Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 03, pp.28708-28712, March, 2016

INTERNATIONAL JOURNAL
OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

AN APPLICATION OF EXTREME VALUE THEORY IN AUTOMOBILE INSURANCE: A NEW APPROACH TO THE CONVEX COMBINATION OF TWO VARIABLES THRESHOLDS MINIMIZING THE VARIANCE OF THIS COMBINATION

*Daouda Diawara

Zhongnan University of Economics and Law Wuhan 182# Nanhu Avenue,
East Lake High-tech Development Zone, Wuhan 430073

ARTICLE INFO

Article History:
Received 25th December, 2015
Received in revised form
29th January, 2016
Accepted 10th February, 2016
Published online 31st March, 2016

Key words:

Literature,
Insurance,
Numerical and graphical methods.

ABSTRACT

In the literature many determinists approaches (numerical and graphical methods), probabilists (the probability law, extreme value theory, Bayesian methods) exist for the detection of extreme claims. In this paper after a reminder of the extreme value theory and its application in the simulated data of a Malian mutual insurance, a contribution of the convex combination method of two threshold variables minimizing the variance is proposed.

Copyright © 2016, Daouda Diawara. *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Citation: Daouda Diawara, 2016. "An application of extreme value theory in automobile insurance: A new approach to the convex combination of two variables thresholds minimizing the variance of this combination", *International Journal of Current Research*, 8, (03), 28708-28712.

INTRODUCTION

In car insurance, classes formed from characteristics of the insured and the vehicle are assumed to be homogeneous in terms of claims. The presences of extreme events disturb the homogeneity of the portfolio and their detection is essential to avoid errors in pricing and interpretation that may have the impact of tariff changes useful or useless opposite. It is very important to know the severity of a disaster in a rate case to ensure stability of damage indicators and thus a match between the reference premium and loss experience. In our article we will firstly make an overview of the extreme value theory. Using this theory to draw the line between ordinary claims and serious claims. One note that this estimate is accurate the threshold should be well chosen. Finally; we propose a contribution from the convex combination of two thresholds method that minimizes the variance, two thresholds obtained from the extreme value theory. Our method is a comparison between the variance of a convex combination of two thresholds and the variance of a threshold. So the reasonable threshold will be one that has the smallest variance.

Law of extreme statistics

In this part we are interested in the limit distributions of order statistics when $n \to +\infty$. Therefore, let us consider random variables $X = (X_1, X_2, ..., X_n)$ with distribution function $F(x) = P(X \le x)$ and the density function f. Let $M_n = \max(X_1, X_2, ..., X_n)$ and $m_n = \min(X_1, X_2, ..., X_n)$. The value $d_n = M_n - m_n$ is called extreme deviation. The law that follows M_n and m_n are:

$$F_{M_n}(x) = P(M_n \le x) = [F_X(x)]^n$$
$$F_{m_n}(x) = 1 - (1 - F_X(x))^n$$

We can therefore conclude that M_n is a random variable distribution function F^n .

$$\begin{cases} F_{M_n}(x) = F^n(x) \\ f_{M_n}(x) = n\, F^{n-1}(x) f(x) \end{cases}$$

So if the distribution function is known then we can find a simple way of maximum

Corresponding author: Daouda Diawara,
Zhongnan University of Economics and Law Wuhan 182# Nanhu Avenue, East Lake High-tech Development Zone, Wuhan 430073.

But the distribution function F of X is generally unknown therefore impossible to determine the distribution of the maximum M_n from this result. So you have to be interested in the asymptotic distribution of the maximum. So you have to be interested in the asymptotic distribution of the maximum by making n tend to infinity. We have:

$$\lim_{x\to\infty} F_{M_n}(x) = \lim_{x\to\infty}\left(F_X(x)\right)^n = \begin{cases} 0 & if \quad F(x)<1 \\ 1 & if \quad F(x)=1 \end{cases}$$

At infinity extreme law $F_{M_n} \in \{0,1\}$, therefore it is called degenerate. This causes a lack of information on the extreme.

