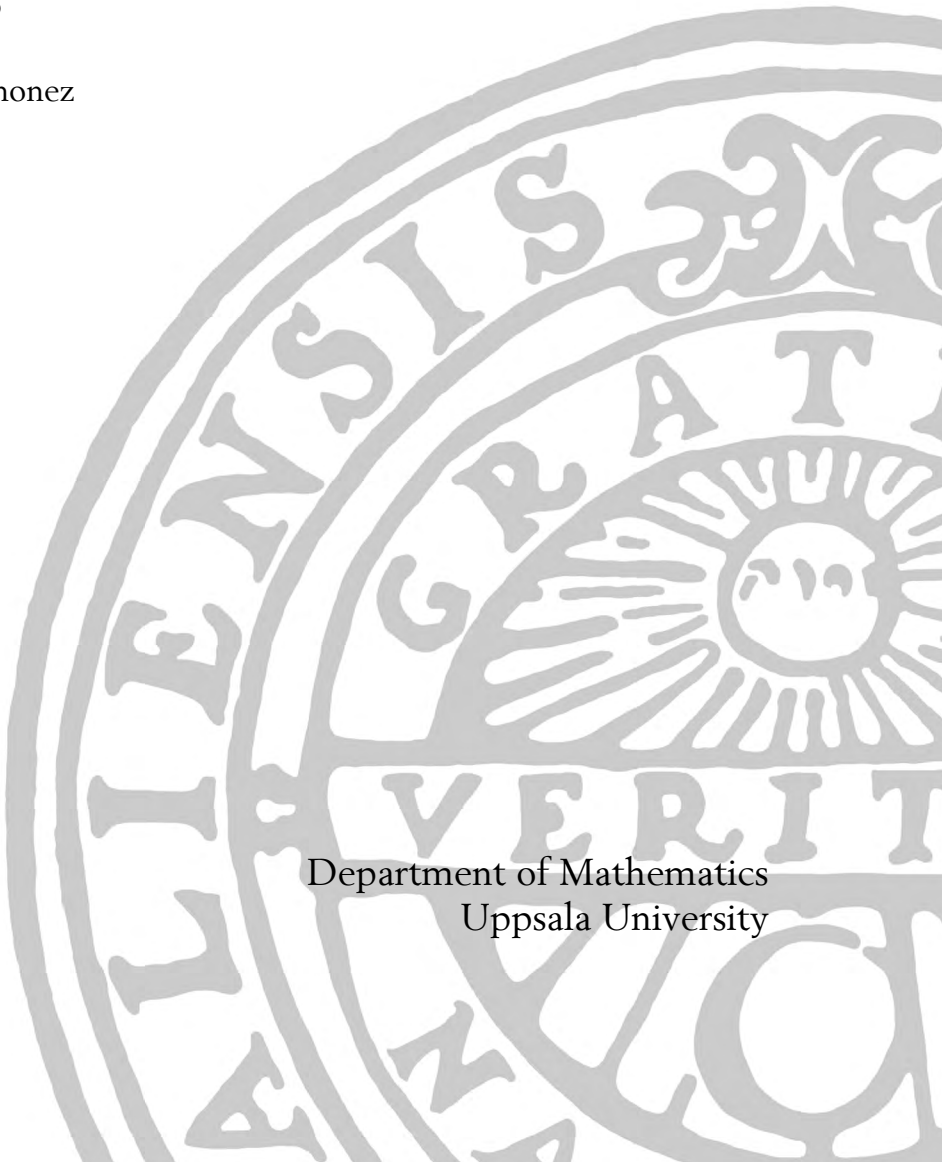
UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2018:32

Application of the Peaks-Over-Threshold Method on Insurance Data

Max Rydman

Examensarbete i matematik, 15 hp
Handledare: Jesper Rydén
Examinator: Veronica Crispin Quinonez
Juni 2018



Department of Mathematics
Uppsala University

Application of the Peaks-Over-Threshold Method on Insurance Data

Max Rydman

June 27, 2018

Contents

1 Introduction

Extreme value analysis is used as a tool to analyze and study statistics on sample values that deviate extremely from the mean of the full sample. This has uses in many different areas of applications such as hydrology, structural engineering, medicine, finance, insurance and many others.

In insurance one application is to model the frequency of heavy damages to cars, houses or other high-value investments. It is important to balance the risk of damage with their potential consequences so as to make a profit.

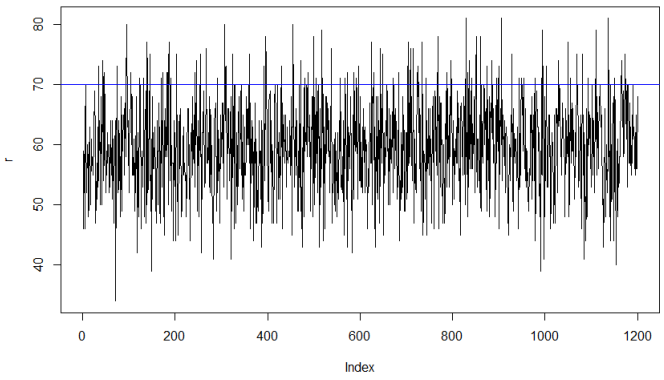


Figure 1: Randomly generated data according to a Poisson distribution with $\lambda = 60$ with a threshold at the 90% quantile.

