

MODELLING RARE EVENTS IN NON-LIFE INSURANCE WITH EXTREME VALUE THEORY

MOUSSA WAJDI

Higher Institute of Management of Tunis

Wajdi.Moussa@isg.rnu.tn

Abstract

This work introduces the extreme value theory and its application in insurance. This statistical theory allows the quantification of the behaviour and the impact of the extreme accidents in an insurance portfolio. It is used to calculate the index of extreme values as well as the adjusted premium with its confidence interval. A study by simulation and an algorithmic implementation were carried out to compare, test and validate the results of the theoretical approaches dedicated to extreme values. This methodology was applied to a real problem, which involves the data of automobile loss ratio. It also constitutes an assistance strategy to price setting in the face of extreme accidents.

Keywords: Extreme value theory, modelling, estimate, queue index, adjusted premium.

JEL Classification: C24; G22.

1. INTRODUCTION

In the face of a rapidly changing world, insurers must “keep pace” (Kwiecień et al. (2020)). The perception of the risks facing insurance companies today has changed dramatically. These risks are characterized by a great diversity of dangerousness, from the simple risk with low consequence to the very dangerous risk whose repercussions are considerable and threaten the stability of the company's results. (Denuit and Charpentier (2004), Deelstra and Plantin (2006), Ohlsson and Johansson (2010), Partrat (2005), Taylor and Weinkle (2020)). The occurrence of major events causes significant human and material damage and disrupts the homogeneity of the insurance portfolio. The insurer is led to set up the most homogeneous possible mutual societies for which statistical and actuarial techniques make it possible to determine a reasonable risk coverage system.

However, data mining, in the extreme case, is of particular interest for the management and prevention of exceptional claims in insurance (Johannsdottir and Cook (2019)). In addition, extreme risks pose real limitations for insurers to constitute a very significant potential loss that could lead them to ruin and are both very uncertain and ambiguous (Jiang and Faure (2020)).

For these reasons, the creation and implementation of adequate financial coverage for these events is increasingly of particular interest. However, statistical and technical tools must be considered when developing measures to identify, quantify and price major risks in insurance. Unfortunately, classic large-scale claims modelling, which pays limited attention to tail end modelling, can prove to be penalizing because they lead to too low a representation of extreme values (Mancini et al. (2014)).

To do this, there are two distinct approaches to modelling the tail of the distribution: the first consists in using a statistical law which reproduces the tails of the empirical distribution, without being concerned with the distribution of the mean values, and the second characterizes the extremes from the global distribution of a phenomenon; this is the object of the extreme value theory (Kotz and Nadarajah (2002), Embrechts, Klupperberg and Mikosch (2003), Coles (2004), Celano and Castagliola (2020)). This modelling provides a global grid for actuaries, to deal with the problems of coverage of exceptional risks beyond a threshold (Liu et al. (2020)). From a statistical point of view, this comes back to modelling and analysing the tail of the distribution by seeking to estimate the extreme quantiles exceeded with a very low probability.

In this context, we propose to select the model that makes it possible to explore extreme claims and results in an adequate modelling in order to determine a good estimation of quantiles extremes and to constitute a pricing assistance strategy in the event of extreme claims.

The article proceeds as follows. Section 2 explains the model to be tested. Section 3 presents the description of the data set. Section 4 comments on the results, and Section 5 concludes.

2. ECONOMETRIC METHODOLOGY

2.1. EXTREME VALUE THEORY

In a sample during an experiment, the observations are distributed approximately according to their distribution law. Thus, observations are rarer in the extremes of this law.

However, we want to study the behaviour at the level of the "tails" of the distribution of observations, and that is why we are interested in the extremes of the observations, which are called extreme values. We describe these extreme values first by characterizing the law of these extreme, in particular of the maximum, and then we generalize by considering the largest values of the observations. We give a general form for the law of extreme which will be valid for the observations coming from any initial distribution. The law of extreme values, when it exists, is indexed by a parameter called the index of extreme values.