Asymptotic distribution of the maximum

The distribution function of maximum obtained at infinity leads to degenerate law, we look for a non-degenerate law for maximum .This non-degenerate limit law is provided by the following theorem (Theorem of Gnedenko, 1943).

Theorem 1

Whether $X_1, X_2, ..., X_n$ is a sequence of random variables i.i.d with distribution function F and $M_n = \max(X_1, X_2, ..., X_n)$. If a sequence $(a_n)_{n\geq 1} > 0$ of positive terms exists, a real sequence $(b_n)_{n\geq 1}$ and a non-degenerate distribution function G, so that:

$$\lim_{x\to\infty} P\left(\frac{M_n - b_n}{a_n}\right) = \lim_{x\to\infty} F^n(a_n x + b_n) = G(x)$$

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = P(M_n \leq a_n x + b_n)$$
$$= F^n(a_n x + b_n) \xrightarrow{d} G(x)$$

Then the only possible forms of G are the Gumbel, Fréchet or Weibull distributions, also called type I, II and III distributions respectively. The variable $\alpha_n = \frac{M_n - b_n}{a_n}$ is called normalized maximum. Jenkinson (1955) which gives a unique general form of limit laws: This family can be described by a single term to three parameters:

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

Here $x \in □$, $\mu \in □$ and $\sigma > 0$. The case $\xi > 0$ corresponds to the Fréchet distribution, $\xi < 0$ corresponds to the Weibull distribution and $\xi = 0$ corresponds to the Gumbel distribution (see 1, 2, 5, 10).

Conditional excess distribution

The second part of the extreme value theory called POT (Peaks Over Threshold) is to choose an appropriate threshold and use observations that exceed this threshold, called excess.

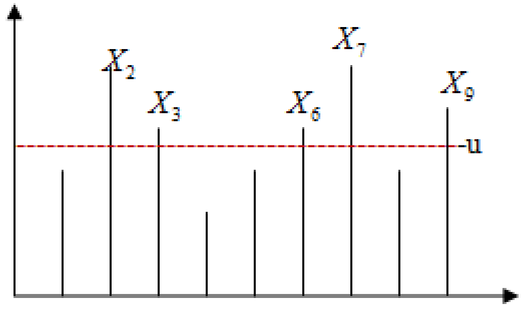


Fig. 1. Threshold

From Figure 1 we see that u is the threshold and the numbers X_2, X_3, X_6, X_7 and X_9 are in excess of the threshold. The random variables X_2, X_3, X_6, X_7 and X_9 are extreme values. The Insurance companies have a tradition of using a threshold to differentiate the two types of claims. Suppose that $X_1, X_2, ..., X_n$ are n independent realizations of a random variable X with a distribution function $F(x)$. Let $x_F = \sup\{x : F(x) < 1\}$ be the finite or infinite right endpoint of the distribution F. The distribution function of the excesses over certain (high) threshold u (F_u) is given by:

$$F_u(y) = P(X - u \leq y / X > u) = \begin{cases} \dfrac{F(y+u) - F(u)}{1 - F(u)} & \text{if } y \geq 0 \\ 0 & \text{if } \ y < 0 \end{cases}$$

The purpose of the POT method is defined by what probability distribution can be approached .Balkema and de Haan (1974), Pickands (1975) proposed a theorem states that the conditional distribution of excess (see 5,10).

Theorem 2

The Pickands-Balkema-de Haan theorem (Balkema & de Haan 1974; Pickands 1975) states that if the distribution function $F \in DA(G_\xi)$ then $\exists$ a positive measurable function $\sigma(u)$ such that:

$$\lim_{u\to x_F} \sup_{y\in[0, x_F - u]} \left|F_u(y) - G_{\xi,\sigma(u)}(y)\right| = 0$$

and vice versa, where $G_{\xi,\sigma(u)}(y)$ denote the Generalized Pareto distribution. The Generalized Pareto Distribution (GPD) is given:

$$G_{\xi,\sigma(u)}(y) = \begin{cases} 1 - \left(1 + \xi\dfrac{y}{\sigma(u)}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - \exp\left(-\dfrac{y}{\sigma(u)}\right) & \text{if } \xi = 0 \end{cases}$$

