Les sinistres graves en assurance automobile : Une nouvelle approche par la théorie des valeurs extrêmes

Noureddine Benlagha (*), Michel Grun-Réhomme (*), Olga Vasechko (**)
(*) Université Paris 2, ERMES-UMR7181-CNRS, 12 place du Panthéon, 75005 Paris, France
E-Mail: blnouri2002@yahoo.fr, E-Mail: grun@u-paris2.fr
(**) Research Institute of Statistics, 3 Shota Rustaveli str., 01023 Kyiv, Ukraine
E-Mail : O.Vasechko@ukrstat.gov.ua

Résumé

La construction des classes de risque en assurance automobile est stratégique pour que le principe de mutualisation soit fonctionnel dans cet environnement concurrentiel. Ces classes, constituées à partir de caractéristiques de l'assuré et du véhicule, sont supposées homogènes en termes de sinistralité.

La présence de sinistres graves (rares) dans une classe vient perturber cette hypothèse d'homogénéité des classes et de stabilité des indicateurs de risque comme la prime pure. En général, face à de tels événements, les assureurs effectuent des écrêtements et répartissent la charge sur l'ensemble du portefeuille, mais la question se pose de déterminer ce qu'est un sinistre grave pour une classe de risque donnée afin d'assurer une certaine stabilité des indicateurs de sinistralité et donc une adéquation entre la prime de référence et la sinistralité.

La théorie des valeurs extrêmes permet d'obtenir des estimations fiables d'événements rares. Trois méthodes classiques (valeurs record, moyenne des excès, approximation par la loi de Pareto généralisée) sont proposées pour la détermination d'un seuil d'écrêtement au-delà duquel un événement est considéré comme atypique. A priori, cette démarche n'a pas été utilisée en assurance automobile.

Ces méthodes sont comparées sur des données du portefeuille d'une mutuelle d'assurance française, puis une nouvelle méthode basée sur une combinaison convexe des précédentes, qui minimise la variance, est proposée, permettant ainsi de décider entre ces différentes stratégies.

Mots clés : Gestion des risques, valeurs extrêmes, écrêtement, distribution de Pareto généralisée, assurance automobile.

JEL: C1, C16, G22

ABSTRACT

In a competitive environment, the development of rate classes in car insurance is strategic, so that the principle of mutualisation is more functional. These classes, based on different characteristics of the policyholder and the vehicle, are supposed homogeneous in term of claim costs.

The presence of large claim costs (rare) in a group comes to disrupt this hypothesis of homogeneity of the classes and the stability of the indicator and thus an adequacy between the reference premium and the claims.

The extreme value theory allows obtaining reliable estimations of rare events. Three classic methods (record values, mean of excesses, and the generalized Pareto estimate) are proposed for the determination of a threshold from which an event is considered atypical.
This step has never been used before in car insurance studies.

These methods are compared on a French mutual insurance portfolio. Then, we propose a new method for the determination of a threshold based on a convex combination of these methods, which minimizes the variance.

Key words: Risk management, Extreme values, data trimming, Generalized Pareto Distribution, car insurance.

1. Introduction

Les Assurances doivent faire face à des flux financiers inter temporels et individualisés, et de ce fait très nombreux. Face à cet environnement incertain, lié à la sinistralité, différents mécanismes de gestion des risques peuvent être proposés. Le principe de mutualisation des risques en assurance automobile est en général retenu, il permet de diminuer les contraintes individuelles sur les assurés, mais peut provoquer un phénomène d'aléa moral. Cette mutualisation s'avère opérationnelle puisque les risques sont dispersés.

La construction des classes de risque est stratégique pour que le principe de mutualisation soit fonctionnel dans un environnement concurrentiel. Ces classes, constituées à partir de caractéristiques de l'assuré et du véhicule, possèdent donc des caractéristiques similaires et sont supposées homogènes en terme de sinistralité. Cette homogénéité s'entend donc relativement à un ensemble de variables explicatives convenablement sélectionnées, en général à l'aide de modèles linéaires généralisés. Cette homogénéité peut se mesurer par des indicateurs de risque, à savoir la fréquence et le coût moyen des sinistres qui permettent de déterminer la prime pure (produit de ces deux indicateurs).

D'une manière générale, les risques des individus d'une classe homogène de risque dépendent de deux variables aléatoires indépendantes et équidistribuées: une variable structurelle qui caractérise l'hétérogénéité interindividuelle acceptée au sein de la classe et une variable endogène qui correspond au risque collectif de la classe. C'est cette dernière que l'assureur cherche à prévoir.

Une certaine stabilité temporelle de ces indicateurs est donc nécessaire pour avoir une bonne adéquation entre la sinistralité et la tarification.

La présence de sinistres graves dans une classe vient perturber cette hypothèse d'homogénéité des classes et de stabilité des indicateurs. En général, face à de tels événements, les assureurs répartissent la charge sur l'ensemble du portefeuille. Toutefois la question se pose de déterminer ce qu'est un sinistre grave dans une classe de risque pour assurer la stabilité des indicateurs et donc la hiérarchie de ces classes.

Des approches par la théorie des valeurs extrêmes ont été appliquées en finance pour le calcul de mesures de risque comme la Value at Risk (Fernandes, 2003) et dans d'autres domaines (Reiss, 1997, 2001).

A notre connaissance, cette démarche n'a pas été utilisée en assurance automobile.

Après avoir rappelé les fondements de la théorie de valeurs extrêmes et en particulier la méthode basée sur la distribution de Pareto généralisée (GPD), trois méthodes d'estimation (valeurs record, moyenne des excès, approximation GPD) sont proposées pour la détermination d'un seuil à partir duquel un événement est considéré comme atypique. Ces méthodes sont comparées sur des données du portefeuille d'une mutuelle d'assurance française. Elles permettent de prévoir des sinistres graves pour une probabilité d'occurrence donnée (très faible) et un intervalle de confiance fixé.

Une analyse comparative est effectuée entre ces méthodes et une nouvelle méthode qui minimise la variance d'une combinaison convexe des seuils obtenus par les méthodes précédentes. Cette démarche permet de décider entre ces différentes stratégies.

2. De la théorie classique des valeurs extrêmes à la méthode de dépassement de seuil

La théorie des valeurs extrêmes est à la fois ancienne et moderne, elle concerne l'étude du maximum d'une distribution et de sa loi. L'œuvre mathématique d'Emile Gumbel (1891-1966), résumée dans ses travaux de 1954 et 1958, est souvent identifiée à la théorie statistique des valeurs extrêmes. Mais il faut aussi citer les travaux de Bortkiewicz (1922), Fréchet (1927), Fisher et Tippett (1928), Finetti (1932).

Les applications de la théorie des valeurs extrêmes, ainsi que les avancées théoriques, sont de plus en plus nombreuses et touchent des domaines variés, comme la météorologie (Jenkinson, 1955), l'hydrologie (Gumbel, 1955, Bernardara *et al.*, 2005, Guillou *et al.*, 2006), la finance (Boulier *et al*., 1998), l'assurance (Reiss, 1997, Embrechts *et al.*, 1997)...
On peut aussi consulter les ouvrages généralistes récents suivants : Embrechts *et al.* (1999), Kotz *et al.* (2000), Coles (2001), Beirlant *et al*. (2004), de Haan et Ferreira (2006). La revue « Extremes » a été créée en 1999 et traite des travaux sur ces questions.
La théorie des valeurs extrêmes, développée pour l'estimation de la probabilité d'occurrence d'événements rares, permet d'obtenir des estimations fiables des valeurs extrêmes, pour lesquelles les observations sont peu nombreuses. L'utilisation des lois des valeurs extrêmes repose sur des propriétés des statistiques d'ordre et sur des méthodes d'extrapolation. Plus précisément, elle repose sur les convergences en loi des maxima ou des minima de variables aléatoires indépendantes convenablement re-normalisées. Les lois limites sont connues et sont appelées les lois des valeurs extrêmes.

2.1 Distribution des extrema dans le cas d'un échantillon de taille finie

On définit la variable aléatoire réelle M_n, qui traduit le maximum d'un *n*-échantillon d'une variable aléatoire réelle *X* (les variables aléatoires X_i sont indépendantes et suivent la même loi que *X*), par :

$$M_n = \max(X_i)_{1\le i\le n} \qquad (1)$$

On pourrait aussi s'intéresser au minimum et utiliser la relation :

$$\min(X_i)_{1\le i\le n} = -\max(-X_i)_{1\le i\le n} \ .$$

En théorie des valeurs extrêmes, le but visé est de déterminer la loi limite normalisée que suit le maximum (ou le minimum) en fonction de celle de la variable aléatoire X . Notons F_X la fonction de répartition de la variable X de loi de probabilité P , à savoir $F_X(x) = P(X < x)$.
La fonction de répartition de M_n est alors définie par:

$$\begin{aligned} F_{M_n}(x) = P(M_n < x) &= P(X_1 < x, \cdots, X_n < x) \\ &= P(X_1 < x)\cdots P(X_n < x) \\ &= \left[F_X(x)\right]^n \end{aligned} \qquad (2)$$

De ces résultats, nous tirons la conclusion que le maximum M_n est une variable aléatoire dont la fonction de répartition est égale à F_X^n .

La fonction de répartition de X étant souvent inconnue, il n'est généralement pas possible de déterminer la distribution du maximum à partir de ce résultat. Comme les valeurs extrêmes se trouvent, à droite et à la fin du support de la distribution, intuitivement le comportement asymptotique de M_n doit permettre de rendre compte de la fin de la distribution.

Notons $x_F = \sup\{x \in R : F_X(x) < 1\}$ le point terminal à droite (right-end point) de la fonction de répartition F_X. Ce point terminal peut être infini ou fini (Embrechts *et al*., 1997, exemple 3.3.22, p.139).

On s'intéresse alors à la distribution asymptotique du maximum en faisant tendre *n* vers l'infini. On a :

$$\lim_{n\to\infty} F_{M_n}(x) = \lim_{n\to\infty}[F_X(x)]^n = \begin{cases} 0 & si\ x < x_F \\ 1 & si\ x \geq x_F \end{cases} \qquad (3)$$

On constate que la distribution asymptotique du maximum donne une loi dégénérée, une masse de Dirac en x_F, puisque pour certaines valeurs de x, la probabilité peut être égale à 1 dans le cas où x_F est fini. Ce fait ne fournit pas assez d'informations, d'où l'idée d'utiliser une transformation afin d'obtenir des résultats plus exploitables pour les lois limites des maxima M_n. La transformation la plus simple est l'opération de standardisation.

2.2 Distributions asymptotiques du maximum

Comme la loi limite de la fonction de répartition précédente est dégénérée, on recherche une loi non dégénérée pour le maximum de $(X_1,...,X_n)$. La théorie des valeurs extrêmes permet de donner une réponse à cette problématique. Les premiers résultats sur la caractérisation du comportement asymptotique de la loi du maximum normalisé ont été obtenus par Fisher et Tippett en 1928. Le théorème suivant explicite ces résultats.

Définition 1

Soit $X_1,.....,X_n$ une suite de n variables aléatoires réelles iid, de somme S_n. On dit que *X* (ou F_X) appartient au domaine d'attraction d'une fonction de répartition non dégénérée *H* s'il existe une suite de réels $(a_n)_{n\geq 1}$ et une suite $(b_n)_{n\geq 1}$ à termes strictement positifs telles que :

$\frac{S_n - a_n}{b_n}$ converge faiblement (ou en distribution) vers *H*.

