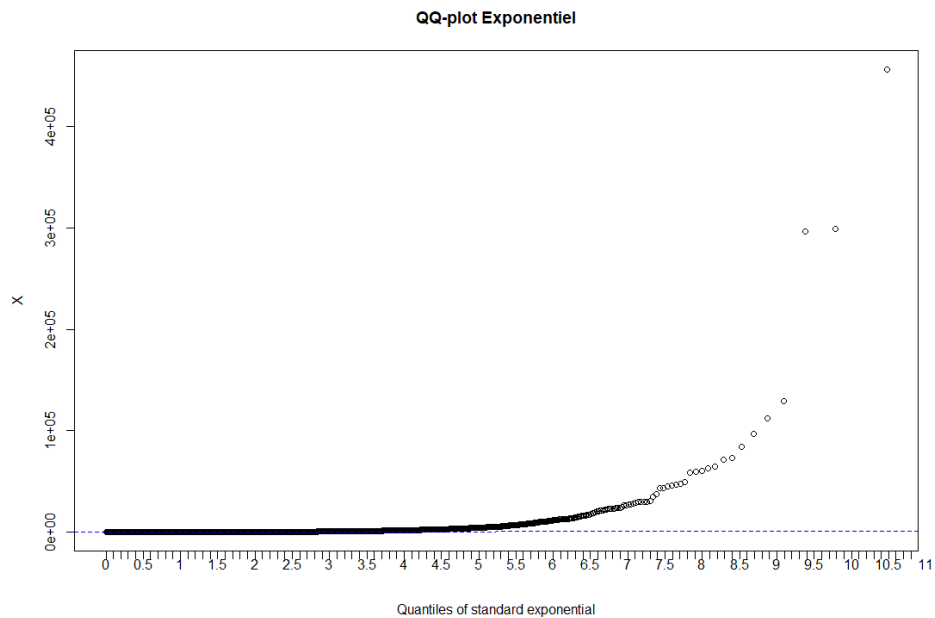


Figure 5.4: Graphe du quantile-quantile exponentiel des *pertes encourues* simulées selon le *scénario 1*

Aussi, en introduisant la dépendance entre les délais de règlement, les délais de déclaration et les montants des pertes encourues en responsabilité civile dans le *scénario 2*, la Figure 5.5 du tracé de la fonction d'excès moyen empirique présente toujours une tendance à la hausse. De même, le graphique du quantile-quantile exponentiel à la Figure 5.6 nous montre clairement que la forme du graphe est convexe comparée à la droite de pente $1/\lambda$. Ainsi la distribution des *pertes encourues* simulées selon le *scénario 2* indique également une queue épaisse.

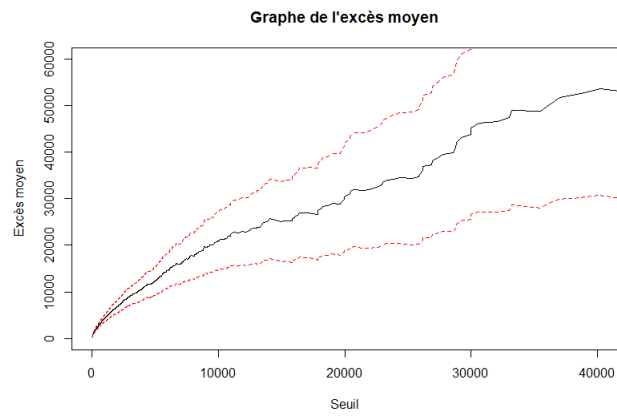


Figure 5.5: Graphe de la fonction d'excès moyen empirique des *pertes encourues* simulées selon le *scénario 2*