With $y \geq 0$ for $\xi > 0$ *and* $0 \leq y \leq -\dfrac{\sigma(u)}{\xi}$ for $\xi < 0$. In other words well chosen for a threshold, the excesses of law can be approximated by a GPD whose extreme index is the same as that of the GEV law. The main difficulties of this model are the choice of the threshold and the estimation method parameters (see 5).

Threshold Selection

The extreme value theory provides different methods for estimating a threshold above observation will be considered extreme. If one chooses a low threshold, certain non-outliers will be declared as extreme and involve an under the pure premium. If the threshold should be big but not too high to have enough data beyond this threshold (enough data to a good estimate of the model). A threshold selecting tools is the graph of the sample mean excess function $e_n(u)$ (ME-plot).

Definition: The ME-plot is defined as follows:

$\{(u;e_n(u));X_1<u<X_n\}$ X_1 and X_n are respectively the minimum and the maximum of the sample is given by the formula:

$$e_n(u)=\frac{\sum_{i=1}^{n}(X_i-u)^+}{\sum_{i=1}^{n}\left(I_{(X_i>u)}\right)}=\frac{1}{N_u}\sum_{i=1}^{n}(X_i-u)^+$$

In other words the sum of the excess over the threshold u divided by the number N_u of data that exceeds .The sample mean excess function $e_n(u)$ is an empirical estimate of the mean excess function: $e(u)=E(X-u\mid X>u)$.

The mean excess function of GPD is:

$$e(u)=\frac{\sigma_{(u)}+\xi u}{1-\xi}$$

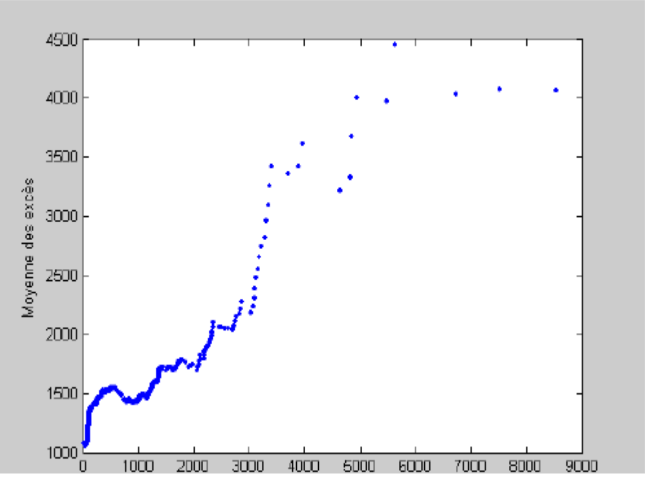


Fig. 2. Mean Excess Function of GPD

This is a linear function at u .Then the choice of threshold corresponds to the beginning of the linearity observed on the graph defined by $(u,e(u))$. The threshold u is determined from the time when the graph of the function has an affine part unchanged (see 2,3 and 10). From figure 2 we have an representation of the empirical mean excess function of a GPD distribution with a positive parameter ξ. It is found that this feature is stable for a threshold of the order of 5500. So we can choose as a threshold beyond which any observation (Loss) is considered extreme.

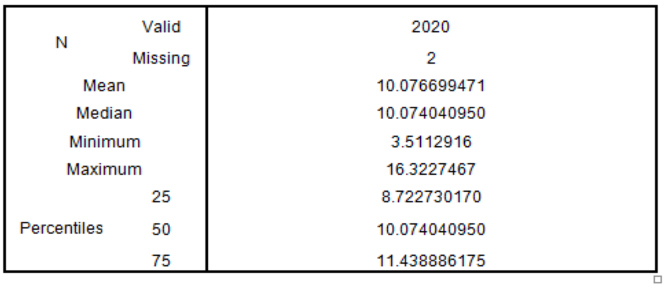
Here the mean excess function is stable of a threshold of the order of 5 500. So we can choose as a threshold $u=5500$ beyond which any observation (Loss) is considered extreme (see 1, 4 and 5).