Dans ce cas, on note : $\frac{S_n - a_n}{b_n} \xrightarrow{d} H$

Théorème 1 (Fisher et Tippett, 1928, Gnedenko,1943)

Soit $X_1,.....,X_n$ une suite de n variables aléatoires réelles iid de loi continue P et $M_n = \max(X_1,...,X_n)$.

S'il existe deux suites réelles $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ avec $b_n > 0$, et une fonction de répartition non-dégénérée G telle que, $\frac{M_n - a_n}{b_n} \xrightarrow{d} G$ lorsque n tend vers l'infini, alors G est nécessairement de l'un des trois types suivant :

$$G_0(x) = \exp(-\exp(-x)) \qquad -\infty < x < +\infty$$
$$G_{1,\alpha}(x) = \begin{cases} 0 & x < 0 \\ \exp(-x^{-\alpha}) & x \geq 0,\ \alpha > 0 \end{cases} \qquad (4)$$
$$G_{2,\alpha}(x) = \begin{cases} \exp(-(-x)^{-\alpha}) & x < 0,\ \alpha < 0 \\ 1 & x \geq 0 \end{cases}$$

On appelle G_0 *loi de* ***Gumbel****,* $G_{1,\alpha}$ *loi de* ***Fréchet*** *et* $G_{2,\alpha}$ *loi de* ***Weibull.***

On dit alors que X (ou F_X *) appartient au domaine d'attraction maximum de G ou au max-domaine d'attraction et on note* $X \in MDA(G)$.
Une démonstration détaillée de ce théorème est donnée dans Resnick (1987) et avec des développements dans Embrechts *et al.* (1997, p. 152).
Les suites $(a_n)_{n\geq 1}$ *et* $(b_n)_{n\geq 1}$ dépendent des paramètres de la loi de *X*.

Par exemple, pour une suite (X_i) de variables aléatoires qui suivent la loi exponentielle standard, la fonction de répartition est donnée par $F(x) = 1 - e^{-x}$. Pour $a_n = \log(n)$ et $b_n = 1$ (Embrechts *et al.*, 1997, p. 155), on obtient par application de ce théorème :
$P(M_n - \log(n) \leq x) = P(M_n \leq x + \log(n))$
$= (P(X \leq x + \log(n)))^n = (F(x + \log(n)))^n = \left(1 - \frac{1}{n}e^{-x}\right)^n$ qui tend vers $\exp(-\exp(-x))$ quand n tend vers l'infini.
On a donc ici un résultat de même type que celui du théorème central limite.

Jenkinson (1955) a proposé une écriture unifiée des trois types de distribution limite du maximum. En introduisant les paramètres de localisation μ et de dispersion σ, qui dépendent des suites (a_n) *et* (b_n), dans l'expression des distributions extrêmes, on obtient la forme la plus générale de la distribution des valeurs extrêmes, notée GEV (Generalized Extreme Value Distribution).

Elle correspond à :

$$G_{\mu,\sigma,\xi}(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \qquad (5)$$

où ξ est un paramètre de forme (shape parameter) encore appelé indice des valeurs extrêmes ou indice de queue qui ne dépend que de la loi de *X*. Plus cet indice est élevé en valeur absolue, plus le poids des extrêmes dans la distribution initiale est important. On parle alors de distributions à « queues épaisses ».

Pour ξ non nul, on a $\xi = \frac{1}{\alpha}$ et le signe de ξ nous renseigne sur le type de la loi asymptotique du maximum. La fonction de répartition $G_{\mu,\sigma,\xi}$ est dans le domaine d'attraction maximum, respectivement, de Fréchet, Gumbel ou Weibull selon que $\xi > 0$, $\xi = 0$ ou $\xi < 0$ (Galambos, 1987, pp.53-54, Embrechts *et al.*, 1997, p.152).

Pour simplifier, on note $G_\xi = G_{0,1,\xi}$.

La connaissance de ξ permet, à elle seule, de caractériser à un changement d'échelle près, le comportement asymptotique du maximum normalisé.

2.3 Distribution conditionnelle des excès

Ce second volet de la théorie des valeurs extrêmes, appelé méthode à dépassement de seuil ou méthode POT (Peaks Over Threshold) consiste à utiliser les observations qui dépassent un certain seuil suffisamment élevé et plus particulièrement les différences entre ces observations et le seuil, appelées excès (cf. Davison et Smith, 1990). Cette méthode a été d'abord développée pour le traitement de données hydrologiques (par exemple, le volume et la durée des déficits en eau) afin d'établir des bases théoriques des modèles de dépassement ou non dépassement d'un certain seuil. Les premiers développements systématiques se trouvent dans les travaux de Todorovic et Zelenhasic (1970) et Todorovic et Rousselle (1971). L'approche basée sur la loi de Pareto généralisée a été introduite par Pickands (1975) et reprise par de Haan et Rootzen (1993). On peut aussi se référer aux travaux sur les fonctions à variations régulières (Bingham *et al.*, 1987). Pour une comparaison plus fine entre ces deux approches on peut se référer aux travaux de Mc Neil et Frey (2002), Katz (2002) qui recommandent l'utilisation de la méthode à dépassement de seuil pour l'estimation des quantiles extrêmes.

En effet, l'approche classique de la théorie des valeurs extrêmes a été fortement critiquée dans le sens où l'estimation des paramètres de la distribution GEV est une estimation dite « *Block component-wise* », ce qui veut dire qu'à partir des données initiales, on extrait des blocs (sous populations) de même taille, puis on considère la distribution formée par les maxima de chacun de ces blocs (une seule valeur maximale par bloc). Cette méthode implique donc une perte d'informations. En particulier, certains blocs peuvent contenir plusieurs valeurs extrêmes pour la distribution initiale, alors que d'autres peuvent ne pas en contenir. Le problème a été résolu, en considérant toutes les valeurs au-delà d'un seuil donné.

Définition 2

Soit $X_1,...,X_n$ n variables aléatoires réelles iid de fonction de répartition commune F et u un réel suffisamment grand, inférieur au point terminal x_F, appelé seuil.
Notons $E_u = \{j \in \{1,2,...n\} / X_j > u\}$.
On définit les excès au-delà du seuil u comme l'ensemble des variables aléatoires Y_j définies par : $Y_j = X_j - u$ pour $j \in E_u$.

On cherche à partir de la distribution F de X à définir une distribution conditionnelle F_u par rapport au seuil u pour les variables aléatoires dépassant ce seuil.

Définition 3

La distribution conditionnelle F_u des excès au-delà du seuil u est définie par :

$$F_u(y) = P(X - u \le y / X > u) = \frac{F(u+y) - F(u)}{1 - F(u)} \text{ pour } 0 \le y \le x_F - u \,. \qquad (6)$$

Ce qui équivaut à :

$$F_u(x) = P(X < x / X > u) = \frac{F(x) - F(u)}{1 - F(u)} \quad \text{pour} \quad x \ge u \,. \qquad (7)$$

L'objectif de la méthode POT est de déterminer par quelle loi de probabilité on peut approcher cette distribution conditionnelle. Balkema et de Haan (1974), Pickands (1975), ont proposé le théorème ci-après qui précise la distribution conditionnelle des excès lorsque le seuil déterministe tend vers le point terminal. Le paragraphe 3.3 illustre l'intérêt de cette méthode dans l'estimation des quantiles extrêmes.

Théorème 2 (Pickands, 1975; Balkema et de Haan, 1974)

Soit F_u la distribution conditionnelle des excès au-delà d'un seuil u (cf. Théorème 1), associée à une fonction de répartition inconnue F .
Cette fonction F appartient au domaine d'attraction maximum de G_ξ (cf. théorème 1) si et seulement s'il existe une fonction positive σ telle que

$$\lim_{u \to x_F} \; \underset{0<y<x_F-u}{Sup} \left| F_u(y) - F^{GPD}_{\xi,\sigma(u)}(y) \right| = 0,$$

où $F^{GPD}_{\xi,\sigma}$ est la fonction de répartition de la loi de Pareto généralisée (GPD), définie par :

$$F^{GPD}_{\xi,\sigma}(y) = \begin{cases} 1 - (1 + \xi\, y/\sigma)^{-1/\xi} & si \;\; \xi \neq 0, \\ 1 - \exp(-y/\sigma) & si \;\; \xi = 0, \end{cases}$$

$$pour \;\; y \in [0, (x_F - u)] \; si \;\; \xi \ge 0 \; et \; y \in [0, Min(-\sigma/\xi, x_F - u)] \quad si \;\; \xi < 0 \,.$$

Ce théorème établit le lien entre le paramètre de la loi du domaine d'attraction maximum et le comportement limite des excès au-delà d'un seuil assez grand. En particulier, l'indice de queue ξ , obtenu dans l'étude du maximum, est identique au paramètre de la loi de Pareto généralisée.
Ceci permet de distinguer les lois à queue épaisse qui appartiennent au domaine d'attraction de Fréchet des lois à queue fine ou légère qui appartiennent au domaine d'attraction de Gumbel. Une queue de distribution est d'autant plus épaisse qu'elle se distingue de la loi normale en s'étalant plus lentement, et donc avec un coefficient d'aplatissement (kurtosis) supérieur à celui de la loi normale (égal à trois).

La loi de Pareto généralisée peut aussi s'écrire sous la forme :

$$F^{GPD}_{\xi,\sigma}(y) = 1 + \log G(y) \qquad (8)$$

où $G(y)$ correspondant à la loi GEV (Generalized Extreme Value) avec $\mu = 0$ (cf. Théorème 1). En effet, pour les excès, l'effet du paramètre de localisation est pris en compte dans les suites (a_n) (cf. Théorème 1).

Dans la littérature, différentes méthodes ont été proposées pour estimer les paramètres (σ, ξ) de la loi GPD : la méthode du maximum de vraisemblance, la méthode des moments, la méthode

bayésienne, l'estimateur de Pickands (1975) et l'estimateur de Hill (1975). Notons que ces deux derniers estimateurs ne sont utilisables que pour l'indice de la queue de la distribution (ξ). Les travaux de Hosking et Wallis (1987) donnent des comparaisons entre les différentes méthodes d'estimation. Lors de l'estimation des paramètres de la loi GPD, on se trouve devant un problème d'estimation du seuil *u*.
Plusieurs méthodes de détermination du seuil au-delà duquel les observations sont extrêmes ont été proposées.

3. Les méthodes d'estimation du seuil

La théorie des valeurs extrêmes, selon l'approche retenue, propose différentes méthodes pour estimer un seuil au-delà duquel une observation sera considérée comme valeur extrême. On peut distinguer les valeurs record, la fonction moyenne des excès et l'approximation GPD (par la distribution de Pareto généralisée). Ce seuil doit être suffisamment grand pour pouvoir utiliser les résultats précédents, mais pas trop afin de disposer d'un nombre suffisant d'observations pour obtenir des estimations de qualité. Présentons rapidement ces différentes méthodes qui seront utilisées dans les classes de risque.

