Afrika Statistika

Afrika Statistika
Vol. 16 (3), 2021, pages 2883 - 2909.
DOI: http://dx.doi.org/10.16929/as/2021.2883.188

ISSN 2316-090X

Applying of the Extreme Value Theory for determining extreme claims in the automobile insurance sector: Case of a China car insurance

Daouda Diawara [(1)], **Ladji Kane** [(1)], **Soumaila Dembele** [(1,2,*)] **and Gane Samb Lo** [(2,3,4)]

[(1)] Faculté des Sciences Économiques et de Gestion (FSEG) Bamako, Mali
[(2)] LERSTAD - Université Gaston Berger,Saint-Louis, SENEGAL
[(3)] African University of Sciences and Technology (AUST), Abuja, NIGERIA
[(4)] LASTA - Université Pierre et Marie Curie, Paris, FRANCE (Affiliated to)

Received on August 20, 2021; Accepted on November 20, 2021

Abstract. . According to the Chinese Health Statistics Yearbook, in 2005, the number of traffic accidents was 187781 with total direct property losses of 103691.7 (10000 Yuan). This research aims to fill the gap in the literature by investigating the extreme claim sizes not only for the entire portfolio. This empirical study investigates the behavior of the upper tail of the claim size by class of policyholders.

Key words: China car insurance ; extreme value theory ; threshold ; POT.
AMS 2010 Mathematics Subject Classification Objects : 62-07; 60G70; 62G32.

[(*)] Corresponding author: Soumaila Dembele (soumailadembeleussgb@gmail.com, dembele.soumaila@ugb.edu.sn)
Daouda Diawara: btddiawara@yahoo.fr
Ladji Kane : fsegmath@gmail.com
Gane Samb Lo : gane-samb.lo@ugb.edu.sn, gslo@aust.edu.ng

Résumé. Selon l'annuaire statistique de la santé pour la Chine en 2005, le nombre d'accidents de la circulation était de 187781 avec des pertes totales directes de biens de 103691,7 (10000 Yuan). Cette recherche vise à combler le vide dans la littérature en étudiant les montants extrêmes des sinistres pour l'ensemble du portefeuille. Cette étude empirique examine le comportement de la queue supérieure de la taille des sinistres par catégorie d'assurés.

The authors.

Daouda Diawara, Ph.D., is rank and discipline at Faculté des Sciences Économiques et de Gestion (FSEG), Université des Sciences Sociales et de Gestion de Bamako (USSGB)

Ladji Kane, Ph.D., is rank and discipline at Faculté des Sciences Économiques et de Gestion (FSEG), Université des Sciences Sociales et de Gestion de Bamako (USSGB)

Soumaila Dembele, Ph.D., is is rank and discipline at Faculté des Sciences Économiques et de Gestion (FSEG), Université des Sciences Sociales et de Gestion de Bamako (USSGB)

Gane Samb Lo, Ph.D., is full professor of Mathematics and Statistics at LERSTAD, Gaston Berger University, Saint-Louis, SENEGAL and at African University of Sciences and Technology [AUST], NIGERIA. He is affiliated to LSTA, Pierre and Marie Curie University, Paris VI, France. He is the head and founder of the virtual Imhotep Mathematical Center (IMC), imhotepsciences.org

1. Introduction

1.1. Context of the study

China has been experiencing a high-speed urbanization and motorization related to rapid economic growth. The growing urbanization and motorization levels result in increasing frequencies of road traffic accidents, injuries and deaths. As reported by the world Health Organization, in 2015, the total fatalities in china are among the highest in the world. According to the Chinese Health Statistics Yearbook, in 2005, the number of traffic accidents was 187781 with total direct property losses of 103691.7 (10000 Yuan). The Chinese car insurance and reinsurance companies mainly support these losses. Moreover, statistics highlight that 322 events are considered as serious accidents and generate very large losses for the insurers. The appearance of these excessively large claims leads to think about claims resulting from severs accidents as with massive pile up of vehicles or light-truck accidents. In other instances, an insurer might be confronted with large claims coming from a policy involving very valuable items such as concentrated risks (e.g. a large number of luxury cars located in the same area).

This research is a contribution to the statistical investigating of the claims. We do not focus, as in many studies, on the entire portfolio. Rather, we will investigate the behavior of the upper tail of the claim size by class of policyholders.

Since we use the univariate extreme value theory, we feel obliged to give a brief account of such a theory

1.2. An easy introduction of the univariate extreme value theory

.
The theory is usually presented in tow main point of views: the max-stability apporach (MSA) and the Peak Over Threshold (POT) Method.

1.2.1. The max-stability approach

Let X, X_1, X_2, $\cdots$ be a sequence independent real-valued randoms, defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with common cumulative distribution function F, which has the lower and upper endpoints, the first asymptotic moment function and the generalized inverse function respectively defined by

$$lep(F) = \inf\{x \in \mathbb{R},\ F(x) > 0\},\ uep(F) = \sup\{x \in \mathbb{R},\ F(x) < 1\}$$

$$R(x, F) = \frac{1}{1 - F(x)} \int_x^{uep(F)} (1 - F(y))\ dy,\ x \in]lep(F), uep(F)[$$

and

$$F^{-1}(u) = \inf\{x \in \mathbb{R},\ F(x) \geq y\}\ for\ u \in]0, 1[\ and\ F^{-1}(0) = F^{-1}(0+).$$

The main task the *UEVT*, originally, was the study of the max-stability of iid sequence. Namely, the essential problem is finding real and nonrandom sequences $(a_n > 0)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ such that the centered and normalized sequence of partial maxima

$$\frac{\max(X_1, \cdots, X_n) - b_n}{a_n} =: \frac{X_{n,n} - b_n}{a_n}$$

(with $X_{n,n} = \max(X_1, \cdots, X_n)$) weakly converges to some *cdf* M, which is equivalent to: for any continuity point x of the *cdf* G of M:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = M(x). \tag{1}$$

The *UVT* goes back to the earlier years of the 1900's with many contributors as Fisher and Tippet (1928), Gumbel (1955), etc. for statistical purposes. On the

mathematical ground, Gnedenko (1943) (see Lo et al. (2018) [Theorem 2, page 13] and Resnick (1987)) have final characterizations of max-stable random variables as stated in Formula (7) below. First important accounts of the theory are given in Galambos (1985), de Haan (1970), etc. The *UEVT* uses regularly varying functions, a concept deeply described in Feller (1968) (page 275) and Loève (1997) (page 354). Later several authors, on the basis of these fundamental texts, provided other exposition of the theory, some of them focusing on what is now called *statistics of extremes*. Let use cite a few of them de Haan and Feireira (2006) (who significantly developed his pioneering work in de Haan (1970)), Resnick (1987), Embrechts *et al.* (1997), Beirlant *et al.* (2004), Lo et al. (2018). Other important references can be found in the cited books.

It is showed that the convergence in (1) is a convergence in type, meaning that a change of the coefficients $(a_n > 0)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ to $(\alpha_n > 0)_{n\geq 1}$ and $(\beta_n)_{n\geq 1}$ and a change of M to M_1 with *cdf* G_1 necessarily leads to the following facts:

$$\alpha_n/a_n \to A > 0 \ \ and \ \ (\beta_n - \alpha_n)/a_n \to 0 \ \ as \ \ n \to +\infty \tag{2}$$

and

$$\forall x \in \mathbb{R},\ G_1(x) = G(Ax + B), \tag{3}$$

(see Lo et al. (2018) [Lemma 42, page 12] and Resnick (1987)) The later fact says that G_1 and G are equal in type. Gnedenko's theorem establishes that if (1) holds and if M is not concentrated on a single point, the only three possibilities are the following:

$$Fr_\alpha(x) = \exp(-x^{1/\gamma})\ 1_{(x\geq 0)}, \quad (Type\ I) \tag{4}$$

$$W_\alpha(x) = \exp(-(-x)^{1/\gamma})\ 1_{(x\geq 1)} + 1_{(x>0)}, \quad (Type\ II) \tag{5}$$

and

$$\Lambda(x) = \exp(-e^{-x})\ 1_{\mathbb{R}}(x), \quad (Type\ III) \tag{6}$$

Let us apply Formulas (2) and (3) to (4) with $A = \gamma = \alpha$ and $B = 1$, next Formulas (2) and (3) to (4) with $A = -\gamma = -1/\beta$ and $B = -1$ and finally by interpreting

$$\left[\exp(-(1+\gamma x)^{-1/\gamma})\right]_{\gamma=0} = \exp\left(-\exp(-x)\right),\ x \in \mathbb{R}.$$

to get that any weak M in (1) has a *cdf* in the family of the Generalized Extreme Value (GEV) *df* :

$$H_\gamma(x) = \exp(-(1+\gamma x)^{-1/\gamma}),\ 1+\gamma x \geq 0, \tag{7}$$

parametrized by $\gamma \in \mathbb{R}$, with $H_0(x) = 1 - \exp(-e^{-x})$, $x \in \mathbb{R}$, for $\gamma = 0$. The parameter γ is called the extreme value index.

