SELECTED PROBLEMS ASSOCIATED WITH MODELING THE BIGGEST COMPENSATIONS IN NON-LIFE INSURANCE

Ewa Poprawska, Anna Jędrzychowska

Abstract*:Modelling the highest compensations are very important for insurance companies, as more and more policies now can generate high claims (for example third party liability insurance, property, especially connected to catastrophic events). In modelling extreme values of compensations paid by insurance company it is possible to use different models (exponential, gamma, Pareto distribution are most common) for typical values and different ones for the highest compensations (for example Generalised Pareto distribution). In this case however a problem of choice of the point, in which the change of the way of modelling is taking place appears. The paper is devoted to methods useful in choosing this point, that can be found in literature. Using simple example of the data, these methods can be analysed to find their advantages and disadvantages. All of the presented methods are mainly based on the analysis of graphs of selected parameters, which makes the results obtained are not strict guidelines, however, provide indicative results.*

Keywords: *Extreme value modeling, Insurance claims.*

JEL Classification: *C16, , C46, C65.*

Introduction

From the standpoint of insurance proper premium calculation is crucial in ensuring financial balance. The base pure premium calculation is the expected value of random variable, with which is described by the amount of compensation. Therefore, proper modeling of the random variable is important for insurance.

In case of damage of a typical size insurance companies generally have a sufficiently large amount of data that the selection of the appropriate distribution is not a problem. Furthermore, it is also possible to use an empirical distribution.

However, in the case of insurance, in which the terms of the contract allow the occurrence of exceptionally high compensation (the need to pay such compensation may occur eg in connection with the occurrence of events that can be described as natural disasters), there is a problem associated with the occurrence of a small number of observations, based on which can be requested on the form and distribution parameters. At the same time it is precisely these claims very strongly affect the expected value of damages.

Also from the viewpoint of reinsurer the largest claims are of particular importance. This mainly applies to the excess of loss reinsurance, in which the reinsurer assumes liability for damages, which amount exceeds the value of the contract of reinsurance.

Thus, modeling is particularly important damage from the right tail of the probability distribution of random variable describing the amount of damages.

1 Modelling extreme losses

1.1 Approaches to modeling extreme losses

The modeling of extreme values can distinguish several approaches (for [3]). The first approach is based on the distributions of extreme values. Based on the Fisher-Tippett theorem can be concluded that the maximum compensation can be modeled using the generalized extreme value distribution (GEV).

The second approach is to match the distribution for all observations or for values that exceed a fixed value (using the censored distributions and conditional). Problems with this approach will focus on the further part of the paper. If the distribution is matched to all observations, it may well describe a typical value, while for higher values may be poor fit to empirical data. Thus, often it is reasonable to separate modeling of typical damage and the values derived from the tail distribution - the use of censored distributions and mixtures of distributions. In this case, however, there is the problem of choosing the point at which a change in modeling.

If the amount of compensation from the tail of the distribution are modeled separately, we introduce the distribution function of the excesses over threshold *u* (conditional excess distribution) defined as:

$$F_u(x) = P(X - u \le x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)} \tag{1}$$

Of course, knowing the conditional excess distribution we can express a probability distribution for the original value $x \ge u$ as follows:

$$\hat{F}(x) = (1 - F_n(u)) F_u(x-u) + F_n(u) \tag{2}$$

where $F_n(x)$ is a distribution function for typical values, or ot can be empirical distribution either.

On the basis of Pickans-Balke-de Haan statements (see [5]) for a wide class of distributions of the conditional excess distribution for sufficiently high values of u can be approximated by a generalized Pareto distribution (GPD):

$$G_{\xi,\sigma}(x) = \begin{cases} 1 - (1 + \xi x / \sigma)^{-1/\xi} & \text{dla } \xi \ne 0 \\ 1 - \exp(-x/\sigma) & \text{dla } \xi = 0 \end{cases} \tag{3}$$

Additional paremeter (μ) can be introduced as follows:

$$G_{\xi,\mu,\sigma}(x) = G_{\xi,\sigma}(x-\mu) \tag{4}$$

ξ can interpreted as a shape parameter, β as a scale parameter.