3.1 Les valeurs record

La méthode consiste en une comparaison entre les valeurs record et les valeurs anticipées de ces records issues d'une suite de variables aléatoires réelles *iid*. Soit X_i une suite de variables aléatoires réelles *iid*, un record X_n (comme dans les compétitions sportives, Einmahl et Magnus, 2006) est atteint si $X_n > M_{n-1}$ avec $M_{n-1} = \max(X_1, X_2, ..., X_{n-1})$.
Le processus de comptage (9) est défini comme suit ;

$$N_1 = 1, \quad N_n = 1 + \sum_{k=2}^{n} I_{\{X_k > M_{k-1}\}}, \quad n \geq 2 \tag{9}$$

où la fonction indicatrice $I_{\{X_i > M_{k-1}\}}$ est définie par : $I_{\{X_i > M_{k-1}\}} = \begin{cases} 1 & si \ X_i > M_{k-1} \\ 0 & \text{sinon} \end{cases}$

Avec la méthode des valeurs record, le seuil correspond à une valeur de la distribution.
Il est possible de calculer la moyenne et la variance du processus de dénombrement (9) des records N_n (Embrechts *et al.*, 1997, p. 307) :

$$E(N_n) = \sum_{k=1}^{n} \frac{1}{k}, \qquad Var(N_n) = \sum_{k=1}^{n} \left(\frac{1}{k} - \frac{1}{k^2} \right). \tag{10}$$

La démonstration se trouve dans l'annexe 1.
Le Tableau 1 présente les valeurs de $E(N_n)$ et $V(N_n)$ calculées en fonction de diverses valeurs de *n*, à partir de (10).

n	$E(N_n)$	$V(N_n)$
1000	7,48	5,84
2000	8,17	6,53
5000	9,09	7,45
10000	9,78	8,14
20000	10,48	8,83
50000	11,37	9,75
100000	12,09	10,44
150000	12,49	10,85
200000	12,78	11,13

Tableau 1 : Calcul de l'espérance et de la variance d'un processus de comptage en fonction de la taille de l'échantillon

Pour n grand, l'espérance de N_n varie comme $\ln(n)$, plus précisément $E(N_n) = \sum_{k=1}^{n} \frac{1}{k} \approx \ln(n) + \gamma$, où γ est la constante d'Euler, à savoir $\gamma \cong 0{,}577$. Pour $n = 1000$, l'erreur est inférieure au centième.
L'espérance de N_n permet donc d'estimer le nombre de valeurs extrêmes, puis d'en déduire le seuil au-delà duquel les observations sont extrêmes à partir de la statistique d'ordre.

3.2 La fonction moyenne des excès

La fonction moyenne des excès (FME), qui permet de décrire la prédiction du dépassement du seuil u lorsqu'un excès se produit, est définie par :

$$e_n(u) = E\left[X - u / X > u\right]. \qquad (11)$$

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

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n} (X_i - u)^+}{\sum_{i=1}^{n} I_{\{X_i > u\}}}. \qquad (11a)$$

où $(X_i - u)^+ = Sup(X_i - u, 0)$

Cette approche pratique visant à choisir le seuil *u* consiste à tracer l'estimateur empirique de l'espérance résiduelle de *X* et à choisir *u* de manière à ce que $\hat{e}_n(u)$ soit approximativement linéaire pour tout $x \geq u$. En effet, comme la fonction d'espérance résiduelle d'une loi GPD de paramètre $\xi < 1$ est affine (propriété connue des réassureurs depuis le début du siècle, on peut consulter Schirmacher, 2005), on cherchera un *u* aussi petit que possible sous contrainte que l'estimateur empirique de l'espérance résiduelle des excès au-delà de *u* soit approximativement linéaire.
La fonction moyenne des excès empirique sous la forme affine s'écrit, pour $u < x_F$, comme suit (Davison et Smith, 1990, p. 396, Embrecht *et al.*, 1997, p.165):

$$\hat{e}_n(u) = \frac{\hat{\sigma} + \hat{\xi} u}{1 - \hat{\xi}}, \text{ avec } \hat{\sigma} + \hat{\xi} u > 0. \qquad (11b)$$

Le seuil u est donc déterminé à partir du moment où le graphe de la fonction $\hat{e}_n(u)$ présente une partie affine stable (Figure 1).
Trois cas peuvent alors se présenter :
Si à un certain seuil, la fonction moyenne des excès empirique est marquée par une pente positive, alors les données suivent une distribution de Pareto généralisée avec un paramètre ξ positif.
Un deuxième cas peut se présenter, lorsque la fonction moyenne des excès empirique présente une pente horizontale, les données suivent une distribution exponentielle.
Un dernier cas peut se rencontrer lorsque la FME empirique est décroissante, on a une distribution bornée à droite. Les données suivent alors une distribution à queue légère.

La figure suivante (Figure 1) présente la FME empirique d'une distribution GPD avec un paramètre ξ positif. On constate que la FME devient stable pour un seuil de l'ordre de 5500.

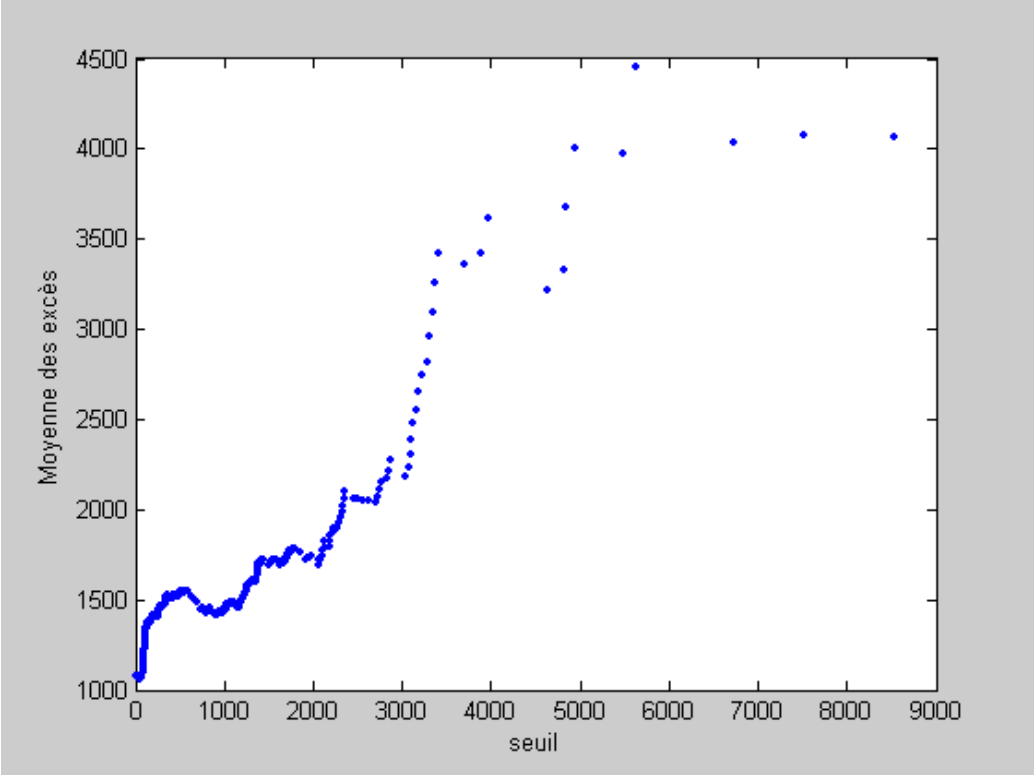


Figure 1 : La fonction moyenne des excès

L'estimation graphique du seuil $\hat{u}$ consiste à repérer les plages de linéarité. Nous avons une première plage de $\hat{u} = 3000$ à $\hat{u} = 4500$, cette plage n'est pas acceptable car par la suite, on observe un changement de pente alors que le nombre d'observations reste important.
Une deuxième plage de linéarité concerne le segment $[5500, 9000]$ dont la pente est légèrement positive. C'est une estimation graphique qui se base sur la visualisation des points qui se présentent sous la forme d'une droite. Cette seconde plage de linéarité, qui donne une stabilité à la fonction moyenne des excès de l'échantillon étudié, permet d'estimer le seuil, $\hat{u} = 5500$.

La valeur de $\hat{u}$ pour chaque classe est déterminée à l'aide du graphique de la fonction moyenne des excès. La valeur de $\hat{u}$ correspond au point à partir duquel le graphe devient stable (proche d'une droite).
Cette méthode est parfois délicate, car le graphe de l'estimation de $E\left[X-u/X>u\right]$ peut présenter des aspects irréguliers et donc difficiles à interpréter.

3.3 La fonction GPD et l'estimation de la queue de distribution

La méthode se base sur l'estimation du quantile extrême présenté dans la formule (14), au-delà de ce quantile les observations sont jugées extrêmes et le nombre de valeurs extrêmes est égal au nombre d'observations qui dépassent ce quantile estimée par l'approximation GPD.
Cette troisième méthode d'estimation du seuil utilise donc la distribution de Pareto généralisée $F_{\xi,\sigma}^{GPD}$ et s'appuie sur la loi asymptotique des excès (Théorème 2) pour produire un estimateur de quantile extrême. Nous utilisons pour cela un seuil aléatoire u_n comme étant le quantile d'ordre $1-\frac{m_n}{n}$ de la loi des données : $u_n = F^{-1}\left(1-m_n/n\right)$ où F^{-1} est l'inverse généralisé de la fonction de répartition *F* défini par : $F^{-1}(y)=\inf\left\{x: F(x)\geq y\right\}$.
Le nombre d'excès m_n est un entier choisi; il doit tendre vers l'infini avec la taille n de l'échantillon mais rester petit devant n pour que le seuil u_n soit suffisamment grand : $\lim_{n\to\infty} m_n = +\infty$ et $\lim_{n\to\infty}\frac{m_n}{n}=0$.

Des solutions pour fixer m_n de manière à obtenir un estimateur asymptotiquement sans biais ont notamment été proposées par Goldie et Smith (1978) puis De Haan et Peng (1998). Dans notre application (section 5), nous avons utilisé la même démarche que celle proposée par Embrechts *et al.* (1997) qui préconisent de tracer les estimateurs en fonction de m_n/n et de choisir m_n dans un intervalle où le graphe des estimations est approximativement linéaire.

Pour plus de détails sur cette question de l'estimation d'un seuil, on peut aussi consulter le livre récent de De Haan et Ferreira, 2006, chapitre 4, pp. 128-154.

Le seuil doit être suffisamment élevé pour vérifier le caractère asymptotique du modèle, mais pas trop grand pour garder un nombre suffisant d'observations qui dépassent ce seuil afin de pouvoir estimer les paramètres du modèle. En pratique, on estime le seuil par la $(n-m_n)^{\text{ième}}$ observation ordonnée ; c'est-à-dire $\hat{u}_n = x_{(n-m_n)}$, qui correspond à la statistique d'ordre des observations situées au niveau de la queue de la distribution. Puis on retient celui pour lequel les estimations apparaissent stables pour un modèle bien ajusté aux données.

L'adéquation des excès au-delà du seuil par la fonction GPD

Etant donné un échantillon $\left(x_1,.....,x_n\right)$, nous voulons vérifier si un modèle paramétrique permet d'obtenir une bonne approximation de la loi des données, particulièrement en queue de distribution. Les tests d'adéquation classiques de Cramer-Von Mises ou d'Anderson-Darling, permettent de mesurer l'adéquation aux données de la partie centrale de l'intervalle mais ils ne sont pas applicables au niveau de la queue de distribution, même si le test d'Anderson-Darling est plus sensible aux valeurs extrêmes.