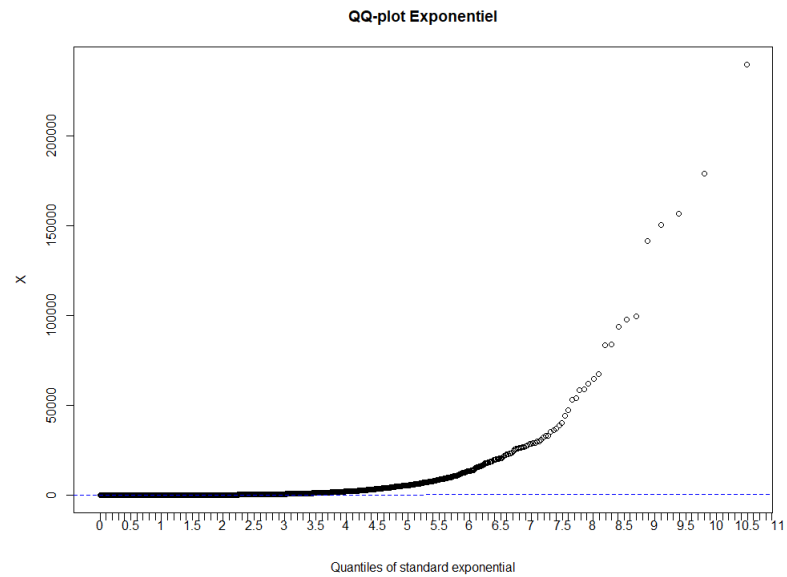


Figure 5.6: Graphe du quantile-quantile exponentiel des *pertes encourues* simulées selon le *scénario 2*

5.2.1.2 Le choix du seuil des valeurs extrêmes

Après avoir conclu que les distributions des *pertes encourues* simulées semblent toutes avoir une queue épaisse pour tous les scénarios, nous allons déterminer un seuil optimal à partir duquel les montants qui dépassent ce seuil sont des valeurs extrêmes. Une fois le seuil déterminé, cela nous permettrait de modéliser les *pertes encourues* simulées avec une loi de Pareto Généralisée.

Dans un premier lieu, nous allons déterminer un seuil optimal pour le *scénario 1*. Lorsque nous observons le graphe de quantile-quantile exponentiel des *pertes encourues* simulées selon le *scénario 1*, à la Figure 5.4, nous remarquons que le comportement du graphe commence à se détacher de la droite de pente $1/\lambda$ autour du point quantile de 4,5, ce qui correspond à un seuil de 2 772.

Ensuite, nous étudions le comportement de α , qui est l'inverse de l'index de queue ξ , comme suggéré dans (Charpentier, 2014) pour une loi de Pareto Généralisée avec différents seuils. L'idée est de trouver un comportement stable du graphe de l'index autour d'un seuil optimal. La Figure 5.7 nous montre que l'index $\alpha = 1/\xi$ de la loi Pareto Généralisée issue de la distribution des *pertes encourues* simulées selon le *scénario 1* est plutôt stable avec une valeur autour de 1,14 pour des seuils qui varient entre 2 710 et 3 287.

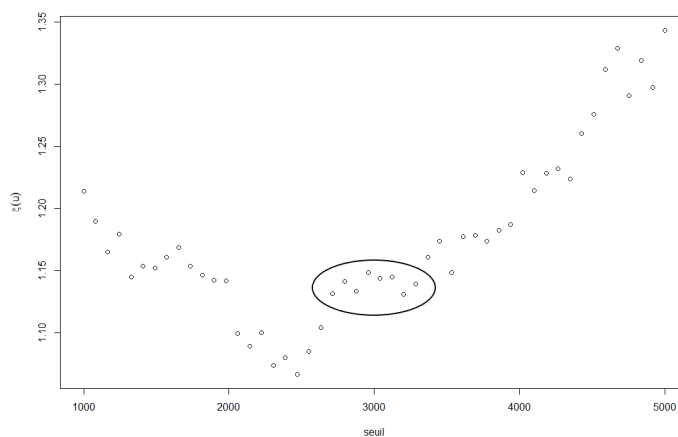


Figure 5.7: Graphe de l'indice de queue $1/\xi$ des *pertes encourues* simulées selon le *scénario 1*

Nous sommes sur le point de conclure que le seuil optimal pourrait être autour de 3 000 à partir duquel nous pouvons appliquer la distribution de Pareto Généralisée. Pour se conforter avec le choix du seuil de 3 000, nous allons modéliser la distribution des *pertes encourues* simulées selon le *scénario 1* avec une loi de Pareto Généralisée pour des niveaux de seuils différents, soit 1 000, 2 000, 3 000 et 4 000. Les résultats résumés dans la Table 5.6 nous montrent que pour les seuils entre 2000 et 3 000, les valeurs de probabilités issues du test d'Anderson-Darling sont les plus élevées. Aussi, nous constatons que dans les quatre situations, l'hypothèse H_0 3.2 n'est pas rejetée, ce qui ne nous donne pas un consensus sur la valeur de 3 000 comme seuil optimal, mais nous conforte que la valeur de 3 000 peut être choisie comme seuil optimal.