It can be shown that if the conditional distribution excess is approximated generalized Pareto distribution, then (2) also has a generalized Pareto distribution with the same parameterξ and $\tilde{\sigma} = \sigma\left(1-F_n(u)\right)^{\xi}$, $\tilde{\mu} = \mu - \tilde{\sigma}\left(\left(1-F_n(u)\right)^{-\xi} - 1\right)/\xi$. Thus there are strong theoretical arguments supporting the modelling the probability of extreme values of compensation using of the generalized Pareto distribution.

In case of modeling the distribution of values that exceed the threshold *u* using GPD threshold selection plays an important role. If this value is too low, then the approximation of distribution is not justified, and if too high, the number of observations on which made the distribution of estimates of parameters, it will be too small.

1.2 Choice of the optimal value of threshold u

In literature you will find various hints to help you determine the value of *u*, but none deal with the problem in an unambiguous manner. One of the simplest suggestions is to set the threshold at a level level of empirical quantile.

Another method is based on the analysis of the shape of mean excess function – MEF:

$$e(x) = E(X - x | X > x) = \frac{E(X) - E(X \wedge x)}{1 - F(x)} \tag{5}$$

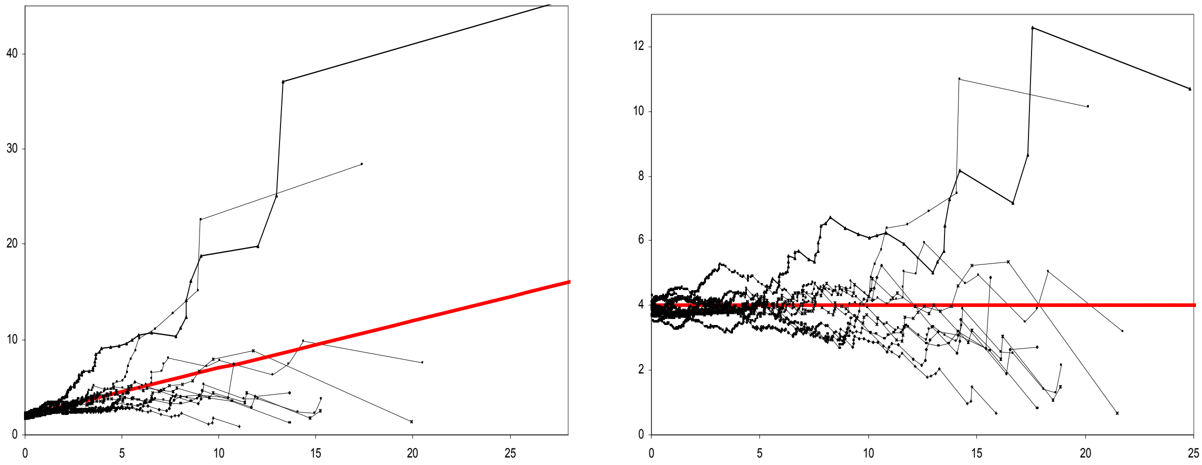
with empirical function (see [5]):

$$e_n(x) = \frac{\sum_{i=1}^{n} (X_i - x)^+}{\sum_{i=1}^{n} I_{\{X_i > x\}}} \quad \text{where} \quad I_{\{X_i > x\}} = \begin{cases} 1 \text{ dla } X_i > x \\ 0 \text{ dla } X_i \le x \end{cases} \tag{6}$$

For heavy tailed distributions this fuction is increasing, otherwise decreasing, while for exponential distribution - constant. For Pareto distribution, also for GPD, the is increasing and linear.

It can therefore be concluded that *u* can be should be set at a level from which the graph of the function of MEF is approximately linear. This criterion may be useful for preliminary analysis of empirical data, but it should me mentioned that for a small number of observations MEF, even for observations generated from a particular distribution, may differ significantly from the model run, as illustrated by the following charts. The largest differences occur for the highest value, which is in the area's most interesting from the perspective of modeling extreme values.