The most famous and used estimator's extreme values are the Hill and Pickands estimators.

2.1.1. HILL ESTIMATOR

We consider variables $(X_i)_{1 \leq i \leq n}$ independent and identically distributed by a "heavy tailed" law, that the distribution function F satisfies:

$$1 - F(x) \approx x^{-\alpha} l(x) \quad x \rightarrow +\infty \text{ where } l \text{ checking } \lim_{t \rightarrow +\infty} \frac{l(tx)}{l(t)} = 1. \tag{1}$$

For t large enough, we have:

$$\frac{1 - F(tx)}{1 - F(t)} = \frac{P(X_1 > tx)}{P(X_1 > t)} = \frac{P(X_1 > tx, X_1 > t)}{P(X_1 > t)} = P\left[\frac{X_1}{t} > x / X_1 > t\right] \approx x^{-\alpha} \tag{2}$$

Conditionally to $X_{n-k,n}, (X_{n,n}, X_{n-1,n}, \dots, X_{n-k+1,n})$ are the order statistics of a sequence of k random variables of density having for support $[X_{n-k,n}, \infty[$.

Hill's estimator (1975) is given by the following formula:

For $\gamma > 0$ we have:

$$\hat{\gamma}_n^{(H)} = H_{k,n} = k_n^{-1} \sum_{i=1}^{k_n} \ln(x_{n-i+1,n} / x_{n-k_n,n}). \tag{3}$$

We define the extreme quantile by:

$$x_p = F^{-1}(1 - p_n) \text{ for } p_n \rightarrow 0 \text{ when } n \rightarrow \infty \tag{4}$$

The extreme quantile estimator is given by:

$$\hat{x}_p = x_{n-k_n,n} (np_n / k_n)^{-\hat{\gamma}_n^{(H)}} \tag{5}$$

We must choose $k_n = [n^\alpha]$ or $[n\alpha]$ $\forall 0 < \alpha < 1$

2.1.2. PICKANDS ESTIMATOR

The Pickands estimator is also constructed using the m largest observations of a sample. The advantage of this estimator is that it is valid regardless of the sign of the extreme value index. Pickands demonstrates the weak consistency of his estimator. Strong convergence as well as asymptotic normality have been demonstrated by Dekkers and de Haan (1989). We define the Pickands estimator by:

$$\hat{\gamma}_{k,n}^p = \frac{1}{\ln 2} \ln \frac{X_{k,n} - X_{2k,n}}{X_{2k,n} - X_{4k,n}} \tag{6}$$

for

$$k \rightarrow \infty, k/n \rightarrow 0$$

$$\hat{\gamma}_{k,n}^p \xrightarrow{p} \gamma, n \rightarrow \infty$$

The extreme quantile estimator is given by:

$$\hat{x}_{k,n}^p = X_{n-k+1,n} + \frac{\left(\frac{k}{(n+1)p}\right)^{\hat{\gamma}_{k,n}^p} - 1}{1 - 2^{\hat{\gamma}_{k,n}^p}} (X_{n-k+1,n} - X_{n-2k+1,n}) \tag{7}$$

2.2. ESTIMATE OF THE RISK PREMIUM ADJUSTED FOR TAIL RISKS

In determining the adjusted premium for major claims, we are faced with the problem of the distribution of extreme values.

In insurance a few large claims that hit a portfolio usually represent the largest portion of claims paid by the company. These extreme events (or risk) are, therefore, of primary interest to actuaries. An example of such a problem has been discussed by Cebrián, Denuit and Lambert (2003) for the Large Medical Insurance Claims Database. To determine an adequate price for the risk premium we mainly use an appropriate pricing, knowing that the premiums cannot be too low because this would result in unacceptable large losses for the insurers.

On the other hand, the premium cannot be too high because of the competition between insurance companies. This main premium corresponds to the equivalent certainty of the double theory of expected utility developed by Yaari (1987).

In what follows we consider a main approach to calculating the premium to derive an estimator of the premium adjusted for the largest demands. In addition, the normality of the asymptotic of such estimator is given.