Reiss et Thomas (1997) ont développé un test d'adéquation de queue de distribution aux données. Le modèle doit s'adapter aussi bien à l'ensemble de la distribution des données qu'à la queue de la distribution.

En effet, pour $x \geq u$ (des points appartenant à la queue de la distribution),

$$F(x) = P\{X \leq x\} = (1 - P\{X \leq u\})F_u(x-u) + P\{X \leq u\}. \quad (12)$$

Dans la relation précédente, on peut estimer $F_u(x-u)$ par $F_{\hat{\xi},\hat{\sigma}}^{GPD}(x-u)$ pour u élevé.

Cette approximation est l'application directe du Théorème 2. Nous pouvons aussi estimer $P\{X \leq u\}$ à partir des données par la fonction de répartition empirique $F_n(u)$ évaluée au point u :

$F_n(u) = 1 - \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i > u\}}$.

Cela veut dire que pour $x \geq u$ nous pouvons utiliser l'estimation de la queue,

$\hat{F}(x) = (1 - F_n(u))F_{\hat{\xi},\hat{\sigma}}^{GPD}(x) + F_n(u)$, pour approximer la fonction de répartition $F(x)$.

Il est facile de montrer que $\hat{F}(x)$ est aussi une distribution GPD, avec le même paramètre de forme ξ, mais avec un paramètre d'échelle $\tilde{\sigma} = \sigma(1 - F_n(u))^{\xi}$ et un paramètre de localisation $\tilde{\mu} = u - \tilde{\sigma}\left((1 - F_n(u))^{-\xi} - 1\right)/\xi$.

L'estimateur du quantile extrême $q_{GPD,n}$, issu de l'approximation par la loi de Pareto généralisée et qui sera utilisé par la suite pour la détermination des seuils par classe de risque se présente comme suit :

$$\hat{q}_{GPD,n} = \hat{u}_n + \frac{\hat{\sigma}_n}{\hat{\xi}_n}\left[\left(\frac{np_n}{m_n}\right)^{-\hat{\xi}_n} - 1\right]. \quad (13)$$

Cette expression figure, par exemple, dans Davison et Smith (1990, p. 397) et Embrechts *et al.* (1997, p. 354).

Pour trouver un estimateur de $q_{GPD,n}$ il faut donc trouver les estimateurs de u_n, σ_n, ξ_n. Dans notre application les paramètres de la loi GPD sont estimés par la méthode de maximum de vraisemblance.

3.4 Méthode mixte de minimisation de la variance

Les différentes méthodes précédentes proposent un seuil, souvent différent d'une méthode à l'autre. Le problème des valeurs atypiques ou aberrantes est bien connu en statistique pratique, et il garde toujours une part d'indétermination qui empêche une formalisation complètement rationnelle. Bien sûr, il est possible de consulter des experts ou d'utiliser de l'information auxiliaire (si disponible), mais cette démarche prend beaucoup de temps. Le travail pratique du statisticien, comme du gestionnaire, est de trouver un compromis entre la qualité (précision des estimations) et le coût pour y parvenir.

Dans cette voie, nous proposons cette nouvelle méthode pour déterminer un seuil minimisant la variance (critère de qualité) d'une combinaison convexe des seuils obtenus par ces méthodes issues de la théorie des valeurs extrêmes.

Théorème 3

Soient U_i des variables aléatoires réelles ($i = 1,..., p$) et Z une variable aléatoire définie comme combinaison convexe des U_i, à savoir : $Z = \sum_{i=1}^{p} \alpha_i U_i$, où $\alpha_i > 0$ et $\sum_{i=1}^{p} \alpha_i = 1$.

On note V_i la variance de U_i, V est la matrice de variance-covariance des U_i et $\alpha = (\alpha_1,..., \alpha_p)$.

On suppose que les variables U_i ont toutes des moments d'ordre deux et que V est inversible.

La variance de Z est alors minimale pour $\alpha = \frac{V^{-1}}{\sum_{i,j=1}^{p} v_{ij}^{-1}} A$, où V^{-1} est l'inverse de la matrice V, v_{ij}^{-1} est le terme général de V^{-1} *et A est la matrice uni-colonne d'ordre p dont tous les coefficients sont égaux à 1.*

La preuve se trouve dans l'Annexe 2.

Dans cette démarche, qui permet d'obtenir un seuil de référence, les variables aléatoires U_i correspondent aux seuils estimés par les méthodes précédentes. Mais dans cette approche, les variances de ces seuils ne sont pas connues explicitement, il est donc nécessaire de les estimer.

La méthode du bootstrap introduite à la fin des années 70 par Bradley Efron (Efron, 1979) permet d'estimer la précision des estimateurs de paramètres, comme la variance, sans connaître la vraie valeur de cette précision et même une expression de cette précision. Cette méthode procède par ré-échantillonnage (Efron et Tibshirani, 1993, Ardilly, 1994).

Puisque les seuils correspondent à des quantiles, la variance a donc été estimée par la méthode du bootstrap qui consiste à tirer, avec remise, *K* échantillons de même taille que l'échantillon considéré (classe de risque). On obtient ainsi une suite de quantiles.
Les covariances sont également estimées par la méthode du bootstrap (Beran et Srivastava, 1985) : pour chaque échantillon, on détermine un seuil par chacune des méthodes, obtenant ainsi des p-uples (ou des paires dans la méthode mixte du paragraphe 5.3). On a choisi *K*=1000. Les temps de mise en œuvre et de calcul sont assez longs. Cette démarche est appliquée pour chaque variable aléatoire U_i.
Plus précisément si $\hat{u}_k$ désigne l'estimation de u_k (seuil) dans l'échantillon *k* ($k = 1,..., K$), on a :

$$\bar{u} = \frac{1}{K} \sum_{k=1}^{K} \hat{u}_k \qquad (14)$$

et la variance de $\hat{u}$ est estimée par :

$$\hat{V}(\hat{u}) = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{u}_k - \hat{u})^2 \qquad (15)$$

Si *v* est un estimateur du seuil *u* obtenu par une autre méthode, on définit de façon analogue aux expressions (14) et (15), $\hat{v}$ et l'estimateur de la variance de $\hat{v}$, l'estimation de la covariance est alors donnée par :

$$\hat{Cov}(\hat{u},\hat{v}) = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{u}_k - \hat{u})(\hat{v}_k - v_k) \qquad (16)$$

où le couple $(\hat{u}_k, \hat{v}_k)$ est. obtenu lors du $k^{ème}$ tirage bootstrap.
Le calcul des $\hat{V}(\hat{u})$ permet d'estimer α, puis d'obtenir une estimation du seuil pour cette méthode mixte de minimisation de la variance.

On peut aussi utiliser une méthode de « balanced resampling » (Davison *et al.*, 1986) dans laquelle chaque observation de l'échantillon initial apparaît le même nombre de fois dans la réunion des *K* échantillons bootstrap créés.

Avant de passer à l'application de ces méthodes issues de la théorie des valeurs extrêmes à la détection des sinistres graves par classe de risque, il est nécessaire de présenter les indicateurs de risque utilisés pour cette étude.

4. Les outils de mesure de la sinistralité

Le risque est souvent considéré comme mesurable s'il est possible de calculer un risque moyen qui caractérise la tendance de sinistralité du phénomène étudié. Dans chaque classe, le risque est mesuré en termes de fréquence et de coût moyen, puis la prime pure est déterminée comme produit de ces deux indicateurs. La prime pure correspond à l'espérance de la charge de sinistres à laquelle devra faire face l'assureur. La loi des grands nombres permet d'utiliser cette démarche. Le calcul de la prime pure a pour objectif d'évaluer pour chaque assuré (selon ses caractéristiques) le montant attendu des sinistres pour une période d'assurance donnée. Cette prime pure théorique est commune à tous les assurés d'une même classe.
Plus précisément, notons *k* une classe de risque (*k=1,...K*) et n_k le nombre de véhicules dans la classe *k*. La prime pure dans la classe *k* est alors définie, sur une période donnée (en général 1 an), par :

$$P_k = \frac{\sum_{i=1}^{n_k} c_{k,i}}{\sum_{i=1}^{n_k} w_{k,i}} \qquad (17)$$

où $c_{k,i}$ correspond au coût total des sinistres sur la période pour le véhicule *i* de la classe *k*. De nombreux coûts sont nuls. Dans cette expression de la prime pure, $w_{k,i}$ correspond au poids du véhicule assuré *i* de la classe *k*. En effet, au cours d'une année, le nombre d'assurés dans une classe varie, certains arrivent, d'autres résilient leur contrat ou changent de véhicule. Chaque observation *i* est donc pondérée par $w_{k,i} = \frac{1}{12} \times$ (nombre de mois où l'assuré *i* est présent dans la classe *k* pour une année donnée). Un changement de véhicule peut impliquer un changement de classe. Un assuré qui contracte une assurance seulement sur 3 mois dans l'année, aura donc un poids égal à 0.25 (sa cotisation ne correspond qu'à trois mois d'assurance). Par conséquent, $\sum_{i=1}^{n_k} w_{k,i} \leq n_k$.

Ces indicateurs sont en général normés (en divisant chaque indicateur par la prime pure de l'ensemble du portefeuille) et multipliés par 100. Ainsi la prime pure du portefeuille est égale à 100 et les primes pures des classes sont ainsi facilement interprétables par rapport à la moyenne du portefeuille.

A travers ces indicateurs, on recherche une tendance qui dépend des risques pour lesquels le contrat est souscrit, de la régularité de l'environnement et du portefeuille, de la structure de l'entreprise, ainsi que de sa gestion et de sa stratégie commerciale.

Cet indicateur de prime pure permet d'une part de hiérarchiser les classes. En pratique, on constate un décalage entre le moment où les indicateurs de risque sont calculés dans les classes et le moment où l'étude est prise en compte au niveau de la tarification. Au meilleur des cas, il faut compter un délai de trois ou quatre ans.
D'une manière générale les risques des individus d'une classe homogène de risques dépendent de deux variables aléatoires indépendantes et équidistribuées : une variable structurelle qui caractérise l'hétérogénéité interindividuelle acceptée au sein de la classe et une variable endogène qui correspond au risque collectif de la classe. C'est cette dernière que l'assureur cherche à prévoir. La présence de sinistres graves vient perturber cette hypothèse de différenciation du risque collectif d'une classe à l'autre et la stabilité temporelle des indicateurs (Grun-Rehomme *et al*., 2007).

La pratique actuelle des assureurs mutualistes consiste souvent à écrêter la distribution des coûts des sinistres uniformément selon les classes de risque (ou avec quelques aménagements) et à répartir cette charge supplémentaire sur l'ensemble du portefeuille. Cette démarche a l'avantage d'être rapide, simple et de garder un équilibre entre la sinistralité et les cotisations (ou primes payées) dans le ratio : sinistres/cotisations. Mais elle présente l'inconvénient d'une part de ne pas prendre en compte la particularité des classes (plus ou moins risquées) et d'autre part cette troncature uniforme trop grossière peut venir perturber la hiérarchie des classes de risque et donc l'adéquation entre la prime de référence et la sinistralité.