Numerical applications

The data base provides a sample of 2020 observations for 4 wheel vehicle for personal use during the year 2013.Les data come from a Malian insurance company and concern the amounts of claims caused by the insured of a risk class. This file contains only the amounts of claims during the insurance year. Risk boxes are constructed from the vehicle characteristics and other variables. For confidentiality reasons, the company would not give us the other variables.

Statistical analysis of the data

Table 1. Statistics of data



N	Valid	2020
	Missing	2
Mean		10.076699471
Median		10.074040950
Minimum		3.5112916
Maximum		16.3227467
Percentiles	25	8.722730170
	50	10.074040950
	75	11.438886175

The Table 1 shows that the amounts of claims are increasing on average with the number of disaster.

Detection of extreme value

We will initially use the boxplot which allows a simple reading to detect the presence of extreme values. This figure shows the upper limit and lower limit of the simple boxplot.

If our data do not contain extreme values then all data will be between the upper limit and lower limit, this is not the case of Figure 3, therefore we can clearly see the presence of extreme values. After having detected the presence of extreme values, the average excess function allows to predict the approximate threshold (beginning of linearity).

The mean excess function of all data plotted below (Figure 4) gives a threshold beyond which a claim can be considered serious. This choice must satisfy criteria: keeping a large enough data up for a good estimate of the model.

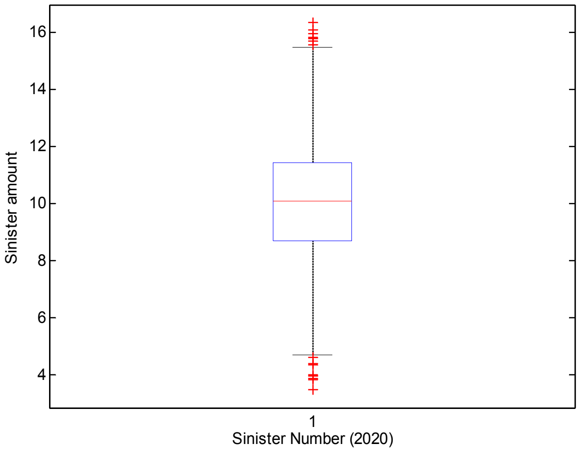


Fig. 3. Boxplot of data

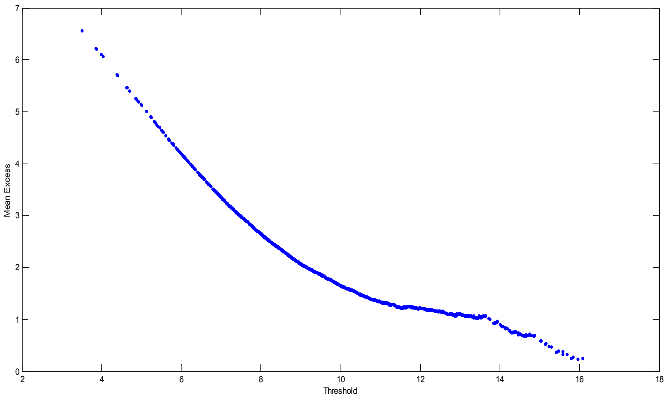


Fig. 4. Mean Excess Function of Data

From this figure we can retain as a threshold value 12. There is a stability curve from this value (see 2 and 10). Plotting the Mean Residual Life helps support the choice of threshold (Figure 5)

