Modelling Large Claims in Property and Home Insurance - Extreme Value Analysis

Henrik Paldynski

Master's thesis
2015:E8

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

ABSTRACT

It is of paramount interest for insurance companies to have an estimate of the probability of being exposed to extremely large claims that could render them directly insolvent or decrease the size of their regulatory capital to the point of non-viability. The difficulty with finding such an estimate is that extreme events are by definition rare and therefore difficult to model. This study approaches the problem by utilizing methods developed in extreme value theory, a branch of statistics dedicated to the study of such extreme events.

The purpose of this study was to construct a model for the property and home insurance claim process for a specific insurance company, Folksam Skadeförsäkring Ab, based in Helsinki, Finland. The aim was to fit the data to the models proposed by extreme value theory and see whether these would describe the actual observations in a meaningful way. The quantiles of these fitted distributions and the associated confidence intervals would serve as a quantified guideline of the risks the company is exposed to. Furthermore, the distributions could be used to price simple types of reinsurance contracts, used as hedging tools by insurance companies.

Two sets of data were analysed, one containing the daily maxima and the other containing the summed daily claims. These were fitted using the maximum likelihood method to four interlinked, but separate models: the General Extreme Value distribution model for the block maxima and three threshold models, the General Pareto distribution, the Poisson-GPD model and the point process approach. Standard statistical tools were deployed to determine the goodness of fit for the difference models.

The first set of data was fairly well modelled by both the block maxima and threshold approaches, both severity and frequency. In addition to the range of quantiles and return levels, a conditional probability distribution was estimated to model the behaviour of claims given that they are larger than a predefined amount. Additionally a simulation study was performed, which gave an estimate of the distribution of aggregated daily maxima exceeding a certain threshold over a range of years.

The models did not provide a sufficient goodness of fit for the second data set. This is possibly explained by the weak autocorrelation found in the summed daily claims. The large confidence intervals ended up being the largest deficiency in the study, stemming from the relatively short period the data was collected from.

FOREWORD

" Il est impossible que l'improbable n'arrive jamais."
Emil Gumbel

This Master's thesis was written at the department of Mathematical Statistics at Lund Institute of Technology. It is the conclusion of the author's rather lengthy process of obtaining a Master of Science degree in Industrial Engineering and Management.

The author would like to thank Associate Professor Nader Tajvidi for supervising the thesis, giving valuable ideas for the analysis and originally giving the lecture which inspired the thesis. I'd also like to thank Associate Professor Magnus Wiktorsson for his support throughout my studies.

Furthermore, I would like to express my gratitude to Mr. Fredrik Nygård and his colleagues at Folksam Skadeförsäkring Ab for their cooperation, for supplying the data used in the thesis as well as help and consultation with the analysis.

Helsinki, 21.12.2014

TABLE OF CONTENTS

ABSTRACT I
FOREWORD II
TABLE OF CONTENTS III
1 INTRODUCTION 1
1.1 Background 1
1.2 Aim of Thesis 1
1.3 Background on Folksam Skadeförsäkring Ab 2
1.4 On Reinsurance 2
1.5 Outline of the Analysis 3
2 DESCRIPTION OF DATA 5
3 THEORY 7
3.1 Background 7
3.2 The Generalized Extreme Value Distribution 8
3.2.1 Inference for the GEV Distribution 9
3.3 Generalized Pareto Distribution and the Threshold model 13
3.3.1 Inference for Threshold Model 14
3.4 Poisson-GPD Model and the Point Process Characterization 17
3.4.1 Poisson-GPD Model 17
3.4.1.1 Inference on Poisson-GPD Model 18
3.4.2 Point Process Approach 19
3.4.2.1 Inference on Point Process Model 20
4 ANALYSIS AND RESULTS 21
4.1.1 Stationarity of Data 22
4.2 Modelling the Daily Maximum Claim 26
4.2.1 Generalized Extreme Value 26
4.2.2 Gumbel 28
4.2.3 Generalized Pareto Distribution 29
4.2.4 Poisson-GPD model 32
4.2.4.1 Simulation Study Using the Estimated Poisson-GPD Model 33
4.2.5 Point Process 35

4.3 Modelling the Sum of Claims per Day 37
4.3.1 GEV 37
4.3.2 Gumbel 39
4.3.3 Generalized Pareto Distribution 40
4.3.4 Poisson-GPD 42
4.3.4.1 Simulation Study Using the Estimated Poisson-GPD Model 43
4.3.5 Point Process 46
5 DISCUSSION AND CONCLUSIONS 47
5.1 Summary of the Results 47
5.2 Possible Sources of Uncertainty and Bias 48
5.3 Suggestions for Improvements and Future Research 50
6 APPENDIX 53

1 INTRODUCTION

1.1 Background

The basic proposition for a non-life insurance company is that the costs of expected claims are covered with paid insurance premiums and the return on capital for those. Additionally, the operational expenses have to be covered by the incoming cash flow as well. The insurance company needs to keep a certain amount of capital buffers, or initial reserves, for prudency and regulatory reasons in case the cash flows from the premiums momentarily do not cover unexpectedly large claims. These capital buffers consist of equity, accumulated retained earnings and certain forms of debt called hybrid debt that partly counts as equity. It is not merely enough to retain buffers that are large enough to withstand catastrophically large claims, but the remaining buffers after these must still be above certain regulatory requirements. These rules will be updated soon by the European Union directive called Solvency II, scheduled to come in to effect in the near future. Another possibility for an insurance company to withstand larger, improbable claims is to sign re-insurance agreements.

For the company to remain solvent in cases of extremely large insurance claims, a combination of capital buffers and reinsurance must be able to absorb the unexpected losses. As these possible capital shortfalls affect the very existence of the company, the probability of such large losses is obviously of paramount interest. The company could of course insure this tail risk away completely or keep very large capital buffers, but both of these solutions would sincerely harm the return on equity. In order to optimize the use of such cautionary measures, which are expensive to maintain, but provide a certain degree of security, an estimate for the occurrence of abnormally large or extreme claims, needs to be maintained.

1.2 Aim of Thesis

The purpose of the study is not to look at the average behaviour of the claims the company faces, but at the stochastic claim process's behaviour at its extremes. The challenge is to estimate what levels of claims might occur over the next 10 or 20 years, when data is only attainable for a shorter period.

The aim of this thesis is to construct such an estimate for a specific insurance company, Folksam Skadeförsäkring Ab. This estimate can be characterized by the return level over a certain period, i.e. the claim size which would be exceeded on average once during that period and the level of uncertainty in that figure. A second characterization would take the form of the quantiles of the estimated distribution function, that is the claim size which will be surpassed with a defined small probability during a relevant time

period. A further result which is of interest is the conditional distribution of excess losses over a certain threshold level.

It is not only the distribution of the size of single claims that is of interest, but also what the cumulative amount of claims might be under a certain time period. Thus also a measure for this will be produced.

These results can be relevant for the company in order to prepare for capital shortfalls with adequate capital buffers and/or determine the fair value of reinsurance contracts.

1.3 Background on Folksam Skadeförsäkring Ab

Changes in the ownership of the company whose data is studied in this thesis occurred during the writing process of this thesis. The majority owner of Aktia Skadeförsäkring changed from Aktia Group to the Sweden-based insurance company Folksam Sak in 2012. Aktia Group and Veritas Pensionsförsäkring continue as minority owners. The name was subsequently changed to Folksam Skadeförsäkring Ab (in the future referred to as Folksam in this thesis).

Folksam's premium income was 74 million euro in 2013, it had approximately 85 000 customers and 233 000 underwritten insurances. Folksam's customers are mainly located in the coastal areas of Finland plus a number of growth areas (Om oss: Folksam Skadeförsäkring Ab, 2014). The geographical distribution of their customers coincides with the footprint of the former majority owner. Aktia mainly caters to the Swedish speaking population of Finland, which is concentrated on the coastal areas.

Folksam is a non-life insurance company, which means that its product offering includes e.g. home, property, farm and forest insurance, but not life insurance, i.e. savings products. The product line consists of both statutory and voluntary insurances. The client base consists of natural persons and companies, especially in farming.

1.4 On Reinsurance

It is typical for insurance companies to buy insurance for risks in their portfolio they do not wish to retain. These outsourced risks can be for example hedging against catastrophic weather in the geographical area where the insurer is focused, sharing the risks of policies underwritten for key customers that incur too big possible losses for the insurer to bear on its own, or decreasing the variance of the underwriting results. It is also a way to increase the geographical presence in a regulated industry, e.g. if a multinational customer wants an insurance policy in a jurisdiction where the insurer isn't legally certified, the insurance company can sell reinsurance to a local, certified insurer who then underwrites the insurance. In some ways reinsurance caters to similar risk transfer needs as the Credit Default Swap in finance or the packaging of loans to

deal with balance sheet limitations in banking. Reinsurers are often large multinationals as a certain size is needed to achieve a critical level of risk differentiation and decrease credit risks to the insurance buyer.

There is a wide variety of reinsurance types, where the main distinction is between proportional, where a certain percentage of all policies is reinsured, i.e. the reinsurer effectively buys a portion of the premiums and liabilities from the cedant; and non-proportional contracts, where the reinsurer only assumes the risks beyond a certain threshold. Excess-of-loss (XL) reinsurance is a non-proportional contract where the reinsurer pays the excess over a certain fixed limit, up to a certain maximum level, for individual claim sizes. The lower limit is called the retention level, and the difference between the upper limit the reinsurer is liable to pay and this level is called the cover (Reiss & Thomas, 1997). Stop-loss reinsurance is similar to XL reinsurance but covers the excess of a certain limit for the total claim amount of an insurer's portfolio. Catastrophe excess of loss is conceptually in between the two former ones; a homogenous portfolio is reinsured against multiple losses stemming from one event (e.g. natural disasters) (Munich Reinsurance America, Inc., 2010). To formalize, the reinsurer's share of the claims, Z can be expressed as

$$Z = \begin{cases} 0 & X \leq R \\ (1-c)(X-R) & R < X < R+L, \\ (1-c)L & X \geq R+L \end{cases} \qquad (1.1)$$

where R is the retention level, L is the limit of the reinsurer's liability and c is the share of the excess that the cedant itself covers. X is the total claim amount in the contract period in the stop loss type, and the individual claim amount in the XL case (Rytgaard, 2006).

Net premium is the premium necessary to cover only anticipated losses, before loading to cover other expenses (Gulati, 2009). A simple pricing model for the expected value of Z, $E[Z]$, or net premium required by the reinsurer, could be constructed based on Equation (1.1) if the distribution of X was known.

1.5 Outline of the Analysis

Only parametric models are used in the thesis, largely due to their predominance in literature.

The block maxima, i.e. the maximum observation for a certain time period, will be fitted to the standard Generalized Extreme Value distribution using the Maximum Likelihood Method. The delta and profile likelihood methods will be used to determine confidence intervals for the estimated parameters and return levels. Probability, Quantile, Return Level and Density plots will be used for model checking as well as a likelihood ratio test for the viability of a simpler model, that is the Gumbel distribution.

As this “classical” approach to extreme values only use the block maxima for inference, and thus leave out possibly interesting data, so called threshold models will also be used. Here observations with value over a certain threshold are fitted to the Generalized Pareto distribution. The same procedure as for the GEV-distribution will be applied, with the added study of an appropriate threshold. Mean residual life plots will be used and the variability of the scale and shape parameter when the threshold is altered will be studied. Hopefully, smaller confidence bounds will be obtained than for the previous method.

Thirdly, a Poisson-GPD model will be fitted to the data, the rationale being that even more of the data (threshold exceedance rate) is used. Similar tools for fitting will be used as above. A probability distribution condition on the exceedance of a certain high level of claims will also be computed. Furthermore, a simple simulation study will be performed. Finally, a point process approach for modelling the data will be tried as well.

2 DESCRIPTION OF DATA

The raw data was drawn from the home and property insurance claims paid out by the company during the years 2005-2009. The date, type of insurance, reason for accident or damage and size of payment are listed for each single claim, along with other information which will not be used in this thesis. The reasons for the claims range from theft to natural phenomena, electrical failures, fires and leakages. It is clear that the larger claims stem almost solely from fire or leakage accidents. The rest of the large claims result from a natural phenomena, storm or vandalism. The data consists of 11 476 individual claims and the sum of all claims exceeds 40 million euro. The four largest claims made up approximately 10 % of the sum of all claims and the largest individual claim was 1,3 million euro. There were two faulty dates in the data: 20080931 and 20070431. These were changed to the previous day's date. The date represents the date of the accident that led to the claim; the claim could have been filed later and especially settled much later.

The dates refer to the date of the event that led to the claim. This can in some cases be an approximation or best guess, e.g. if the accident was not discovered immediately. This can in some cases distort the identification of related or clustered events such as storms. Another general difficulty in analyzing insurance data is that all claims are not immediately settled; the time to settle varies and the process can be fairly long in some cases. The data in question was however drawn sufficiently after the last claim was filed so that all claims from this period were settled and included in the data.

The insurance portfolio that generates the claims has stayed fairly constant in size through the five years in question. The number of individual insurance contracts has grown by under 1 % yearly. The number of claims per year has decreased by on average under a percentage per year. Also, the premium income generated from the portfolio has increased. However, according to discussions with a company employee knowledgeable in the subject, the composition of the insurance portfolio has not changed dramatically during the time period. That is the client base, product specifications and underwriting standards have been practically constant. The increase in premium income stems simply from a price increase. It is assumed that there is no significant trend, during the five years the data is from, in the occurrence of claims stemming from an increase in the amount of underwritten insurance contracts or the composition of the underlying portfolio.

Although the claim frequency will be assumed to be constant in this study, there can still be a trend in the claim severity. If the trend is a constant increase in claim severity, it is called claim inflation. Underlying reasons for this include a general increase in the price level (inflation) and/or more specific price increases like legal costs, fluctuations in

currency rates, changes in legislation, health care advances, car design, traffic density, changing weather patterns (frequency also) and above average price increases in property values. Claims inflation is a very important factor in the modelling of future liabilities, especially for long ones. This trend is also notoriously difficult to measure with any degree of certainty according to Sheaf, Brickman and Forster (2005). UK practitioners have put it somewhere around 3% to 5% yearly according to Gesmann, Rayees & Clapham (2013).

The harmonized consumer price inflation in Finland between 2005 and 2009 ranged between 0-4 percent on a yearly basis (Finlands officiella statistik, 2014). The yearly median claim size was used as a proxy for the whole data in order to study if any claims inflation might be visible and significant. The median did grow in three out of four cases. However the growth was only significant during one year as can be seen Table 1 below. Also, the correlation between the change of the median claim from year to year and the yearly consumer price inflation in Finland is low, below 50 %. This is natural since the HCPI lacks some components of claims such as legal costs and social inflation. Although the differences are perhaps accentuated for casualty insurance rather than property and home insurance, which is modelled in this paper, it is still relevant. The correlation to CPI varies e.g. in the US from 50 % (Auto Bodily Injury) to 80 % (Homeowners insurance). The correlation is also stronger during high inflation periods (Stephan, 2013). The average CPI in Finland during the period in question was only 1,8 %.