As to the choice of the sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, the following ones

$$a_n = F^{-1}\left(1 - \frac{1}{n}\right) \quad and \quad b_n = 0,$$

$$a_n = F^{-1}\left(1 - \frac{1}{n}\right) - uep(F) \quad and \quad b_n = -uep(F),$$

(with $uep(F) < +\infty$ necessarily) and

$$a_n = F^{-1}\left(1 - \frac{1}{ne}\right) - F^{-1}\left(1 - \frac{1}{n}\right) \quad and \quad b_n = F^{-1}\left(1 - \frac{1}{n}\right),$$

for $n \geq 1$, leads to the the limits (4), (5) and (6) respectively.

Now, we are recalls useful criteria for *cdf*'s to belong to the whole domain

$$\mathcal{D} = \{G_\gamma,\ \gamma > 0\}$$

of extreme attraction and functional representation of *cdf*'s and their quantile functions in $\mathcal{D}$.

First, we have:

Proposition 1. *(Main criteria for $F \in \mathcal{D}$)*
Let F be a cdf on $\mathbb{R}$. We have two general cases and in each of them, we consider two sub-cases.

(A) *- Let $uep(F) = +\infty$.*

(A1) *- $F \in D(G_{1/\gamma})$, $\gamma > 0$, if one of the following assertions hold.*

$$\forall \lambda > 0,\ \lim_{x\to+\infty} \frac{1 - F(\lambda x)}{1 - F(x)} = \lambda^{-1/\gamma}. \quad (A11)$$

$$\forall \lambda > 0, \lim_{u\to 0} \frac{F^{-1}(1 - \lambda u)}{F^{-1}(1 - u)} = \lambda^{-\gamma}. \quad (A12)$$

$$\lim_{x\to+\infty} \frac{x\ F'(x)}{1 - F(x)} = 1/\gamma. \quad (A13)$$

(A2) *- $F \in D(G_0)$ if one of the following assertions hold.*

(a) There exists a slowly varying function at zero that we denote as $s(u)$ of $u \in]0,1[$ such that

$$\forall \lambda > 0, \lim_{u \to 0} \frac{F^{-1}(1-\lambda u) - F^{-1}(1-u)}{s(u)} = -\log \lambda. \quad (A21)$$

(b) The Γ-variation formula holds (due to de Haan (1970)).

$$\forall t \in \mathbb{R}, \lim_{x \to uep(F)} \frac{1 - F(x + tR(x)}{1 - F(x)} = e^{-t} \quad (A22)$$

(c) Upon the twice differentiability of F on the left neighborhood of $uep(F)$, the following Von-Mises condition holds:

$$\lim_{x \to +\infty} \frac{F''(x)F(x)}{(F'(x))^2} = -1. \quad (A23)$$

(d) Upon the twice differentiability of F on the left neighborhood of $uep(F)$, the function (Lo (1986)'s criteria)

$$s(u) = -u \left(F^{-1}(1-u)\right)'$$

is slowly varying at zero.

(B) *- Let $uep(F) < +\infty$.*

(B1). *$F \in D(G_\gamma)$, $\gamma < 0$ if one of the following assertions holds.*

$$\forall \lambda > 0, \lim_{x \to uep(F)} \frac{1 - F(uep(F) - (\lambda x)^{-1})}{1 - F(uep(F) - x^{-1})} = \lambda^{1/\gamma}. \quad (B11)$$

$$\forall \lambda > 0, \lim_{u \to 0} \frac{uep(F) - F^{-1}(1-\lambda u)}{uep(F) - F^{-1}(1-u)} = \lambda^{-\gamma}. \quad (B12)$$

$$\lim_{x \to uep(F)} \frac{(upe(F) - x)\ F'(x)}{1 - F(x)} = -1/\gamma. \quad (B13)$$

(B2) *To test whether or not $F \in D(G_0)$, we re-use the criteria of Sub-case (A2).*

Next, we have the main representations of F^{-1} for $F \in \mathcal{D}$.

Proposition 2. *(Representations of quqntile functions whithin the extreme attraction domain) We have the following characterizations for the three extremal domains.*

(a) $F \in D(H_\gamma)$, $\gamma > 0$, *if and only if there exist a constant* c *and functions* $a(u)$ *and* $\ell(u)$ *of* $u \to u \in]0,1]$ *satisfying*

$$(a(u), \ell(u)) \to (0,0) \text{ as } u \to +\infty,$$

such that F^{-1} *admits the following representation of karamata (1030)*

$$F^{-1}(1-u) = c(1+a(u))u^{-\gamma} \exp\left(\int_u^1 \frac{\ell(t)}{t} dt\right). \tag{8}$$

(b) $F \in D(H_\gamma)$, $\gamma < 0$, *if and only if* $uep(F) < +\infty$ *and there exist a constant* c *and functions* $a(u)$ *and* $\ell(u)$ *of* $u \in]0,1]$ *satisfying*

$$(a(u), \ell(u)) \to (0,0) \text{ as } u \to +\infty,$$

such that F^{-1} *admit the following representation of karamata (1030)*

$$uep(F) - F^{-1}(1-u) = c(1+a(u))u^{-\gamma} \exp\left(\int_u^1 \frac{\ell(t)}{t} dt\right). \tag{9}$$

(c) $F \in D(H_0)$ *if and only if there exist a constant* d *and a slowly varying function* $s(u)$ *such that*

$$F^{-1}(1-u) = d + s(u) + \int_u^1 \frac{s(t)}{t} dt, 0 < u < 1, \tag{10}$$

and there exist a constant c *and functions* $a(u)$ *and* $\ell(u)$ *of* $u \to u \in]0,1]$ *satisfying*

$$(a(u), \ell(u)) \to (0,0) \text{ as } u \to +\infty,$$

such that s *admits the de Haan (1970) representation*

$$s(u) = c(1+a(u)) \exp\left(\int_u^1 \frac{\ell(t)}{t} dt\right). \tag{11}$$

Moreover, if $F^{-1}(1-u)$ *is differentiable for small values of* s *such that* $r(s) = -s(F^{-1}(1-s))' = udF^{-1}(1-s)/ds$ *is slowly varying at zero, then 10 may be replaced by*

$$F^{-1}(1-u) = d + \int_u^{u_0} \frac{r(t)}{t} dt, \ 0 < u < u_0 < 1, \tag{12}$$

which will be called a Lo (1986)'s representation or areduced de Haan (1970) representation of F^{-1}.

Remark. Let us remark that any $F \in \mathcal{D}$ is associated to a paire of functions $(a(u), b(u))$ of $u \in]0,1[$. For $\gamma \neq 0$, these functions directly appear in the representations (8) and (9) in Proposition . For $\gamma = 0$, the representation uses $s(\circ)$ which, in turn, uses the function $(a(u), b(u))$ of $u \in]0,1[$. In that sense each $F \in \mathcal{D}$ can be represented as $F \equiv (a, b)$.

Let us finish by stating a rule for differentiable *cds*'s.

Rules of working . In the domain of extremal attraction, most of the *cdf*'s which are used in applications are differentiable in a left-neighborhood of the upper endpoint. In such a case, we may take $a \equiv 0$ in Representation (8) and (9) in Proposition 1.2.1. In that case, we need only the function $b(\circ)$ and by solving easy differential equations, we may take

$$b(u) = -u(G^{-1}(1-u))' - \gamma,\ u \in (0,1)\ \ and\ \ a \equiv 0 \tag{13}$$

for $\gamma > 0$ and

$$b(u) = -\gamma - \frac{u}{F'\Big(F^{-1}(1-u)\Big)\Big(uep(F) - F^{-1}(1-u)\Big)},\ u \in]0,1[. \tag{14}$$

for $\gamma < 0$, whenever we have $b(u) \to 0$ as $u \to 0$. Consequently, we may drop a_n in the rates of convergence to reduce them to $O_{\mathbb{P}}(b_n \vee c_n)$.