The peak over threshold-method (POT-method) is one of many methods that falls under extreme value analysis and is based around looking at extreme values of some sample that exceeds a certain threshold value. Pictured above is a randomly generated time series $x_1, ..., x_{1200}$ with a threshold u set at the 90% quantile of the data. Using the POT-method all sample points that exceeds the threshold $y_1, ..., y_{120}$ are of interest and separated from the rest of the data for further analysis. It has been proven that the values above the threshold can be modeled as a generalized Pareto distribution, for further discussion see Section 2.

The scope of this paper is to briefly introduce the POT-method and apply it to insurance data belonging to a property insurance company from the U.S.A. using several common ways of finding and setting an appropriate threshold as well as analyzing and modeling the sample. Section 2 will contain a brief introduction to the theory behind extreme value distributions. In Section 3 some methods used to get an appropriate threshold will be introduced. Section 4 will introduce means to estimate parameters for a model based on the POT-method. Finally in Section 5 we will apply the information discussed in the earlier sections to a dataset of insurance data.

2 Peaks-Over-Threshold

The Peak Over Threshold-method (POT-method) is one way to model extreme values. The main concept of the method is to use a threshold to seclude values considered extreme to the rest of the data and create a model for the extreme values by modeling the tail of all the values the exceeds this threshold. This is done in practice by setting a threshold u to be some value defined on $\mathbb{R}$ that exceeds most but not all values defined in some time series or some other vector of collected values. Furthermore it can be shown that for some sufficiently large threshold u the distribution of the values exceeding the threshold approximate to a General Pareto Distribution with some Shape and Scale parameter.
Below, all this will be shown by using theorems and definitions presented by Franke, Härdle & Hafner (2008)[1].

2.1 Extreme Value Distributions

Extreme value distributions are a family of distributions related to extreme value theory. Let's assume we have some independent, identically distributed random variables $X_1, ..., X_n$ with distribution $F(x)$. Let $M_n = \max(X_1, ..., X_n)$ be the maximum, then the cumulative distribution function (cdf) of M_n is

$$P(M_n \leq x) = \prod_{t=1}^{n} P(X_t \leq x) = F^n(x). \tag{1}$$

Given that all $F(x) < 1$ are unbound for all $x < \infty$ we get that $F^n(x) \to 0$ for every x when we let $n \to \infty$. Thus to avoid degenerate limits we need to introduce some suitable sequences $c_n > 0$ and d_n we see that when $n \to \infty$

$$\frac{M_n - d_n}{c_n} \xrightarrow{\text{L}} G. \tag{2}$$

There is only three distributions that are considered asymptotic to the limit distribution of M_n These are the Frechet, Gumbel and Weibull distributions:

$$G_{1,\alpha}(x) = \exp(-x^{-\alpha}), \quad x \geq 0 \quad \text{for} \quad \alpha > 0,$$

$$G_0(x) = \exp(-e^{-x}), \quad x \in \mathbb{R}$$

$$G_{2,\alpha}(x) = \exp(-|x|^{-\alpha}), \quad x \leq 0 \quad \text{for} \quad \alpha < 0,$$

Now let u be a threshold and $F(x)$ the distribution function of an unbounded r.v X. Then the excess distribution above u is:

$$F_u(x) = P(X - u \leq x | X > u) = (F(u + x) - F(u))/\bar{F}(u), \quad 0 \leq x < \infty \tag{3}$$

where $\bar{F}(u)$ is the exceedance probability. It follows that the average excess function is $e(u) = E(X - u \leq x | X > u)$.

2.2 The Generalized Pareto Distribution

The Generalized Pareto Distribution (GPD) is a variation of the Pareto distribution. Given a shape parameter $\gamma \in \mathbb{R}$ and a scale parameter $\beta > 0$, which in the case of modeling POT-models will both be dependent on the choice of u.

$$G(x|u,\beta,\gamma) = P_{\beta,\gamma}(X < x|X > u) = 1 - \left[1 + \frac{\gamma x}{\beta}\right]^{-1/\gamma} \tag{4}$$

$$\text{for} \quad \begin{cases} x \geq 0 & \text{if} \quad \gamma > 0 \\ 0 \leq x \leq -\frac{\beta}{\gamma} & \text{if} \quad \gamma < 0 \end{cases}$$

And for the special case when $\gamma = 0$ we have:

$$G(x|u,\beta,0) = P_{\beta,0}(X < x|X > u) = 1 - \exp(-\frac{1}{\beta}x) \tag{5}$$

i.e the conventional exponential distribution.