Table 1. Changes in portfolio and average inflation in Finland

Year	Change in median claim size from previous year	Average inflation between year and previous year
2006	-10,5 %	0,8 %
2007	16,6 %	2,0 %
2008	0,7 %	3,2 %
2009	3,3 %	2,0 %

Even though it would be computationally straightforward to correct the claims data for inflation with the assumption that the consumer price inflation would be a good (and best readily available) approximation for claims inflation, the data was not modified at all. The data series stems from a relatively short time period (five years) so the effect of inflation is not as crucial as for a long data series. Furthermore, due to the abovementioned difficulties with estimating claims inflation, the usefulness of modifying the data with CPI is unclear.

3 THEORY

3.1 Background

Extreme value theory takes its foundations from the work by Fisher and Tippet (1928). The basic ideas and notations are presented below. Much of the theoretical background and inference methodology used in this thesis can be found in Coles (2001). The aim of the theory section is to present the basic theory and inference methods as thoroughly as possible to the reader, in a step by step fashion.

Let $X_1, \dots, X_n$ be a sequence of independent and identically distributed random variables with distribution function F and define

$$M_n = \max\{X_1, \dots, X_n\}$$

The distribution of M_n can then be derived:

$$Pr\{M_n \leq z\} = Pr\{X_1 \leq z, \dots, X_n \leq z\}$$

$$= Pr\{X_1 \leq z\} \times, \dots, \times Pr\{X_n \leq z\}$$

$$= \{F(z)\}^n$$

The problem is that the function F is not known in practical applications. Although F could be estimated using standard statistical methods and then inserting the estimate into the above equation, very small variations in the estimate of F lead to large variations in F^n. The *extremal types theorem* allows for an alternative approach, where the characteristics of an asymptotic distribution for the maxima are defined. F^ncan be approximated by the limiting distribution as $n \to \infty$. The problem is that if $F(z) < 1$, then $F(z)^n \to 0$, as $n \to \infty$. This can be avoided by seeking the limit distributions for

$$M_n^* = \frac{M_n - b_n}{a_n},$$

for suitable sequences of constants $\{a_n > 0\}$ and $\{b_n\}$. The range of possible limiting distributions for M_n^* is given by the extremal types theorem.

If there exists such constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\left\{\frac{(M_n - b_n)}{a_n} \leq z\right\} \to G(z) \text{ as n} \to \infty,$$

where G is a non-degenerate distribution function, then G belongs to one of the following families:

$$I: G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, -\infty < z < \infty$$

$$II: G(z) = \begin{cases} 0, & z \le b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b \end{cases}$$

$$III: G(z) = \begin{cases} \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z < b \\ 1, & z \ge b \end{cases}$$

for parameters $a > 0$ and b and, in the case of families II and III, $\alpha > 0$. These distributions are called extreme value distributions and the different types are known as Gumbel, Fréchet and Weibull families respectively. What the theorem implies is that if M_n can be stabilized with suitable sequences $\{a_n > 0\}$ and $\{b_n\}$, the extreme value distributions are the only possible limiting distributions for the normalized maxima M_n^*, regardless of the distribution F for the population.

3.2 The Generalized Extreme Value Distribution

The three different families give quite different characteristics to the tail behaviour of the model. The families can however be combined into a single family of models having distribution functions of the form

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \qquad (3.1)$$

defined on the set $\{z: 1 + \xi(z-\mu)/\sigma > 0\}$ and the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. This is called the Generalized Extreme Value (GEV) distribution with location parameter μ, scale parameter σ and shape parameter ξ. The subset with $\xi = 0$ is interpreted as the limit of the Equation 3.1, as $\xi \to 0$, resulting in the Gumbel family with distribution function

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, -\infty < z < \infty \qquad (3.2)$$

The shape parameter is dominant in describing the tail behaviour of the distribution function F for the original random variable X_i. $\xi > 0$ corresponds to the Fréchet case where the density of G decays polynomially that is, it has a "heavy" tail and the upper end point is infinite. The heavy tail often fits e.g. economic impacts and precipitation. In $\xi = 0$, the Gumbel case, the density decays exponentially, it has a "light" tail and the upper end point is also infinite. $\xi < 0$ corresponds to the Weibull case where the distribution is finite, which describes well many real-world processes, e.g. temperatures, wind speeds, sea levels and insurance claims (Katz, 2008).

3.2.1 Inference for the GEV Distribution

The limiting distribution is in practical inference considered as an approximation for large values of n, i.e. for maxima of long sequences. In practice, the sequences of normalizing constants do not have to be determined. Assume:

$$\Pr\left\{\frac{(M_n - b_n)}{a_n} \leq z\right\} \approx G(z)$$

for large enough n. Equivalently,

$$\Pr\{M_n \leq z\} \approx G\left(\frac{(z - b_n)}{a_n}\right) = G^*(z)$$

where G^* is a GEV distribution with different parameters. So if the distribution of M_n^* can be approximated by a member of the GEV family for large n, then the distribution of M_n itself can be approximated by a different member of the same family (Coles, An Introduction to Statistical Modeling of Extreme Values, 2001).

Likelihood methods can be used to estimate the parameters of the limiting distribution, but caution must be used. Maximum-likelihood estimators can lack asymptotic normality if the set of data values which has positive probability (or positive probability density) depend on the unknown parameter (Coles & Davidson, Statistical Modelling of Extreme Values, 2008). In the case of the GEV-model, for example when $\xi < 0$ the upper end-point is $\mu - \sigma/\xi$, and thus dependent on the parameter values. According to Smith (1985) likelihood methods have the usual asymptotic properties when $\xi > -0{,}5$.

Choosing the appropriate block size where the maxima are taken from (e.g. quarterly or yearly maxima) involves a trade of between the variance of the model and the systematic bias. The problem is similar than for choosing the threshold for threshold models described later. Taking large block sizes means that the maxima are drawn from many underlying observations, thus making the asymptotic argumentation more valid. At the same time large block sizes lead to fewer data points that can be used in the inference leading to larger variance in the estimation. The opposite then holds for choosing smaller block sizes.

If the block maxima $Z_1, \ldots, Z_n$ are independent, having the GEV distribution, the likelihood for the GEV-distribution when $\xi \neq 0$ is

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{m} \frac{1}{\sigma}\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1-\frac{1}{\xi}} \exp\left\{-\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

As is it usually more convenient to work with the log-likelihood function

$$\ell(\mu,\sigma,\xi) = m\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\log\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-1/\xi} \quad (3.3)$$

with the condition $1+\frac{\xi(z_i-\mu)}{\sigma} > 0$, for $i = 1, \dots, m$

When $\xi = 0$, the log-likelihood is

$$\ell(\mu,\sigma) = -m\log\sigma - \sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma}\right) - \sum_{i=1}^{m}\exp\left\{-\left(\frac{z_i-\mu}{\sigma}\right)\right\} \quad (3.4)$$

The maximum likelihood estimator maximizes both the log-likelihood and likelihood functions as the logarithmic function is monotonic. If the usual asymptotic are valid, the distribution of the estimated parameters, $(\hat{\mu},\hat{\sigma},\hat{\xi})$ is multivariate normal with mean (μ,σ,ξ) and variance-covariance matrix equal to the inverse of the observed information matrix evaluated at the maximum likelihood estimate (MLE).

As we are studying the possibility of very large claims, it is the quantiles of the estimated distribution that are of special interest. If we define the return level, z_p, as the value that is exceeded with probability p, that is $G(z_p) = 1-p$, then the inverse of the cumulative distribution function is

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\{-\log(1-p)\}^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma\log\{-\log(1-p)\}, & \xi = 0 \end{cases} \quad (3.5)$$

Then maximum likelihood estimate of the return level is $\hat{z}_p$, and is obtained by inserting the maximum likelihood estimator $(\hat{\mu},\hat{\sigma},\hat{\xi})$ into the above equation. z_p is called the return level associated with the return period $1/p$, since the level is expected to on average be exceeded once every $1/p$ periods.

The uncertainty in z_p can be obtained by the delta method and the profile likelihood. The variance using the delta method is

$$\mathrm{Var}(z_p) \approx \nabla z_p^T V \nabla z_p \quad (3.6)$$

where V is the variance-covariance matrix of $(\hat{\mu},\hat{\sigma},\hat{\xi})$ and

$$\nabla z_p^T = \left[\frac{\partial z_p}{\partial\mu}, \frac{\partial z_p}{\partial\sigma}, \frac{\partial z_p}{\partial\xi}\right]$$

evaluated at the maximum likelihood estimator.

As stated above, the profile likelihood can also be used to obtain confidence intervals for the return levels and also for the maximum likelihood estimations of the parameters of the GEV distribution. First the deviance function needs to be defined

$$D(\theta) = 2\{\ell(\hat{\theta}_0) - \ell(\theta)\} \qquad (3.7)$$

If $x_1, \ldots, x_n$ are independent realizations from a parametric family distribution $\mathcal{F}$ and $\hat{\theta}_0$ denotes the maximum likelihood estimator of a d-dimensional model parameter θ_0, then for large n and under suitable regularity conditions the deviance function asymptotically follows

$$D(\theta_0) \dot{\sim} \chi^2_d$$

Then an approximate confidence region is given by

$$C_\alpha = \{\theta : D(\theta) \leq c_\alpha\}$$

where c_α is the $(1-\alpha)$ quantile of the chi-squared distribution. This approximation tends to be more accurate than the one based on asymptotic normality of the maximum likelihood estimator.

The profile log-likelihood for $\theta^{(1)}$ is defined as

$$\ell_p(\theta^{(1)}) = \max_{\theta^{(2)}} \ell(\theta^{(1)}, \theta^{(2)}),$$

where $\theta^{(1)}$ is the k-dimensional vector of interest and $\theta^{(2)}$ comprises the remaining $(d-k)$ parameters.

Again if $x_1, \ldots, x_n$ are independent realizations from a parametric family distribution $\mathcal{F}$ and $\hat{\theta}_0$ denotes the maximum likelihood estimator of a d-dimensional model parameter $\theta_0 = (\theta^{(1)}, \theta^{(2)})$. Then under suitable regularity conditions and for large n

$$D_p(\theta^{(1)}) = 2\{\ell(\hat{\theta}_0) - \ell(\theta^{(1)})\} \dot{\sim} \chi^2_d$$

The above result can not only be used for determination of confidence intervals for single parameters or combinations of those in the maximum likelihood estimation, but also for model testing in the form of a likelihood ratio test.

Let $\mathcal{M}_0$ with parameter $\theta^{(2)}$ be a sub model of $\mathcal{M}_1$ with parameter $\theta_0 = (\theta^{(1)}, \theta^{(2)})$ under the constraint that the k-dimensional sub vector $\theta^{(1)} = 0$. Also let $\ell_0(\mathcal{M}_0)$ and $\ell_1(\mathcal{M}_1)$ be the maximized values of the log-likelihood for the models. Then model $\mathcal{M}_0$ can be rejected at the significance level α in favour of $\mathcal{M}_1$ if

$$D = 2\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\} > c_\alpha$$

The inverted GEV distribution function above is expressed as a function of the return level instead

$$\mu = \begin{cases} z_p + \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], & for\ \xi \neq 0 \\ z_p + \sigma \log\{-\log(1-p)\}, & for\ \xi = 0 \end{cases} \quad (3.8)$$

and then the resulting likelihood function $\ell(\mathrm{z_p}, \sigma, \xi)$ is maximized with respect to the new parameters for a range of return levels. The confidence interval is then computed as $\ell^{-1}\left(\ell(\hat{\theta}_0) - \frac{1}{2}c_\alpha\right)$, where c_α is the $(1-\alpha)$ quantile of the χ_1^2 distribution. Since the exact computation of the inverse log-likelihood function proved laborious, the mean of the two closest z_p was taken instead.

A wide range of graphical techniques can be employed for goodness of fit purposes. Given an ordered sample of independent observations

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

from a population with estimated distribution function $\hat{F}$. Then the probability plot consists of the points

$$\left\{\left(\hat{F}(x_{(i)}), \frac{i}{n+1}\right) : i = 1, \ldots, n\right\}$$

The estimated distribution should coincide with the empirical distribution function $\frac{i}{n+1}$ to a reasonable level and thus the points in the plot should lie around the unit diagonal. The quantile plot consists of the points

$$\left\{\left(\hat{F}^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right) : i = 1, \ldots, n\right\}$$

Again, if $\hat{F}$ is not a valid representation of F, the points will not gather around the unit diagonal.

The problem with the probability plot for the GEV model is that both the empirical distribution function $\tilde{G}(z_{(i)})$ and the estimated one, $\hat{G}(z_{(i)})$ tend to approach 1 for large values of $z_{(i)}$. This is unfortunate, as it is exactly for large values of $z_{(i)}$ where the correctness of the estimate is interesting. The quantile plot consists of empirical and modelled estimates of the $\frac{i}{n+1}$ quantiles of $F$and does not have the aforementioned problem. In extreme value applications the goodness of fit is also obviously most interesting around the higher quantiles.

The return level plot consists of the maximum likelihood estimates of the return level, $\hat{z}_p$, against $-log(1-p)$, $0 < p < 1$, on a logarithmic scale. The shape parameter defines the shape and of the plot and upper bound of the return level. The plot is linear in the case $\xi = 0$ and concave if $\xi > 0$, both lacking an upper bound. The plot is convex if $\xi < 0$, with an asymptotic limit as $p \to 0$ at $-\sigma/\xi$. The return level plot with confidence

intervals coupled with empirical estimates for the return level also functions as a model testing tool. The model based curve and empirical estimates should coincide to some level for the model to be accepted. The delta method was used to calculate the confidence intervals for the return level in the return level plot.

Lastly, the modelled probability density function is plotted together with a histogram of the data. Unfortunately, as there is no objective way of choosing the grouping intervals for the histogram, and this largely subjective choice having a large impact on the histogram, the plot is rendered less useful than the three previous ones mentioned.

3.3 Generalized Pareto Distribution and the Threshold model

Perhaps the greatest drawback of the block maxima approach is that it is terribly wasteful with data, especially considering that extreme events are rare by definition. Often more observations than only the maxima are available, but these data points, possibly informative about the extreme behaviour of the process in question, are disregarded completely in the classical extreme value approach. One improvement is to take into consideration not only the maxima, but more of the most extreme observations. This is called the r largest order statistics model, where a certain number, r, of the largest observations in each block are used. As more data is used, the variance should be decreased, but a certain extent of bias might be introduced into the model, as the asymptotic argument becomes weaker. Although more data is now incorporated into the inference useful data is still omitted, as only a preset number of observations are considered extreme, many useful data points are disregarded in periods with unusually many large observations. The following is very much based on Coles (2001).

A different method for choosing what data to incorporate into the analysis for greater precision is the so called threshold method, where an appropriate threshold, u, is selected, and all the points above that threshold are used in the analysis.