For $\gamma = 0$, Representation (12) in Proposition 1.2.1 holds for

$$s(u) = -u(F^{-1}(1-u))',\ 0 < u < 1,$$

whenever it is slowly varying at zero and the rate of convergence a_n becomes useless. In such cases, the rate of convergence reduces $O_{\mathbb{P}}(d_n \vee c_n)$.

1.2.2. The POT approach

That approach relies on the Generalized Pareto Distribution (GPD) with two parameters $\lambda \in \mathbb{R}$ and $b > 0$, defined as follow

$$G_{\lambda,b}(x) = \Big(1 - (1 + \lambda x/b)^{-1/\lambda}\Big)1_{(\lambda \neq 0)} + \Big(1 - \exp(-x/b)\Big)1_{(\lambda = 0)},$$

where $b > 0$, $x \geq 0$ for $a \geq 0$. $0 \leq x \leq -b/a$ for $a < 0$. The parameter λ and b are called shape and scale parameters respectively. The link of the *GPD* with *UEVT* relates to

the excess distribution over threshold u for a cdf F associated to a random variable X, which is is defined by

$$F_u(x) = \mathbb{P}(X - u \leq x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}, \quad 0 \leq x \leq uep(F) - u.$$

The mathematical expectation of that conditional law is called the mean excess function is defined by

$$e(u) - \mathbb{E}(X - u | x > u), \ u \in \mathbb{R}.$$

We have the following important theorem

Theorem 1. *There exists a function $b(u)$ of $u \in \mathbb{R}$ such that*

$$\lim_{u \to uep(F)} \sup_{0 \leq x \leq uep(F) - u} |F_u(x) - G_{\lambda, b(u)}(x)| = 0$$

if and only if $F \in G_\lambda$

It is known that for $0 \leq \lambda < 1$, the mean of $G_{\lambda,b}$ is $b/(1-\lambda)$. In the frame of Theorem 1

$$e(u) \approx \frac{b}{1-\lambda} + \frac{\lambda}{1-\lambda} u$$

for u large enough. So the empirical methodology suggested for estimation a heavy tail consists in observing the empirical excess function from the data

$$e_n(u) = \frac{\sum_{1 \leq i \leq n} (X_i - u) 1_{X_i > u}}{\sum_{1 \leq i \leq n} 1_{X_i > u}},$$

and to check whether or not the curve $(u, e_n(u))$ is approximately a straight line for u near *uep(F)*. (see the paper by Ba *et al.* (2004) for empirical estimation of the mean excess function by confidence bounds).

After this quick round up of the *UEVT*, we are going to use some keys elements of it in our empirical studies.

1.3. Scope of the study and organization of the paper

As outlined before, this research is concerned with the Chinese insurance market and aims to identify the best distribution in modeling extreme claim sizes stock returns by using the peak over threshold method (POT) to identify the best distribution in modeling extreme claim sizes.

The first classification is based on the gender and the second is the experience of the driver.

This new approach has many practical implications for both, the insurer and the reinsurer. The main implication is that differentiating by characteristics of the policyholders allows a fair premium paid by each category of insured. Moreover, it allows an accurate estimation of the distribution of extreme losses and a better determination of the limit for individual claim size by the reinsurer in the case of the application of the excess-of-loss reinsurance strategy.

The remainder of this paper is organized as follows: The Section 1 is a general introduction. The section 2 describes the data and sample statistics. The research methodology is discussed in Section 3. Section 4 presents the empirical results. Section 5 discusses the implications of the findings for the Chinese insurance market. Summary and concluding remarks are given in Section 6.

2. Data and sample statistics

In this empirical investigation, the dataset consists of a sample of 405 177 observations for 4-wheeled vehicles and motorcycles from an insurance company of the Hubei Province headquartered in Wuhan (the insurance company covers the entire province of Hubei). The name of the insurance company is omitted for security reasons. The data covers five calendar years between 2012 and 2017. The data contains information about the characteristic of the policyholders, the insured car and variables related to the claims. In this paper, we mainly focus on the claim sizes, the gender and the experience of the drivers.

In this section, we particularly describe the claim sizes. We proceed by a logarithm transformation to allow a better investigation of the distribution of the claim amount. The logarithm transformation is also recommended for the study of the extreme distribution. A complete description of the portfolio is upon request in the supplementary documents.

Note. Table 1 displays the main summary statistics, the mean, the minimum, the maximum, the Skewness and the Kurtosis. The P 90%, P 95% and P 99% indicate the 90^{th}, 95^{th} and 99^{th} percentiles.

Table 1 presents descriptive statistics of the claim sizes after logarithm transformation for the five samples. The mean of the claim sizes is likely to be equal for the five cases. The skewness value is positive for all datasets, indicating that

variable	Mean	Min	Max	SD	Skewness	Kurtosis	P 90%	P 95%	P 99%
Portfolio	7.661	0.559	12.923	1.184	0.776	4.308	9.176	9.867	11.342
Young	7.671	6.214	11.608	0.927	1.531	6.965	8.829	9.230	11.289
Exper	7.875	3.737	12.923	1.198	0.784	4.072	9.220	9.956	11.283
Male	7.660	0.559	12.923	1.182	0.782	4.327	9.168	9.867	11.342
Female	7.662	0.559	12.923	1.922	0.963	4.218	9.199	9.885	11.342

Note. This table displays the main summary statistics, the mean, the minimum, the maximum, the Skewness and the Kurtosis. The P 90%, P95% and P 99% indicate the 90th, 95th and 99th percentiles.

Table 1. Summary statistics for the log claim sizes

distributions of the claims sizes are right skewed indication the presence of an upper tail. Moreover, Kurtosis exceeds the reference value of the Gaussian distribution (equal to 3) for all cases. These preliminary statistics corroborate the non-normality of the studied distributions. The examination of the different percentile shows large gaps between the calculated quantiles and the mean for all datasets. This gap may be considered as a serious problem for the insurer. Having estimated the pure premium using the mean of the claim sizes, the insurer may support additional losses related to the presence of these gaps. Accordingly, the insurer has to examine in more in-depth the claim sizes in the upper tail of the distributions.

The non-normality of the claim sizes justifies using other distributions characterized by the presence of a right tail as the skewness indicates.

Figure 1 clearly shows the presence of extreme values for the entire portfolio as well as the different studied groups of insureds. The Boxplot helps only to show and detect the presence of the extreme values in a dataset. However, it did not allow for modeling the distribution of the detected extreme observations.

3. Methodology

In this paper, we assume that the sequence $X_1, X_2, ..., X_n$ of successive claim sizes consists of independent and identically distributed random variables generated by distribution F_X of generic random variable X. The n^{th} moment of the claim size distribution is usually denoted by $\mu_X^{(n)} = E(X^n) = \int_0^\infty x^n \, dF_X(x)$. For $n = 1$ we write $\mu_X = \mu_X^{(1)}$. The variance of the claim sizes is written as $Var(X) = \mu_X^{(2)} - (\mu_X)^2$. The aggregate claim amount at time t can be written as $Y(t) = \sum_{i=1}^{N(t)} X_t$ and has the distribution function $F_{Y(t)}(y) = P(Y(t) \leq y)$. In any case, more specific information on the interdependencies within and between the two processes $N(t)$ and $X(t)$, describing the number and the size of the claims, is crucial.

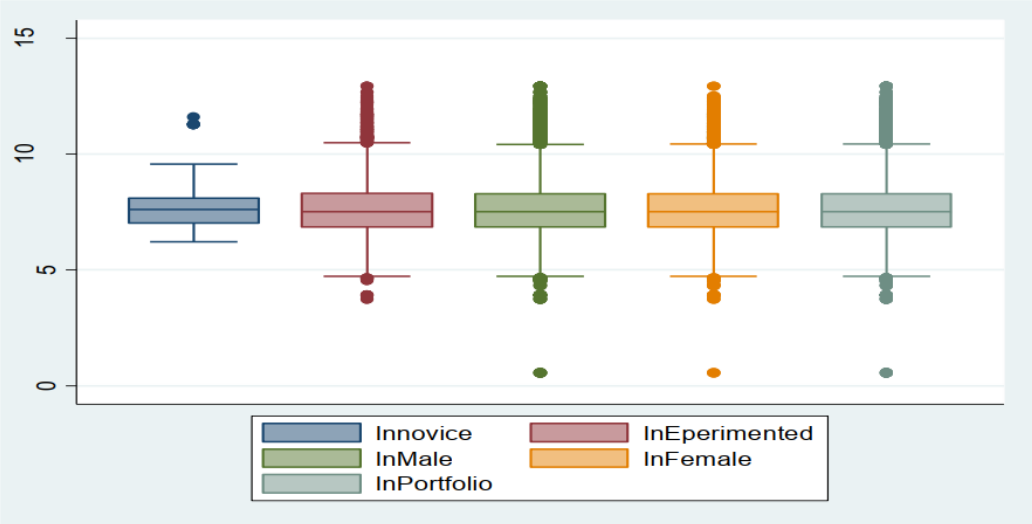

Fig. 1. The Box-plot of the claim sizes distributions

In this paper we will focus only on the claim amount distribution. Thus, there is no impact of the dependence between the claim size and claim number on our modeling. However, this study will investigate the claim amounts, in particular, those reflecting the dangerous risks.