Fig. 1: MEF for observation generated from a) Pareto and b) exponential distribution compared with theoretical value of MEF



Source: own studies

Another way of preliminary data analysis is to analyze the shape of the quantile plots (empirical quantiles compared with theoretical - derived from the fitted to the empirical data distribution). This is a graph of points:

$$\left\{ \left(X_{k,n}, F^{-1}\left(\frac{n-k+1}{n+1} \right) \right) : k = 1, \ldots, n \right\} \tag{7}$$

where $X_{k,n}$ is k-th empirical quantile.

If the matched observations well describes distribution, then the quantile plot is approximately linear. Distributions characterized by heavier tails the plot deviates from a straight line up for the points describing the highest quantile. The threshold should be set at a level where the plot begins to deviate from straight line.

Another way (cf. [1]) is a graph showing dependency between p-th quantile estimated using GPD and the threshold, which is a chart of points:

$$\left\{ \left(u, \hat{x}_p \right) : u \geq 0 \right\} \tag{8}$$

where $\hat{x}_p = u + \frac{\hat{\sigma}}{\hat{\xi}}\left(\left(\frac{n}{N_u}(1-p)\right)^{-\hat{\xi}} - 1\right)$ is maximum likelihood estimator of p-th quantile.

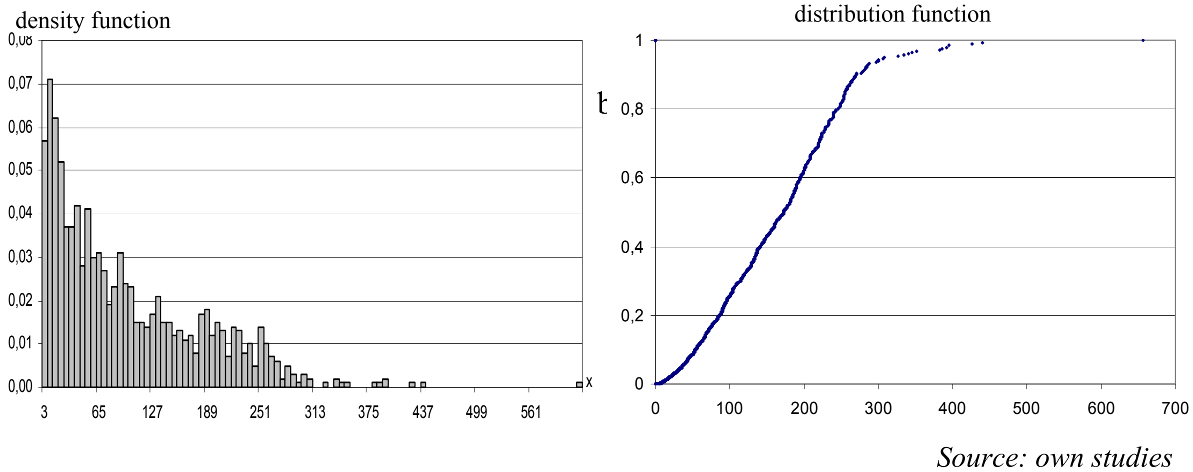
The threshold *u* should provide a stabilization in the values of a selected quantile.

Finally, when choosing an optimal threshold value may be helpful to analyze the value of the estimators of parameters of the generalized Pareto distribution, depending on u. The parameter which most affects the thickness of the tail distribution is the shape parameter ξ. A graph of the values of the threshold amount, together with marked confidence intervals, allow to find a compromise between error and stability of parameter estimation.

2 Problem solving – empirical example

The purpose of this article is the comparison of proposed in the literature criterion for selection threshold. Calculations used in the analysis are based on 1000 observations generated from the Pareto distribution. At the same time in a better way to bring the behavior of data from heterogeneous portfolio of policies, mixing variable was introduced. Expected value is not constant, but in its place a mixing variable normally distributed N(100,40) was introduced. Similarly the standard deviation - N (80, 20) - distribution of parameters describes the structure of risk in the portfolio.