Let $X_1, \dots, X_n$ be a sequence of independent and identically distributed random variables with distribution function F and denote an arbitrary term in the sequence X. It follows that

$$Pr\{X > u + y \mid X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \qquad y > 0 \qquad (3.9)$$

Similarly to the block maxima case, if F was known, the distribution of the threshold exceedances could be calculated exactly, but in practical applications this is often not the case. Therefore asymptotic arguments are used, similar to the GEV-model in the block maxima case. With the same assumption for X as above and let

$$M_n = \max\{X_1, \dots, X_n\}$$

If for large n

$$Pr(M_n \leq z) \approx G(z)$$

for $\mu, \sigma > 0$ and ξ, where $G(z)$ is the GEV-distribution function. Then, with an appropriate choice of u, the distribution function of $(X - u)$, conditional on $X > u$, is approximately

$$H(y) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi} \qquad (3.10)$$

defined on the set $\{y: y > 0 \text{ and } (1 + \xi\, y/\sigma)\}$, where

$$\sigma = \sigma_{GEV} + \xi(u - \mu).$$

The above distribution is called the Generalised Pareto Distribution (GPD). The limit $\xi \to 0$ can be considered for finding out the distribution in the special case $\xi = 0$, resulting in

$$H(y) = 1 - exp\left(-\frac{y}{\sigma}\right), \qquad y > 0 \qquad (3.11)$$

This means that if the block maxima have an approximating distribution belonging to the GEV-distribution family, then the threshold exceedances have an approximating distribution belonging to the GP-distribution family. The parameters of this GPD are also determined by the parameters of the GEV-distribution for the corresponding block maxima. The shape parameter ξ equals the one in the corresponding GEV-model and hence also determines to a large extent the properties of the GPD, as it does for the GEV-distribution.

3.3.1 Inference for Threshold Model

The threshold selection obviously plays a paramount role in creating a useful and viable model. A threshold that is low means more data points are used in the analysis, thus decreasing the variance of the result. On the other hand the asymptotic argument for the model is weakened at the same time as the definition of an extreme event is relaxed in a way, leading to bias in the result. Two methods for facilitating the choice of threshold are used.

The first method is based on the mean residual plot and can be used prior to parameter estimation. It is based on the following argument. Provided that the GPD is a valid model for the exceedances over u_0 generated from the series $X_1, \dots, X_n$ and that $\xi < 1$. Then based on the mean of the GPD we get

$$E(X - u_0 \mid X > u_0) = \frac{\sigma_{u_0}}{1 - \xi}$$

If the GPD is valid for the threshold u_0, it should also be valid for thresholds $u > u_0$. So

$$E(X - u \mid X > u) = \frac{\sigma_u}{1 - \xi} = \frac{\sigma_{u_o} + \xi u}{1 - \xi} \qquad (3.12)$$

where $\sigma_u = \sigma_{u_0} + \xi(u - \mu)$ from above is used. So for $u > u_0$, $E(X - u \mid X > u)$ should be a linear function of u. The sample mean of the threshold exceedances provides an empirical estimate for $E(X - u \mid X > u)$ and so the following locus of points can be used for threshold choice.

$$\left\{\left(u, \frac{1}{n_u}\sum_{i=1}^{n_u}\left(x_{(i)} - u\right)\right) : u < x_{max}\right\}$$

The plot should be approximately linear in u above the threshold where the generalized Pareto distribution should be a valid approximation of the distributions of the exceedances. Apart from these points, confidence intervals are added based on the approximate normality of sample means.

The second method is based on choosing u as the smallest value for a range of thresholds that give rise to roughly constant estimated parameters (sampling variability will lead to some differences). More exactly, if a GPD is a reasonable model for excesses of a threshold u_0, the excesses of a higher threshold u should also be distributed like a GPD. The shape parameter, ξ, is invariant of choice of threshold. The scale parameter changes with the threshold unless $\xi = 0$ as seen below for $u > u_0$.

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0) \qquad (3.13)$$

This problem can be circumvented with the reparameterization of the scale parameter as

$$\sigma^* = \sigma_u - \xi u$$

So the estimates of both σ^* and ξ should be roughly constant against for the threshold values that are suitable. Confidence intervals for $\hat{\xi}$ are obtained directly from the variance-covariance matrix. Confidence intervals for $\hat{\sigma}^*$ are obtained by using the delta method.

Maximum likelihood is again used for parameter estimation for the generalized Pareto distribution. If the k exceedances over a threshold u are marked $y_1, \ldots, y_k$, then the log-likelihood for $\xi \neq 0$ is

$$\ell(\sigma, \xi) = k \log\sigma - (1 + 1/\xi)\sum_{i=1}^{k} \log\left(1 + \xi\, {y_i}/{\sigma}\right) \qquad (3.14)$$

given that $(1 + \xi\, y_i/\sigma) > 0$ for $i = 1, \ldots, k$. If $\xi = 0$ the log-likelihood is

$$\ell(\sigma) = k \log\sigma - \frac{1}{\sigma}\sum_{i=1}^{k} y_i \qquad (3.15)$$

Again, the quantiles of the distribution is more interesting than the parameters by themselves. As in the block maxima approach, a model for the return level is derived. So suppose again that the GPD is a suitable model for the exceedances over a threshold u, then if $\xi \neq 0$

$$Pr(X > x \mid X > u) = \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}$$

Assuming $x > u$, then

$$Pr(X > x \mid X > u) = \frac{Pr(X > x \cap X > u)}{Pr(X > u)} = \frac{Pr(X > x)}{Pr(X > u)}$$

It follows that

$$Pr(X > x) = \zeta_u\left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi} \qquad (3.16)$$

where $\zeta_u = \Pr\{X > u\}$. So the level x_m that is on average exceeded once every m observations is given by

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right] \qquad (3.17)$$

If $\xi = 0$ then x_m is given by

$$x_m = u + \sigma\log(m\zeta_u) \qquad (3.18)$$

The return levels are often more easily understood on an annual scale, i.e. the N-year return level is expected to be exceeded once every N years. So if there are n_y observations per year, m is replaced by $N \times n$ in order to get the N-year return level z_n.

The natural estimator of the probability that an individual observation exceeds the threshold u is $\hat{\zeta}_u = \frac{k}{n}$, i.e. the proportion of sample points exceeding the threshold.

Confidence intervals for x_m can be derived using the delta method again. Since the number of exceedances over u follows the binomial Bin(n, ζ_u)-distribution, the variance of the estimator can be derived to be $\mathrm{Var}(\hat{\zeta}_u) \approx \hat{\zeta}_u\left(1 - \hat{\zeta}_u\right)/n$. The complete variance-covariance matrix for $\left(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi}\right)$ can then be approximated.

Profile likelihoods give again rise to better estimates for the uncertainty in the parameters and the return levels.

Similar plots are constructed for model checking as in the block maxima analysis.

3.4 Poisson-GPD Model and the Point Process Characterization

The models introduced so far only take into account the magnitudes of the extreme events. The times at which the events occur can also provide important information to make the inference more exact. Two methods of incorporating a time-dimension into the model are considered, where the first, the Poisson-GPD model is a special case of the second, the point process characterization of extreme events and thus mathematically equivalent. The Poisson-GPD model is closely connected to insurance industry, as it is similar to the claim process in the Cramér–Lundberg model (Lundberg, 1903). This model figures in early ruin theory, which is the study of how insurance companies are exposed to insolvency risks.

3.4.1 Poisson-GPD Model

The number of exceedances is distributed according to a binomial distribution that can be replaced by the Poisson distribution under certain conditions (Reiss & Thomas, 1997) Given random variables $X_1, \dots, X_n$, we may write $K = \sum_{i<n} I(X_i > u)$, where $I(X_i > u)$ is an indicator function with $I(X_i > u) = 1$ if $X_i > u$ holds and zero if it doesn't. If the X_i are i.i.d. random variables, then

$$P\{K = k\} = \binom{n}{k} p^{k(1-p)^{n-k}} =: B_{n,p}\{k\}, \;\; k = 0, \dots, n \qquad (3.19)$$

where $B_{n,p}$ is the binomial distribution with parameters n and $p = 1 - F(u)$. The mean number of exceedances over u is

$$\Psi_{n,F}(u) = np = n(1 - F(u)), \qquad (3.20)$$

which is a decreasing mean value function.

Let $X_1, \dots, X_n$ be a sequence of independent and identically distributed random variables and the indices i of the exceedances of $X_i > u$ are observed and rescaled to points i/n. These points can be viewed as a process of rescaled exceedance times on [0,1]. If $n \to \infty$ and $1 - F(u) \to 0$, such that $n\big(1 - F(u)\big) \to \lambda \;\; (0 < \lambda < \infty)$, the process converges weakly to a homogenous Poisson process on [0,1] with intensity λ. The model is constructed based on a limiting form of the joint point process of exceedance times and exceedances over the threshold. The number of exceedances in let us say one year, N, follows a Poisson distribution with mean λ and the exceedance values $Y_1, \dots, Y_n$ are i.i.d from the GPD (Smith R. , Statistics of Extremes, with Applications in Environment, Insurance and Finance, 2004). So, supposing $x > u$, the probability of the annual maximum being less than x for the GPD-Poisson model is

$$Pr\left\{\max_{1 \le i \le N} Y_i \le x\right\} = Pr\{N = 0\} + \sum_{n=1}^{\infty} Pr\{N = n,\, Y_1 \le x, \dots, Y_n \le x\}$$

$$= e^{-\lambda} + \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \cdot \left\{ 1 - \left(1 + \xi \frac{x-u}{\sigma} \right)_+^{-1/\xi} \right\}^n$$

$$= exp\left\{ -\lambda \left(1 + \xi \frac{x-u}{\sigma} \right)_+^{-1/\xi} \right\} \qquad (3.21)$$

If the following substitutions are made

$$\sigma = \sigma_{GEV} + \xi(u - \mu) \text{ and}$$

$$\lambda = \left(1 + \xi \frac{u-\mu}{\sigma_{GEV}} \right)^{-1/\xi}$$

the distribution reduces to the GEV-form. Hence, the two models are consistent with each other above the threshold u.

Much of the following analysis is based on the work of Rootzén & Tajvidi (1997). According to the authors, the GPD-Poisson model is stable under an increase of the level, i.e. if the excesses over the level u occur as a Poisson process and the sizes of the excesses are distributed according to a GPD and are independent, then the excesses over a higher level $u + v$ for $v > 0$ have the same properties. The distribution function of excesses over $u + v$ can be shown to be

$$H(y) = 1 - \left(1 + \xi \frac{y}{(\sigma + v\xi)} \right)_+^{-1/\xi} \qquad (3.22)$$

3.4.1.1 Inference on Poisson-GPD Model

Suppose we have N observations above the threshold in time T. The log-likelihood function of the Poisson-GPD model is then

$$l_{N,Y}(\lambda, \sigma, \xi) = N \log \lambda - \lambda T - N \log \sigma - \left(1 + {}^1/_{\xi} \right) \sum_{i=1}^{N} \log \left(1 + \xi \frac{Y_i}{\sigma} \right) \qquad (3.23)$$

The quantile Probable Maximum Loss, PML, for the risk level p and time N can be derived in the much the same way as the N-year return level in the GEV and GP cases. Using the same notation it is found out to be

$$PML_{N,p} = u + \frac{\sigma}{\xi} \left[\frac{(\lambda N n)^{\xi}}{\left(-log(1-p) \right)^{\xi}} - 1 \right] \qquad (3.24)$$

where n is the number of observations per year N. The N-year return level x_N can then be computed using the above formula as well.

Rootzén and Tajvidi also point out that the median of the excesses over the level $u + v$ is given by the formula

$$m(u+v) = \frac{\sigma + v\xi}{\xi}\left(2^{\xi} - 1\right) \qquad (3.25)$$

So the median of the excess over the limit increases with v.

Equation 3.25 can also be interpreted in a second way according to the authors. It also gives the median of the distribution of the size of the next claim which is larger than the largest claim so far, $m_{X_{maxnew}}$

$$m_{X_{maxnew}} = X_{max} + \frac{\sigma}{\xi}\left(2^{\xi} - 1\right) + (X_{max} - u)\left(2^{\xi} - 1\right) \qquad (3.26)$$

3.4.2 Point Process Approach

The times of exceedance occurrences and the magnitude of the exceedances are considered as two separate processes in the Poisson-GPD model. These are combined into a single process in the point process approach. This process behaves like a non-homogeneous Poisson process under suitable normalisations, according to the asymptotic theory of threshold exceedances. It is not in the scope of this paper to go thoroughly through the limit arguments.

A general non-homogeneous Poisson process on domain $\mathcal{D}$ is defined by an intensity density function $\lambda(x), x \in \mathcal{D}$, such that if A is a measurable subset of $\mathcal{D}$ and $N(A)$ denotes the number of points in A, then $N(A)$ has a Poisson distribution with mean

$$\Lambda(A) = \int_A \lambda(x)dx$$

where $\Lambda(A) = E\{N(A)\}$ is called the intensity measure.

The likelihood function is

$$exp\left\{-\int_{\mathcal{D}} \lambda(t,y)dt\,dy\right\}\prod_{i=1}^{N}\lambda(T_i, Y_i) \qquad (3.27)$$

for a process observed on the domain $\mathcal{D}$, with intensity $\lambda(t,y)$ and $(T_1, Y_1) \ldots (T_N, Y_N)$ being the N observed points of the process.

We again interpret a limiting model as a reasonable approximation for large sample behaviour. So if $X_1, \ldots, X_n$ are iid random variables and let

$$N_n = \{(i/(n+1), X_i): i = 1, \ldots, n\}$$

For sufficiently large u, on the regions of form $\mathcal{D} = (0,T) \times [u, \infty)$, N_n is approximately a Poisson process. The intensity density function on $A = [t_1, t_2] \times (y, \infty)$ is then

$$\lambda(y,t) = \frac{1}{\sigma}\left(1 + \xi\frac{y-\mu}{\sigma}\right)^{-1/\xi-1} \quad (3.28)$$

and the intensity measure

$$\Lambda(A) = (t_2 - t_1)\left(1 + \xi\frac{y-\mu}{\sigma}\right)^{-1/\xi} \quad (3.29)$$

defined on $\{1 + \xi(y-\mu)/\sigma\} > 0$.

The point process model unifies all the three models mentioned before, as they can all be derived from it. Although not utilized in this paper, this characterization is especially beneficial when non-stationarity is introduced into the model.

3.4.2.1 Inference on Point Process Model

Using the above likelihood function for Poisson processes (3.27) the log likelihood is

$$l(\mu,\sigma,\xi; y_1,\dots,y_n) = -n_y\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi} + \sum_{i=1}^{N(A)} \log\left\{\frac{1}{\sigma}\left(1 + \xi\frac{y_i-\mu}{\sigma}\right)^{-1/\xi-1}\right\} \quad (3.30)$$

when $\{1 + \xi(y_i-\mu)/\sigma\} > 0$ for all $i = 1,\dots,k$ (Smith R. , Statistics of Extremes, with Applications in Environment, Insurance and Finance, 2004). n_y is the number of periods of observations, the adjustment made to change the estimated parameters into a more comprehendible model. That is if omitted and data has been observed for e.g. m years the parameters of the point process model will correspond to the GEV-distribution of the m-year maximum. Usually distributions of annual maxima are preferred.