This result provides an asymptotic confidence interval for the net premium of risks that exceed a high threshold and reopens the discussion on utility including the greatest demands in the decision-making process.

Let X_1, X_2, \dots, X_n a series of independent and identically distributed risk random variables with a common distribution function $F(x) = P(X \leq x), x \in \mathfrak{R}$.

Wang (1996) proposes to calculate the risk premium adjusted for a risk variable X :

$$\pi(X) = \int_0^\infty (1 - F(s))^{1/p} ds, \quad \text{for } p \geq 1 \tag{8}$$

The parameter $p \geq 1$ is called the distortion coefficient. The adjusted risk premium corresponds to the loss that has been defined, for a large threshold $u > 0$, such as:

$$\pi_u(X) = \int_u^\infty (\mathbf{1} - F(s))^{1/p} ds. \tag{9}$$

Necir and Boukhetala (2004) propose an estimator for $\pi_u(x)$

For a suitable economic interest, the threshold u must be very large and depends on a sample size $n \geq 1$ income demands. It is for this reason that we must assume This leads us to write:

$$\pi_u(X) = \pi_{u_n}(X) \tag{10}$$

Let $k = k_n$ a sequence of integers that satisfy $1 \leq k \leq n$, $k \rightarrow \infty$ and $k/n \rightarrow 0$ such as $u_n = Q(\mathbf{1} - k/n)$, where $Q(s) = \inf\{t \in \mathfrak{R}, F(t) \geq s\}$, $0 \leq s < 1$, this is the generalized quantile of the inverse function of F , which means that

$$\pi_{u_n}(X) = \int_{Q(\mathbf{1}-k/n)}^\infty (\mathbf{1} - F(s))^{1/p} ds \quad \forall n \geq 1 \tag{11}$$

Let $X_{1,n} \leq \dots \leq X_{n,n}$ an order statistic based on X_1, \dots, X_n , replacing $Q(\mathbf{1} - k/n)$ and $F(s)$ by a suitable estimator, this leads us to the construction of an estimator for an adjusted risk premium of loss which concerns a sample X_1, \dots, X_n

$$\text{We have : } \hat{\pi}_{u_n} = \int_{X_{N-k,n}}^\infty (\mathbf{1} - F_n(s))^{1/p} ds \tag{12}$$

where $F_n(x) = n^{-1} \# \{X_i \leq x : 1 \leq i \leq n\}$ for $x \in \mathfrak{R}$, (with $\# \Omega$ which means cardinal of Ω), it is An empirical distribution function. The quantile that corresponds to this empirical function is defined as follows:

$$\begin{aligned} Q_n(s) &= \inf\{t \in \mathfrak{R}, F_n(t) \geq s\}, \quad 0 \leq s \leq 1 \\ &= X_{i,n}, \quad (i-1)/n < s \leq i/n, \quad i = 1, \dots, n \end{aligned} \tag{13}$$

with $Q_n(0) = X_{1,n}$, now let's observe:

$$\hat{\pi}_{u_n} = - \int_0^{k/n} s^{1/p} dQ_n(\mathbf{1}-s) \tag{14}$$

By integrating by parts, we find:

$$\begin{aligned} \hat{\pi}_{u_n} &= p^{-1} \int_0^{k/n} s^{1/p-1} Q_n(\mathbf{1}-s) ds - (k/n)^{1/p} X_{n-k,n} \\ &= p^{-1} \sum_{i=1}^k \left(\int_{(i-1)/n}^{i/n} s^{1/p-1} ds \right) X_{n-i+1,n} - (k/n)^{1/p} X_{n-k,n} \\ &= \sum_{i=1}^k \left\{ \left(\frac{i}{n} \right)^{1/p} - \left(\frac{i-1}{n} \right)^{1/p} \right\} X_{n-i+1,n} - (k/n)^{1/p} X_{n-k,n} \\ &= \sum_{i=1}^k \left(\frac{i}{n} \right)^{1/p} \{ X_{n-i+1,n} - X_{n-i,n} \}, \quad p \geq 1 \end{aligned} \tag{15}$$

For $p=1$, the statistics $n\hat{\pi}_{u_n}$ corresponds to the reinsurance treaty by ECOMOR (excess of the relative average cost) introduced by Thépaut (1950).