D'autre part, cet indicateur de prime pure sert de base au calcul de la prime de référence. L'assureur élimine au préalable les coûts des sinistres extrêmes. Il effectue une opération d'écrêtement. Les sinistres extrêmes sont alors affectés à un compte spécial, puis répartis entre les différentes classes selon, par exemple, le nombre de véhicules ou le nombre de sinistres ou sur les primes (Moreau, 1984). La prime payée par l'assuré est égale à la prime de référence multipliée par le coefficient réduction majoration (bonus-malus) de l'assuré (Grun-Réhomme, 2000). Les primes payées par les sociétaires d'une même classe de risque vont donc varier, mais il importe que le montant total encaissé par l'assureur reste constant Le système bonus-malus, qui est légal obligatoire et forfaitaire, a pour but de différencier les « bons » conducteurs des « mauvais » conducteurs, sur la base de la sinistralité observée. L'hétérogénéité au sein d'une classe est prise en compte par ce système bonus-malus. Le nombre de sinistres déclarés dans le passé est souvent une variable très pertinente pour prédire le nombre futur de sinistres. C'est une façon pour l'assureur de récupérer de l'information non accessible a priori. Le système bonus-malus est aussi un mécanisme permettant de limiter le risque moral (la tendance des assurés à prendre moins de précautions lorsqu'ils sont couverts par une police d'assurance). Le livre de Lemaire (1985) est une référence incontournable pour le système bonus-malus.

Par définition, le montant d'un sinistre inclut l'indemnisation directe des victimes, les frais de gestion internes à la mutuelle ou la compagnie d'assurance ainsi que les frais externes (expertise, frais judiciaires) afférents à ce sinistre. Les frais d'acquisition du contrat ne sont pas inclus dans ce montant. Le montant de sinistre n'est autre qu'une variable aléatoire positive ou nulle. Il est à noter

que le montant du sinistre utilisé dans cet article est différent de la charge de sinistre qui est aussi une variable aléatoire du montant de sinistre mais diminuée de la franchise.

Une remarque à propos des coûts des sinistres : tous les sinistres ne sont pas réglés financièrement l'année où ils surviennent. On peut estimer qu'environ un tiers d'entre eux sont clos la même année, un petit tiers l'année suivante et de 10 à 15% la troisième année. Les règlements sont en général plus longs pour les accidents corporels. Pour répondre aux besoins de provisionnements, imposés par la réglementation, la compagnie d'assurance doit effectuer des estimations des coûts des sinistres non réglés l'année en cours. L'assureur, en s'appuyant sur son expérience passée, établit des provisions pour frais au cas par cas. Dans le calcul de la prime pure de l'application numérique proposée par la suite, il s'agit donc des coûts réels ou de coûts estimés, en l'absence des coûts réels définitifs.

Pour les accidents corporels, l'assureur peut connaître le montant total de ses dépenses seulement après plusieurs mois, voire plusieurs années, d'où cette nécessité de faire une prévision de ses dépenses. Le coût des sinistres corporels comprend plusieurs composantes : indemnités pour les personnes physiques, soins, tierce personne, préjudices personnels et économiques.

5. Application numérique

La base de données concerne un échantillon aléatoire de 50000 observations pour des véhicules 4 roues de tourisme durant l'année 2004, issu du portefeuille d'une mutuelle d'assurance française. Les données sont individuelles et anonymes. Un enregistrement correspond à un véhicule assuré (si un assuré possède deux véhicules, on aura deux enregistrements dans le fichier).

Ce fichier contient un groupe de variables caractéristiques des assurés (sexe, âge, ancienneté de permis, type de conducteur, zone d'habitation,…) et des véhicules (puissance, ancienneté, marque,…), ainsi que les facteurs relatifs à la sinistralité (fréquence, coût, coût estimé, date de survenance du sinistre).

Les applications numériques ont été réalisées avec les logiciels Matlab et SAS.

5.1 Analyse exploratoire des données

Le fichier contient 53% d'hommes et 47% de femmes. Dans 30% des cas, l'assuré n'est pas le conducteur principal (variable : type de conducteur).

Quelques statistiques sur les variables numériques :

Variables	Min	Q_1 (25%)	Médiane	Moyenne	Q_3 (75%)	Max
Coût des sinistres non nuls (€)	37	128	468	967	1168	45972
Age du conducteur	18	35	48	47	57	98
Ancienneté du véhicule*	0	3	7	7.5	11	63
Puissance du véhicule**	24	60	76	83	100	485

Tableau 2 : Caractéristiques de quelques variables

* L'ancienneté du véhicule est exprimée en années.

** La puissance (réelle) du véhicule exprime la puissance du moteur en chevaux Din (Deutsch Industrie Normen). Cette mesure donne une vision plus réaliste de la puissance effective au niveau des roues (1 ch. Din = 0,735 Watt), ce qui n'est pas le cas de la puissance fiscale.

Les coûts augmentent en moyenne de façon logarithmique avec la puissance du véhicule et ils diminuent de façon parabolique (concave) avec l'ancienneté du véhicule. Les véhicules puissants et les véhicules de marque étrangère sont plus fréquents chez les hommes que chez les femmes.

Une droite de Henry permet d'étudier la forme asymétrique de la distribution des coûts des sinistres.

QQ-plot des données

Un graphique QQ-plot est un outil convenable pour examiner si la distribution d'une variable dans un échantillon provient d'une distribution théorique spécifique. Il donne les quantiles de la distribution empirique en fonction des quantiles de la distribution théorique envisagée. Si l'échantillon provient bien de cette distribution théorique, alors la forme du graphique QQ-plot sera linéaire.

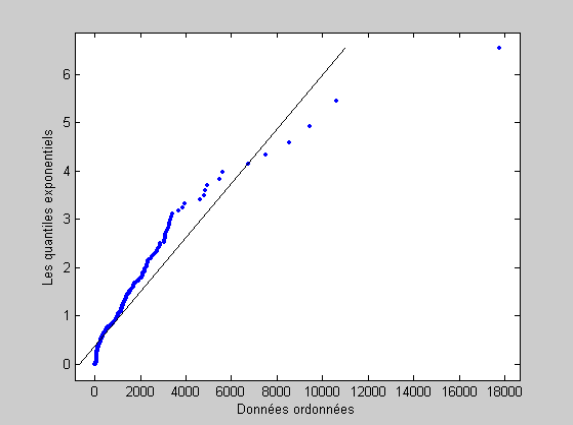
Dans la théorie des valeurs extrêmes, le QQ-plot est basé sur la distribution exponentielle.
Ce graphique est l'ensemble des points tel que :

$$\left\{\left(X_{k,n} , G_{0,1}^{-1}\left(\frac{k}{n+1}\right)\right)\right\} \qquad (18)$$

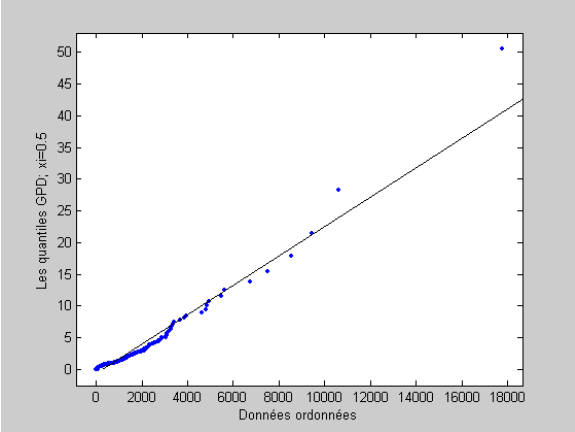
où $X_{k,n}$ représente la statistique d'ordre k dans l'ordre croissant de la distribution X, à savoir $X_{1,n} \leq ... \leq X_{n,n}$ et $G_{0,1}^{-1}$ est la fonction inverse de la distribution exponentielle.

Le graphique QQ-plot permet d'obtenir la forme de la queue de la distribution.
Trois cas de figure sont possibles :
- Les données suivent la loi exponentielle : la distribution présente une queue très légère, les points du graphique présentent une forme linéaire.
- Les données suivent une distribution à queue épaisse « *fat-tailed distribution* »: le graphique QQ-plot est concave, cela revient à la présence d'un grand nombre de valeurs extrêmes au niveau de la queue de la distribution.
- Les données suivent une distribution à queue légère « *short-tailed distribution* » : le graphique QQ-plot a une forme convexe. Le nombre de valeurs extrêmes est faible.

Les deux graphiques suivants (Figure 2) représentent le QQ-plot du montant des sinistres pour l'ensemble des observations, et pour deux valeurs de l'indice de queue ξ.



QQ-plot du coût des sinistres ($\xi = 0$)



QQ-plot du coût des sinistres ($\xi = 0{,}5$)

Figure 2 : QQ-plot du montant des sinistres

Le QQ-plot sous l'hypothèse d'une distribution exponentielle se traduit graphiquement par une représentation linéaire des données et la valeur de ξ est nulle, ce qui n'est pas le cas en visualisant le graphe de gauche de la Figure 2. Les données ne suivent donc pas la loi exponentielle, le graphique présente une incurvation du QQ-plot vers le bas, la queue de la distribution est donc épaisse. On peut penser que la loi des données est de type Pareto généralisée.
Le QQ-plot pour une valeur de $\xi = 0{,}5$ est linéaire, on peut en déduire que l'adéquation des données à la loi de Pareto généralisée semble convenir.

Les classes de risque

Dans la constitution des classes de risque, où l'information doit être disponible et fiable, un équilibre doit être trouvé entre la granularité et la robustesse (Grun-Rehomme *et al*., 2007). Si la granularité (ou la segmentation) est trop grossière, certes la robustesse temporelle des indicateurs de sinistralité est assurée, mais la mutualisation est trop large et un concurrent peut très bien attirer les bons risques de cette classe en proposant une cotisation plus faible grâce à une segmentation plus fine. A l'inverse une granularité trop fine ne permet pas d'avoir cette robustesse. Au sein d'une mutualisation des risques, il existe une volatilité résiduelle.

Trois facteurs principaux permettent d'expliquer une variance importante du montant des sinistres dans une classe de risque:
- La présence d'une ou de quelques valeurs extrêmes,
- La présence d'une petite sous population plus risquée, d'une niche dans ce segment qu'il est nécessaire de suivre avec attention, pour éventuellement envisager une segmentation plus fine,
- Un manque d'homogénéité structurelle de la classe qui peut provenir de variables non retenues ou non observables.

Dans notre application, les classes de risque sont construites à partir de l'information disponible : des caractéristiques du conducteur (ancienneté de permis, type de conducteur), des caractéristiques du véhicule (ancienneté, puissance) et du lieu d'habitation. Le type de conducteur correspond au

fait que l'assuré est, ou n'est pas, le conducteur principal du véhicule assuré; cette distinction est pertinente pour les jeunes conducteurs (mais cette dimension est déjà prise en compte par l'ancienneté de permis et le bonus-malus) et dans une moindre mesure pour les conjoints. Nous disposons aussi de la période de couverture : période (exprimée en mois) au cours de laquelle l'assuré est couvert par la police qu'il a souscrit, le plus souvent cette période est d'une année.