The R-package *in2extRemes* uses a diagnostic plot called the z-plot. The Z-plot was originally proposed by Smith & Shively (1995). The times when then observations exceed the threshold, $T_k, k = 1,2\dots$ are obtained, beginning at time T_0 and the Poisson parameter (which may be time dependent) is integrated from exceedance time $k-1$ to k so that the random variables Z_k are obtained.

$$Z_k = \int_{T_{k-1}}^{T_k} \lambda(t)dt = \int_{T_{k-1}}^{T_k} \{1 + \xi(t)\big(u-\mu(t)\big)/\sigma(t)\}^{-1/\xi(t)}dt, k \geq 1 \quad (3.31)$$

The random variables Z_k should by construction be independent and exponentially distributed, with mean 1. Thus the quantile-quantile plot of Z_k against the expected quantile values from an exponential distribution function with mean one functions as a diagnostic tool. If the threshold is e.g. too high so that the frequency of occurrences cannot adequately be modelled, the plot will deviate from the unit line (Gilleland & Katz, extRemes 2.0: An Extreme Value Analysis Package in R, 2014).

4 ANALYSIS AND RESULTS

Basic formatting of the raw data and rudimentary calculation were performed in Microsoft Excel. The bulk of the analysis was implemented with scripts written in the programming language R, specifically for this thesis. RGUI was used as an editor and graphical user interface. Later on also RStudio was used. The ready built package *in2extRemes* was used for fitting the data to the Gumbel distribution and the point process model.

The sums of the daily claims was plotted on a logarithmic axis against time in Figure 1, where the days with larger claim sizes clearly stand out, but no clear pattern is noticeable here. The cumulative number of claims and cumulative amount of claims are also visible in Figure 1. The claims seem to arrive at a fairly constant state, as there is only one larger jump in the graph, roughly in the end of 2007. The cumulative amount of the claims is much more prone to jumps, as the result of single larger claims.

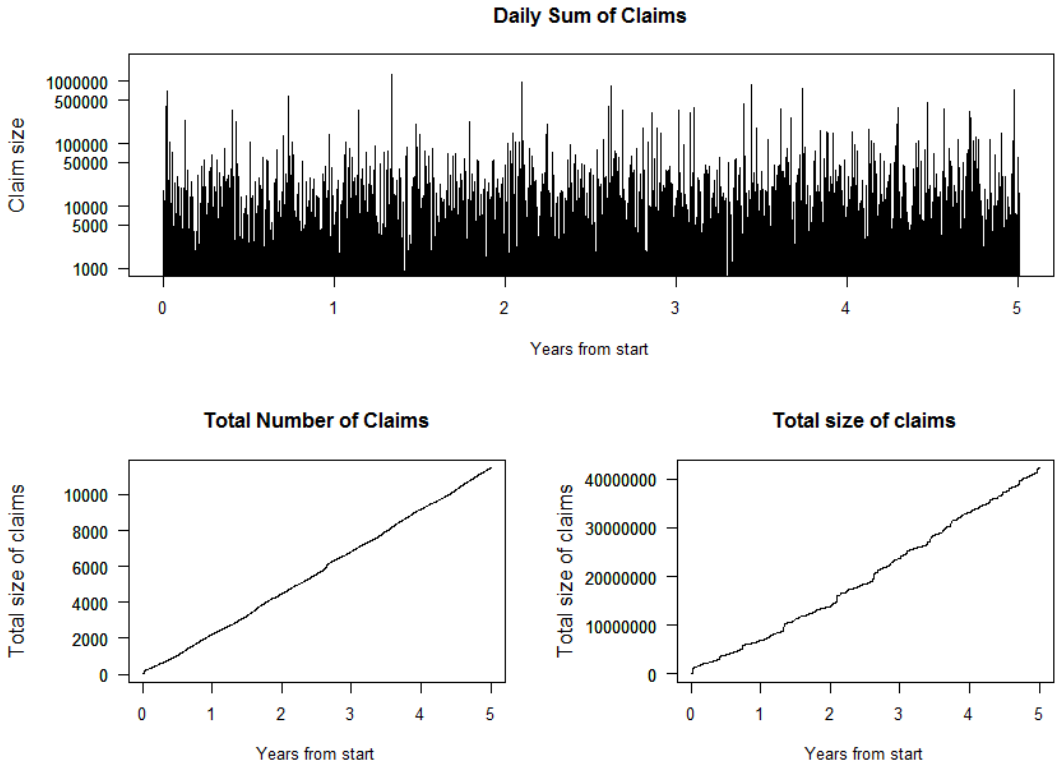

Figure 1. Plots of daily claim amounts, the cumulative number of claims and cumulative amount of claims for the five years of data between 2005 - 2009

4.1.1 Stationarity of Data

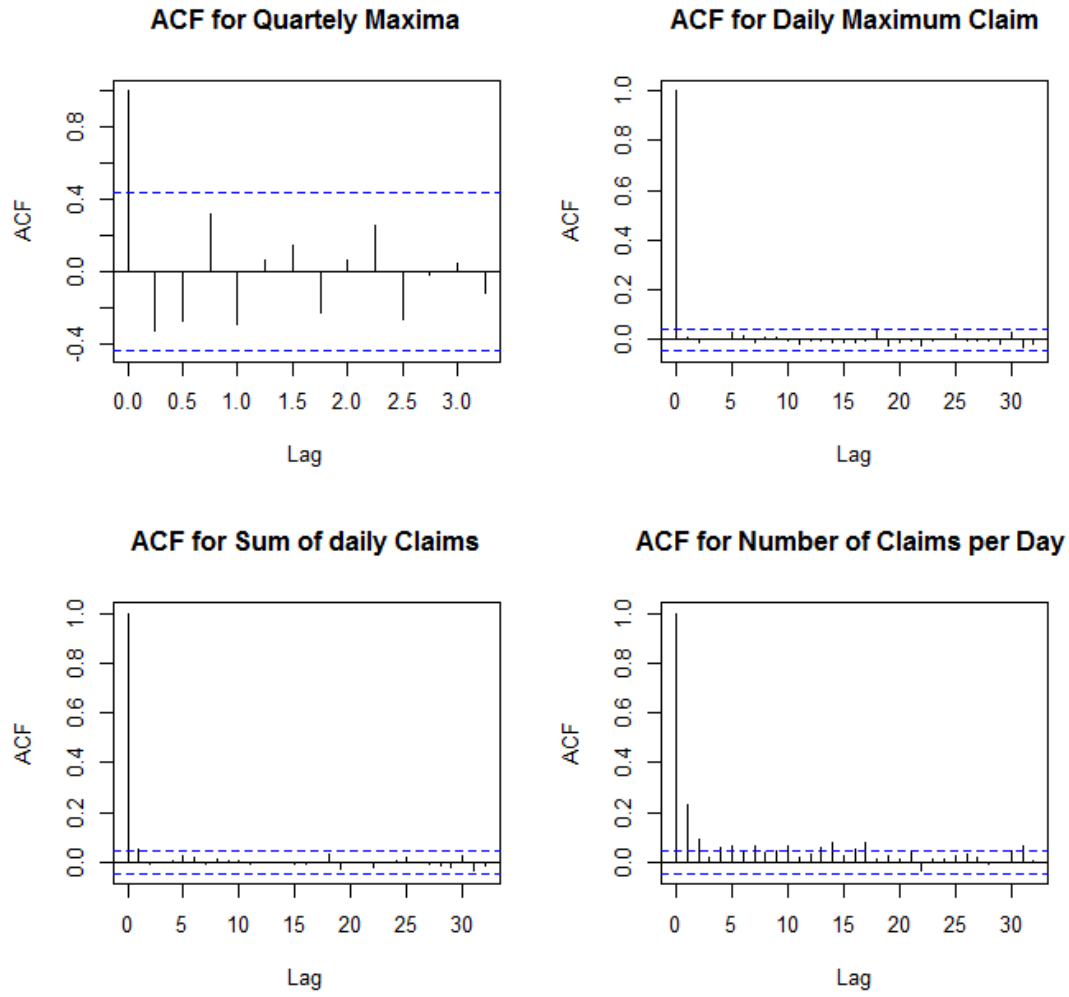

Figure 2. Autocorrelation function plot with 95 % confidence intervals for quarterly maxima, the daily maxima, the sums of all daily claims and finally number of claims per day.

There is one larger event that resulted in significant, related claims during the time period the data was collected. It occurred in January of 2005 and caused claims adding up to approximately 950 000 € that were settled. The reason in the data for the claims was "Storm" and according to Mr. Nygård most of the damages were due to floods. There are several other occurrences of storms, where several claims with the reason "Storm" occur on the same or adjacent days, but the sums of all the claims in these events are still negligible.

These storms and other events contribute to the fact that there is a statistically significant autocorrelation for the number of claims per day, as can clearly be seen in the

bottom right autocorrelation function plot in Figure 2. Autocorrelation function plots show the level of autocorrelation between consecutive observations in a time series. The null hypothesis is that there is no auto-correlation in the data. If the bars in the plots in Figure 2 are below the blue line, the null hypothesis that there is no auto-correlation cannot be rejected at the significance level 95 %. In that case it is assumed that the data is indeed stationary (Shumway & Stoer, 2011). As opposed to the number of individual claims, the quarterly and daily maxima seem to be independent as there is no autocorrelation on the significance level 95 % among these, as can be seen in the two top plots in Figure 2. So even the storms do not generate sufficiently large claims on consecutive days that tomorrow's largest claim could be predicted in a meaningful way from today's largest claim. But if there are many claims today, it is more likely that there are many claims tomorrow as well. The autocorrelation of the summed daily claims is a borderline case. The null-hypothesis that there is no autocorrelation can be rejected, but only very barely. In any case, the autocorrelation is very weak, if visible at all, as can be seen in the bottom left plot in Figure 2. The result is noted, but for practical purposes, this effect will not be explicitly modelled in this study.

The raw data does not include information if the claimant is a natural person or if the claim was filed by a company. An interesting pattern is visible in the data that might be related to this type of distinction between claimants. The dates in the data were divided into holidays and business days, i.e. Saturdays, Sundays and public holidays were separated from the rest. A few ad hoc tests were made to see if there was any difference in the sizes of the daily maxima between business and non business days. The median for the business days turns out to be 55 % larger than for non-business days. The difference is even bigger if the median for the 20 largest claims is compared, amounting to a result 82 % larger for the business days. Also, the frequency of larger claims seems to be higher on business days, a threshold of 45 000 € was exceeded in 7 % of the business days and on 5 % of non-business days. These differences might result from e.g. leaks and fires being more probable in business locales when they are actively used. Larger industrial plants are constantly operational, but smaller warehouses and manufacturing works should stand still on most holidays. Another reason for the possible larger claims on business days is that damage might not be noticed until a holiday is over and people come back to work, thus making the exact date of the accident more difficult to identify. This would be a reason why also weather related accidents might to some extent discriminate against non-business dates in the data. A more formal study of this possible tendency was not performed as it was outside the scope of this thesis, but it might result in interesting results and improvements to the models currently suggested.

Another seasonality that could come into play is that of differences in behaviour of the claim process depending on the time of the year. Some ad hoc tests were made on data

on monthly, quarterly and semi-annual levels. The quarterly level or calendar season seemed to be the most informative and only those results are presented. It seems that only the number of claims is heavily seasonal by looking at the boxplots in Figure 3 and line graphs in Figure 4 (that essentially give the same information, only in slightly different formats). There are only a few data points for each quarter, five, but the difference in number of claims is still quite clear as the smallest number of claims in any Q3 (1.7-30.9) is still larger than the number of claims in any other quarter. However, the two other quantities studied, the sums of all claims per quarter and the largest total daily claim amount per quarter, do not show any clear seasonal pattern. As it is principally the maxima of claims and sums of claims that that are the object of study, no seasonal models need to be adapted. Even though there are many more claims during the third quarter than other quarters, these apparently tend to be small enough not to affect the aggregate claim amounts, i.e. we're not talking about severe storms. The reasons for the higher number of claims during Q3 are not completely clear. The concept of the summer season is highly subjective and relative in Finland, but Q3 usually includes two of the three warmest months in the year and also the time when most Finns are on vacation. These two factors could contribute to people being more active and thus being more accident prone, perhaps visible in the growth in the number of claims.

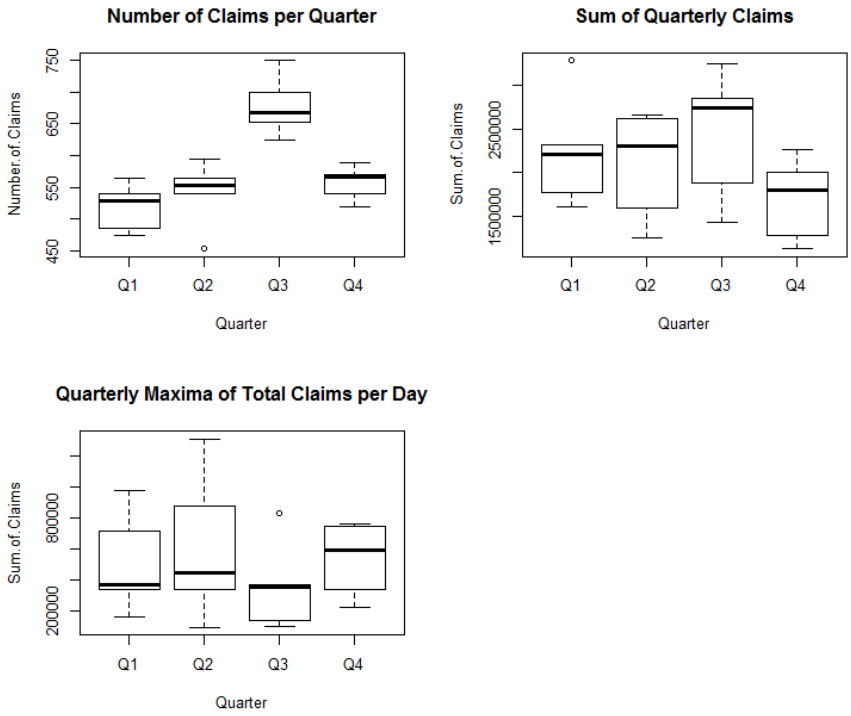

Figure 3. Seasonality study box plots with quarterly number of claims, sum of claims and maximum claims.