Dans un deuxième lieu, nous allons déterminer le seuil optimal pour le *scénario 2*. Ainsi, l'étude du comportement de l'inverse l'index de queue ξ d'une distribution de Pareto Généralisée avec différents seuils, nous montrons à la Figure 5.8 que l'index $\alpha = 1/\xi$ de la loi Pareto Généralisée issue de la distribution des *pertes*

Table 5.6: Paramètres de la loi Pareto Généralisée pour différents seuils pour les *pertes encourues* simulées selon le *scénario 1*. (*AD* pour le test d'Anderson-Darling)

Seuil	$\hat{\beta}$	s.e ($\hat{\beta}$)	$\hat{\xi}$	s.e ($\hat{\xi}$)	valeur p (AD)
1 000	1 114	63	0,82	0,05	0,374
2 000	1 795	145	0,88	0,08	0,525
3 000	2 772	282	0,88	0,10	0,516
4 000	3 953	468	0,83	0,11	0,303

encourues simulées selon le *scénario 2* est plutôt stable avec une valeur autour de 1,4 pour des seuils qui varient entre 2 795 et 3 204.

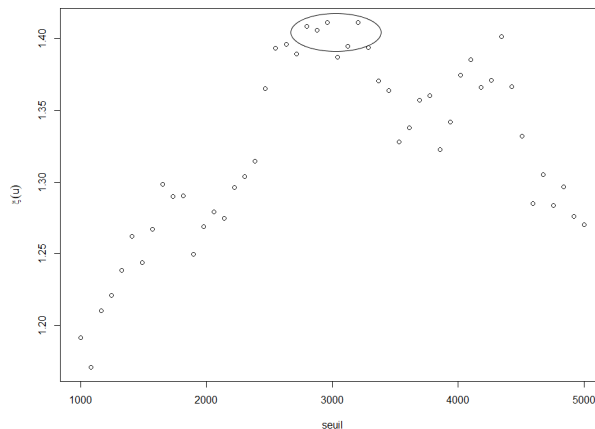


Figure 5.8: Graphe de l'indice de queue $1/\xi$ des *pertes encourues* simulées selon le *scénario 2*

Aussi la modélisation des *pertes encourues* simulées selon le *scénario 2* avec une loi de Pareto Généralisée pour des niveaux de seuils différents, nous donne des résultats qui confirment que le seuil de 3 000 est un choix optimal. Comme on le

voit dans la Table 5.7, le seuil de 3 000 a la valeur de probabilité la plus élevée même si pour les quatre seuils l'hypothèse H_0 3.2 n'est pas rejetée. Ces résultats nous amènent à conclure que malgré l'introduction dans le *scénario 2*, le seuil optimal reste toujours le même, soit 3 000.

Table 5.7: Paramètres de la loi Pareto Généralisée pour différents seuils pour les *pertes encourues* simulées selon le *scénario 2* (AD pour le test d'Anderson-Darling)

Seuil	$\hat{\beta}$	s.e ($\hat{\beta}$)	$\hat{\xi}$	s.e ($\hat{\xi}$)	valeur p (AD)
2 000	2 030	149	0,89	0,07	0,213
3 000	3 184	284	0,83	0,09	0,704
4 000	3 576	371	0,90	0,10	0,138
5 000	4 006	473	0,96	0,12	0,138

5.2.2 Modèle basé sur les lois standards

Dans cette partie, nous allons présenter les résultats des paramètres des modélisations des deux scénarios avec les lois Gamma (modèle 2) et Lognormale (modèle 3), qui sont des lois très utilisées dans le milieu de l'assurance pour analyser les sinistres. En plus, puisque nous ferons notre application dans le milieu de la réassurance en excédent de sinistres, nous nous intéressons aux paramètres qui s'adaptent le mieux aux données au-delà d'un certain seuil. Par souci de continuité, nous utiliserons le même seuil de 3 000, qui est celui déterminé avec la loi de Pareto Généralisée. Ainsi, dans la Table 5.8, sont résumées les paramètres des distributions de sévérité qui ont été déterminés par maximum de vraisemblance.