The following theorem is presented by Franke, Härdle & Hafner[1]:
Theorem 1: The distribution F contained in the Maximum domain of attraction of the GEV distribution $G\gamma$ with the shape parameter $\gamma \geq 0$, exactly when for a measurable function $\beta(u) > 0$ and a GP distribution $W_{\gamma,\beta(u)}(x)$ it holds that[1]:

$$\sup_{x \geq 0} |F_u(x) - W_{\gamma,\beta(u)}(x)| \to 0 \quad \text{for} \quad u \to \infty. \tag{6}$$

This is a very important theorem for the POT-method and was first proven by Pickands (1975)[6].

2.3 Excess and exceedances

Given a variable X consisting of randomly distributed observations X_1, ..., X_m we can define an exceedance as any $X_i > u$ where u is some threshold. Let $K_n(u)$ be the index of observation n that exceeds u and $N(u)$ be the number of exceedances in X, then we define the excess Y_j, $j = 1, ..., N(u)$ of X as by Franke, Härdle & Hafner (2008)[1] to be:

$$\left\{Y_1, ..., Y_{N(u)}\right\} = \{X_i - u; i \in K_n(u)\} = \left\{X^{(1)} - u, ..., X^{(N(u))} - u\right\} \tag{7}$$

$Y_1, ..., Y_{N(u)}$ are thus i.i.d random variables with distribution F_u from Equation (3) given that the amount of variables $N(u)$ are completely random. Thus for a GPD $W_{\gamma,\beta(u)}$ it holds due to Theorem 1 that $F_u(x) \approx W_{\gamma,\beta(u)}(x)$ for all sufficiently large thresholds u[1].

The excess will help us estimate the exceedance probability $\bar{F}(x)$ for large values of x. An estimator is the empirical distribution function $\hat{\bar{F}}_n(x)$. It needs to be taken into consideration that when x gets very large the empirical distribution function will depend on a handful of very large values and will likely vary largely. A better

estimator can be found by using a relationship between $\bar{F}(x)$, $\bar{F}(u)$ for some large but not extreme threshold u and the excess distribution seen in Equation (4) in Section 2.1.

The excess distribution can be rewritten as:

$$\bar{F}(x) = \bar{F}(u) \cdot \bar{F}_u(x-u) \tag{8}$$

for $x > u$. Replacing the right hand side with their approximations a POT-estimator can be found. Replacing $F(u)$ with the empirical distribution function and using the approximation for $\bar{F}_u(x-u)$ according to the results of Equation (7) we get the POT-estimator as defined by Franke, Härdle & Hafner (2008)[1]:

Definition 1 *The POT estimator $\hat{\bar{F}}(x)$ for the exceedance probability $\bar{F}(x)$, for large x, is*

$$\hat{\bar{F}}(x) = \frac{N(u)}{n}\{1 + \frac{\hat{\gamma}(x-u)}{\hat{\beta}}\}^{-1/\hat{\gamma}} \tag{9}$$

for $u < x < \infty$ where $\hat{\gamma}, \hat{\beta}$ are suitable estimators for the GPD.

3 Setting a threshold

The main difficulty of modeling with the POT-method is setting an appropriate threshold. If it's too large there will be too few values to model the tail of the distribution correctly as the variance is likely to be large due to only very extreme observations remaining. On the other hand a low threshold will include too many values giving a high bias. It is thus of importance of finding a good balance in setting the threshold to find a suitable balance between the variance and the bias of the model.

There are several ways of setting a threshold, and they all carry some positive as well as some negative properties. Below some more common methods will be discussed.

3.1 Rule of thumb

One way to approach setting a threshold is by using a rule of thumb to choose the k largest observations and modeling. Commonly used is the 90th percentile, but others have also been proposed, such as $k = \sqrt{n}$ and $k = n^{(2/3)}/\log(\log(n))$ all of which are practical but to some level theoretically improper[2].

It is the fastest method of setting a threshold but due to the difference in behavior between different data it is not necessarily a reliable way of setting a good threshold due to the inevitable difference between most data. From the view of an insurance company it is possible that the information of interest may be the distribution of claims above some certain value, or the size of some upper quantile of claims. It is therefore sometimes of interest to use a certain threshold to get information even if it depreciates the theoretical analysis of the data.

3.2 Graphical approach

Another way to approach the problem is by using graphical tools to perform diagnostics and draw conclusion based on the results. Scarrot and MacDonald (2012) speaks of how a graphical approach hold some benefits in that it requires more involvement with the data set when setting a threshold and analyzing the final model.

The threshold level set will be more dependent on the data itself and allows for more parameters when deciding relative to using a predetermined rule. These methods are however far more time-consuming than using a rule of thumb and when working with multiple data sets it may be more effective to use a rule of thumb at the cost of accuracy. Below the theory behind some graphical tools defined by Scarrot and MacDonald (2012)[2] are presented.