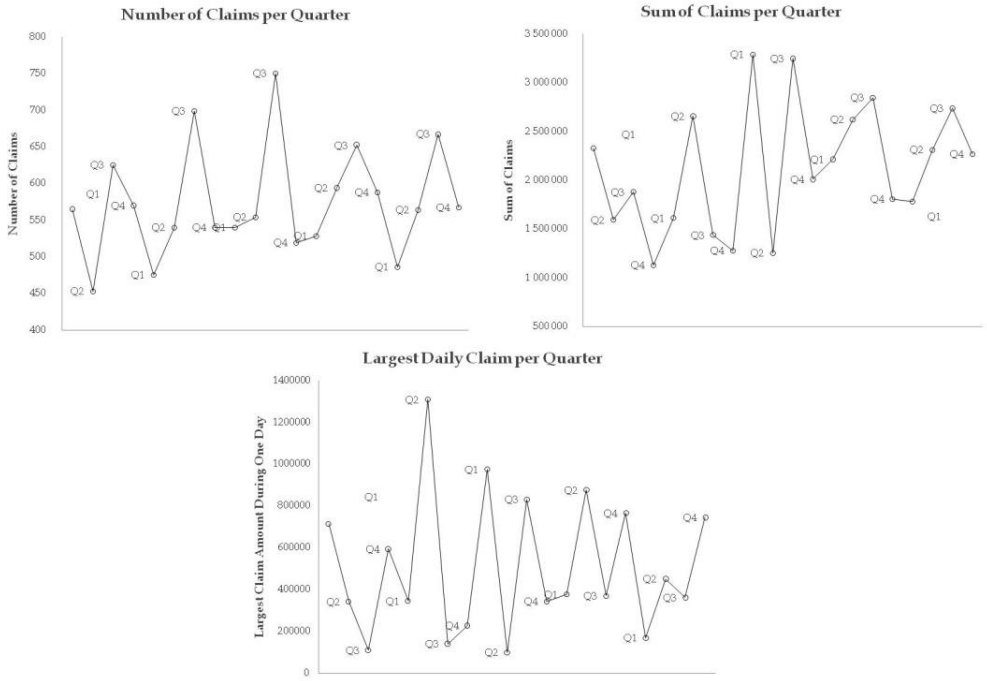

Figure 4. Seasonality study time series plots with quarterly number of claims, sum of claims and maximum claims.

What data to model is not self-evident, since we have a random number of observations per day. The large claims dwarf the median claim in size and thus the smaller claims are of less importance to the total, even though they are plentiful. However, some days contain more than one larger claim and thus if only the daily maximum is considered, useful info is disregarded. Another option would be to model the sum of all the claims during one day. An analogous case from weather related usage of extreme value methods would be the difference between heavy rain and extreme windstorms; the intensity of the rain can vary during the day, but eventually the cumulative amount of rain is what is usually of interest. When preparing for strong wind speeds on the other hand, it is the maximum presumable speed that is typically of interest. A third approach, used by Smith & Goodman (2000), would be to try to aggregate the simultaneous claims arising from the same event into a single claim for that day. The reason for this is to avoid clustering effects from claims from the same cause. This was not performed for the data in question, as there were an insignificant amount of such claims (arising from e.g. storms), and the data lacked a geographical dimension, making it more difficult to evaluate whether simultaneous claims actually originated from the same event. At the same time, the policy holders are located within a fairly compact geographical area, so it is unlikely that there would be several simultaneous and independent storms in that area. In the end, the analysis was first performed using the daily maxima as an observation and then the same analysis was made using the sum of all claims during one day as an observation.

4.2 Modelling the Daily Maximum Claim

For the first part of the analysis the largest claim for each date was used as the single observation for that day.

4.2.1 Generalized Extreme Value

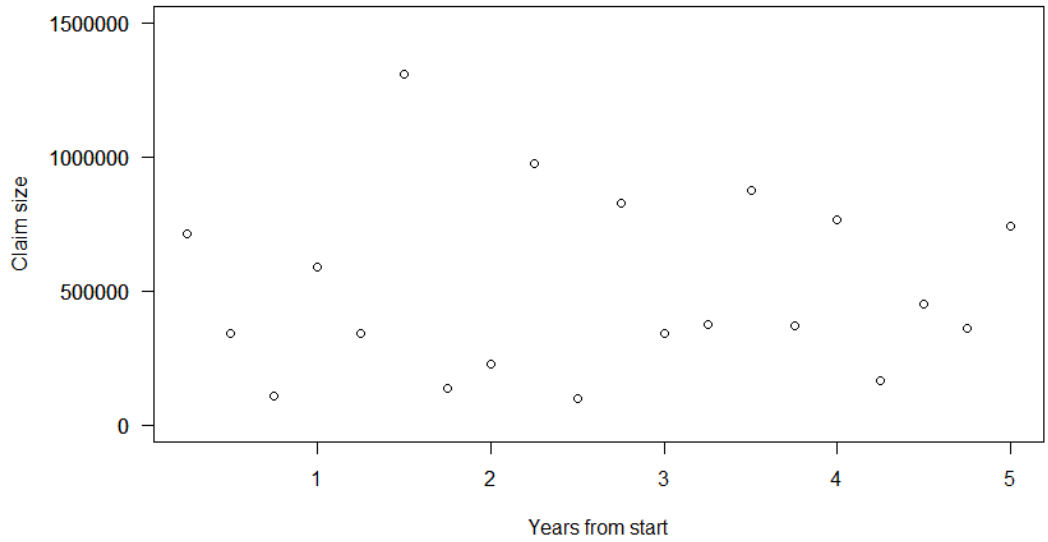

Figure 5. Quarterly maximum claims.

Weekly, monthly and quarterly maxima were fitted to a GEV-distribution, the best fit and balance between variance and bias was obtained by fitting the quarterly data. It seems that that the number of observations from where the maximum is drawn is not large enough for weekly or monthly maxima and so the asymptotic argument is not valid. Only the results from the quarterly maxima are shown, but the difference in the goodness of fit was clear as the fit for the maxima from other time periods was very poor. Numerical optimization was used in the maximization of the likelihood functions, using the R-function **optim( )** throughout the study, except for the fitting of the data to the Gumbel distribution and the point process model, where the ready built package *in2extRemes* was used.

Table 2. Estimated GEV parameters and return levels with 95 % confidence intervals obtained with delta method

	Log Likelihood	Location ($\hat{\mu}$)	Scale ($\hat{\sigma}$)	Shape ($\hat{\xi}$)	5-year Return Level
Estimate	278.8	269 000	183 000	0,442	1 400 000
95 % CI		[165 000; 373 000]	[87 400; 279 000]	[-0.212; 1.1]	[122 000; 2 670 000]

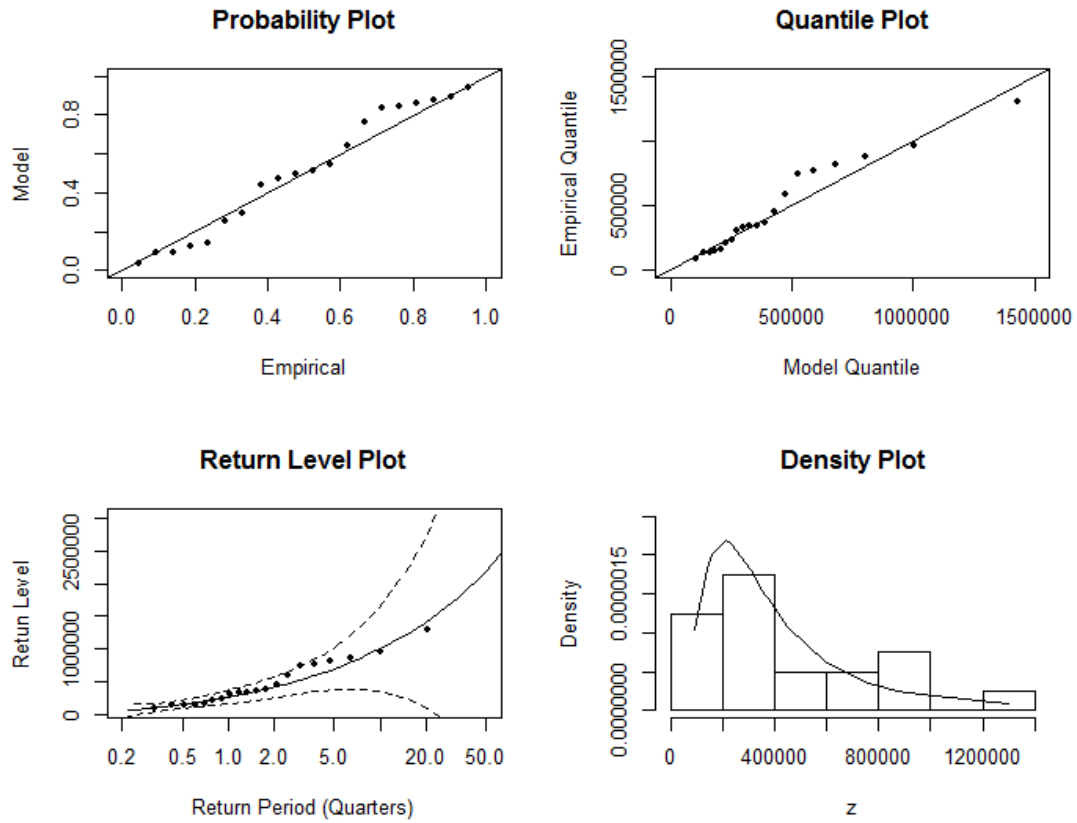

Figure 6. Goodness of fit diagnostic plots for GEV fit for the quarterly maxima data.

It is difficult to interpret the density plot as there are only 20 observations (five years) to construct the histogram. Generally, the fit seems to be fairly good. The dots in the quantile and probability plots coincide well to the unit line. The 95 % confidence intervals in the return level plot are rather large, but the empirical dots are quite well placed on the full line which represents the return level deduced from the model. The five year return level of 1 400 000 coincides well with the empirical value of approximately 1 300 000. It is also noteworthy that the estimate of the shape parameter is positive, but the confidence interval includes zero.

Table 3. Estimated GEV parameters and return levels with 95 % confidence intervals obtained with profile likelihood.

	Log Likelihood	Shape $(\hat{\xi})$	5-year Return Level
Estimate	278,8	0,442	1 400 000
95 % CI		[-0,12; 1,2]	[800 000, 6 705 000]

The confidence interval for the return level is more skewed to the right, as expected. The difference to the delta method is quite large and especially the lower bound is much more credible.

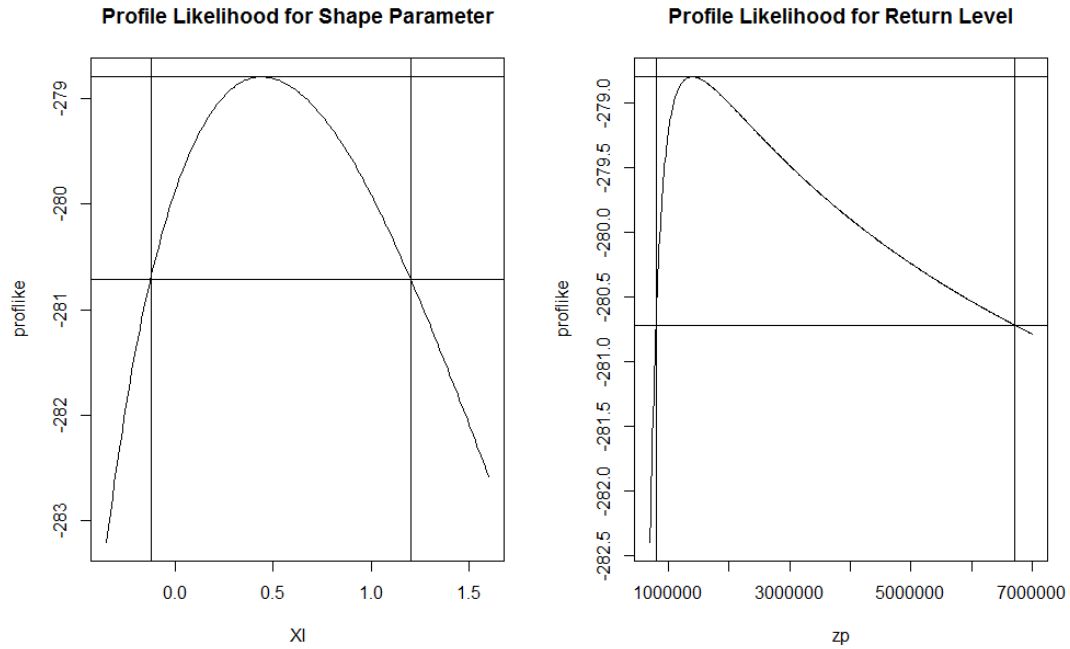

Figure 7. Profile likelihood for ξ, the shape parameter and for 5 year return level for GEV fit for quarterly maxima using the daily maxima data.

4.2.2 Gumbel

As the confidence interval for the shape parameter above includes zero and the null hypothesis could not be rejected in favour of M1 at chosen significance level, the quarterly maxima was fitted to a Gumbel distribution as well. The R-package in2Extremes was used (Gilleland & Katz, 2011).

Table 4. Estimated Gumbel parameters with 95 % confidence intervals obtained with delta method.

	Log Likelihood	Location ($\hat{\mu}$)	Scale ($\hat{\sigma}$)	5-year Return Level
Estimate	280,0	293 000	233 000	984 000
95 % CI		[185 000; 401 000]	[146 000; 319 000]	[692 000; 1 276 000]

The fit does not seem to be as good as in the GEV-model with $\xi \neq 0$. The return period plot gives perhaps the strongest evidence of a weak fit, where almost all of the empirical estimates of return level are considerably above the model and at times even above the 95 % confidence interval. The five year return period estimate and the upper confidence interval bound are also below the largest value observed in the five year long data, casting the model validity into serious doubt.

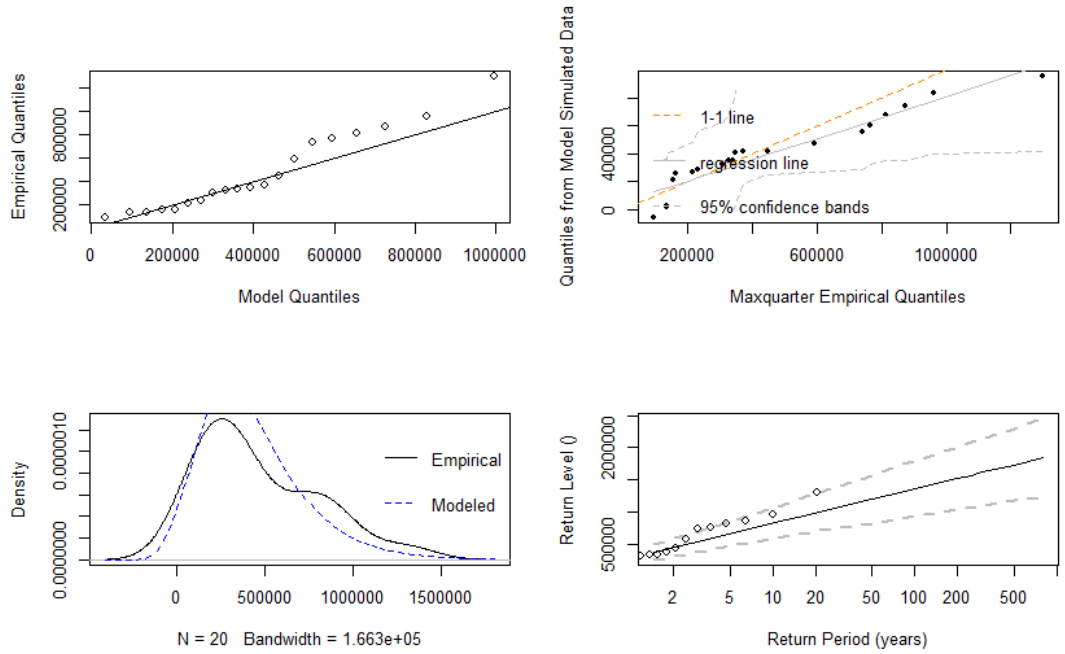

Figure 8. Goodness of fit diagnostic plots from in2extRemes package for Gumbel fit for the quarterly maxima data. From left to right, Quantile plot, Simulated quantile plot, Density plot and Return level plot

4.2.3 Generalized Pareto Distribution

A variety of thresholds were tested based on the mean residual life plot and the Figure 10 below, depicting the estimates of the shape and scale parameters as a function of the threshold. In the end the threshold 45 000 was chosen, again as it seems to provide the best balance between number of observations and the validity of the asymptotic argument. The number of daily maximum claims above the threshold was 120.