3.1. Concepts of risks and dangerous risks

The risk, in its most general form, can be defined as uncertainty associated with a future outcome or event. To apply this more specifically to insurer activity, we can say that risk is the expected variance in losses or in the claim amounts. Statistically, as suggested by Rolski *et al.*(1999), any non-negative random variable or its distribution is frequently called a risk. Therefore, any distribution, concentrated on the non-negative half-line, can be used as a claim size distribution. However, it will be worthwhile to make a distinction between "well-behaved" distributions and dangerous distributions with a heavy tail. The concepts of well-behaved or heavy-tailed distributions belong to the common vocabulary of actuaries. Rolski *et al.*(1999) formalize these two concepts in a mathematically sound definition. We also find an interest introduction to extreme value see Lo, (2017).

The class of well-behaved distributions consists of those distributions with an exponentially bounded tail. This condition means that large claim sizes are not impossible, however their occurrence probabilities decreases exponentially fast to zero as the fixed threshold becomes larger and larger. In contrast, for the heavy tailed distributions there is no proper exponential bound and huge or extreme claims are getting more likely. A natural nonparametric class of heavy-tailed claim size distributions is the class of sub-exponential distributions (e.g. lognormal,

Pareto and Weibull distributions with shape parameter smaller than 1). For more details about the statistical proprieties of sub-exponential distributions see, Rolski *et al.*(1999), Omey(2006) or more recently Lu and Bin Zhang(2016) how provide some asymptotic results of the ruin probabilities in renewal risk model with some strongly sub-exponential claims and they obtain the asymptotic upper and lower bounds the studied distributions.

The sub exponential distributions are used to detect and predict the large claim amounts. However, a number of sub-exponential distributions are not defined in the domain of attraction of an extreme value distribution, Goldie and Resnick (1988).

Since we are interested in modeling the claim sizes in the tail, we will focus only on the distributions defined in the domain of attraction of an extreme value distribution.

3.2. Modeling large claim amounts

It is obvious that the detection of large or extreme claim size distributions is one of the main concerns of the practicing actuary in insurance and reinsurance companies. The Extreme Value Theory is a tool for estimating the tails of a distribution. Two types of extreme value theory are used to this purpose, the classical extreme value theory (EVT) and the peak over threshold method (POT). In this paper, we will focus on the second method, which is suitable to our purposes. The next section gives a brief overview of the used extreme value methodology; first we will introduce the concepts of VaR and the expected shortfall. Then we will describe the peak over threshold method.

3.2.1. Value-at-Risk and Expected Shortfall

A very common risk measure in the financial world is the Value-at-Risk (VaR). This is in fact nothing else but at 95%. For a distribution function F, VaR is the q^{th} quantile of F: $VaR_q = F^{-1}(q)$.

For some q, typically in the domain $q \in (0,95;1)$ VaR provides an upper bound for a loss that is only exceeded on a small proportion of occasions, sometimes referred to as a confidence level. VaR has been criticized as a risk measure, since it is not necessarily subadditive: There are cases where a portfolio can be split into sub-portfolio such that the sum of the VaR for the sub-portfolios is smaller than the VaR for the total portfolio. Further, VaR gives no information of the potential size of the loss exceeding VaR.

Therefore, it has been proposed to use the expected shortfall (ES) or " tail conditional expectation" instead of VaR. ES is the expected size of a loss exceeding VaR : $ES_q := E[X \mid X > VaR_q]$.

3.2.2. Peaks-Over-Threshold (POT)

The most common group of models are the Peaks-Over-Threshold (POT) models. These are models for all large observations that exceed a high threshold. The POT models are generally considered to be the most useful for practical applications. With the POT class of models, one may further distinguish two styles of analysis. There are semi-parametric models, built around the so called Hill estimator (and its relatives), and the fully parametric models, based on the generalized Pareto distribution (GPD).Both classes are theoretically justified and empirically useful when used correctly.

The distribution of exceedances

Let $X_1, ..., X_n$ be identically distributed (not independent) random variables with unknown distribution function $F(x) = \mathbb{P}(X \leq x)$. Given a high threshold u_n, we index each observation exceeding u_n and obtain another sample $\{Y_1, ..., Y_{N_u}\}$ $N_u \leq n$.

Consider the $i.i.d$ case. Each point has the same chance to exceed the threshold with success probability $\mathbb{P}(X_i > u_n)$, $i = 1, ..., n$. Hence, the number of exceeding observations is:

$$N_{u_n} := \#\{i : X_i > u_n, i = 1, ..., n\} = \sum_{i=1}^{n} 1_{X_i > u_n}.$$

N_{u_n} follows a binomial distribution with parameters n and $\mathbb{P}(X_i > u_n)$. Now a limit process can be derived by letting the sample size n tend to infinity and, simultaneously, increasing u_n in the correct proportion: If for some $\tau > 0, n\mathbb{P}(X_i > u_n) \longrightarrow \tau, n \longrightarrow \infty$. Then, by a classical theorem $N_{u_n} \xrightarrow{d} \mathbb{P}_o(\tau)$,

where $\mathbb{P}_o$: Poisson's variable. If $X_1, ..., X_n$, $i = 1, ..., n$ come from an absolutely continuous distribution, a suitable series u_n can be found for every $\tau > 0$. (See Embrechts *et al.* (1997), chapter 3). Indexing all points $\{i : X_i > u_n, i = 1, ..., n\}$ in the interval $[0, n]$, this interval will grow larger whereas the indexed points will become sparser and sparser as u_n increases with n.

The distribution of exceedances

We are not only interested in when and how often the exceedances occur, but also in how large the excess $X - u|X > u$ is. Consider the conditional CDF of the excess observations $X - u$,

$F_u(x) = \mathbb{P}(X - u|X > u)$. Or in terms of the underlying F as $F_u(x) = \frac{F(x+u)-F(u)}{1-F(u)}$.

An important result in EVT is that for a very large class of distributions, it can be shown that:

$\lim_{u \longrightarrow \infty} \underset{x \geq 0}{Sup} \left| F_u(x) - G_{\xi,\beta(u)}(x) \right| = 0$, where $G_{\xi,\beta}$ is the CDF of the Generalized Pareto Distribution (GPD) :

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{-1/\xi}, & \xi \neq 0 \\ 1 - e^{-x/\xi}, & \xi = 0 \end{cases}$$

If $\xi \geq 0$, the support of this distribution is $[0, \infty)$, for $\xi < 0$, the support is a compact interval. This distribution is generalized in the sense that it subsumes other distributions under a common parametric form. ξ is the important shape parameter. The case $\xi = 0$ corresponds to the exponential distribution.The case $\xi < 0$ is known as a Pareto II distribution. If $\xi > 0$, then $G_{\xi,\beta}(x)$ is a reparameterized version of the ordinary Pareto distribution, that has a long history in actuarial mathematics. This is because the GPD is heavy-tailed when $\xi > 0$. Whereas a normal distribution has finite moments of all orders, a heavy-tailed distribution does not, as already mentioned, possess a complete set of moments. In the case of a GPD with $\xi > 0$, it is found that $E\left(X^k\right) = \infty$ for $k \geq 1/\xi$. When $\xi = 1/2$, the GPD has an infinite second moment, variance. When $\xi = 1/4$, the GPD has an infinite fourth moment. Empirically, as mentioned in the section Empirical properties of financial time series, it is often found that our series have an infinite fourth moment. The normal distribution cannot model these phenomena, but the GPD can be used to capture exactly this behavior.

The summarise, the excess distribution converges to a GPD. We have not defined the very large class of distributions for which this is true, but for our purposes it is enough to say that this one holds for all the common parametric continuous distributions (such as normal, lognormal, χ^2, t, F, gamma...). Hence, the GPD is the natural model for the unknown excess distribution.