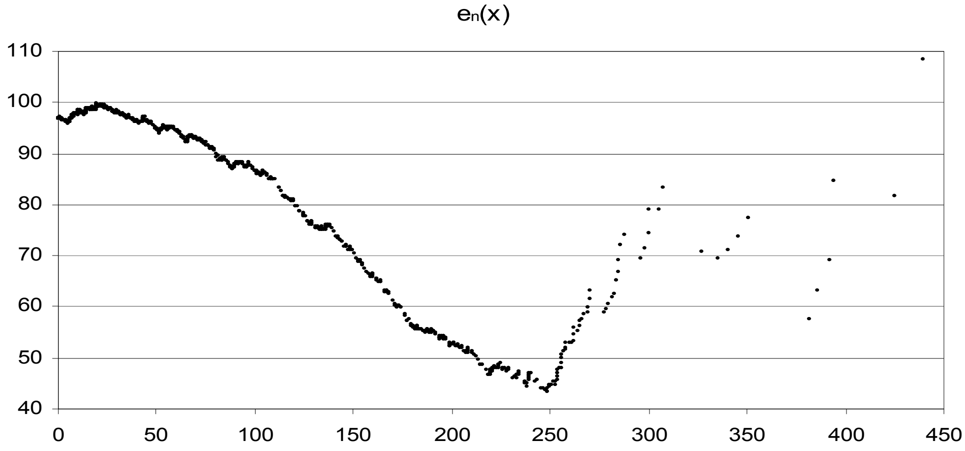
Fig. 2: Empirical density (left chart) and distribution function (right chart) of data used in an example



Source: own studies

The first way of finding optimal value of threshold *u* is to analise MEF function (Fig 3). A clear change in the nature of the graph is visible for values close to 250. The points corresponding to values greater than 250 are beginning to increase. As it is obvious it is quite difficult to give one good answer to the problem of finding threshold u in this case.

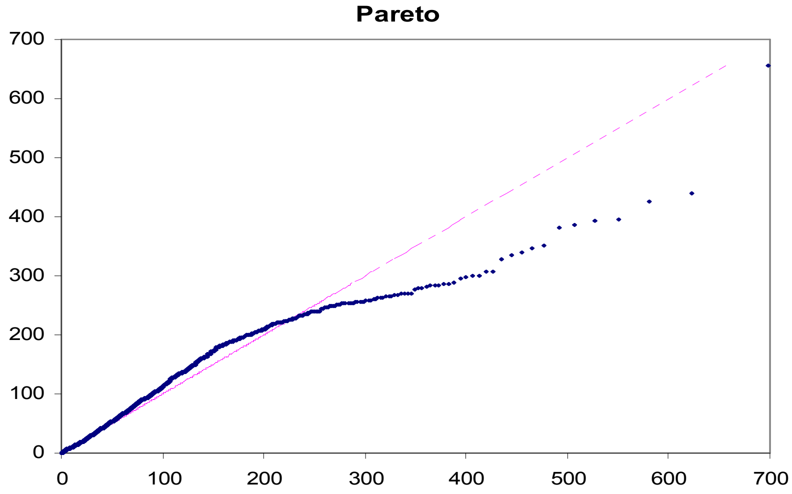
Fig. 3: Empirical MEF for the data used in an example



Source: Own studies

Second way of finding u is to use quantile plot (Fig 4 – compared to Pareto distribution. Similar deviations from the straight line appear near the values of 200-250. Data compared with distributions such as gamma, Weibull, Burr, GPD gave very similar conclusions.