Table 5. Estimated GP parameters and return level with 95 % confidence intervals obtained with the delta method.

	Log Likelihood	Scale ($\hat{\sigma}$)	Shape ($\hat{\xi}$)	5-year Return Level
Estimate	1508	57 100	0,619	1 740 000
95 % CI		[38 200; 76 000]	[0,319; 0,918]	[239 000; 3 230 000]

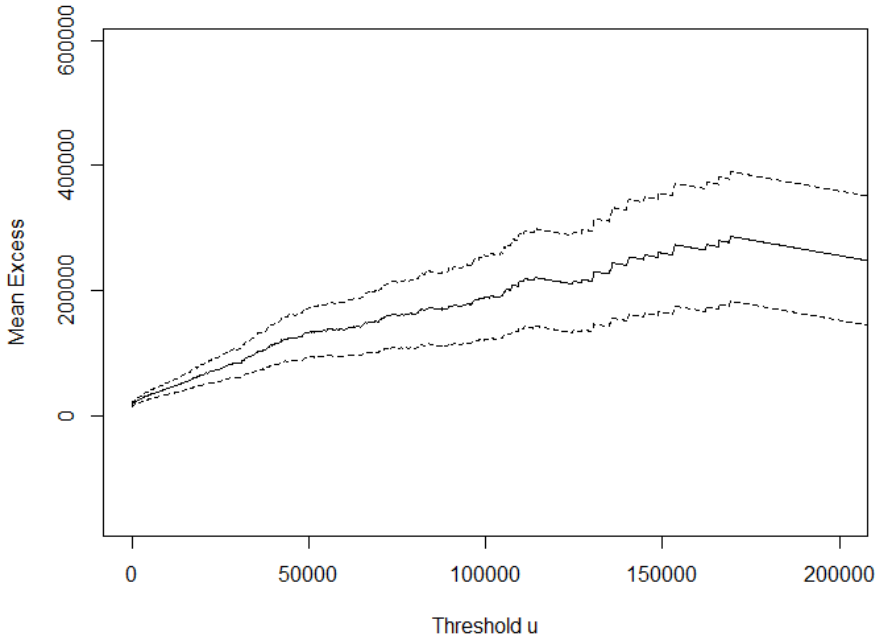

Figure 9. Mean residual life plot for daily maxima data.

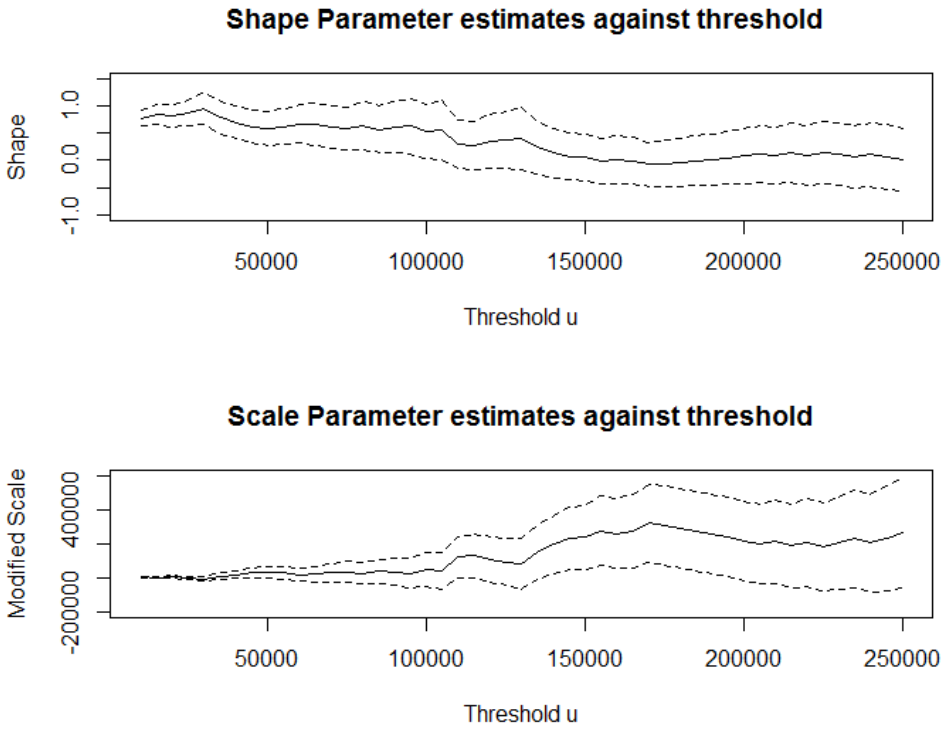

Figure 10. Estimates of parameters for generalized Pareto model fit against threshold for daily maxima data.

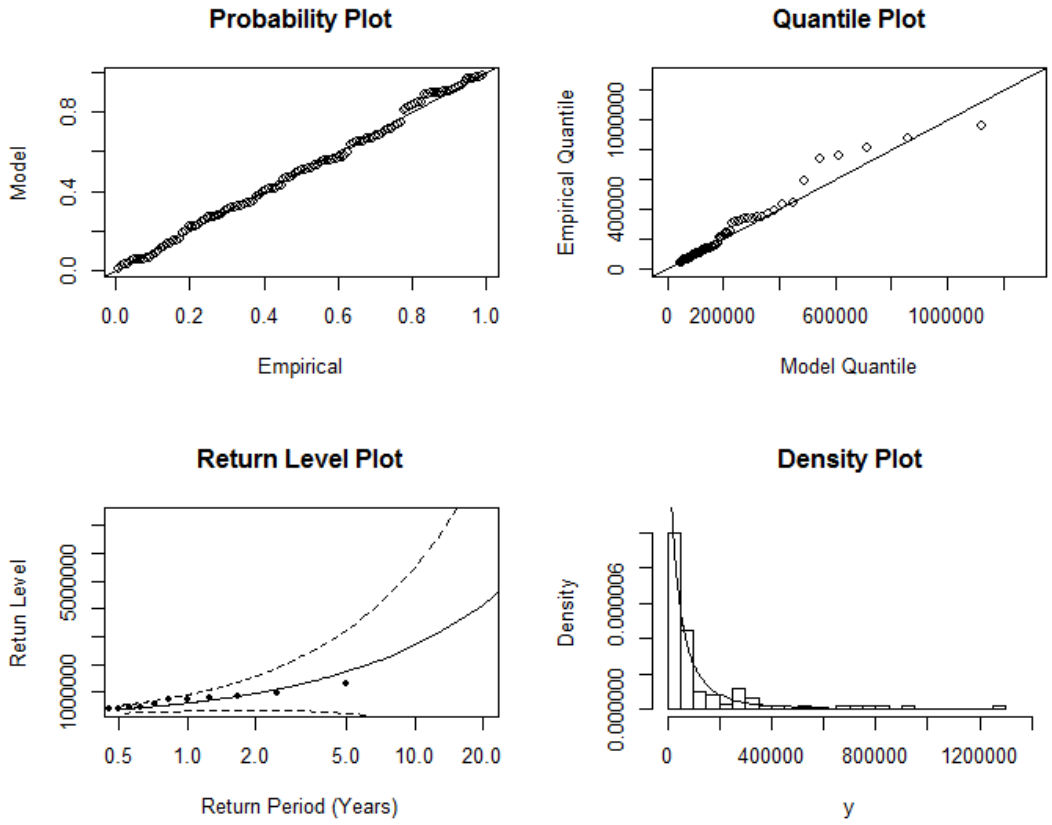

Figure 11. Goodness of fit diagnostic plots for GPD fit for the daily maxima data.

The model induces a little bit of conservatism into the estimation of larger quantiles, as it gives somewhat larger estimates in the tail of the claim process. This can be seen especially in the quantile and return level plots. The shape parameter was clearly above zero, and M0 could be rejected in favour of M1 at the 95 % significance level. The estimated return levels where slightly higher than in the GEV case and the confidence intervals were smaller. The magnitude of the return level seems again reasonable compared to the data.

Table 6. Estimated GP parameters and return level with 95 % confidence intervals obtained with profile likelihood.

	Log Likelihood	Shape $(\hat{\xi})$	5-year Return Level
Estimate	278.8	0,619	1 740 000
95 % CI		[0,369; 0,967]	[889 000; 5 230 000]

As expected, the confidence interval derived using profile likelihoods is more skewed to the right, and at least the lower bound again seems more realistic. The shape parameter seems to be confidently above zero.

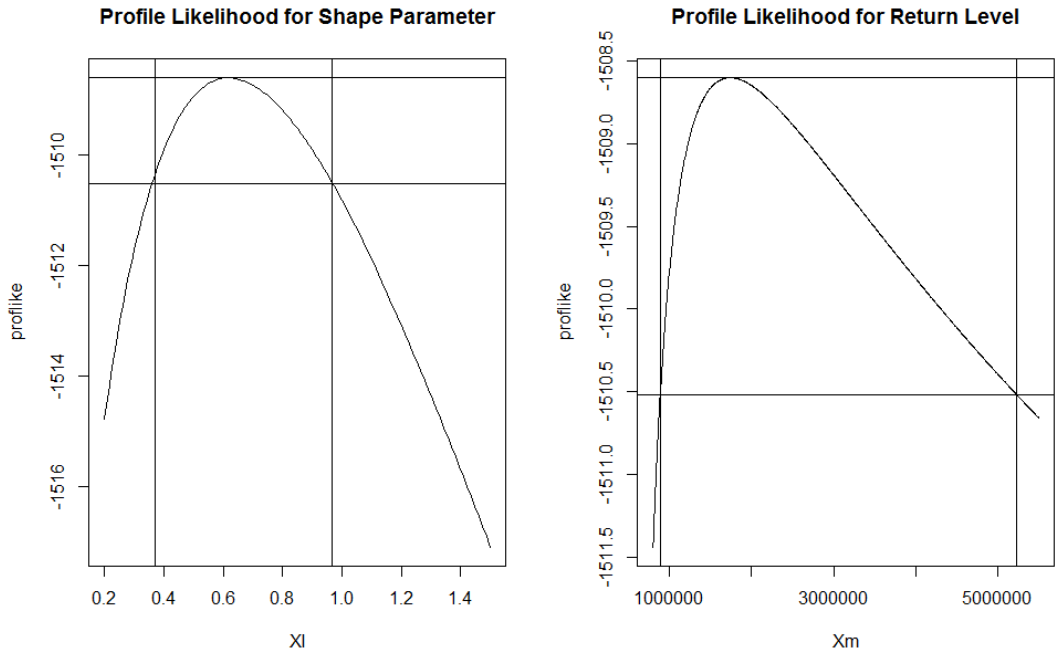

Figure 12. Profile likelihood for ξ, the shape parameter and for 5 year return level for the generalized Pareto distribution fit for the daily maxima data.

4.2.4 Poisson-GPD model

The same threshold of 45 000 was used as in the pure GPD-model. To analyze how the probabilities in the very far tail behave, the distribution of the claims conditioned upon an exceedance of 15 million euro was computed. This result is especially of interest when pricing reinsurance, i.e. what is the probability of exceeding a certain limit, and how much bigger than the limit that exceeding claim is likely to be. The probability of exceeding 15 million during a period of five years was estimated to be 3 %. The estimated conditional distribution of exceedances over 15 million is shown in Figure 13. As laid out in Rootzén & Tajvidi (1997) the median of the next excess over the limit was estimated to be 8 080 000 and the median of the next largest loss to be 2 020 000. The 5-year return level was estimated to be approximately 1 630 000.

Table 7. Estimated parameters and 95 % confidence intervals obtained with delta method for the Poisson-GPD model

	Log Likelihood	Intensity $(\hat{\lambda})$	Scale $(\hat{\sigma})$	Shape $(\hat{\xi})$
Estimate	1955	0,0657	57 100	0,619
95 % CI		[0,0544; 0,0771]	[38 200; 76 000]	[0,0544; 0,918]

Table 8. Probable Maximum Losses for different time periods and confidence levels.

Confidence degree	1 year	5 years	10 years
10%	2 600 000	7 100 000	11 000 000
5%	4 100 000	11 000 000	17 000 000
1%	11 000 000	31 000 000	47 000 000

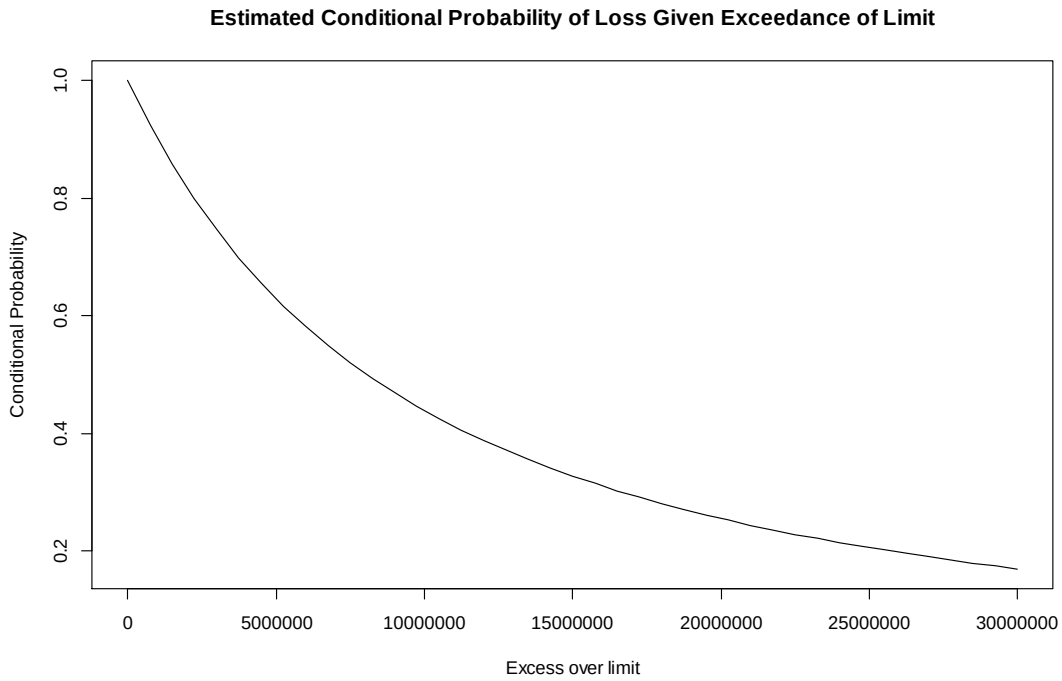

Figure 13. Estimated conditional probability that an excess loss over 15 000 000 € is larger than x.