Parameter estimation

Finally, the parameters ξ and β must be estimated. A standard method is Maximum Likelihood (ML), where the joint PDF is maximized. However, in practice, this might be numerically troublesome if the data set is small, and one cannot rely on the asymptotic optimality properties of the ML-estimators. Recall, that only the excess fraction of the set is used, and the used data set depends of course on the choice of threshold u. In our case, we typically have a very large number of observations, which is considered as enough to estimate our parameters.

For the choice of threshold u, the mean excess functions a useful tool.

$$e(u) = E(X - u | X > u)$$

It can be estimated by the empirical function: $e(u) = \frac{1}{\#\{i : X_i > u_n, i = 1, \ldots, n\}} \sum_{i=1}^{n} (X_i - u)^+$.

For fat tails, $e(u)$ tends to infinity. For the GPD with $\xi > 0$ it can be shown that $e(u)$ is a linearly increasing function. Hence, a possible choice of u is given by the value for which $e(u)$ is approximately linear. In practice, this often gives values such that the GPD is a good model for, very roughly, half of the sample.

Parameter estimation

Now, these results can be used to estimate tails and quantiles. Denote the tail of F by $\overline{F} = 1 - F$ These yields

$\overline{F}_u(y) = \mathbb{P}\{X - u > y | X > u\} = \frac{\overline{F}_u(u+y)}{\overline{F}(u)}$ or $\overline{F}_u(u+y) = \overline{F}_u(y)\,\overline{F}(u), y \geq 0.$

Hence, an estimator of the tail $\overline{F}_u(y)$ (for values greater than u) can be obtained by estimating the tails $\overline{F}_u(y)$ and $\overline{F}(u).\overline{F}(u)$ can be estimated by its empirical counterpart.

$\overset{-}{F}^*(u) = \frac{1}{n}\sum_{i=1}^{n} I(X_i > u) = N_u/n$ and $\overline{F}_u(y)$ by the GPD, where the scaling function $\beta(u)$ has to be taken into account. This gives $\overset{-}{F}^*(y) \approx (1 + \xi^* y/\beta^*)^{-1/\xi^*}$.

The two parameters ξ and β have to be estimated, which is (theoretically) best done using ML.

Now, for a given u this gives the tail estimator: $\overset{-}{F}^*(u+y) \approx \frac{1}{n}\sum_{i=1}^{n}(1 + \xi^* y/\beta^*)^{-1/\xi^*}$.

For a given $q \in (0,1)$, this function can be inverted to give an estimator of the q-quantile:

$x_q^* = VaR_q^* = u + \frac{\beta^*}{\xi^*}\left(\left(\frac{n}{N_u}(1-q)^{\xi^*}\right) - 1\right).$

Hence, for a given probability $q > F(u)$, VaR estimate VaR_q^*.

4. The empirical results

4.1. Q-Q plot against the exponential distribution

Before applying the extreme value theory to investigate the behavior of the right tailed distributions it will be worthwhile assuming that the data are exponentially distributed. A commonly used methodology is to plot a quantile to quantile graphic against an exponential distribution.

The five plots presented in figure 2 show that a concave departure from the straight line in the QQ-plot. This behavior indicates a heavy tailed distribution of the claim sizes for the studied claim sizes series. The QQ-plot for a positive value of ξ is linear; we can deduce that the adequacy of the data to the Generalized Pareto law seems to agree. A small percentage of insured persons cause responsible accidents with

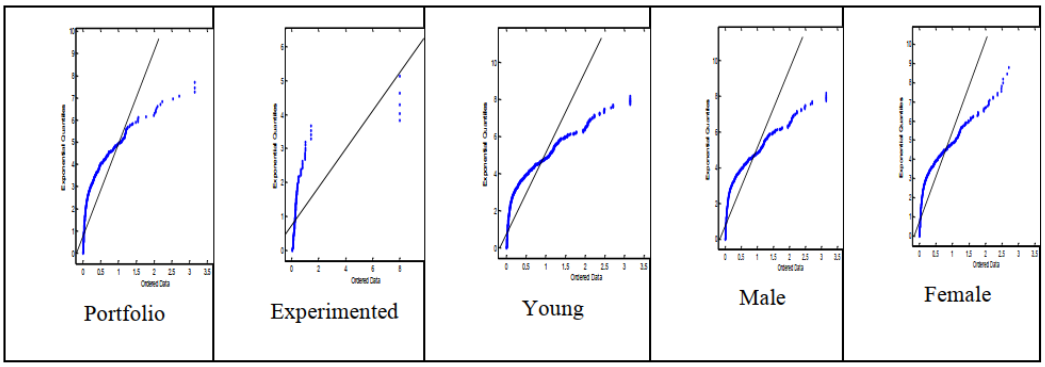

Fig. 2. QQ-plot of the claim sizes data against standard exponential quantiles

very high costs that the insurance company must bear and to which the insurance company applies special treatment.

4.2. Peak over threshold results

In this section, the estimation results of the GPD following the peak over threshold approach are presented. This investigation was conducted in order to compare the extremal behavior of the claim losses for the different studied groups of policyholders.

4.2.1. Threshold choice

As a first step for estimating the GPD, a threshold needs to be selected from the maximum insurance claim sizes. To select the optimal threshold an assessment of mean residual life plot a threshold Choice plot and a L-moment plot were carried out following Coles (2001), Fersi *et al.*(2011) Farah and Azevedo, (2017).

Mean residual life plot: To represent the mean residual life plot we use the theoretical mean of the GPD. For X a random variable distributed as Generalized Pareto with parameters μ, σ and ξ. Theoretically we have :

$E[X] = \mu + \frac{\sigma}{1-\xi}$ for $\xi < 1$.

Empirically, if X represents excess over a threshold μ_0, and if the approximation by a GPD is good enough, we obtain: $E[X - \mu_0 | X > \mu_0] = \frac{\sigma_{\mu_0}}{1-\xi}$. For all updated threshold μ_1 such as $\mu_1 > \mu_0$, excesses above the new threshold are also approximate by GPD after new parametrization. Thus,

$E[X - \mu_1 | X > \mu_1] = \frac{\sigma_{\mu_1}}{1-\xi} = \frac{\sigma_{\mu_0} - \xi\mu_1}{1-\xi}$.

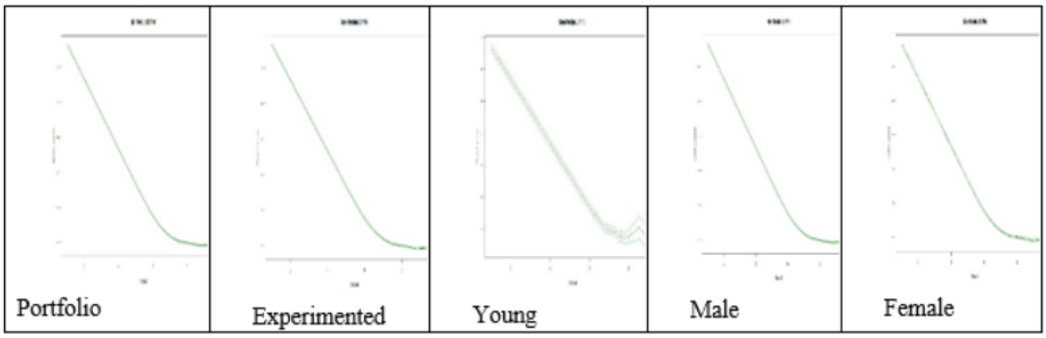

Fig. 3. The mean residual life plot for the portfolio and the different insurance classes

The quantity $E[X-\mu_1|X>\mu_1]$ is linear in μ_1 and corresponds to the mean of excesses above the threshold μ_1 which can easily be estimated using the empirical mean. Finally, a mean residual life plot consists in representing points :

$$\left\{\left(\mu, \frac{1}{n_\mu}\sum_{i=1}^{n_\mu} x_{i,n_\mu} - \mu\right), \mu \leq x_{\max}\right\}.$$

Where n_μ is the number of observations x above the threshold μ, x_{i,n_μ} is the i^{th} observation above the threshold μ et $x_{\max}$ is the maximum of the observations x. Graphically, a threshold has to be selected when the mean residual plot is practically linear and the modified scale and shape estimates become constant.

The figure 3 indicate that the mean residual life plot of the maximum loss sizes thresholds, for the entire portfolio, is linear starting from a threshold of 8, where the line becomes more stable until about 9.5. For the experienced insureds, the figure 3 indicates that the mean residual life plot is linear starting from 7.9 to 9.8. For the young drivers the mean residual life plot is linear starting from a threshold 8 and becomes more stable until about 9.5. Finally, for the males and females insureds, figure 3 indicate that the mean residual life plot is linear from thresholds, respectively, 7.8 and 7.4 and becomes more stable until about 9.8 for the both classes.