3.2.1 Mean Residual Life Plot

One graphical tool that can prove helpful to setting a good threshold is the mean residual life plot (MRL-plot) that Davison and Smith (1990) introduced. The MRL-plot used the expectation value of the GPD excesses. For a threshold u the expectation of the excesses will be $E(X - u|X > u) = \beta_u/(1 - \gamma)$, where β_u is the scale

parameter given threshold u and γ is the shape parameter which needs to be defined for $\gamma < 1$ to ensure that the mean exists. Now given the linear property shown in Section 2 the expected value of the excesses for any threshold $v > u$ will be

$$E(X - v|X > v) = \frac{\beta_u + \gamma v}{1 - \gamma} \tag{10}$$

which can be shown is linear in v with the gradient $\gamma/(1-\gamma)$ and intercept $\beta_u/(1-\gamma)$[2]. From this point the mean excesses can be plotted and the assumption is that when the plot starts showing a linear behavior a suitable threshold can be estimated. The plot is likely to lose its linear behavior when the threshold gets too large due to the variance of the few extremes left will cause the plot to jump. There may occur cases where for some data the plot is completely or never linear and little to no information may be collected from observing the plot.

3.2.2 Parameter Stability Plot

The Parameter stability plot, called Threshold stability plot by Scarrot and MacDonald (2012)[2], looks at the estimates of the shape and scale parameters to find a suitable threshold. Over some range u and the assumption of a constant shape parameter $\hat{\gamma_u} = \hat{\gamma}$ the modified scale parameter is estimated to be $\hat{\beta}_u - \hat{\gamma}u$. Two parameter stability plots can then be made by plotting these estimators against the interval u with the 95% confidence interval for the estimators. Looking at the plots one can find a suitable threshold at the lowest value where the plots are approximately constant. Depending on the data there may be cases where the parameter stability plot won't shed any further information on the situation just like with the mean residual life plot.

4 Parameter estimation

Once a threshold is set, parameters can be estimated. When it comes to modeling extreme value distributions there are a couple of methods. The most commonly used variant is the ML-method (Maximum likelihood-method) and it will be further explored below. There exists some problems with the ML-estimation which will also be expanded upon, as well as the PWM-method (probability weighted moments) which can be a better alternative in certain cases.

4.1 Maximum Likelihood method

The ML-method is based around finding a likelihood function and maximizing it to find a suitable estimator. We want to do this for the GPD and so from Equation (4) we derive a log-likelihood function for some exceedances $Y_1, ..., Y_{N(u)}$ over a threshold u for the case when $\gamma \neq 0$:

$$l(\beta, \gamma | Y_1, ..., Y_{N(u)}) = -N(u) \log \beta - (1 + \frac{1}{\gamma}) \sum_{j=1}^{N(u)} \log(1 + \frac{\gamma}{\beta} Y_j) \tag{11}$$

And for $\gamma = 0$ the log-likelihood is derived from Equation (5) to be:

$$l(\beta | Y_1, ..., Y_{N(u)}) = -N(u) \log \beta - (1/\beta) \sum_{i=1}^{N(u)} Y_i. \tag{12}$$

It has shown by Smith (1985)[7] showing that the ML estimates exists for all scale parameters $\gamma > -1$ and show regularity in cases where the parameter is $\gamma > -0.5$.

4.2 Probability Weighted Moments method

There exists a few other ways to estimate the parameters of an extreme value distribution. One of these is the method of using probability wighted moments. The basis behind this theory was proposed by Landwehr et al. (1979)[9] with further expansion on the subject summarized by Coles & Dixon (1999)[4]. The idea is to try and equate the theoretical moments to the sample moments in an attempt to find an alternate method to the ML-method similarly to the method of moments.

For $r = 0, 1, 2, ...$ a variable X with distribution function F has theoretical probability weighted moments defined as

$$\beta_r = E[X\{F(X)\}^r]. \tag{13}$$

Further, given a sample $x_1, ..., x_n$ we get an estimator for β_r that can be used as sample moments to equate with the theoretical moments:

$$\hat{\beta}_r = \frac{1}{n} \sum_{i=1}^{n} x_i \{\tilde{F}(x_i)\}^r \tag{14}$$

where $\tilde{F}$ is an estimate for the distribution function F of $x = x_1, ..., x_n$. Further an analytical expression for the PWM:s of the generalized extreme value distribution can be obtained from Equation (4):

$$\beta_r = (r+1)^{-1}[\mu - \beta\{1 - (r+1)^{\gamma}\Gamma(1-\gamma)\}/\gamma]. \tag{15}$$

Now estimators for the shape, scale and location parameters can be found by equating Equations (13) and (14).

There occurs problems with estimations using the PWM:s when the shape parameter is too large. This is similar to the method of moments which requires the shape parameter to be $\gamma < 1/3$ as proven by Cohen & Whitten (1982)[8]. In the case of PWM:s the efficiency decreases as $|\gamma|$ increases. It is through simulation apparent that for $|\gamma| < 0.4$ the probability weighted moments-method is more efficient when it come to bias and mean square error compared to the ML-method for smaller sample sizes[4].