Table 5.8: Paramètres de l'ajustement des *pertes encourues* simulées, avec les lois standards

Scénario	Loi	Paramètres 1	Paramètres 2
Scénario 1	Gamma (modèle 2)	$\alpha : 0,9$	$\beta : 67,6$
	Lognormale (modèle 3)	$\mu : 8,9$	$\sigma : 0,9$
Scénario 2	Gamma (modèle 2)	$\alpha : 0,8$	$\beta : 49,2$
	Lognormale (modèle 3)	$\mu : 9,0$	$\sigma : 0,9$

5.3 Application de la théorie des valeurs extrêmes dans le milieu de la réassurance

Dans cette partie, nous allons montrer de façon pratique comment les résultats de l'application de la théorie des valeurs extrêmes sont interprétés dans le milieu de la réassurance. Nous allons par la suite, pour chaque scénario, comparer les résultats du calcul des niveaux de retour (EML) et la prime pure de réassurance, basés sur les trois modèles de sévérité ajustés.

5.3.1 Calcul des niveaux de retour pour différentes périodes de retour

Dans le milieu de la réassurance, cette notion est très populaire et souvent assimilée à la perte maximum probable qui peut atteindre une compagnie d'assurance selon la ligne d'affaires. Elle permet de déterminer aussi le niveau d'appétit au risque de cette compagnie d'assurance. Par exemple, au Canada, la plupart des compagnies d'assurance doivent s'assurer de gérer au mieux leur risque de tremblement de terre selon une période de retour de 1 sur 500 ans, soit une probabilité de non-réalisation de 99,8 %.

5.3.1.1 Les pertes encourues simulées selon le scénario 1

La Table 5.9 nous présente le niveau de retour calculé selon nos trois modèles. En comparant les modèles 2 (Gamma) et 3 (Lognormale) au modèle 1 (Pareto Généralisée), on constate que le modèle 1 a une tendance plus conservatrice. Par exemple, on a une probabilité de 99 % (période de retour de 1 sur 100 ans) que les pertes ne dépassent pas le montant (EML) de 571 626 pour le modèle 1, 3 697 pour le modèle 2 et 247 474 pour le modèle 3. Aussi, plus on s'éloigne dans la queue de la distribution, l'écart relatif devient de plus en plus grand lorsqu'on compare le modèle 1 aux deux autres modèles, et plus accentué avec le modèle 2. De plus, l'observation de la Figure 5.9, confirme que le modèle 1 (en rouge) est le mieux adapté aux données empiriques (en points noirs), car les deux autres modèles (modèle 2 en bleu et modèle 3 en vert) suivent des trajectoires différentes des données empiriques. On observe également que le modèle 3 surestime le montant (EML) jusqu'à une période de retour autour de 1 sur 30 ans. Cela peut s'expliquer par le fait que pour les périodes de retour en dessous de 30 ans, cette tranche est une tranche travaillante avec la présence de sinistres comme le montre les points noirs dans la Figure 5.9. De toutes ces observations, on conclut que les modèles basés sur les lois standards ont tendance à sous-estimer les niveaux de retour plus nous nous éloignons dans la queue de distribution. Cela ne serait pas avantageux pour une compagnie d'assurance qui souhaiterait céder une partie des sinistres potentiels qui, lorsqu'ils surviendraient, mettraient en péril sa situation financière. Par exemple dans notre situation, si la compagnie d'assurance doit céder une partie de son risque qui correspond à une période de retour de 1 sur 100 ans, elle devra avoir une rétention plus basse pour les modèles 2 et 3 (soit 3 697 et 247 474), et céder aux réassureurs plus de montants qu'il en faut.