Threshold Choice plot

The threshold choice plot is constructed using a random variable X distributed as Generalized Pareto with parameters σ_0 and ξ_0. Let $\mu_1 > \mu_0$. The random variable $X|X > \mu_1$ is also GPD with parameters $\sigma_1 = \sigma_0 + \xi_0(\mu_1 - \mu_0)$ and $\xi_1 = \xi_0$. Let $\sigma_* = \sigma_1 + \xi_1\mu_1$, with this updated parameterization σ_* is independent of μ_1. Thus, estimates of σ_* and ξ_1 are constant for all $\mu_1 > \mu_0$ is a suitable threshold for the asymptotic approximation. Then, the threshold choice plots represent the points defined by: $\{(\mu_1, \sigma_*) : \mu_1 \leq x_{\max}\}$ where $x_{\max}$ is the maximum of the observations

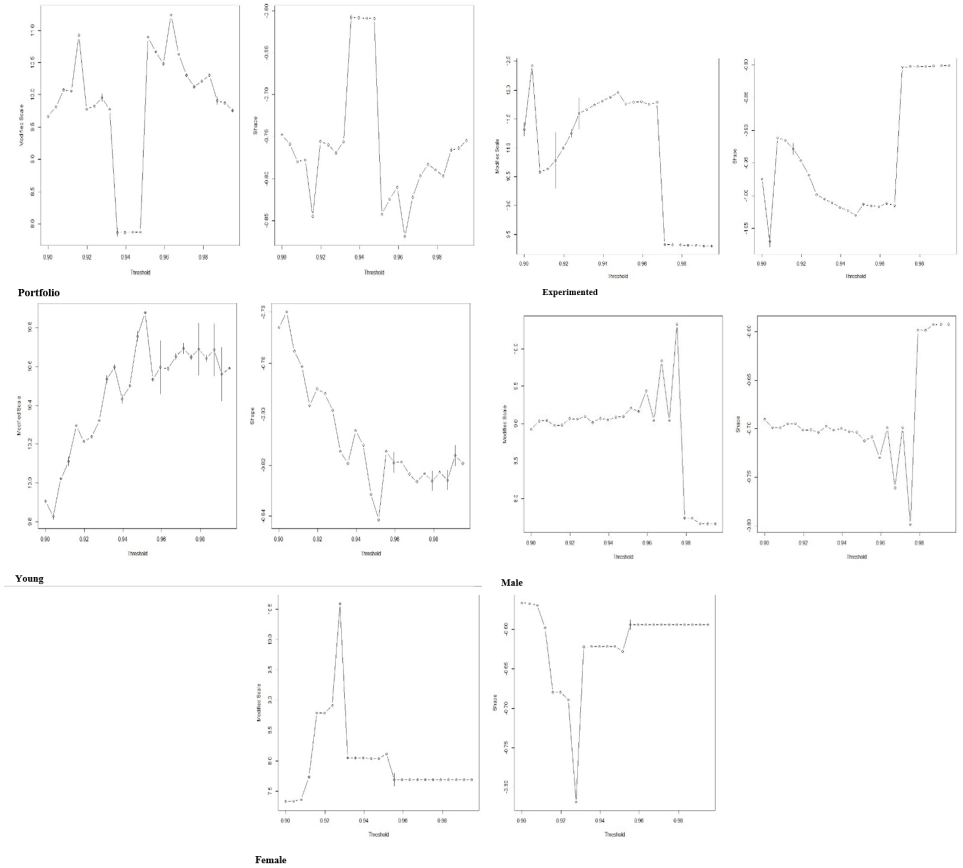

Fig. 4. Results of the threshold choice plot function

X.

Figure 4 displays results of the threshold choice plot function. We can see that thresholds can be contained in an interval with a minimum value of 0.90 and a maximum value of 0.98. For example, a threshold of 0.97 is a reasonable choice for males and a threshold of 0.95 is a reasonable choice for females. For the entire portfolio, a threshold of 0.95 would be a suitable choice. We must notice that in practice decision or threshold choices are not so clear-cut using the threshold choice plot function.

L-moments plot:

L-moments are the summary statistics for probability distributions and data samples. They provide measures of location, dispersion, skewness, kurtosis and other aspects of the shape of probability distributions or data sample. These measures are calculated from linear combinations of the ordered data values. For

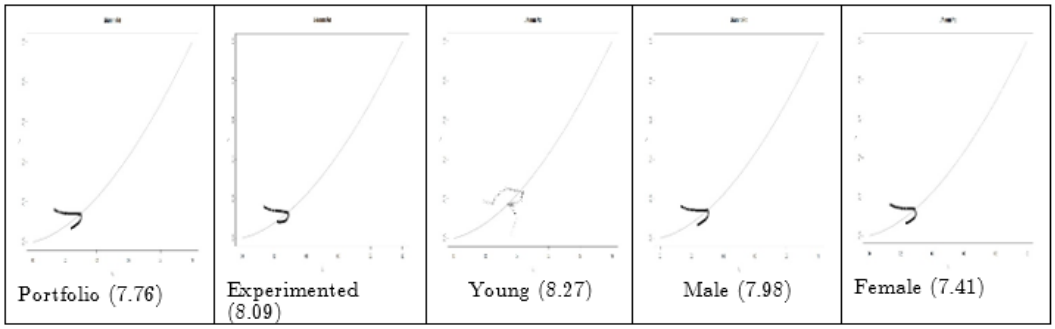

Fig. 5. Results of the L-moments plots

the Generalized Pareto distribution, the relation can be written as: $\tau_4 = \tau_3 \frac{1+5\tau_3}{5+\tau_3}$

where τ_4 is the L-Kurtosis and τ_3 is the L-Skewness. The L-Moment plot represents point defined by: $\{(\widehat{\tau}_{3,u}, \widehat{\tau}_{4,u}) : u \leq x_{\max}\}$.

Where $\widehat{\tau}_{3,u}$ and $\widehat{\tau}_{4,u}$ are estimations of the L-Kurtosis and L-Skewness based on excesses over threshold u and $x_{\max}$ is the maximum of the observations X.

The figure 5 displays the L-Moment plots. The L-moments plots provide the value of the observations that will be considered as a threshold. The detected values are in log scales. For example, 7.76 is the threshold detect for the entire portfolio. For the experimented policyholders, the detected threshold is about eight.

4.2.2. Fitting the GPD

After identifying a threshold above which the points are considered as extremes, we use a maximum likelihood method to estimate the two GPD parameters (ξ and σ). In this section the estimation results of the GPD implementing the POT approach are presented. This indicates that the modeling of the distribution of the claim size by a law limited from the right is more adequate. Our results corroborate with those obtained by Pisarenko and Rodkin(2010).

For the different threshold, the estimated shape parameters are negative $(\xi < 0)$, then the extreme claim sizes have a distribution on a bounded interval $[0, \beta/\xi]$. As highlighted previously, we select a threshold allowing for a stability of the estimated shape and scale parameter. The maximum stability is obtained for a threshold $u_3 = 10.5$. For various used thresholds, the two estimated parameters are statistically significant at 5% significance level. Since, the estimated shape parameter is stable and its value, for the different studied samples, $\widehat{\xi} > -0.5$ the estimators from ML are reliable Smith, (1985).

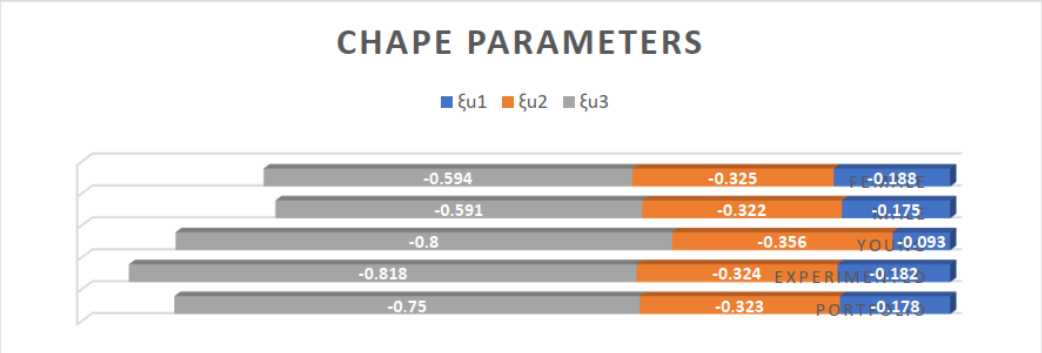

Fig. 6. The estimated shape parameters for the different studied insurance groups

The shape parameter estimates for the insurance claim sizes reveal some interesting facts. To compare between the different insurance groups it will be advisable to plot the different estimated shape parameters.