4.3 Estimation in practice

A problem in practice could be that even with a high threshold the amount of exceedances that we will model from will be in the hundreds. This would be too time consuming to do by hand. Luckily there exists computational programs that allows computations to be completed significantly faster.

R is one of the most prominent software languages in academics. Being open-source and not needing any licensing it is very handy as it allows free sharing of packages that help streamline the methodology and efficiency in applied statistics. It is also one of the best environments for working with extreme values available open-source.

There are several packages that can be downloaded for R which handles POT-methods and parameter estimation for GPD distributions given some threshold. Among these are the *evd, evir, extRemes, fExtremes* and *POT*, to name a few[5].

The oldest of these packages is *evd*, designed by Stephenson in 2002 as a tool for analyzing extreme value distributions. It is modeled around maximum likelihood estimation of parameters and subsequently has good estimation for GPD parameters for the POT method for larger samples of exceedances. Being a very expanded package, *evd* also contains several graphical tools, as the medium residual life plot and parameter stability plots defined in Section 3.2.1 and Section 3.2.2 respectively.

For handling smaller samples of extreme values, estimation using probability weighted moments will be more accurate than using the MLE. *evd* does not support parameter estimation using PWM so it may be inaccurate, especially for particularly small samples. The *fExtremes* package is a modified version of *evd* and *evir* that allows estimation of stationary models using the PWM-method. Due to being a modified version of *evd* it carries over a lot of functionality and can still estimate variables for GEV and GPD distributions using the MLE.

5 Automobile insurance claims

To showcase the different methods of threshold setting and methods of modeling, analysis will be performed on a data set which can be found in the *insuranceData* R-package [3]. The data is taken from an insurance company based in the Midwest of the U.S.A. Each of the 6773 elements in the dataset represents a private person who is claiming their insurance due to some property damage to a vehicle and the amount the insurance company agreed to pay to cover the damages in U.S. dollars. The dataset further contains information about which state the customer lives in, giving them each a randomly assigned numerical code for anonymity. A risk class which uses an internal grading system which is not disclosed, but is based on several parameters such as age, marital status and the amount of usage of the vehicle. As well as basic information as the age and gender of the insured customer.

The values for the payouts vary between just below ten U.S. dollars up to sixty thousand dollars. No further information is given about the claims but it can be assumed that the payed amount depends on several variables such as the value of the car, the risk class of the costumer and the severity of the damage to the car. To model the extreme cases where the insurance company have to settle claims for larger sums of money the model will be based on all cars rather than separating them into classes, since the extremes among more expensive cars won't stand out as much relatively to other cars in the same price range.

Something to note is that the data only contains the claims which are closed with a payout from the insurance company. Some claims can be closed without a payout due to several reasons, so all inference done will be relative to the payout for approved claims and not all claims in general.

5.1 Setting a threshold for the insurance data

The central part of the POT-method is setting the threshold to produce exceedances from which to model. All methods introduced in Section 3 have been used to set a threshold so that they can be compared for further analysis. We still want to make sure that a balance is kept such as to limit both bias and variance from getting too large.

In Table 1 below, thresholds have been set based on the rules of thumb discussed in Section 3.1. There is a noticeable gap between using the 90th percentile and the rule $k = \sqrt{n}$ as can be observed best by looking at the number of exceedances. $k = \sqrt{n}$ has 82 exceedances which corresponds approximately to the 98.8th percentile, compared to the 677 exceedances of the 90th percentile. This leads to a much higher threshold for $k = \sqrt{n}$ as well as much larger variance due to the limited amount of exceedances.

.

Table 1: Values for threshold, the number of values that exceeds the threshold, as well as the ML estimates for the shape and scale parameters when using the discussed rules of thumb to determine threshold value. Parameters were estimated by using functions from the *evd*-package in R

	90th percentile	$k = \sqrt{n}$	$k = n^{2/3}/\log(\log(n))$
Threshold	4171.5	11390	8877
Number of exceedances	677	82	164
Scale estimate	3601	5543	4688
Shape estimate	0.097	0.157	0.134

Figure 2 features the mean residual life plot of the data. In it we can observe some traits described in Section 3.2.1. The plot looks somewhat linear up to a threshold of roughly 15000, but when the threshold reaches the range between 15000 and 20000 U.S. Dollars the behavior of the plot starts to shift. At that threshold there will be roughly 20 exceedances causing very high variance as the most extreme payouts in the data are far larger than that of the average claim.

Since the information of the MRL as the threshold gets large is not of interest we look at Figure 3 in which the x-axis has simply been restrained to a smaller range of thresholds. From this point the theory suggests that there should exist a point at which for all larger threshold the man excess will be a linear function of the threshold and the scale parameter.