Table 5.9: Niveau de retour selon les périodes de retour pour les *pertes encourues* simulées selon le *scénario 1*

Période de retour	Modèle 1(GPD)	Modèle 2(Gamma)	Modèle 3(Lognormale)	Écart Modèle 1 vs 2	Écart Modèle 1 vs 3
1	9 854	3 388	85 218	191%	-88%
5	40 997	3 495	128 447	1 073%	-68%
20	138 921	3 588	176 510	3 771%	-21%
30	198 424	3 616	192 742	5 388%	3%
100	571 626	3 697	247 474	15 364%	131%
250	1 278 578	3 758	296 359	33 921%	331%
500	2 350 582	3 805	337 957	61 679%	596%
750	3 356 312	3 832	364 279	87 485%	821%
1000	4 321 284	3 851	383 891	112 099%	1 026%

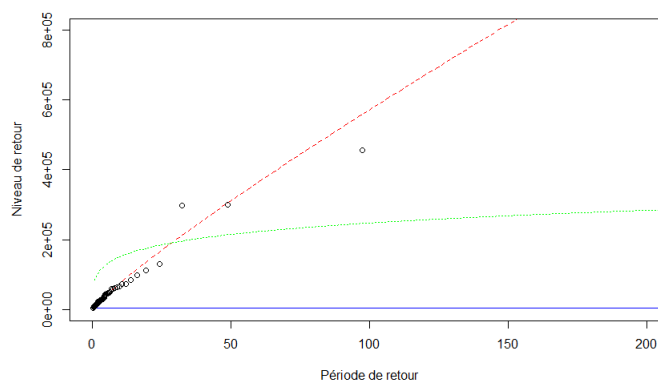


Figure 5.9: Graphe du niveau de retour pour les *pertes encourues* simulées selon le *scénario 1*. Les données empiriques en noir, le modèle 1 en rouge, le modèle 2 en bleu et le modèle 3 en vert.

5.3.1.2 Les pertes encourues simulées selon le scénario 2

Sur la base des *pertes encourues* simulées selon le *scénario 2*, nous avons réalisé les mêmes comparaisons que pour le *scénario 1*. Ainsi, la Table 5.10 nous présente le niveau de retour calculé selon nos trois modèles. On remarque que plus on s'éloigne dans la queue de distribution, plus l'écart relatif s'accroît entre le modèle 1 et les deux autres modèles. De plus, lorsque nous observons la Figure 5.10, malgré quelques données aberrantes, le modèle 1 (en rouge) est celui qui s'adapte le mieux aux données empiriques (en noir). Similaire au *scénario 1*, lorsqu'il y a présence de sinistres dans la base de données empiriques, le modèle 3 a tendance à surestimer plus que les autres modèles, pour des périodes de retour inférieure à 1 sur 30 ans. De toutes ces observations, on s'entend que les lois standards ont toujours tendance à sous-estimer les niveaux de retour, au désavantage des compagnies d'assurance.

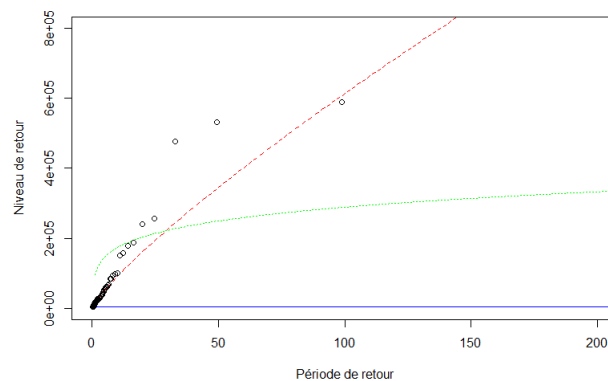


Figure 5.10: Graphe du niveau de retour pour les *pertes encourues* simulées selon le *scénario 2*. Les données empiriques en noir, le modèle 1 en rouge, le modèle 2 en bleu et le modèle 3 en vert.

Table 5.10: Niveau de retour selon les périodes de retour pour les *pertes encourues* simulées selon le *scénario 2*

Période de retour	Modèle 1(GPD)	Modèle 2(Gamma)	Modèle 3(Lognormale)	Écart Modèle 1 vs 2	Écart Modèle 1 vs 3
1	12 830	3 264	95 775	293%	-87%
5	50 865	3 341	146 507	1 422%	-65%
20	161 772	3 408	203 587	4 645%	-21%
30	226 395	3 486	288 785	6 505%	2%
100	613 565	35 530	288 785	17 502%	112%
250	1 308 965	3 564	347 962	36 981%	276%
500	2 321 340	3 564	398 575	65 042%	482%
750	3 245 315	3 583	430 707	90 472%	653%
1000	4 116 192	3 598	454 699	114 332%	805%