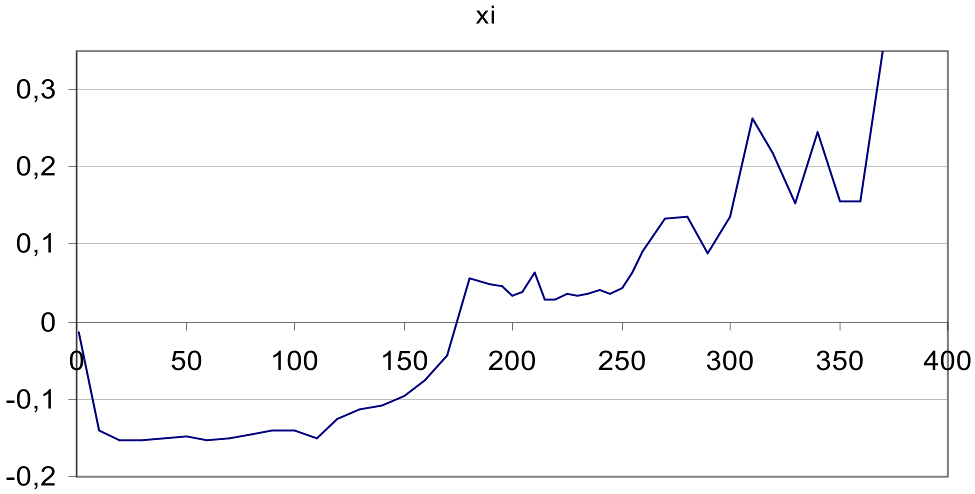
Fig. 4: Quantile plot (data vs Pareto distribution)



Source: Own studies

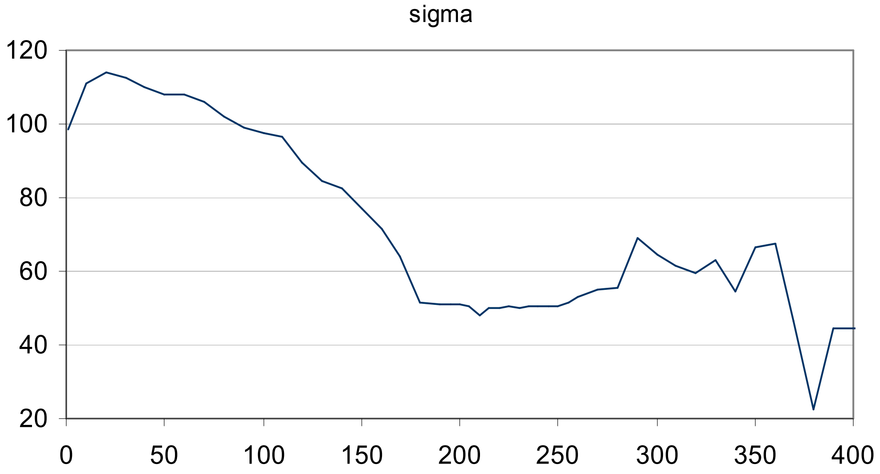
The next two charts (Fig 5 and Fig 6) illustrate the value of the maximum likelihood estimators of parameters of the generalized Pareto distribution, depending on the choice of the threshold u. For the lower thresholds than 170 of the estimators of parameters ξ (Fig 5) and σ (Fig 6) differ significantly from those obtained in case the higher values of u. The values of both estimators stabilizie for a threshold of about 170 to about 250 Above u = 250 the number of observations above the threshold drops below 40, for u> 270 falls below 30.

Fig. 5: Estimator of parameter ξ depending on threshold u



Source: Own studies

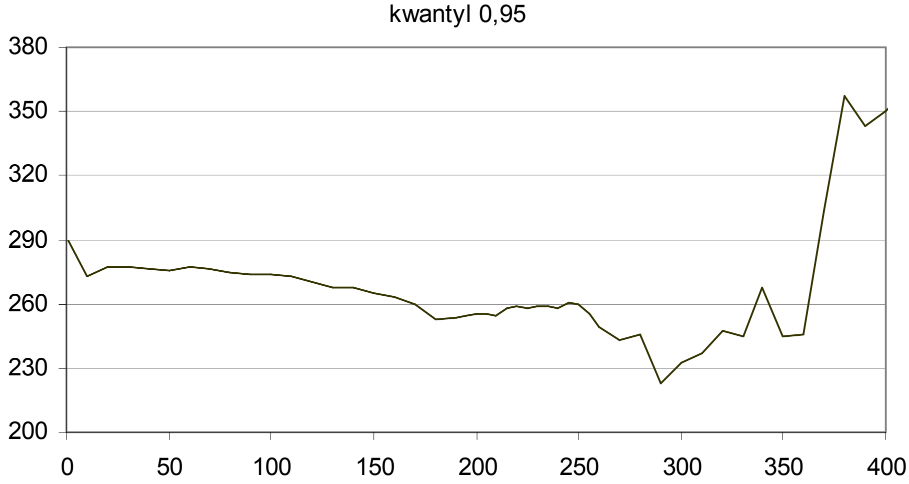
Fig. 6: Estimator of parameter σ depending on threshold u