Since we are interested in rare events, we have to assume that the distribution tail follows a Pareto-like function F , in other words we have to make sure that F is Heavy Tailed.

Namely, there is a constant $\gamma > 0$, such as

$$\lim_{t \rightarrow \infty} \frac{1-F(tx)}{1-F(t)} = x^{-1/\gamma}, \quad \forall x \tag{16}$$

We also have another definition which is equivalent to the previous one:

$$\lim_{t \rightarrow 0} \frac{Q(\mathbf{1}-tx)}{Q(\mathbf{1}-t)} = x^{-1/\gamma}, \quad \forall x \tag{17}$$

3. DATA AND PRELIMINARY ANALYSES

We use a cross-sectional data set from the ‘*National Institute for Transport and Safety Research*¹, a French organization whose purpose is to analyse traffic safety.

The database concerns a random sample of 600 observations for 4-wheel passenger vehicles during 2014, from the portfolio of a French mutual insurance

¹ <https://selectra.info/energie/actualites/politique/fermeture-INRETS>

company. The variable studied corresponds to the costs of the claim. Remember that the amount of a claim includes direct compensation for the victim, the management costs internal to the company as well as the external costs (expertise, legal costs) relating to this claim. It does not include contract acquisition costs. In auto insurance, this amount is unknown. It is therefore the realization of a positive or zero real random variable.

To determine the effects of extreme values on the various dispersion indicators, we study the distribution of claims costs with or without extreme values.

A : All costs

B : Costs without extreme values (extreme values are eliminated).

Table 1. Descriptive analysis of the different samples

Costs	Mean	Median	Standard deviation
A	202.051	0	625.268
B	0	106.854	780.52

This sample corresponds to all observations. The size of this sample is 600 observations. Table 1 shows that on average the costs amount to 202,051 the standard deviation is equal to 625,268.

After eliminating values deemed to be extreme, the sample mean decreases, the median is not sensitive to the presence of extreme values, while the variance (standard deviation) becomes lower.

The following graph is a description of our sample:

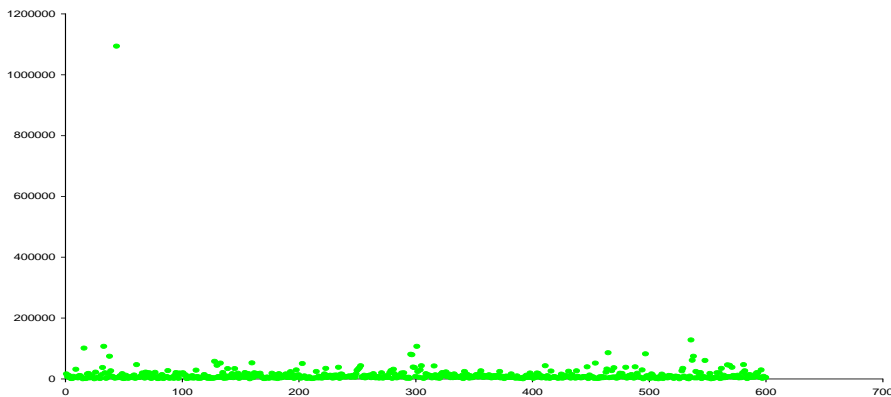


Figure 1. Real sample of size $n = 600$

We notice the existence of extreme values.

We have the number of realizations n_i ($i = 1, \dots, m$) of m eventuality, during n independent identical experiments.

We use the SPSS software which makes it possible to fit the data of the sample with a certain number of laws, from a graphical test which makes it

possible to compare the spacing of the data of a sample compared to a probability law specified represented by a diagonal line. This test is called QQ-plot².