We recall that the larger the value of the shape parameter is, the larger the number of extreme events is. Hence, the tails of the claim size distribution become fatter when the shape parameter increases. In this prospect, the figure 6 shows that the shape parameter for the experimented insureds is the smallest among all the remaining groups. This result is confirmed regardless of the selected threshold. Results also indicate that the class of experimented insureds has lower number of extreme claim sizes than that for young policyholders. Moreover, results reveal that the number of extreme events are greater for male class than for female group of insureds.

Table 2 shows that the scale parameter is statistically significant at 5% significance level for all the studied distributions. It is well known that the scale parameter is related to the volatility. Our results show that the volatility of the extreme claim sizes increases with the experimented and females policyholders. The lowest volatility is highlighted for the young drivers.

Finally, results show that the interval estimates of the scale parameter are reliable since the range between the lower and the upper interval is very low for all the studied distributions.

The comparison between the studied classes is essential for the insurer and actuary since it allows a better assessment of the pure premium that will be played by the insureds. The usual method of calculating the pure premium is to multiply the expected value of the loss amounts by the expected value of the number of accident. However, the expected value or the average is very sensitive to the presence of extreme values.

	u_1=8.5				u_2=9.5				u_3=10.5			
	Scale	Shape	LCL (scale)	UCL (scale)	Scale	Shape	LCL (scale)	UCL (scale)	Scale	Shape	LCL (scale)	UCL (scale)
Portfolio	1.386 (0.009)	-0.178 (0.004)	1.370	1.402	2.258 (0.009)	-0.323 (0.001)	2.243	2.274	8.951 ($^{2.1e-06}$)	-0.750 ($^{2.1e-06}$)	8.951	8.951
Experimented	1.409 (0.024)	-0.182 (0.011)	1.369	1.449	2.280 (0.026)	-0.324 (0.004)	2.237	2.323	9.761 ($^{5.8e-03}$)	-0.818 ($^{2.0e-06}$)	9.752	9.771
Young	1.047 (0.095)	-0.093 (0.057)	0.889	1.204	2.178 (0.151)	-0.356 (0.037)	1.929	2.426	8.493 ($^{9.1e-03}$)	-0.800 ($^{2.0e-06}$)	8.478	8.508
Male	1.380 (0.010)	-0.175 (0.004)	1.363	1.397	2.257 (0.010)	-0.322 (0.001)	2.240	2.275	7.067 ($^{3.6e-03}$)	-0.591 ($^{2.1e-06}$)	7.061	7.072
Female	1.414 (0.017)	-0.188 (0.007)	1.386	1.442	2.265 ($^{5.2e-03}$)	-0.325 ($^{2.0e-06}$)	2.256	2.273	7.088 ($^{2.1e-06}$)	-0.594 ($^{2.0e-06}$)	7.088	7.088

Note. The table displays the scale and the shape of the GPD with various thresholds (8.5, 9.9 and 10.5). LCL and UCL are the lower and upper limits for the interval estimate of the scale parameter. Values between parentheses are the standard deviation of the estimates.

Table 2. Estimation results for the best model the POT approach (u1=8.5, u2=9.5 and u3=10.5)

Thus, the number and the amount of the detected extreme value can be used to adjust the pure premium by class of insureds.

Recall that the main objective of modeling the claim amounts is to estimate the coverage amount for claims. This means to determine a proper estimate of capital that allows insurers to be solvent in future years. Indeed, an insurance company needs to maintain a rational capital adequacy to avoid its insolvency probability. To address this need, the solvency 2 framework requires the use of Value at Risk (VaR) measure at a high level to enhance insured protection. To reach this aim, a useful application will be using the suitable distribution to calculate a value at risk by class and for the entire portfolio. Before calculating the VaR we use deviance and the AIC criteria to select the suitable model for our datasets. The VaR will be calculated only for the entire portfolio as recommended by the capital requirement in solvency II.

The results reported in table 3 show that for all the studied samples the suitable model is a GPD with a threshold equals to 8.5.

In the insurance sector, the main element of sinistrality is described in the database by the amounts of compensation of responsible accidents declared by the insured. As announced in the previous chapters of our research, the skewness and the kurtosis are two very important parameters in modeling the tail of the distribution.

	Threshold	Portfolio	Experienced	Young	Male	Female
Deviance	8.5	77653.78	12885.94	375.4207	65667.7	26647.06
AIC		77657.78	12889.94	379.4207	65671.7	26651.06
Rank		1/3	1/3	1/3	1/3	1/3
Deviance	9.5	135779.4	22379.4	725.1317	113940.3	46535.22
AIC		135783.4	22383.4	729.1317	114944.3	46539.22
Rank		2/3	2/3	2/3	2/3	2/3
Deviance	10.5	237173.5	38878.07	1226.226	206231.1	83427.78
AIC		237177.5	38892.07	1230.226	206235.1	83431.78
Rank		3/3	3/3	3/3	3/3	3/3

Table 3. Selecting the best model

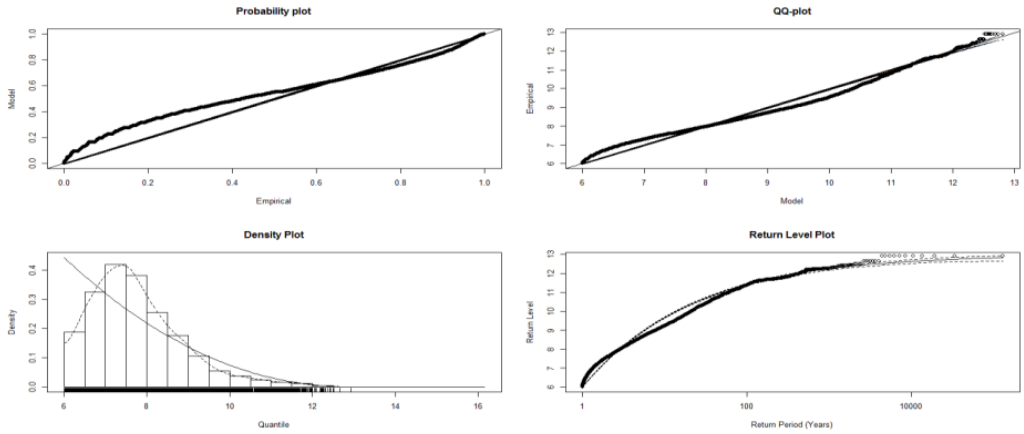

Fig. 7. Diagnostic plots of the fitted generalized Pareto distribution, threshold=8.5, for the portfolio.

4.2.3. Model diagnostic for the portfolio data

After fitting extreme value model to data, next is to assess and interpret the fitted model based on the quantile and return levels computed using the inverse of the distribution function. We calculate and graphically present the return levels the annual scale, so that the 1-year return level is the level that is expected to be exceeded once in every one year. These parameters give example of very valuable information on the tail of distribution.

We use the function implemented in the software R; the result of function gives the name of the estimator, if a varying threshold was used, the threshold value, the number and the proportion of observations above the threshold, parameter estimates, standard error estimates and type, the asymptotic variance-covariance matrix and convergence diagnostic. We assess the adequacy and the validity of the

fitted generalized Pareto distribution by analyzing the diagnostic plots presented in figure 7. Both the probability plot and the quantile plot indicate a reasonable extreme value fit because, the probability plot is almost linear and the quantile plot is practically linear.

The return level plot highlights a slight convexity. Moreover, the return level plot indicates that the extreme values are within the 95% confidence limits and the density plot adequately fit the upper right tail of the distribution. We finally conclude that the diagnostic plots do not raise any problem on the adequacy and the validity of the generalized Pareto fitting.

We also performed the diagnostic to the remaining studied datasets; the results are similar to those obtained for the whole portfolio. These results have important implications. First, having a distribution that fits well the large claim sizes will help the insurer in predicting the possible extreme event and the timing of its occurrence. Secondly, the accurate estimates of the scale and the shape of the large claim sizes will help in calculation of the value at risk and thus a correct estimation of the rational capital adequacy required by the Solvency 2.