5.3.2 Calcul de la prime de réassurance en excédent de sinistre

La formule 1.9 nous permet de calculer une prime pure de réassurance en excédent de sinistre selon les tranches souhaitées par la compagnie d'assurance selon son appétit pour le risque. Pour les *pertes encourues*, simulées selon le *scénario 1*, nous allons former plusieurs tranches à partir des niveaux de retour du modèle 1. Par exemple, on aura une tranche qui va protéger la compagnie d'assurance contre les sinistres entre la période de retour de 1 sur 20 ans et la période de retour de 1 sur 100 ans, soit une rétention de 140 000 avec une limite de 430 000

(430 000 xs 140 000). En observant la Table 5.11, pour le *scénario 1*, on constate que les modèles 2 et 3 ne sont pas adaptés pour tarifier un traité de réassurance en excédent de sinistre, car ils sous-estiment la prime pure de manière considérable. De plus pour une tranche travaillante, comme la tranche 130 000 xs 10 000, le modèle 3 arrive à capturer un peu plus de primes que le modèle 2, mais cela reste toujours très insuffisant pour couvrir un tel contrat de réassurance. Aussi, plus la rétention augmente (dans les tranches non travaillantes), on remarque que les modèles basés sur les lois standards (modèle 2 et 3) ont une prime constante, car la distribution de sévérité a atteint son maximum. Les mêmes constats sont observés avec le *scénario 2*, comme présenté dans la Table 5.12. Ainsi pour les mêmes tranches, les primes du modèle 1 sont plus adaptées pour tarifier un traité en excédent de sinistre. En plus, on constate que l'introduction de la dépendance a pour effet d'augmenter la prime de réassurance, car la queue de la distribution du *scénario 2* est plus allongée que celle du *scénario 1*.

Table 5.11: Prime pure de réassurance en excédent de sinistres pour les *pertes encourues* simulées selon le *scénario 1*

Tranche(1 xs r)	Modèle 1(GPD)	Modèle 2(Gamma)	Modèle 3(Lognormale)	Écart Modèle 1 vs 2	Écart Modèle 1 vs 3
130 000 xs 10 000	236 701	5 500	12 592	4 304%	1 880%
430 000 xs 140 000	782 453	5 500	5 525	14 226%	14 162%
700 000 xs 570 000	1 273 300	5 500	5 501	23 151%	23 151%
1 100 000 xs 1 270 000	2 000 400	5 500	5 500	36 371%	36 371%
1 000 000 xs 2 370 000	1 818 239	5 500	5 500	33 059%	33 059%
950 000 xs 3 370 000	1 727 143	5 500	5 500	31 403%	31 403%

Table 5.12: Prime pure de réassurance en excédent de sinistres pour les *pertes encourues* simulées selon le *scénario 2*

Tranche(1 xs r)	Modèle 1(GPD)	Modèle 2(Gamma)	Modèle 3(Lognormale)	Écart Modèle 1 vs 2	Écart Modèle 1 vs 3
130 000 xs 10 000	290 738	6 750	16 873	4 307%	1 723%
430 000 xs 140 000	961 154	6 750	6 808	14 239%	14 117%
700 000 xs 570 000	1 564 177	6 750	6 751	23 172%	23 172%
1 100 000 xs 1 270 000	2 457 456	6 750	6 750	36 407%	36 407%
1 000 000 xs 2 370 000	2 233 724	6 750	6 750	33 092%	33 092%
950 000 xs 3 370 000	2 121 841	6 750	6 750	31 434%	31 434%