To determine the law that best fits our sample, we restricted ourselves to the main laws which are exponential distribution, Pareto distribution, Log-normal distribution and Student distribution.

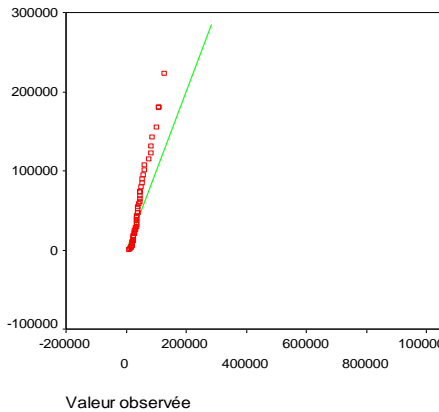


Figure 2. Distribution of claims costs observed in relation to the exponential distribution

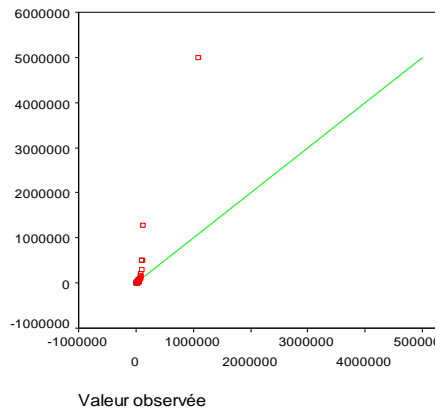


Figure 3. Distribution of claims costs observed compared to the Pareto distribution

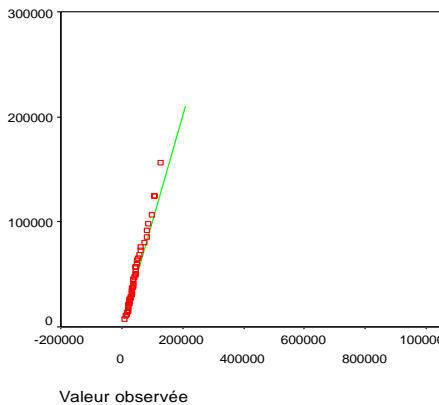


Figure 4. Distribution of loss experience costs observed in relation to the Log-normal distribution

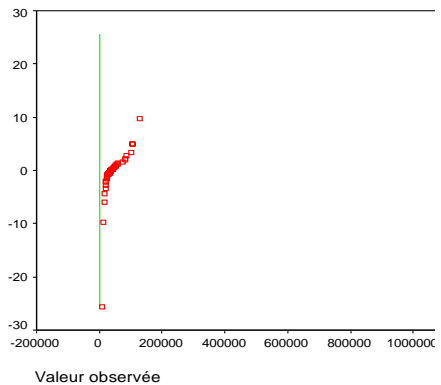


Figure 5. Distribution of claims costs observed in relation to the Student distribution

² We call QQ-plot a quantile versus quantile plot. If the points obtained are approximately along a line with slope 1 and ordinate at the origin 0, meaning that the quantiles of the law found on the ordinate are close to those of the law on the abscissa, we can conclude that the laws are identical. This judgment is made in a purely graphic manner.

The diagonal line represents the cumulative probabilities of the theoretical law, the curve represents the observed probabilities, we notice a very large approximation between the two, in the Log-Normal and exponential test, we can conclude that our sample is heavy tailed therefore heavy tailed.

4. EMPIRICAL RESULTS

Figures (6) and (7) show the graphs of the two estimators Hill and Pickands, for the first estimator we notice that the estimator is stable and has a linear form, which validates the theoretical properties of the estimator, on the other hand the Pickands estimator exhibits a great variability which can be explained by the reduced number of observations which it uses.

The value of Hill's estimator is around 1.5, which shows the consistency of the estimator because the sample is simulated from a Frechet law with parameter $\alpha = 1.5$.