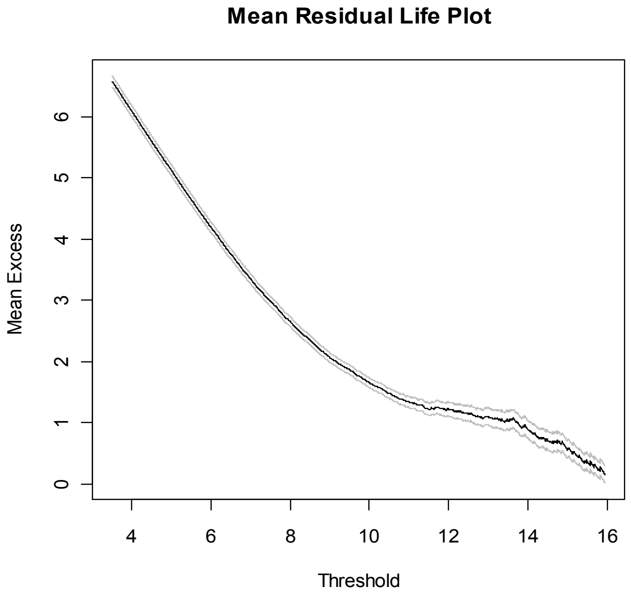


Fig. 5. Mean Residual Life Plot

Looking at the plot (Figure 5), 12 is a reasonable selection for the threshold The choice of this threshold gives the following statistical table:

Table 2. Statistic of data after the choice of threshold

Threshold Call	12
Number Above	323
Proportion Above	0.1599

The statistical data table after choosing a threshold gives 323 observations above (approximately 16% the total of all observations) the threshold. Given the total number of sample observations we can say that the observations beyond the threshold are important to a good approximation of the model. So GPD can be adjusted by the maximum likelihood method from 323 values exceeding the threshold (see 8, 9 and10).

Adjusting the GPD

After choosing our threshold, we create post-calculation diagnostic plots where we desire linearity amongst the fulfillment of other criteria to check the Threshold's suitability. GPD is adjusted by the maximum likelihood method from the 323 values exceeding the threshold $u = 12$, among the 2020 sample values.

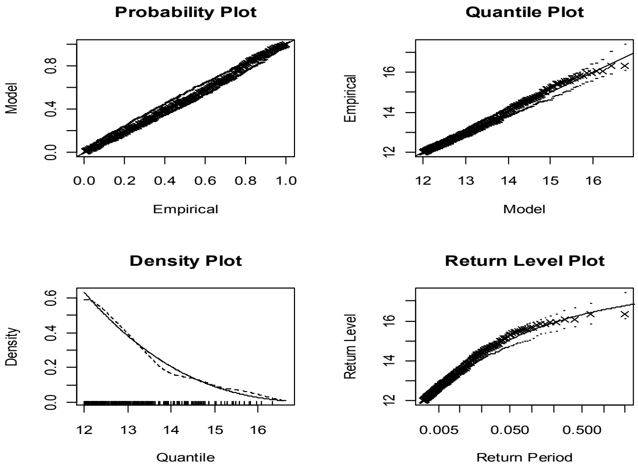


Fig. 6. Post threshold calculation diagnostic plots

We observe that the graph quantile plot is linear, it can be concluded that the excess sample is adequate and approaches a model (GPD) (see 4, 6, and 7). An appropriate threshold is essential for the reliability of the excess sum model in this example:

Table 3. Results of the estimate by the likelihood method of parameters

Varying Threshold	:	FALSE
Threshold Call	:	12
Number Above	:	323
Proportion Above	:	0.1599
Estimates		
scale		shape
1.5805		-0.2762
Standard Error Type:		observed
Standard Errors		
scale		shape
0.12529		0.05852

This table shows that-linearity is characterized by a negative slope so the data belong to MDA (Weibull).

New threshold selection approach

Our method provides a comparison between the variance of a single threshold and that of a convex combination. Acceptable threshold is the threshold which will have a smaller variance. In this approach, which allows to obtain a reference threshold U, the random variables correspond to the thresholds. But in this approach, the variances are not explicitly known, it is necessary to estimate the variances.