4.2.4.1 Simulation Study Using the Estimated Poisson-GPD Model

A simulation study was performed using the estimated Poisson-GPD model. 1000 simulations were drawn for five different time periods. The draws were executed using inverse transform sampling. Two sets of results were obtained. The first set consists of the maxima of the simulated values and the second of the sums of all the simulated values for each time period. The results are plotted as histograms which can be viewed as empirical density functions for the two quantities. The empirical 95 % quantiles were also determined for both the maxima and sums. 1000 simulations for e.g. 15 years results in some quite large values. The ten largest values are shown for graphical purposes in a table format in the appendix, instead of including them in the histogram. The values on the first row in the tables in the appendix can be considered as the 0,5 % percentile of the distribution. The GEV model for quarterly maxima corresponds very well to the simulated values from the Poisson – GPD model with a threshold of 45 000 €, as can be seen in Figure 15, where the pdf of the estimated GEV model is overlaid on the histogram.

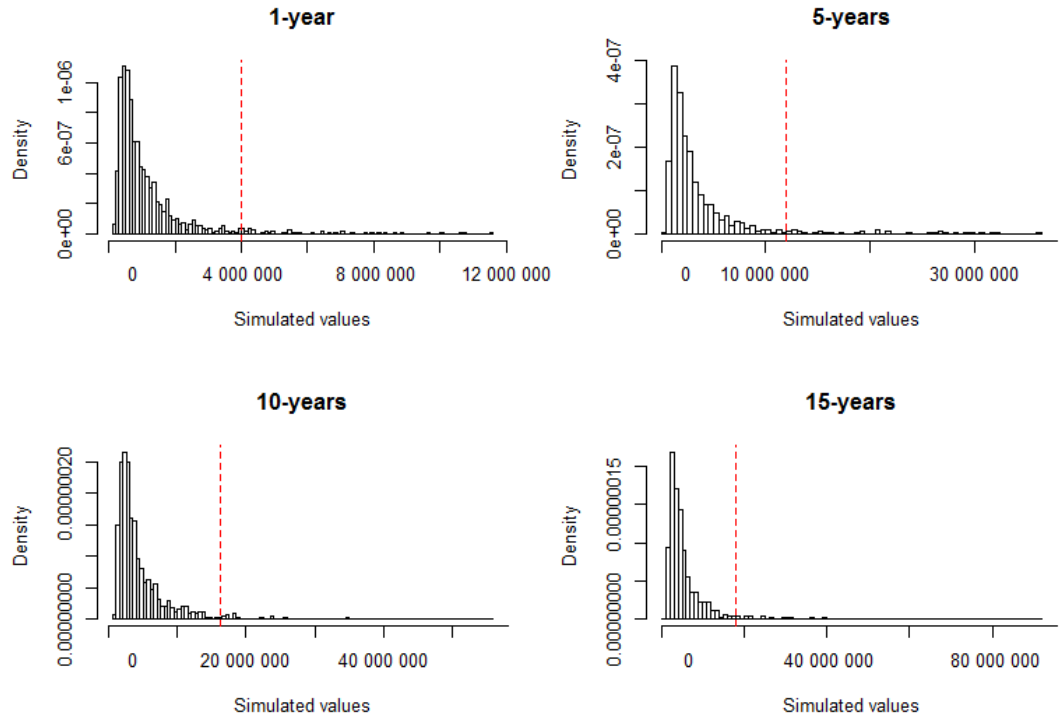

Figure 14. Histograms for 1000 simulated maxima for different time periods.

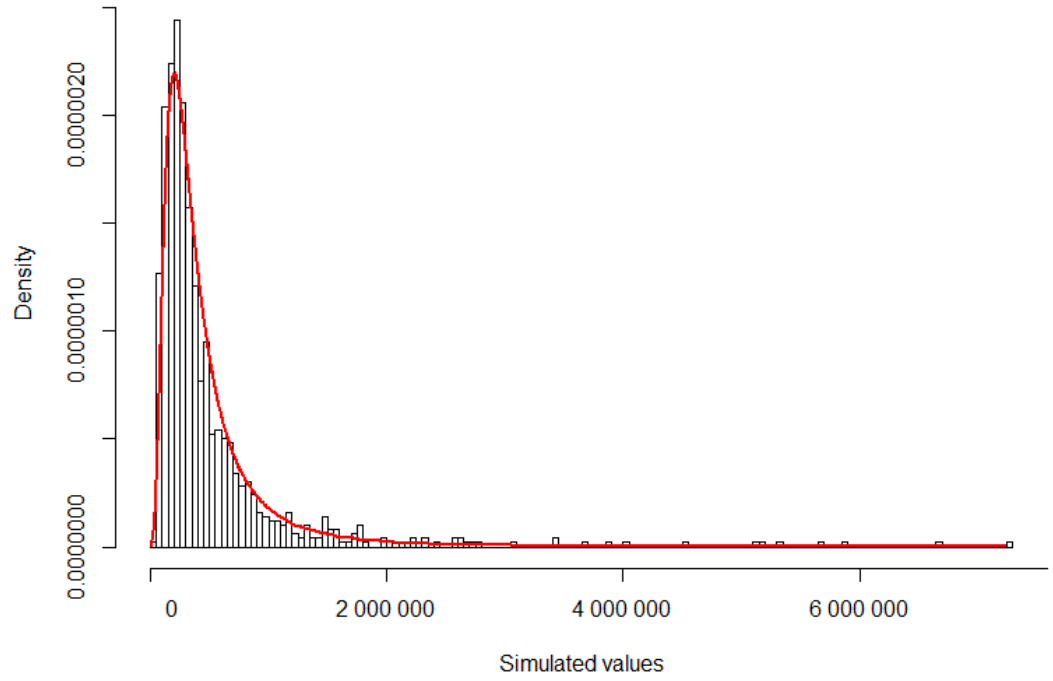

Figure 15. Histogram for 1000 simulated maxima for one quarter. The estimated GEV-pdf is overlaid.

Table 9. 95 % quantiles from 1000 simulated maxima of threshold exceedances for daily maxima claims.

	Quarter	1 year	5 years	10 years	15 years
95 % quantile	1 553 263	4 004 688	12 009 908	16 228 394	18 098 582

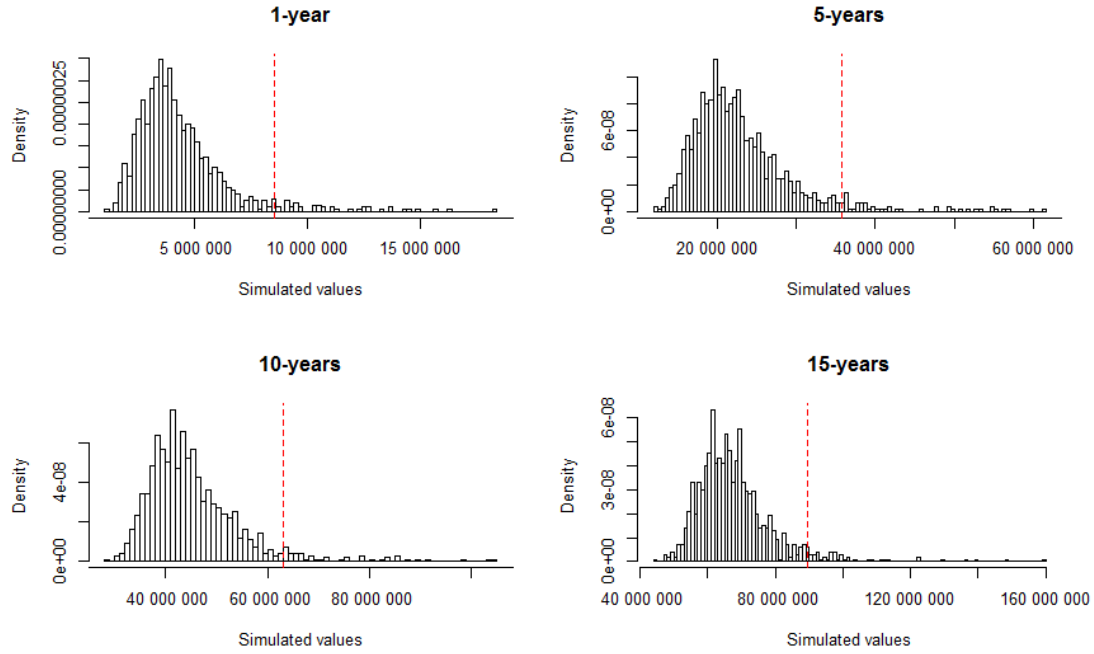

Figure 16. Histograms for 1000 simulated sum of threshold exceedances for different time periods.

Table 10. 95 % quantiles from 1000 simulated sums of threshold exceedances for daily maxima claims.

	Quarter	1 year	5 years	10 years	15 years
95 % quantile	2 812 723	8 525 016	35 832 715	63 085 589	89 388 475

4.2.5 Point Process

The fitting of the data to the model was made with in2extRemes for the point process approach. The same threshold of 45 000 was used. The goodness-of-fit plots give a dual message; the occurrences of the events seem to be well modelled according to the z-plot, but the modelled and empirical densities do not coincide at all. The quantile plot shows a good fit in the tail, but hints at a less good fit in the medium range. Furthermore, the confidence intervals of the return level is considerably smaller than for the other models, especially when comparing to the confidence intervals received through profile likelihoods. As in2extRemes is a fairly new program, there seems to be still some bugs, as for example the profile likelihood confidence intervals could not be computed for the

fitted model. This fit was consequently discarded, based mainly on the considerable difference in the empirical and modelled densities.

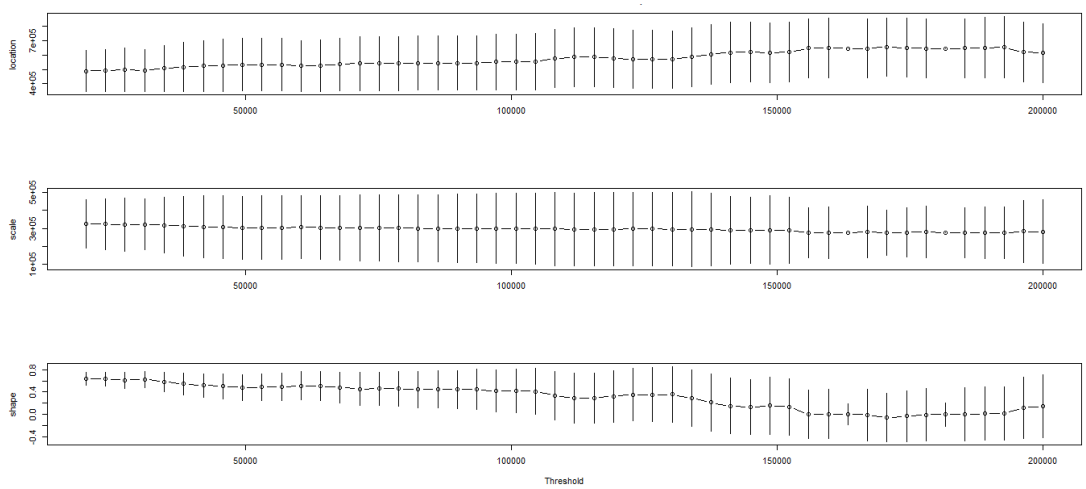

Figure 17. Estimates of parameters for point process model fit against threshold for daily maxima data.

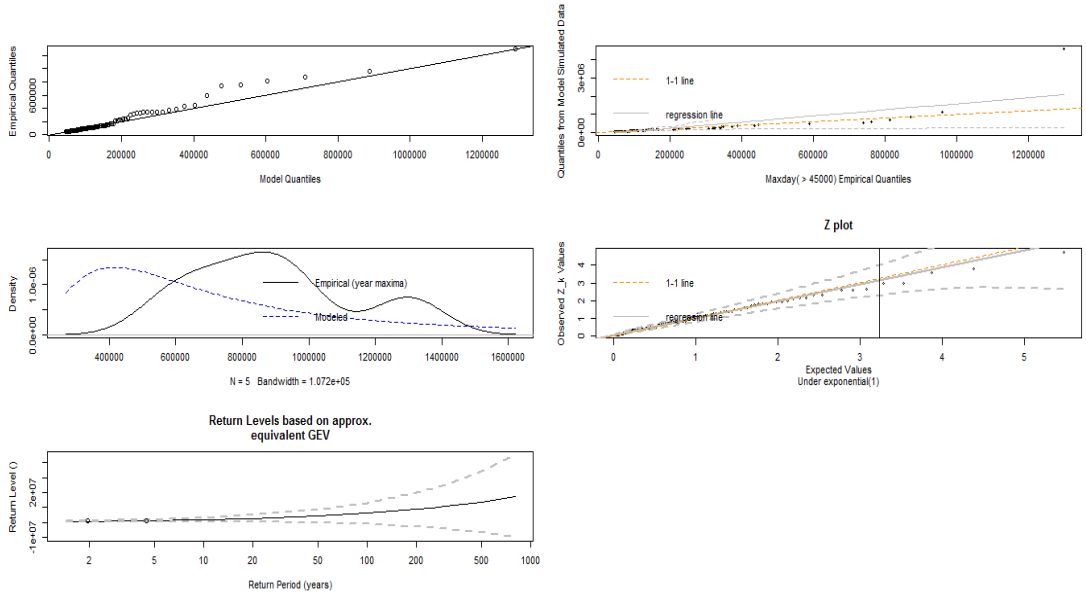

Figure 18. Goodness of fit diagnostic plots from package in2extRemes for point process fit for the daily maxima data.

Table 11. Estimated parameters and return levels for the point process model, with 95 % confidence intervals obtained using the delta method

	Log Likelihood	Location ($\hat{\mu}$)	Scale ($\hat{\sigma}$)	Shape ($\hat{\xi}$)	5-year Return Level
Estimate	1248	529 000	305 000	0,504	1 210 000
95 % CI		[348 000; 710 000]	[129 000; 482 000]	[0.282; 0,726]	[536 000; 1 890 000]

4.3 Modelling the Sum of Claims per Day

This is the so called rain case, i.e. all the claims are summed for each date and used as one observation for that date. Again, weekly, monthly and quarterly maxima were fitted to a GEV-distribution and unsurprisingly the best fit and balance between variance and bias was obtained by fitting the quarterly maxima.