5. Discussion

In this paper, we highlighted the importance of the use of the Peak over threshold in detecting the extreme accident severities. Moreover, this approach helped in comparing between the probability of occurrence of normal and extreme accident severity of several driver's groups. The finding of the empirical assessment of the extreme accident severities that we conducted are mostly stimulating and provides valuable information for road safety policy makers insurance companies as well.

First, our study contributes to develop strategies to improve china's road safety conditions. Indeed, it is well established that for the success of a road safety strategy, realistic quantified road safety targets should be set. In this prospect, our findings contribute to provide accurate quantitative measures that can be used to boost the road safety. Actually, the selection of suitable threshold, beyond which an accident is deliberated as extreme, provides a useful measure allowing separation between low-medium accident severity and extreme ones. This threshold can be considered as a bound between two clusters of accidents that have to be managed differently. Thus, road safety policymakers have to allocate resources and develop various appropriate safety plans according to the nature of severities (Low-medium or extreme). Specifically, the selected limit for the entire dataset exceeds 8.5, whereas the limit for young and experimented drivers are 8 and 7.9, respectively. The similar findings are stressed by gender with 8.5 for the entire dataset versus 7.8 for male and 8.5 the entire dataset versus 7.6 for female. For instance, resources and strategies engaged to improve road safety have to be allocated differently among young and experimented drivers taking into account the difference of probability of occurrence of extreme accidents.

Moreover, our finding provides an extreme value distribution by group of drivers. These distributions provide accurate prediction of extreme accident severity amounts. These predictions can be also exploited in outlining the strategy that should be engaged to reduce the number and the amounts of the extreme accidents.

Further, our results have important implications for the insurance market policymakers. Essentially, threshold estimation helps to identify the retention limits and achieve optimal reinsurance levels. The findings indicate that the retention limits vary among the studied groups and between the classes constructed based on different accident factors. More specifically, considering the experience in driving provides higher retention limit than that when allowing for gender as discrimination factor. Moreover, while the difference between subgroups is not significantly important, the difference between the retention limits of the entire insurance dataset and the different subgroups is suggestively high. Specifically, the selected retention limit of the entire dataset exceeds 8.5, whereas the retention limit for young and experimented drivers are 8 and 7.9, respectively. The similar findings are stressed by gender with 8.5 for the entire dataset versus 7.8 for male and 8.5 the entire dataset versus 7.6 for female.

Overall, these results provide to the insurance company various strategies in selecting the appropriate retention levels and the reinsurance structure as well. Specifically, changing the retention levels based on the entire data results or performing a discriminating factor (experience or age) will modify the risk profile and risk-based target capital for the insurer.

6. Conclusion

The aim of this work was to study the behavior of the extreme car insurance claim sizes by using the peaksover-threshold method. A particular interest was given to the choice of the threshold to perform such analysis. Indeed, the choice of threshold affects the estimation of the GPD parameters. Our studies also aim to compare the extreme distribution for different risk classes. The following findings can be pointed out. There is no large gap in the selected threshold among the studied classes.

References

karamata (1030) Karamata, J.(1930) Sur un mode de croissance régulière des fonctions. *Mathematica (Cluj)*, 4, 38-53

Resnick (1987) Resnick, S.I. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer-Verbag, New-York. (MR0900810)

de Haan (1970) de Haan, L. (1970). *On regular variation and its application to the weak convergence of sample extremes*. Mathematical Centre Tracts, **32**, Amsterdam. (MR0286156)

Galambos (1985) Galambos, J. (1985). The Asymptotic theory of Extreme Order Statistics. Wiley, New-York. (MR0489334)

de Haan and Feireira (2006) de Haan, L. and Feireira A. (2006). *Extreme value theory: An introduction. Springer*. (MR2234156)

Rolski *et al.*(1999) Rolski, T., Schmidt, V., and Teugels, J.(1999). *Stochastic Processes for Insurance and Finance*. John Wiley & Sons.

Omey(2006) Omey, A.M., 2006. Subexponential distribution functions. *Journal of Mathematical Sciences*, 138(1), 5434-5449.

Lu and Bin Zhang(2016) Lu, D. and Bin Zhang, B., 2016. Some asymptotic results of the ruin probabilities in a two-dimensional renewal risk model with some strongly subexponential claims. *Statistics & Probability Letters*, 114, 20-29.

Goldie and Resnick (1988) Goldie, C.M. and Resnick, S., 1988. Distributions that are both subexponential and in the domain of attraction of an extreme-value distribution. *Advances in applied probability*, 20(4), 706-718.

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

CICR (2017) CICR.(2017)*China insurance statistics report 2016*. Available at http://www.circ.gov.cn/web/site0/tab5257/2017.

Lozano-Perez (2012) Lozano-Perez, T.(2012). *Autonomous Robot Vehicles*. Springer Science & Business Media.

Abraham *et al.*(2016) Abraham, H., Lee, C., and Brady, S.(2016). *Autonomous Vehicles, Trust, and Driving Alternatives: A Survey of Consumer Preferences*. AgeLab, Massachusetts Institute of Technology.

Mao(2017) Mao S.(2017). *Vehicles import market 2016 in China Consumption Daily*. Beijing.

Xian and Chiang-Ku (2018) Xian, X. and Chiang-Ku, F.(2018). *Autonomous vehicles, risk perceptions and insurance demand: An individual survey in China*. Elsevier.

Embrechts *et al.* (1997) Embrechts, P., Klűppelberg, C., and Mikosch, T.(1997). *Modelling Extremal Events for Insurance and Finance*. Berlin, Springer, 1997.

Coles (2001) Coles, S.G.(2001). *An Introduction to Statistical Modeling of Extreme Values*.Springer Verlag, New York.

Fersi *et al.*(2011) Fersi, K., Boukhetala, K., and Ammou, S.B.(2011). *Stratégie optimale de réduction de l'intervalle de confiance pour l'estimateur de la prime ajustée. Application en assurance automobile*. Hal.

Farah and Azevedo, (2017) Farah, H. and Azevedo, C.L.(2017). Safety analysis of passing maneuvers using extreme value theory.*IATSS Research*, 41, 12-21.

Pisarenko and Rodkin(2010) Pisarenko, V.F. and Rodkin, M.V.(2010). Estimation of the Probability of Strongest Seismic Disasters Based on the Extreme Value Theory. *Izvestiya, Physics of the Solid Earth*. Vol 50, pp. 311–324 (2014)

Smith, (1985) Smith, R.L.(1985). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). *Statistical Science*, 4, 367-393.

Lo, (2017) Lo, G.S.(2017). *Weak Convergence (IIA) - Functional and Random Aspects of the Univariate Extreme Value Theory*, Spas Textbooks Series. Arxiv :DOI : http://dx.doi.org/10.16929/srm/2016.0009.

Lo (1986) Lô, G.S. (1986). *Sur quelques estimateurs de l'Index d'une loi de Pareto : Estimateur de Hill, de S.Csörgő- Deheuvels-Mason, de de Haan-Resnick et loi limites de sommes de valeurs extrêmes pour une variable aléatoire dans le domaine d'attraction de Gumbel*. Thèse de doctorat. Université Paris VI.

Gumbel (1955) Gumbel, E.J. (1955) *statistical estimation of the endurance limit*,Tchnical report T-3A, Departement of Engineering, Columbia Univ. press.,New-York.

Fisher and Tippet (1928) Fisher, R. and Tippet, L. *Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample* Proceedings of the Cambridge Philosophical Society, 1928, 24, 180-190

Lo et al. (2018) Lo G.S., K. T. A. Ngom M. and Diallo M.(2018). Weak Convergence (IIA) - Functional and Random Aspects of the Univariate Extreme Value Theory. Arxiv : 1810.01625

Loève (1997) Loève,M.,(1997).*Probability theory. Tome 1.* Springer-verlag, 4th Edition.

Feller (1968) Feller W.(1968) *An introduction to Probability Theory and its Applications. Volume 2.* Third Editions. John Wiley & Sons Inc., New-York.

Beirlant *et al.* (2004) Beirlant, J., Goegebeur, Y. Teugels, J.(2004). *Statistics of Extremes Theory and Applications*. Wiley. (MR2108013)

Ba *et al.* (2004) Ba A.D.,Deme E.H, Seck C.T. and Lo G.S.(2016). Consistency Bands for the Mean Excess Function and Application to Graphical Goodness-of-fit Test for Financial Data. *Journal of Mathematical research*, Vol. 8, (1). http://dx.doi.org/10.5539/jmr.v8n1p42, pp. 42-64