Dans cet exemple, les classes (ou strates) sont construites en prenant des limites de ces variables explicatives plus larges que celles utilisées réellement pour l'ensemble du portefeuille. D'ailleurs cette mutuelle d'assurance ne retient pas beaucoup de variables pour construire les classes de risque (ou cases tarifaires). Par exemple, la variable sexe n'est pas prise en compte. Pour des raisons de confidentialité, toutes les variables de construction des classes ne sont pas utilisées et la description précise des classes n'est pas donnée. Il est impossible de donner en même temps le descriptif des classes de risque (cases tarifaires) et la valeur réelle de la prime pure. La connaissance des classes et des primes pures associées permettrait à un concurrent de connaître les ratios sinistres/cotisations de cet assureur, puisqu'il est toujours possible de se renseigner sur le montant de la cotisation.

Cette application numérique n'a valeur que d'exemple puisque l'on travaille sur un échantillon et non sur l'ensemble du portefeuille, mais la démarche méthodologique et informative reste la même.

Le tableau suivant présente la hiérarchie des classes de risque basée sur la prime pure, ainsi que quelques caractéristiques statistiques de la distribution du montant des sinistres (quantile à 90%, écart-type). On constate que globalement le neuvième décile et l'écart-type de la distribution des coûts des sinistres augmentent avec la prime pure.

Classe	Nombre d'observations	Prime pure	Prime pure normée	Quantile 90%	Ecart type
1	3664	55,04	26,06	0 (*)	305
2	6258	90,41	42,81	101	474
3	3098	124,82	59,10	252	605
4	3258	141,59	67,04	341	521
5	3415	165,10	78,17	441	588
6	3893	190,02	89,97	532	603
7	5518	223,36	105,76	656	708
8	6225	226,30	107,15	792	859
9	4544	287,97	136,35	818	1011
10	3253	300,47	142,27	920	907
11	4459	329,39	155,96	947	1071
12	2405	411,99	195,07	922	1551

(*) Ce quantile à 90%, égal à 0 dans la classe 1, vient du fait que plus de 90% des assurés de cette classe n'ont pas déclaré d'accidents responsables.

Tableau 3 : Prime pure selon les classes de risque

5.2 Détection des valeurs extrêmes dans les classes de risque selon les méthodes retenues

Une forte variabilité des coûts dans une classe de risque peut provenir d'une mauvaise construction de la classe (manque structurel d'homogénéité), d'une petite sous population différente ou de la présence de valeurs extrêmes (sinistres graves pour la classe).

Parmi ces coûts très élevés quelles sont les valeurs que nous pouvons qualifier d'extrêmes ?
Pour répondre à cette problématique nous avons recours aux trois méthodes présentées précédemment pour déterminer dans chaque classe le seuil au-delà duquel une valeur sera considérée comme extrême. On peut penser à d'autres méthodes plus classiques d'écrêtement des valeurs extrêmes, comme par exemple, faire en sorte que les coefficients de variation du montant des sinistres soient un peu près égaux dans chacune des classes de risque.

Le choix du seuil

Dans le choix du seuil on se trouve face à l'un des deux problèmes: la présence d'un biais ou d'une variance élevée.
En prenant un seuil faible, le nombre des observations (excès) augmente et l'estimation devient plus précise. Mais le choix d'un tel seuil faible risque de déclarer abusivement des observations comme extrêmes et introduire un biais dans l'estimation de la prime pure en la sous-évaluant.

Le choix du seuil par la méthode d'approximation GPD présente l'avantage de fournir une prévision d'un sinistre extrême pour une probabilité d'occurrence donnée (très faible). Cette estimation peut être ponctuelle ou par intervalles. Dans cette étude nous essayons d'estimer un quantile extrême avec une probabilité de 99,9% d'être une valeur extrême (un sinistre grave) pour la distribution du coût des sinistres avec un niveau de confiance de 95%. Ce quantile a été choisi de façon empirique. Nous avons alors effectué une estimation ponctuelle du quantile extrême puis une estimation par intervalle de confiance. Les paramètres sont estimés par la méthode de maximum

de vraisemblance. Dans le tableau 4, la valeur 17 785 correspond à la perte maximale pour la compagnie d'assurance en cas de survenance d'un sinistre grave avec une probabilité de 0,1%.

Pour l'ensemble des données, on obtient comme seuil prévisionnel :

Limite inférieure	11421
Quantile extrême estimé	17785
Limite supérieure	26631

Tableau 4 : Prévision des quantiles extrêmes

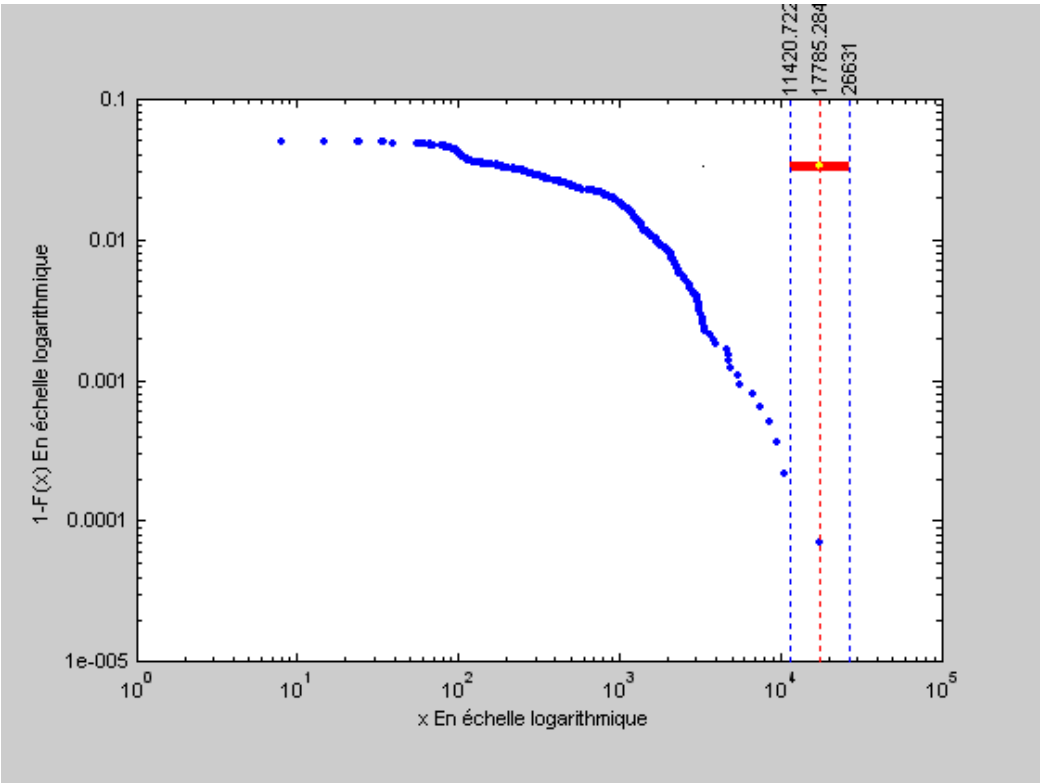


Figure 3 : La prévision d'un coût extrême

La Figure 3 représente la fonction de la queue de distribution, où l'on peut voir le quantile extrême estimé ainsi que les bornes de l'intervalle de confiance associé.

Pour chacune des classes, on calcule le seuil et le nombre de valeurs extrêmes correspondantes avec les trois méthodes. Le tableau ci dessous présente ces résultats, où *N* correspond au nombre de valeurs extrêmes retenu.

Méthode	Record		Moyenne des excès (FME)		GPD	
Classe	**Seuil**	**N**	**Seuil**	**N**	**Seuil**	**N**
1	3140	9	3200	8	4345	3
2	4107	9	4150	7	5249	6
3	3360	9	3900	6	6078	2
4	3895	9	4200	8	5684	2
5	4705	9	5230	5	6204	4
6	4689	9	4650	9	5537	4
7	5904	9	7000	7	7149	7
8	9370	9	9600	8	9700	6
9	10591	9	8500	13	11820	9
10	7010	9	7400	7	9252	2
11	9247	9	9800	6	11961	4
12	13193	8	10100	15	18175	5

Tableau 5 : Comparaison des seuils selon les classes de risque

Comme le montre la Figure 4, la méthode GPD détecte toujours moins de valeurs extrêmes que la FME et les valeurs record, le seuil estimé par la méthode GPD est toujours plus élevé.

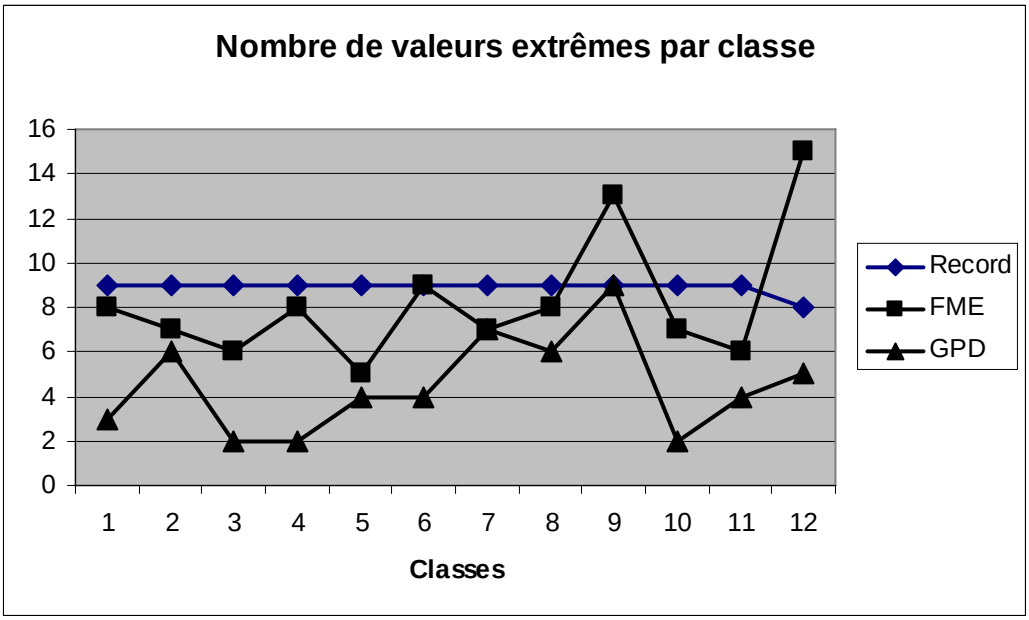


Figure 4 : Nombre de valeurs extrêmes selon les classes et les méthodes

En comparant le nombre de valeurs extrêmes estimé par chacune des trois méthodes précédentes avec le graphique boxplot de la Figure 5, on constate les faits suivants :

- La méthode record propose une méthode maximale (nombre élevé de valeurs extrêmes) dans les classes de risque où la sinistralité est faible ou moyenne par rapport à l'ensemble du portefeuille, c'est à dire dans les classes où le seuil est relativement bas pour l'ensemble du portefeuille (classes de 1 à 7).
- La méthode FME prends moins en compte les ruptures dans la queue de la distribution.

- La méthode GPD offre une méthode de détection des valeurs extrêmes minimale (avec peu de valeurs extrêmes).
- Les trois méthodes donnent des résultats proches sauf dans les cas où la queue de la distribution est constituée de petits groupes de points isolés.
- Pas de corrélation significative entre les trois méthodes (on note simplement un coefficient de corrélation linéaire r = 0,44 entre les méthodes FME et GPD).