CONCLUSION

Grâce à l'outil « Loss Simulator 2.0 », nous avons pu obtenir des bases de données simulées selon deux scénarios. Nous avons obtenu alors des sinistres à développement long avec les bases de données de la branche en responsabilité civile, puis nous avons accentué le développement avec l'introduction de la dépendance entre la sévérité, le délai de déclaration et le délai de règlement. Aussi, à partir des deux bases de données, trois modèles divisés en deux groupes, ont été analysés. Le modèle 1 qui correspond à la théorie des valeurs extrêmes, les modèles 2 et 3, qui correspondent aux distributions standards. Ensuite, nos trois modèles ont été comparés sur les deux scénarios, par l'observation du niveau de retour et le calcul de prime pure d'un traité en réassurance en excédent de sinistre. Ainsi, tous les résultats observés ont déduit que les outils de la théorie des valeurs extrêmes sont les mieux adaptés pour modéliser les queues de distribution comparées aux modèles basés sur des distributions standards. De même, dans les cas d'un traité de réassurance en excédent de sinistre, le modèle 1 nous permet d'obtenir une tarification adéquate pour les tranches non travaillantes. Aussi, plus la dépendance entre la sévérité, le délai de déclaration et le délai de règlement est forte, plus cela aura tendance à accroître la prime pure d'un traité de réassurance en excédent de sinistres. Enfin, dans un objectif d'optimisation d'une structure d'un traité de réassurance en excédent de sinistre, il est important d'utiliser les outils adaptés pour déterminer un seuil optimal des valeurs extrêmes, car son choix a un impact sur la prime pure de réassurance et aussi sur les quantiles extrêmes.

APPENDICE A

COMPOSANTES DES PRIMES PURES DE RÉASSURANCE EN EXCÉDENT DE SINISTRE

Table A.1: Fréquence annuelle pour le *scénario 1*

Type	Modèles 1, 2 et 3
$E[N]$	2 959

Table A.2: Probabilité de dépasser le seuil u pour le *scénario 1*

Type	Modèles 1, 2 et 3
$P(X > u)$	0,062%

Table A.3: Sévérité pour le *scénario 1*

Type	Tranche(l xs r)	Modèle 1(GPD)	Modèle 2(Gamma)	Modèle 3(Lognormale)
$E[X X > u]$	130 000 xs 10 000	129 110	3 000	6 868
	430 000 xs 140 000	426 793	3 000	3 014
	700 000 xs 570 000	694 527	3 000	3 001
	1 100 000 xs 1 270 000	1 091 127	3 000	3 000
	1 000 000 xs 2 370 000	991 767	3 000	3 000
	950 000 xs 3 370 000	942 078	3 000	3 000

Table A.4: Fréquence annuelle pour le *scénario 2*

Type	Modèles 1, 2 et 3
$E[N]$	2 999

Table A.5: Probabilité de dépasser le seuil u pour le *scénario 2*

Type	Modèles 1, 2 et 3
$P(X > u)$	0,075%

Table A.6: Sévérité pour le *scénario 2*

Type	Tranche(l xs r)	Modèle 1(GPD)	Modèle 2(Gamma)	Modèle 3(Lognormale)
$E[X X > u]$	130 000 xs 10 000	129 217	3 000	7 500
	430 000 xs 140 000	427 180	3 000	3 026
	700 000 xs 570 000	695 190	3 000	3 001
	1 100 000 xs 1 270 000	1 092 203	3 000	3 000
	1 000 000 xs 2 370 000	992 766	3 000	3 000
	950 000 xs 3 370 000	943 040	3 000	3 000

BIBLIOGRAPHIE

- Balkema, A. A. et De Haan, L. (1974). Residual life time at great age. *The Annals of probability*, 792–804.
- Cebrián, A. C., Denuit, M. et Lambert, P. (2003). Generalized pareto fit to the society of actuaries' large claims database. *North American Actuarial Journal*, 7(3), 18–36.
- Charpentier, A. (2014). *Computational actuarial science with R*. CRC press.
- Denuit, M. et Charpentier, A. (2005). *Mathématiques de l'Assurance Non-Vie. Tome II : Tarification et Provisionnement*.
- Fisher, R. A. et Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Dans *Mathematical proceedings of the Cambridge philosophical society*, volume 24, 180–190. Cambridge University Press.
- Gerstengarbe, F. et Werner, P. (1989). A method for the statistical definition of extreme-value regions and their application to meteorological time series. *Zeitschrift fuer Meteorologie ;(German Democratic Republic)*, 39(4).
- Gumbel, E. J. (1958). *Statistics of extremes*. Columbia University press.
- Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 1163–1174.
- IIC. (2014). *C47 :La réassurance*. L'Institut d'assurance du Canada.
- Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348), 158–171.
- Mata, A. J. (2000). Pricing excess of loss reinsurance with reinstatements. *ASTIN Bulletin : The Journal of the IAA*, 30(2), 349–368.
- Mises, R. v. (1936). La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalcanique*, 1, 141–160.

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