This is somewhat hard to do without much experience setting thresholds due to the subjectivity required in setting a threshold based on something just by looking at a picture. Looking at Figure 3 a somewhat linear behavior can be seen starting at a threshold of around \$5000 however again around \$7500 it starts to show a linear behavior with a larger gradient. Which one of these that is better isn't easy to determine and both will presumably have advantages over the other. The increased gradient can however be explained by the reduced amount of exceedances at a higher threshold the largest payouts will have a large effect on the mean excess. This would explain an upward trend in the gradient. Information about the two thresholds are collected in Table 2.

.

Table 2: The number of values that exceeds the thresholds taken from the MRL plot, as well as the ML estimates for the shape and scale parameters. Parameters were estimated by using functions from the *evd*-package in R

Threshold	5000	7500
Number of exceedances	512	239
Scale estimate	3813	4345
Shape estimate	0.096	0.124

Figure 2: Empirical mean residual life plot for the insurance data. Plotted using mrlplot from the *evd* package in R. The solid line represents the empirical MRL with the dashed lines representing a 95% confidence interval.

Figure 3: A zoomed in figure of the empirical mean residual life plot for the insurance data plotted across lower threshold values to give better focus before the behavior of the MRL-plot becomes erratic.

Figure 4 and 5 show the parameter stability plots for thresholds up to \$20000 where they depict the estimated modified scale parameter $\hat{\beta}_u - \hat{\gamma}u$ and the estimated

scale parameter $\hat{\gamma}$ respectively. In practice, as mentioned in Section 3.2.2, we want to find a threshold for which the scale and modified shape estimators are approximately constant. Observing the modified scale plot it appears to be constant up until just above \$5000 where it starts to slope downwards breaking. It can be noted that the value of the modified scale around 5000 is contained within the 95% confidence interval of the estimate so there is a possibility that the estimate is constant past some certain point. Looking at the shape estimate plot in Figure 5 it does not behave like a constant at lower threshold values. As with the modified scale parameter plot the confidence interval leaves open the possibility that the estimate is constant after some threshold level. It is however hard to get a good interpretation from the parameter stability plots and other tools provide a more reliable estimation.

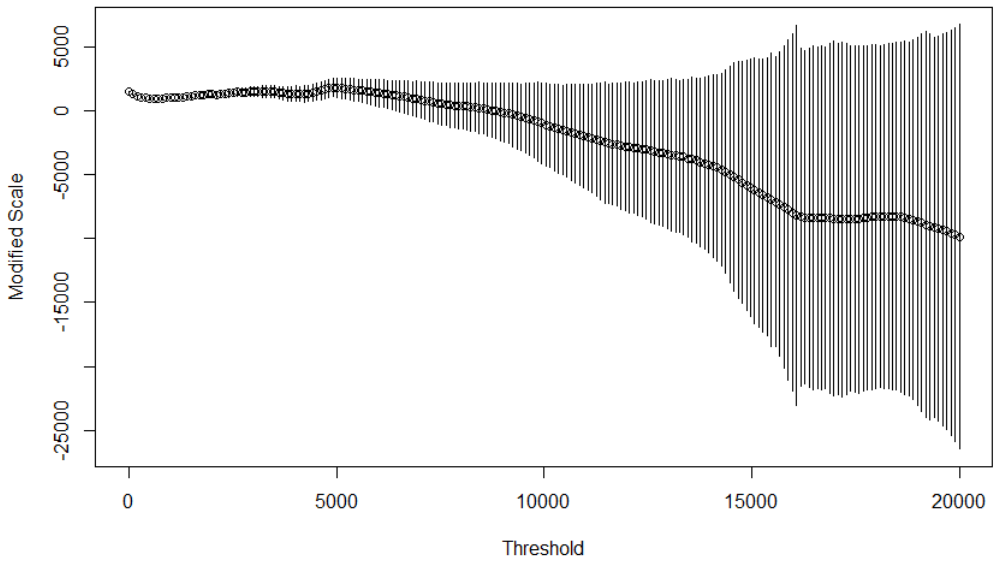


Figure 4: Parameter Threshold Stability Plot of the Modified scale parameter for threshold up to 20000. Dots represent the estimated Modified scale parameter for a GPD model for some threshold. The lines represent a 95% confidence interval. The plot was made with *tcplot* from the *evd* package in R.