La méthode des valeurs record présente deux inconvénients pour notre problématique :

- Le seuil correspond à une valeur de l'échantillon (et non à une valeur estimée comme dans les deux autres méthodes), ainsi la distance comprise entre la première observation atypique et la dernière observation non retenue comme atypique n'intervient pas. Les valeurs seront différentes au cours d'un autre exercice.
- Le seuil retenu ne prend pas en compte la forme géométrique de la queue de la distribution. Il est trop lié à la taille de la classe.

Par conséquent, cette dernière méthode ne sera pas retenue dans la combinaison convexe de la nouvelle méthode.

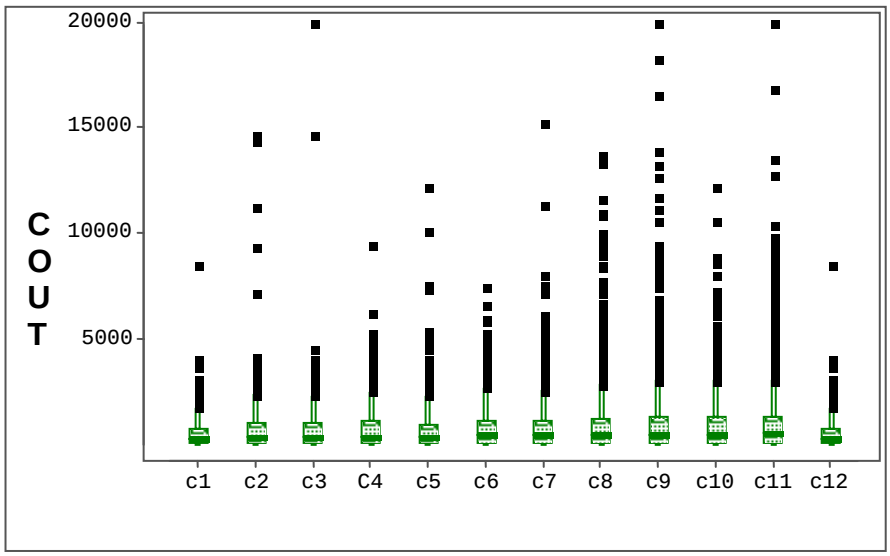


Figure 5 : Graphique boxplot de la distribution du coût des sinistres selon les classes

Rappelons que dans le graphique boxplot, les « moustaches» sont généralement déterminées sur 1,5 fois l'intervalle interquartile. Ce choix du coefficient 1,5 donnerait 99,3% des observations à l'intérieur de la boîte et des moustaches si la distribution était normale. Or la distribution des coûts, comme d'ailleurs de nombreuses distributions économiques (salaires, chiffre d'affaires des entreprises,…), est fortement asymétrique et donc considérer comme extrêmes toutes valeurs à l'extérieur des moustaches n'est vraiment pas raisonnable. La probabilité serait de 0,999 pour un coefficient de 2 (pour une loi normale!).

La courbe représentative de la distribution des coûts étant asymétrique avec une queue plus étalée à droite, on a regardé si son logarithme suit une loi normale. C'est une démarche classique qui permet alors d'utiliser les propriétés simples et bien connues de la loi normale au lieu de chercher une adéquation des données initiales à une distribution qu'il faut trouver, mais cette démarche s'est révélée négative (tests de normalité rejetés ; données traitées avec la procédure capability de SAS).

5.3 Détection des valeurs extrêmes dans les classes de risque selon la nouvelle méthode

On peut toujours considérer que les assurés d'une même classe de risque constituent un échantillon aléatoire de l'ensemble des assurables ayant les mêmes caractéristiques. D'ailleurs d'une année à l'autre, une partie des assurés change dans une même classe de risque, du fait de nouveaux assurés, des départs, des changements de véhicule,…

Dans cette application numérique, on considère une combinaison convexe des deux variables de seuil (u_1 : FME et u_2 : GPD) qui minimise la variance de cette combinaison (théorème 3). Cette démarche est meilleure que de prendre une simple moyenne arithmétique des seuils.

Classe	α	**Seuil**	**Convexe (*N*)**
1	0,283	4021	4
2	0,122	5115	6
3	0,241	5553	3
4	0,327	5114	4
5	0,193	6016	5
6	0,138	5241	5
7	0,186	7121	7
8	0,127	9687	7
9	0,191	11186	10
10	0,236	8815	5
11	0,134	11671	4
12	0,149	16972	7

Tableau 6 : Valeur des seuils selon la méthode mixte de minimisation de la variance

Soient $X = \alpha U_1 + (1-\alpha)U_2$ et $0 < \alpha < 1$. Dans ce tableau, α minimise la variance de *X*. On obtient $\alpha = \dfrac{V_2 - \mathrm{cov}(V_1, V_2)}{V_1 + V_2 - 2\,\mathrm{cov}(V_1, V_2)}$, où V_i correspond à la variance de U_i. *N* désigne toujours le nombre de valeurs extrêmes retenu selon le seuil correspondant à cette combinaison convexe des deux seuils.

5.4 Comparaison des méthodes

Dans chaque classe de risque, la variance estimée du seuil obtenu par la méthode FME est toujours plus grande que celle du seuil obtenu par la méthode GPD et donc la combinaison convexe des seuils est plus proche des seuils retenus par la méthode GPD que ceux déterminés par la FME, comme le montre le graphique suivant.

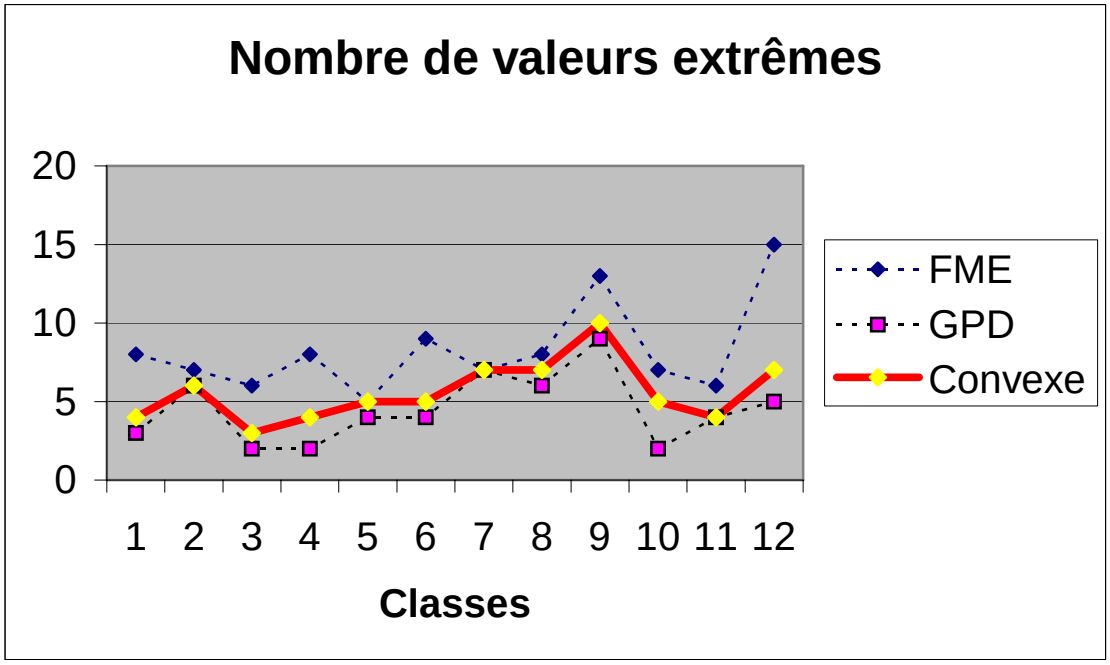


Figure 6 : Positionnement de la méthode convexe par rapport aux méthodes FME et GPD

Cette technique permet d'avoir un compromis entre les deux méthodes, entre une stratégie minimale (GPD) et une stratégie maximale (FME). Elle est davantage corrélée avec la méthode GPD et relativement plus lisse.

6. Conclusion

Les valeurs extrêmes ou sinistres graves en assurance automobile sont rares et leurs prévisions ou estimations doivent souvent être établies avec une grande prudence et en marge des données disponibles. Les modèles doivent être utilisés de façon souple, en restant pragmatique. L'approche doit être ouverte et multiforme, et en ce sens, il n'y a pas une méthode pour un problème. Les modèles ont des limites et les méthodes ne sont pas forcément adaptées à tout problème.

La théorie des valeurs extrêmes classique ou basée sur la loi de Pareto généralisée ne résout pas ces difficultés d'un coup, mais elle fournit des jumelles à travers lesquelles les assureurs peuvent observer les événements extrêmes avec une certaine objectivité afin d'une part de contrôler l'adéquation des primes pures avec la sinistralité et la hiérarchisation des classes de risque et d'autre part, de constituer des provisions pour faire face à ces risques extrêmes.

L'adéquation des données à la loi GPD permet d'estimer un quantile extrême, comme stratégie minimale, sensible à la taille des échantillons, et de prévoir des coûts aléatoires dont la probabilité d'occurrence est très faible, mais le choix du seuil au-delà duquel l'observation sera jugée extrême est un point à manier avec précaution, même si on propose une technique pour quantifier ce seuil. Cette technique, basée sur une diminution de la variance, semble être un bon compromis empirique de qualité entre la méthode FME et celle de la GPD.

La théorie des valeurs extrêmes peut figurer parmi les outils d'analyse pour les assureurs.

Bibliographie

Ardilly P. (1994)
Les techniques de sondage
Ed. Technip

Balkema A., de Haan L. (1974)
Residual life time at great age
The Annals for Probability, vol. 2, n°5, pp. 792-804

Balkema A., de Haan L. (1978)
Limit distribution for order statistics
SIAM Theory of probability and its applications, vol. 23, n. 1, pp. 77-92

Balkema A., de Haan L. (1978)
Limit distribution for order statistics
SIAM Theory of probability and its applications, vol. 23, n. 2, pp. 341-358

Beran R., Srivastava M.S. (1985)
Bootstrap tests and confidence regions for functions of a covariance matrix
The Annals of Statistics, vol. 13, n. 1, pp. 95-115

Bernardara P., Schertzer D., Lang M. (2005)
Intercomparaison of models and estimation techniques of extreme value in hydrology
Geophysical research abstracts Vol. 7

Bingham N.H., Goldie C.M., Teugels J.L. (1987)
Regular Variation
Cambridge University Press, Encyclopedia of Mathematics and its application, vol. 27

Beirlant J., Goegebeur Y., Segers J., Teugels J. (2004)
Statistics of Extreme
Ed. Wiley

Bortkiewicz (1922)
Variantionsbreite und mittlerer fehler
Sitzungber. Berli. Math. Ges., 21, pp. 3-11

Boulier J.F., Dalaud R., Longn F. (1998)
Application de la théorie des valeurs extrêmes aux marchés financiers
Banque et marchés, 32, pp. 5-14

Breiman L., Stone C.J., Kooperberg C. (1990)
Robust Confidence bounds for extreme upper quantiles
Journal of Statistical Computation and Simulation, Vol. 37, pp. 127-149

Butler P., Butler T. (1989)
Driver Record: a Political Red Herring That Reveals the Basic Flaw in Automobile Insurance Pricing
Journal of Insurance Regulation, Vol. 8, No. 2

Callens S. (1994)
Un siècle d'accidents automobile
Risques N° 20

Caussinus H., Hakam S., Riuz-Gazen A. (2002)
Projections révélatrices contrôlées, recherche d'individus atypiques
RSA, L (4), p. 81-94

Coles S. (2001)
An Introduction to Statistical Modelling of Extremes Values,
Springer Series in statistics, Springer-Verlag, London

Davison A.C., Hinkley D.V., Schechtman E. (1986)
Efficient bootstrap methods
Biometrika, 73, pp. 555-566

Davison A.C., Smith R.L (1990)
Models for exceedances over high thresholds
J.R. Statist. Soc. B 52, 3, pp.393-442

De Haan L., Ferreira A. (2006)
Extreme Value Theory: An Introduction
Springer-Verlag

Dekkers A.L.M., Einmahl J.H.L., de Haan L. (1989)
A moment estimator for the index of an extreme value distribution
The Annals of statistics, Vol 17, n. 4, pp. 1833-1855

Dionne G., Gouriéroux C., Vanasse C. (2001)
Testing for Evidence of Adverse Selection in the Automobile Insurance Market
Journal of Political Economy, vol.109, n. 2, pp. 444-453

Efron B. (1979)
Bootstrap methods: another look at the Jackknife
The Annals of Statistics, 7, pp.1-26

Efron B., Tibshirani R. J. (1993)
An introduction to the Bootstrap
New York, Chapman and Hall, p. 436

Einmahl J., Magnus J.R. (2006)
Records in athletics through extreme value theory
Tilburg University, Department of Econometrics and Operations Research, Paper Series n. 2006-83

Embrechts P., Kluppelberg C., Mikosch T. (1997),
Modeling Extremal Events for Insurance and Finance
Springer, Berlin.