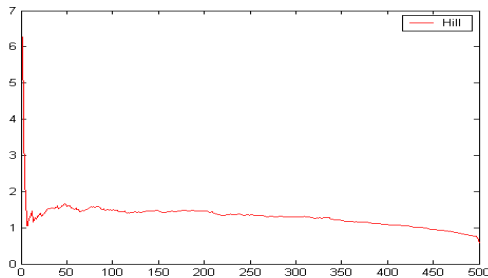


Figure 6. Graph of Hill's estimator as a function of the number of excess

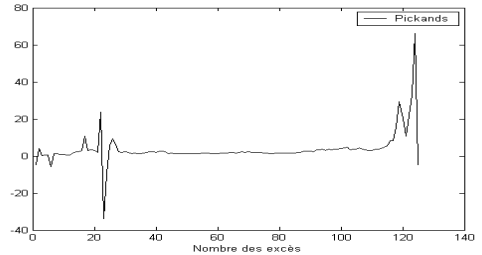


Figure 7. Graph of the Pickands estimator as a function of the number of excess

In Table 2 , the results for the calculation of the premium as a function of P and Alpha (distortion parameter and the number of excess). We fix Alpha and we vary the P to see the change in the premium, the latter as defined is more like the point estimate. We have seen in estimation theory that recourse to confidence intervals is often necessary in order to define a confidence region. Thus, and in order to offer decision-makers some leeway, we define an interval I for the adjusted premium.

Table 2. Result of the variation of the premium and its confidence interval

Alpha	P	Lower bound	Premium	Upper bound
0.5	2	126758.082	149061.843	171365.602
	10	640089.421	689059.057	738028.692
	20	781311.941	835340.138	889368.334

	30	834913.841	890741.789	946569.737
	45	872653.034	929714.002	986774.969
	48	877487.547	934704.545	991921.543
	50	880400.966	937711.787	995022.607
0.9	2	166361.221	174977.267	183593.312
	10	722121.757	732295.726	742469.694
	20	871148.426	881535.978	891923.529
	30	927513.624	937973.362	948433.099
	45	967142.192	977650.333	988158.473
	48	972215.579	982729.786	993243.992
	50	975272.612	985790.461	996308.308

For the calculation of the adjusted premium, we notice that each time we increase P the premium increases then it stabilizes and becomes more linear, see figure 8 below.

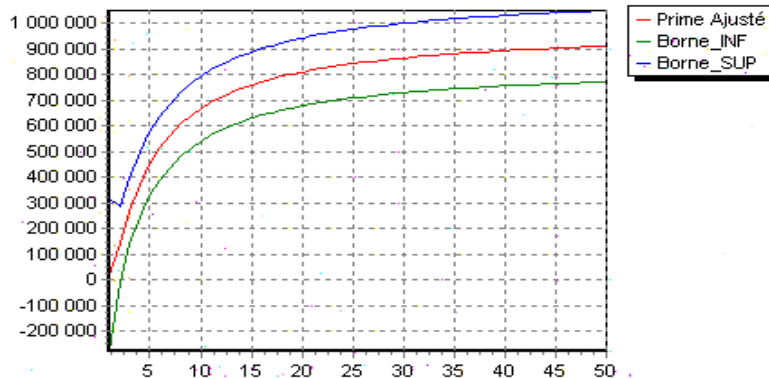


Figure 8. Representation of the linear part that the decision maker can choose a premium

5. CONCLUSION

In this work, we have proposed a quantitative approach to push the limits of risk insurability.

The approach used (the extreme value theory) seems particularly interesting to us for solving the problem of the inequity of pricing. On the one hand, it considers the heterogeneity of the insurance portfolio following the

presence of serious claims. On the other hand, it allows for a rational pricing procedure. So, it is worth mentioning that the standard approach used by insurers for determining the pure premium and which is based on operational principles cannot be used in cases with large value claims. This approach has the effect of introducing a bias which manifests itself in an underestimation or overestimation of the premium sought. It is the inequity of pricing. We can also note that the pricing in the presence of serious claims depends closely on the performance of the method used to calculate the estimates. Another problem here is that the estimates are sensitive to the size of the sample.