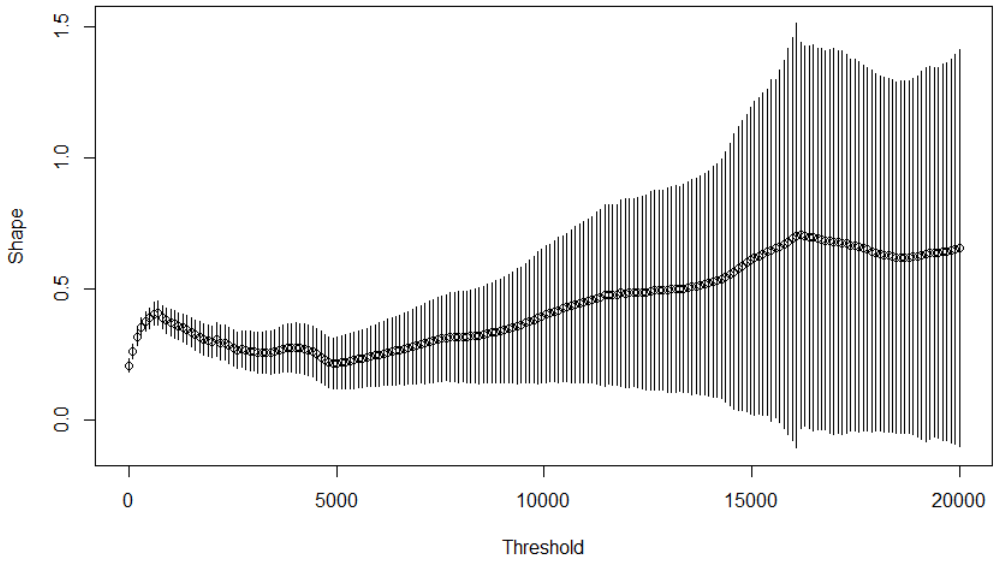


Figure 5: Parameter Threshold Stability Plot of the shape parameter for threshold up to 20000. Dots represent the estimated shape parameter for a GPD model for some threshold. The lines represent a 95% confidence interval. The plot was made with *tcplot* from the *evd* package in R.

5.2 Estimation of the GPD parameters

Having used some rules of thumbs for thresholds as well as looking at the MRL-plot of the sample we now have five potential thresholds to base our model choice on. Picking a suitable model will be done by choosing a threshold and modeling from the exceedances in the data sample. To make a threshold and subsequent model choice we look further into the parameter estimates of the different thresholds.

Table 3 is an extended version of Table 1 with the thresholds based on the three rules of thumb introduced in Section 3.1. Included in Table 3 is the estimated standard error for both the scale and shape parameter as estimated by the *evd*-package mentioned in Section 4.3. These estimates are thus using the ML-method and not the PWM-method. Since the method of PWM is more efficient for small samples but not neccesarily for larger samples we choose to use the ML-method instead since all samples in Table 3 are decently large.

Comparing the largest and smallest thresholds in Table 3, $u = 11390$ and $u = 4171.5$ respectively, we see that the error for the scale parameter is almost five times as large for the larger threshold. The difference between the errors in the shape parameter is not as large but still more than three times greater for the larger threshold. This is presumably due to the increased variance for larger thresholds mentioned in Section 2. Comparing the third threshold to the other two we see the same pattern. Since we want a model that balances both the bias and variation of the parameters we need to take this into account.

.

Table 3: Parameter and standard error estimates for thresholds based on rules of thumb. Estimated with the ML-estimators, using the fpot function in the *evd*-package in R

Threshold	4171.5	11390	8877
Number of exceedances	677	82	164
Scale estimate	3601	5543	4688
Standard error scale estimate	229	1091	657
Shape estimate	0.097	0.157	0.134
Standard error shape estimate	0.037	0.122	0.081

Looking at Table 4 we see the same information about the thresholds that was proposed after looking at the MRL-plot of the sample. Comparing the table with the estimates in Table 3 we see that for the two lowest thresholds the shape parameter is almost the same. We can correlate this with what was mentioned in Section 3.2.2 about the Parameter Stability Plots. The plots were used to find the lowest threshold where the shape parameter $\hat{\gamma}$ and the modified scale parameter $\hat{\beta}_u - \hat{\gamma}u$ are approximately constant. This would then be the threshold we would choose. This could indicate that a suitable threshold is close to the lowest of the proposed thresholds. However, the modified scale parameters for the two thresholds are not particularly close and there may be lower thresholds that still fulfill the criteria for the shape parameter estimate. Taking this as a sign that the proposed lower thresholds are the most suitable might thus be wrong, and we should be careful in making such an assumption.

.

Table 4: Parameter and standard error estimates for thresholds based on the Mean Residual Life plot. Estimated with the ML-estimators, using the fpot function in the *evd*-package in R

Threshold	5000	7500
Number of exceedances	512	239
Scale estimate	3813	4345
Standard error scale estimate	267	521
Shape estimate	0.096	0.124
Standard error shape estimate	0.041	0.062

5.3 Analysis of the insurance data

There have been five proposed thresholds and five models based on parameters estimated from the excesses of each threshold. However, choosing a model isn't necessarily straightforward and does not have to be objectively based around which fit is theoretically better.

The lower thresholds, $u = 4171.5$ and $u = 5000$, both show promising parameter estimates as the error is much smaller than that of the larger thresholds. On the other hand models based on the larger thresholds will be more descriptive of very extreme events as well as having lower bias due to the decreased sample size.