Embrechts P., Resnick S.I., Samorodnitsky G. (1999)
Extreme value theory as a risk management tool
North American Actuarial Journal, vol. 3, n.2, pp.30-41

Fernandes L.B. (2003)
Extreme value theory and value at risk
Revista de Analisis Economico, Vol 18, pp. 57-85

De Finetti B. (1932)
Sulla legge di probabilita degli estremi
Metron, vol. IX, n. ¾, p. 125, Roma

Fisher R.A., Tippett L.H.C. (1928)
Limiting forms of the frequency distribution of the largest or smallest member of a sample
Proc. Cambridge Philos. Soc., 24, pp.180-190

Fréchet M. (1927)
Sur la loi de probabilité de l'écart maximum
Ann. Soc. Math. Polon., vol. 6, pp. 93-116

Galambos J. (1987)
The Asymptotic Theory of Extreme Order Statistics
R.E Krieger publishing Company

Gnedenko B.V. (1943)
Sur la distribution limite du terme maximum d'une série aléatoire
Ann. Math., 44, pp. 423-453

Goldie C., Smith R. (1987)
Slow variation with remainder : a survey of the theory and its applications
Quarterly Journal of Mathematics Oxford, vol. 38(2), pp. 339-349

Grun-Rehomme M. (2000)
Prévision du risque et tarification : le rôle du bonus-malus français
Revue Assurances, Montréal, 1, pp. 21-30

Grun-Rehomme M., Ben Lagha N., Vasechko O. (2007)
Une approche locale de la gestion des sinistres graves en assurance automobile
Revue Assurances et gestion des risques, HEC Montréal, vol 75(3), pp. 409-429

Guillou A., Willems P. (2006)
Application de la théorie des valeurs extrêmes en hydrologie
RSA, LIV (2), 5-31

Gumbel E.J. (1955)

Statistical theory of extremes values and some practical applications
Journal of the Royal Statistical Society, Serie A, vol. 119, n.1, p. 106

Gumbel E.J. (1958)
Statistics of Extremes
Columbia University Press

De Haan L., Rootzen H. (1993)
On the estimation of high quantiles
J. of Statistical Planning and Inference, vol.35, n.1, pp.1-13

De Haan L., Peng L. (1998)
Comparison of tail index estimators
Statistica Neerlandica, vol. 52(1), pp. 60-70

De Haan L., Ferreira A. (2006)
Extreme Value Theory
Springer-Verlag

Hill B.M. (1975)
A simple General Approach to Inference about the Tail of a Distribution
Annals of Statistics, 3(5), pp. 1163-1174

Hosking J.R.M., Wallis J.F. (1987)
Parameter and quantile estimation for the generalized Pareto distribution
Technometrics, vol. 29, Issue 3, pp. 339-349

Jenkinson A.F (1955)
The frequency distribution of the annual maximum (or minimum) of meteorological elements
Quart. J. R. Met. Soc. 81, pp.158-171

Katz R. W. (2002)
Techniques for estimating Uncertainty in climate change Scenarios and impact studies
Climate research, Vol. 20, pp. 167-185

Kotz S., Nadarajah S. (2000)
Extreme Value Distributions
Imperial College Press, London

Kunreutber H. (2001)
Le rôle de l'assurance dans la gestion des événements extrêmes
Risques n° 48

Lemaire J. (1985)
Automobile Insurance Actuarial Models
Kluwer, Amsterdam

McNeil A.J., Saladin T. (1997)

The peaks over thresholds for estimating high quantiles of loss distribution
International ASTIN Colloquium, pp. 70-94.

Mc Neil A.J., Frey R. (2002)
Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach
Journal of Empirical Finance, 7, pp. 271-300

Martin J.L., Derrien Y., Laumon B. (2003)
Estimating relative driver fatality and injury risk according characteristics of cars and drivers using matched-pair multivariate analysis
ESV, Proceedings, Nagoya

Melgar M.C., Ordaz Sanz J.A., Guerrero F.M. (2005)
Diverses alternatives pour déterminer les facteurs significatifs de la fréquence d'accidents dans l'assurance automobile
Revue Assurances et gestion des risques, vol.73-1, 31-54, Montréal, Canada

Moreau L. (1984)
L'écrêtement des sinistres "automobile"
Astin Bulletin, vol. 14, n. 2, pp. 173-181

Pickands J. (1975)
Statistical inference using extreme order statistics
Ann. Statist.3, 119-131.

Prescott P., Walden A.T. (1980)
Maximum likelihood estimation of the parameters of the generalized extreme-value distribution
Biometrika, 67, pp. 723-724.

Reiss R. (1997)
Statistical Analysis of extreme values : from insurance, finance, hydrology and other fields
Birkhauser verlag

Reiss R.D., Thomas M. (1997)
Statistical Analysis of Extreme Values,
Birkhauser Verlag, Boston, MA.

Reiss R., Thomas M. (2001)
Statistical Analysis of extreme values
Birkhauser Verlag

Schirmacher D. (2005)
Stochastic Excess-of-loss pricing with a financial Framework
Casualty Actuarial Society Forum, Spring

Smith R.L., Goodman D.J. (2000)
Bayesian risk analysis in Extremes and integrated risk management,
Risk Books, London 235-251

Todorovic P., Zelenhasic E. (1970)
A stochastig model for flood analysis
Water Resour. Res., 6, pp. 1641-1648

Todorovic P., Rousselle J. (1971)
Some problems on flood analysis, Water Resour. Res., 7, pp. 1144-1150

Annexe 1 : Démonstration de la relation (11)

$E(N_n) = 1 + \sum_{k=2}^{n} P(X_k > M_{k-1})$ et $P(X_k > M_{k-1}) = \frac{(k-1)!}{k!} = \frac{1}{k}$, d'où $E(N_n) = \sum_{k=1}^{n} \frac{1}{k}$

Pour la variance du processus de comptage, on a :

$Var(N_n) = Var(N_n - 1) = E((N_n - 1)^2) - (E(N_n - 1))^2$, puis on calcule ces deux termes

$$E((N_n - 1)^2) = 2\left(\sum_{k>p, p=2}^{n} P(X_k > M_{k-1} \cap X_p > M_{p-1}) \right) + \sum_{k=p=2}^{n} P(X_k > M_{k-1})$$

$$E((N_n - 1)^2) = 2\left(\sum_{k>p, p=2}^{n} \frac{1}{k} \times \frac{1}{p} \right) + \sum_{k=2}^{n} \frac{1}{k}$$

$$(E(N_n - 1))^2 = \left(\sum_{k=2}^{n} \frac{1}{k}\right)^2 = \sum_{k=2}^{n} \frac{1}{k^2} + 2 \sum_{k>p, p=2}^{n} \frac{1}{k} \times \frac{1}{p}$$

d'où $Var(N_n) = \sum_{k=2}^{n} \frac{1}{k} - \sum_{k=2}^{n} \frac{1}{k^2} = \sum_{k=2}^{n} \left(\frac{1}{k} - \frac{1}{k^2}\right)$

Annexe 2: Démonstration du théorème 3

Soient U_i des variables aléatoires réelles ($i = 1,..., p$), on note V_i la variance de U_i et V_{ij} la covariance des variables U_i et U_j. Soit Z la variable aléatoire définie comme combinaison convexe des U_i, à savoir : $Z = \sum_{i=1}^{p} \alpha_i U_i$, où $\alpha_i > 0$ et $\sum_{i=1}^{p} \alpha_i = 1$. On cherche donc à résoudre le problème d'optimisation $\underset{\alpha}{Min}\, Var(Z)$ sous la contrainte $\sum_{i=1}^{p} \alpha_i = 1$, où $\alpha = (\alpha_1,..., \alpha_p)$. La fonction à minimiser et l'ensemble des contraintes étant convexes, la solution du problème vérifie les conditions de Kuhn et Tucker, et réciproquement.

On a $Var(Z) = \sum_i \alpha_i^2 V_i + 2 \sum_{i<j} \alpha_i \alpha_j V_{ij}$. Le lagrangien L de ce problème de minimisation convexe sous contrainte linéaire peut s'écrire : $L(\lambda, \alpha) = Var(Z) - 2\lambda(\sum_i \alpha_i - 1)$, où λ est un multiplicateur de Lagrange qui correspond à la contrainte d'égalité. La solution du problème doit vérifier les conditions de Kuhn et Tucker : $\frac{\partial L}{\partial \alpha_i} = 0$ pour tout $i = 1, 2,..., p$. On a :

$\frac{\partial L}{\partial \alpha_i} = 0 \Leftrightarrow \alpha_i V_i + \sum_{j \neq i} \alpha_j V_{ij} = \lambda \Leftrightarrow \alpha = V^{-1}\Lambda$, où V est la matrice de variance-covariance des U_i, V^{-1} est l'inverse de V, α est la matrice uni-colonne des α_i et la matrice Λ est égale à λ fois la matrice uni-colonne A dont tous les coefficients sont égaux à 1.

Notons v_{ij}^{-1} le terme général de la matrice V^{-1}.

Le système $\alpha = V^{-1}\Lambda$ s'écrit aussi, pour tout $i = 1, 2,..., p$, $\alpha_i = \lambda(\sum_{j=1}^{p} v_{ij}^{-1})$.

De la contrainte $\sum_{i=1}^{p} \alpha_i = 1$, on obtient $\lambda = \frac{1}{\sum_{i,j=1}^{p} v_{ij}^{-1}}$, puis $\alpha = \frac{V^{-1}}{\sum_{i,j=1}^{p} v_{ij}^{-1}} A$. Dans le cas de deux variables, l'obtention de la valeur de α_1 (avec $\alpha_2 = 1 - \alpha_1$) est immédiate (cf. fin du §5.3).