REFERENCES

- Cebrián A., Denuit, M., Lambert, P., (2003). Generalized Pareto Fit to the Society of Actuaries. *Large Claims Database, North American Actuarial Journal*, 7: 18-36.
- Celano, G., Castagliola, P., (2020). On-line Monitoring of Extreme Values of Geometric Profiles in Finite Horizon Processes. *British Association for the Advancement of Science*, 36: 1313-1332.
- Coles, M., Masters, A., (2004). Duration-Dependent Unemployment Insurance Payments and Equilibrium Unemployment. *Economica*, 71: 83-97.
- Deelstra, G., Plantin, G., (2006). Théorie du Risque et de la Réassurance. *Economica, Paris*.
- Dekkers, M., Einmahl, J., Haan, L., (1989). A Moment Estimator for the Index of an Extreme Value Distribution. *Annals of Statistics*, 17: 1833-1855.
- Denuit, M., Charpentier, A., (2004). Mathématiques de l'Assurance Non-Vie. Tome I : Principes Fondamentaux de Théorie du Risque. *Collection Economie et Statistique Avancées, Economica, Paris*.
- Embrechts, P., Klüppelberg, C., Mikosch, T. (2003). Modelling Extremal Events for Insurance and Finance, *Applications of Mathematics*, 33, Springer
- Hill, B M., (1975). A Simple General Approach to Inference about the Tail of a Distribution. *Annals of Statistics*, 3: 1163–1174.
- Jiang, M., Faure, M., (2020). Risk-Sharing in the Context of Fishery Mutual Insurance: Learning from China. *Marine Policy*. <https://doi.org/10.1016/j.marpol.2020.104191>.
- Johannsdottir, L., Cook, D., (2019). Systemic Risk of Maritime-Related Oil Spills Viewed from an Arctic and Insurance Perspective. *Ocean & Coastal Management*, DOI: 10.1016/j.ocecoaman.2019.104853.
- Kotz, S., Nadarajah, S.,(2002).Local Dependence Functions for the Elliptically Symmetric Distributions. *Sankhya Ser*, 65: 207-223.
- Kwieceń, I., Kowalczyk-Rólczyńska, P., Popielas, M., (2020). The Challenges for Life Insurance Underwriting Caused by Changes in Demography and Digitalisation. *Life Insurance in Europe*, 50: 147-163.

- Liu et al. (2020). Estimating Cancer Incidence Based on Claims Data from Medical Insurance Systems in Two Areas Lacking Cancer Registries in China. *EClinicalMedicine*, 100312.
- Mancini, GB., Hartigan, PM., Shaw, L., (2014) Predicting Outcome in the COURAGE Trial. Coronary Anatomy Versus Ischemia. *JACC Cardiovascular Interventions*, 7: 195-201.
- Necir, A., Boukhetala, K., (2004). Estimating the Risk-Adjusted Premium for the Largest Claims Covers. COMPSTAT'2004. *Proceeding in Computational Statistics. Reinsurance ISBN 3-7908-1554-3, Physica-Verlag, Heidelberg, New York. Springer.*
- Ohlsson, E., Johansson, B., (2010). Non-Life Insurance Pricing with Generalized Linear Models. *EAA series.*
- Partrat, C., Besson, J.L., (2005). Assurance Non-Vie, Modélisation, Simulation. *Economica. Paris.*
- Pickands, I., (1975). Statistical Inference Using Extreme Value Order Statistics. *Annals of Statistics*, 3: 119-131.
- Taylor, Z., Weinkle, L., (2020). The Risksapes of Re/Insurance. *Cambridge Journal of Regions, Economy and Society*, 13: 405-422.
- Thépaut, A., (1950). Une nouvelle forme de réassurance. Le traité d'excédent du coût moyen relatif. *Bulletin trimestriel de l'Institut des actuaires français.*
- Yaari, M.E., (1987). The Dual Theory of Choice under Risk. *Econometrica*, 55: 95-115.