It may occur that we are more interested in some threshold depending on what we want to model. It could be that one of the lower thresholds give more accurate models yet still is not chosen. This is entirely dependent on the information we want from the sample. While the lowest proposed threshold might be a better model fit theoretically, we might still be more interested in the distribution of all claims above \$7500 or some larger value.

All parameter estimates in Section 5.2 were made with the *evd* package in R, which estimates using the Maximum Likelihood method. This decision was taken based on the amount of exceedances for all the thresholds. Since using the method of PWM in only proven to be a better estimator for very small samples. Our smallest sample of exceedances is $N(11390) = 82$, for the largest threshold we have chosen. How small a very small sample is has not been very clearly defined, so we have to make a subjective decision on what estimation to use. Thus we choose to only used the MLE to estimate the parameters. Using $fExtremes$ we could still look at the parameter estimates using the method of PWM. Due to some unknown difference in approximation, the estimate $fExtremes$ gives us will vary slightly from those of *evd*. The shape parameter will be slightly larger but still positive and contained within $|\gamma| < 0.4$, and the estimate for the shape parameter will be slightly smaller. This is most likely due to some differences in the baseline approximations between the packages.

Something not covered in this analysis is the relationship between the sample and the undisclosed sample with all claims that were closed without payment. It could have been of interest to further study the similarities between distributions for the samples as well as the combined sample of all claims made to the company.

6 Conclusion

The goal of this paper was to introduce some of the basic theory and concepts behind Extreme Value analysis and the POT-method and then apply it to the insurance data introduced.

Common methods of diagnostics for setting a threshold was introduced and briefly discussed. Since the parameter estimates of the GP distribution is dependent on threshold choice, different diagnostics and techniques was introduced to give a basis for choosing a suitable threshold. Rules of thumb was introduced, a method of threshold choice that does not require particular insight into the data for setting a threshold. Instead focusing on practicality it uses predetermined quantiles to determine a threshold at the cost of not analyzing, and not requiring familiarity with the data as much.

Some insight was given to graphical tools used to determine a suitable threshold. By design, graphical tools need more subjective knowledge of the tools since any result is based on finding some approximate pattern in a plot. This also means that unlike the rules of thumb there may occur cases where no information of value can be extracted from the graphical diagnostics.

Two methods of parameter estimation was introduced. The Maximum Likelihood-method and the method of Probability Weighted Moments was briefly touched upon, introducing some theory behind each method. Mentioning the key difference of the two methods being the better accuracy of the PWM-method for smaller samples. Further introducing practical methods of performing both methods using the software R. Giving some insight in the computational means to estimate the parameters of the GPD distribution given a threshold u for both the MLE-method and the PWM-method.

Lastly the dataset was introduced more broadly and diagnostics was performed on it using the tools provided in the earlier sections. From the graphical tools some approximate estimations are made, giving proposals for thresholds and subsequent model estimates. Followed by discussing the difference in the estimates and uncertainties of the model parameters for the different proposed threshold choices.

There are some further analysis that could be made if the whole sample of claims were released. The analysis is only performed on claims that were settled. Comparing the behavior of models on the data of all claims, settled or not, could have provided more insight regarding the data. More insight in the parameters of the risk class system, that was not touched upon further, could also have led to potentilly further analysis that might have been of interest. Dividing the sample into smaller samples depending on factors such as car type and value, risk class etc. could have given more insight.

7 Bibliography

References

[1] J. Franke, W.K. Härdle & C.M. Hafner, *Statistics of Financial Markets An introduction*, 2nd edition, Springer, 2008.

[2] C. Scarrott & A. MacDonald, A Review of Extreme Value Threshold Estimation and Uncertainty Quantification. *REVSTAT–Statistical Journal*, 10(1):33–60, 2012.

[3] A. Wolny-Dominiak & M. Trzesiok, **Package 'insuranceData'**, September 2014, **URL**, **https://cran.r-project.org/**.

[4] S.G. Coles & M.J. Dixon, Likelihood-Based Inference for Extreme Value Models, *Extremes* 2:1, 5-23, 1999.

[5] E. Gilleland, M. Ribatet & A.G. Stephenson, A software review for extreme value analysis, *Extremes* 16:103–119, 2013.

[6] J. Pickands, Statistical inference using extreme order statistics, *Ann Statist.* 3:119-131, 1975.

[7] R. L. Smith, Maximum Likelihood Estimation in a Class of Non-Regular Cases, *Biometrika*, 72: 67-92.

[8] A.C. Cohen & B.J. Whitten, Modified Maximum Likelihood and Modified Moment Estimators for the Three-Parameter Weibull Distribution, *Comm. Stat. Theor. Meth.* A11: 2631-2656, 1982.

[9] J.M. Landwehr, N.C. Matalas, J.R. Wallis, Probability Weighted Moments Compared with Some Traditional Techniques for Estimating Gumbel Parameters and Quantiles, *Wat. Res. Res.* 15: 1055-1064, 1979.