Source: Own studies

Similarly it can be observed stabilisation of 0,95 quantile (Fig 7) for threshold u lower than 250. For bigger values of u, the values of 0,95 quantile starts to be unstable.

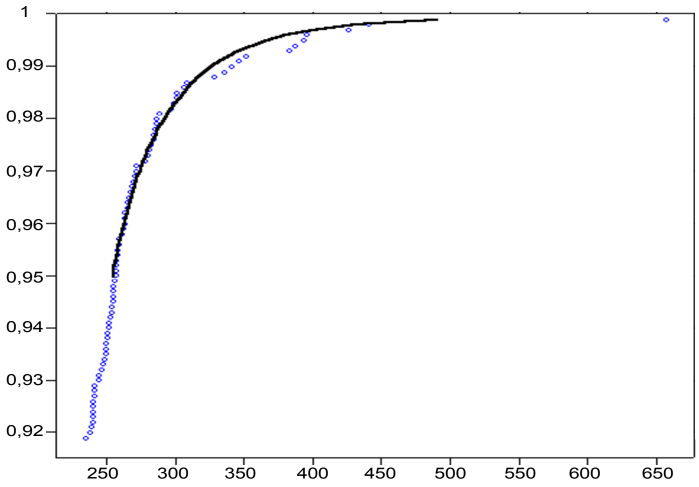
Fig. 7: 0,95 empirical quantile of the data



Source: Own studies

Based on the foregoing, it can be assumed that the optimal threshold value should be between 170 and 250th The following graph (Fig 7) shows a comparison of the empirical distribution function with fitted GPD. The value of the threshold u = 223, the parameters of GPD were estimated on the basis of 100 observations.

Fig. 8: Empirical distribution with GPD for u=223



Source: own studies

Conclusion

All these ways of determining the optimal thresholds are based on visual analysis of graphs. Thus, observations on stabilization of the value estimates may differ from the conclusions of others. Moreover, these methods allow to determine the indicative ranges, which should include a threshold value, the final decision is taken arbitrarily. However, despite the imperfections, they provide information that can make this decision much easier.

References

[1] BASSI F., EMBRECHTS P., KAFETZAKI M. *A survival kit on quantile estimation*, available at WWW: <www.math.edu.ethz.ch/finance>

[2] DAYKIN C.D., PENTIKÄINEN T., PESONEN M. (1996). *Practical Risk Theory for Actuaries*. Chapman & Hall, London.

[3] EMBRECHTS P., KLÜPPELBERG C., MIKOSCH T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer - Verlag, Berlin.

[4] KLUGMAN S., PANJER H.H., WILLMOT G.E.(1998). *Loss models: From Data to Decisions*. John Wiley & Sons, New York.

[5] MC NEIL A.(1997). *Estimating the tails of loss severity distributions using extreme value theory*. ASTIN Bulletin.

Contact address

Ewa Poprawska, PhD
Anna Jędrzychowska, PhD
Wroclaw University of Economics
Department of insurance
ul. Komandorska 118/120
53-345 Wroclaw
E-mail: ewa.poprawska@ue.wroc.pl anna.jedrzychowska@ue.wroc.pl
Phone number: +48 71 36 80 158

Received: 31. 01. 2011
Reviewed: 16. 03. 2011
Approved for publication: 16. 01. 2012