Consequence: Let U_1 and U_2 be two random variables. We note V_1 and V_2 the variance of U_1 and U_2 respectively, V_{12} the covariance of variables U_1 and U_2.Let U the random variable defined as convex combination of U_1 and U_2: $U=\alpha_1 U_1+\alpha_2 U_2$ With $0<\alpha_i<1, i=1,2$ and $\alpha_1+\alpha_2=1$ The variance of U is minimal for $\alpha_1=\frac{V_2-V_1}{V_1+V_2-2V_{12}}$ and $\alpha_2=1-\alpha_1$

Theorem 1

Let U_1 the threshold estimated by the mean excess function method, U_2 the threshold estimated by GPD method. We note V_1 and V_2 the variance of U_1 and U_2 respectively, V_{12} the covariance of variables U_1 and U_2. Let U the random variable defined as convex combination of U_1 and U_2:

$U=\alpha U_1+\alpha U_2$ With $0<\alpha<1$.

1.The variance of U is minimal for $\alpha=\frac{V_2-V_1}{V_1+V_2-2V_{12}}$.

2.The GPD method offers minimal strategy and MEF method offers maximum strategy if and only if $0<\alpha<\frac{1}{2}$ and $V_1>V_2$.

Theorem 2: Let U_1 the threshold estimated by the mean excess function method, U_2 the threshold estimated by GPD method. We note V_1 and V_2 the variance of U_1 and U_2 respectively, V_{12} the covariance of variables U_1 and U_2.

Let U the random variable defined as convex combination of U_1 and U_2:

$U=\alpha U_1+\alpha U_2$ With $0<\alpha<1$.

1.The variance of U is minimal for $\alpha=\frac{V_1-V_{12}}{V_1+V_2-2V_{12}}$.

2.The GPD method offers minimal strategy and FME method offers maximum strategy if and only if $\frac{1}{2}<\alpha<1$ and $V_1>V_2$.

Reinsurance excess of loss covers the portion of each individual claim excess a given priority, limited to a capacity granted by the reinsurer. We place ourselves in the collective risk model. Let N be the number of claims and $X_1, X_2, ..., X_N$ the realizations of X, which is the random variable representing the amounts of loss. As usual we assume mutual independence of random variables. It defines the excess beyond the threshold U or deductible (see Theorem 1, 2) as the set of random variables Y such that: $Y_j=X_j-U$, $X_j>U$.

Lemma

Let C the coverage (capacity) offered by the reinsurer. Then, the portion of each $X_j>U$ dependent the reinsurer is:

$R_j=\min(C,Y_j)=\min(C,X_j-U)$

Conclusion and Recommandation

The method of convex combination minimizing the variance of two thresholds variables may be a very valuable tool for modeling extreme claims in automobile insurance. It is applied after the selection of a number of different thresholds by the methods seen in the literature. Many critics have been formulated in the literature with respect to different threshold selection methods. Our technique, based on a reduction of the variance of the convex combination of two random variables thresholds, even if it seems like a good empirical compromise quality, it must approach the experts in the field for future validation.

REFERENCES

A User's Guide to the POT Package (Version 1.4),Mathieu Ribatet Noureddine Benlagha Michel Grun-Réhomme et Olga Vasechko Université Paris 2, ERMES-UMR7181-CNRS, 12 place du Panthéon, 75005 Paris, France.

Alexander, J., McNeil, Saladin, T. 1997.The peaks over thersholds for estimating high quantiles of loss distribution International ASTIN Colloquium, pp 70-94.

De Haan L. and Ferreira A. Extreme value theory: An introduction.

Embrechts, P, Kluppelberg, C. and Mikosch, T. Modeling extremal events: for insurance and finance.

Méthodes de détection des Unités atypiques : Cas des enquêtes structurelles ukrainiennes.

Noureddine Benlagha & Michel Grun-Réhomme Université Paris 2, ERMES-UMR7181-CNRS, 92 rue d'Assas 75006 Paris, France

Statistiques des extrêmes: Théorie et application,7 Juin 2013

The POT Package October 30,2007

Tukey, J.W. 1977. Exploratory Data Analysis,Ed.Addison-Wesley.

****