4.3.1 GEV

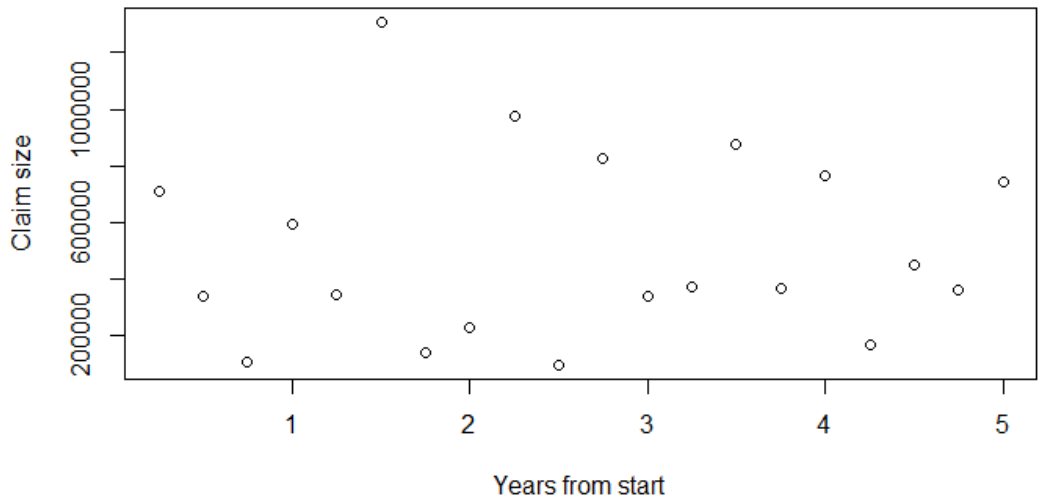

Figure 19. The quarterly maxima of the data where the claims were summed on a daily basis.

Table 12. Estimated GEV parameters and return levels with 95 % confidence intervals obtained with delta method.

	Log Likelihood	Location $(\hat{\mu})$	Scale $(\hat{\sigma})$	Shape $(\hat{\xi})$	5-year Return Level
Estimate	280	343 000	236 000	0,110	1 170 000
95 % CI		[215 000; 471 000]	[134 000; 337 000]	[-0.421; 0,642]	[569 000; 1 770 000]

Table 13. Estimated GEV parameters and return levels with 95 % confidence intervals obtained with profile likelihood.

	Log Likelihood	Shape $(\hat{\xi})$	5-year Return Level
Estimate	278.8	0,110	1 170 000
95 % CI		[-0.31, 0,80]	[840 000, 3 390 000]

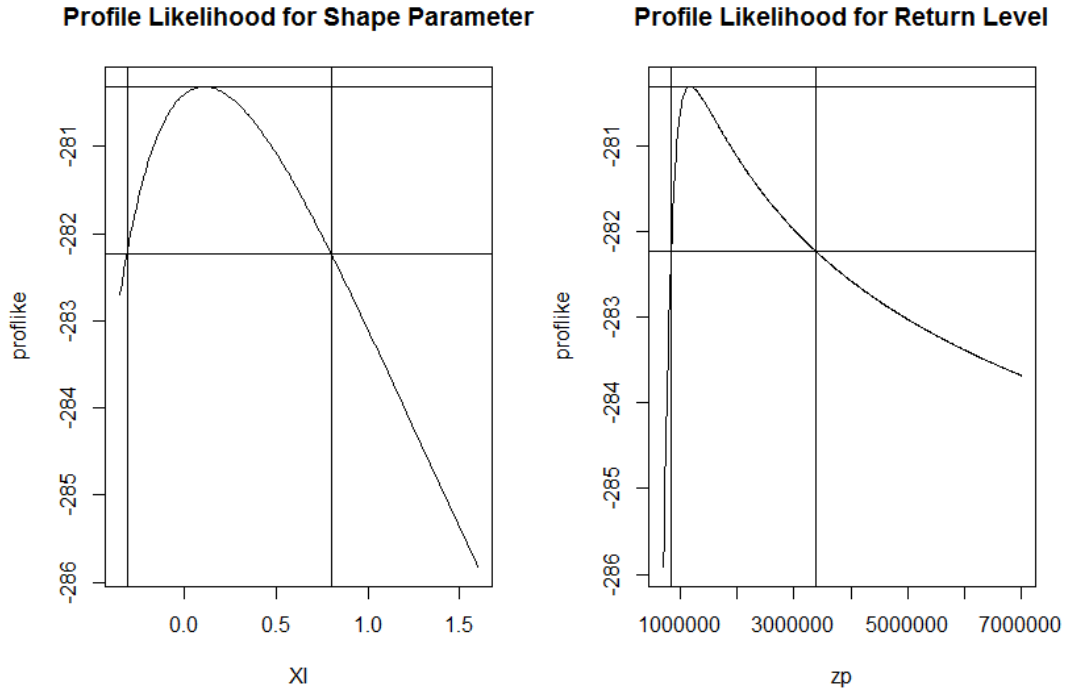

Figure 20. Profile likelihood for ξ, the shape parameter and for 5-year return level for GEV fit for quarterly maxima using the summed daily claims data.

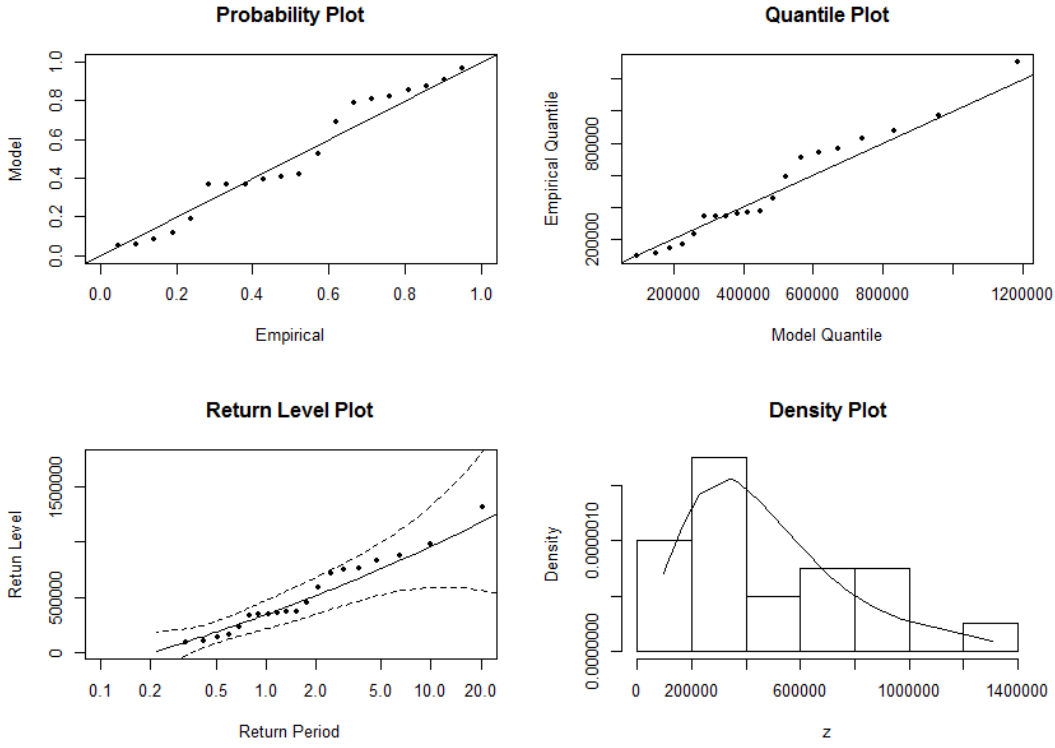

Figure 21. Goodness of fit diagnostic plots for GEV fit for the summed daily claims data.

As in the case for the daily maxima data, the null hypothesis that the shape parameter is zero cannot be rejected in favour of $\mathcal{M}_1$ at the 95 % confidence level. This time the confidence interval is more evenly centred around zero. Therefore also this data was fitted to the Gumbel distribution.

4.3.2 Gumbel

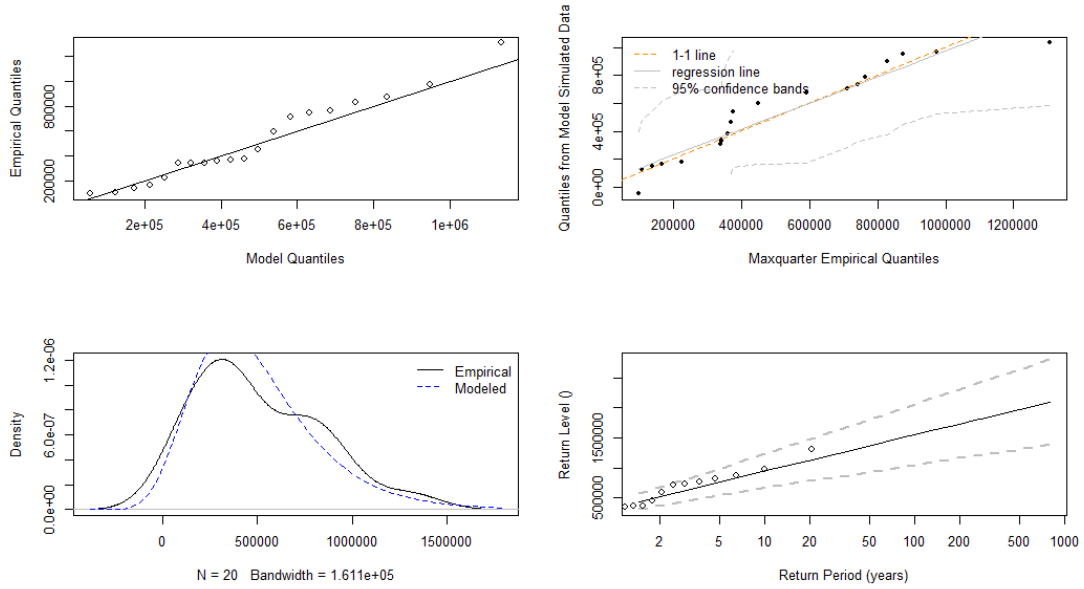

Figure 22. Goodness of fit diagnostic plots from in2extRemes package for Gumbel fit for quarterly maxima with sum of daily maxima. From left to right, Quantile plot, Simulated quantile plot, Density plot and Return level plot

Table 14. Estimated Gumbel parameters with 95 % confidence intervals obtained with delta method.

	Log Likelihood	Location ($\hat{\boldsymbol{\mu}}$)	Scale ($\hat{\boldsymbol{\sigma}}$)	5-year Return Level
Estimate	280	346 000	262 000	1 120 000
95 % CI		[225 000; 467 000]	[161 000; 363 000]	[778 000; 1 470 000]

The goodness of fit seems to be fairly equal for the two fitted models, Gumbel and GEV. What is noteworthy however, is that the summed data seems to be less well modelled by both the GEV- and Gumbel distributions than the data consisting of the daily maxima. The GEV-fit even gives a lower 5-year return level for the summed data than for the daily maxima data, which is of course misleading as the summed data is by construction equal or larger than the daily maxima data. The 5-year return levels computed from the GEV and Gumbel distributions are also both smaller than the largest value in the five year data, which is per se not faulty, but at least noteworthy.

4.3.3 Generalized Pareto Distribution

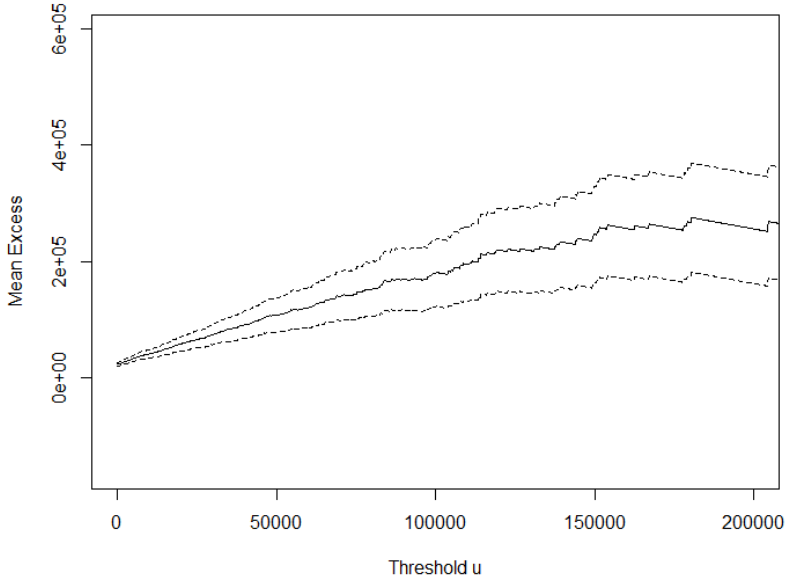

Figure 23. Mean residual life plot for summed daily claims data.

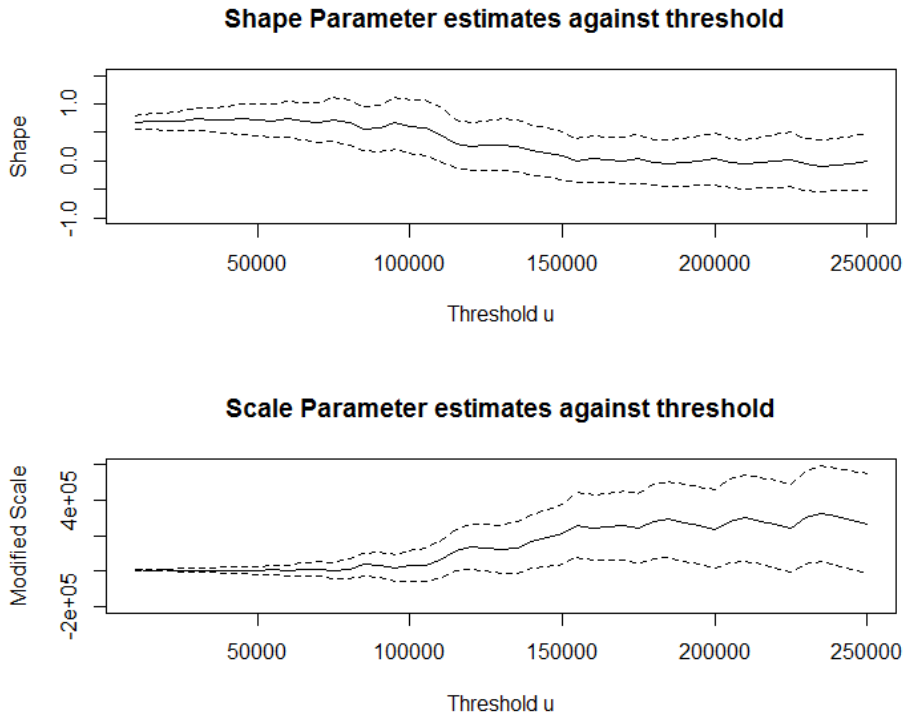

Figure 24. Estimates of parameters for GPD fit against threshold for summed daily claims data.

Again, a wide array of thresholds were tried based on the Figures 23 and 24. The threshold 50 000 was in the end chosen, slightly larger than for the daily maxima case, which is of course not a surprise. This choice resulted in 161 observations above the threshold, compared to 120 in the earlier case. The shape parameter was clearly above zero, and the null hypothesis could be rejected in favour of M1 at the 95 % significance level.

Table 15. Estimated GP parameters and return level with 95 % confidence intervals obtained with the delta method.

	Log Likelihood	Scale ($\hat{\sigma}$)	Shape ($\hat{\xi}$)	5-year Return Level
Estimate	1991	42 100	0,722	2 270 000
95 % CI		[29 700; 54 400]	[0,447; 0,996]	[220 000; 4 330 000]

Table 16. Estimated GP parameters and return level with 95 % confidence intervals obtained with profile likelihood.

	Log Likelihood	Shape ($\hat{\xi}$)	5-year Return Level
Estimate	1991	0,619	2 270 000
95 % CI		[0,48; 1,03]	[1 100 000; 6 850 000]

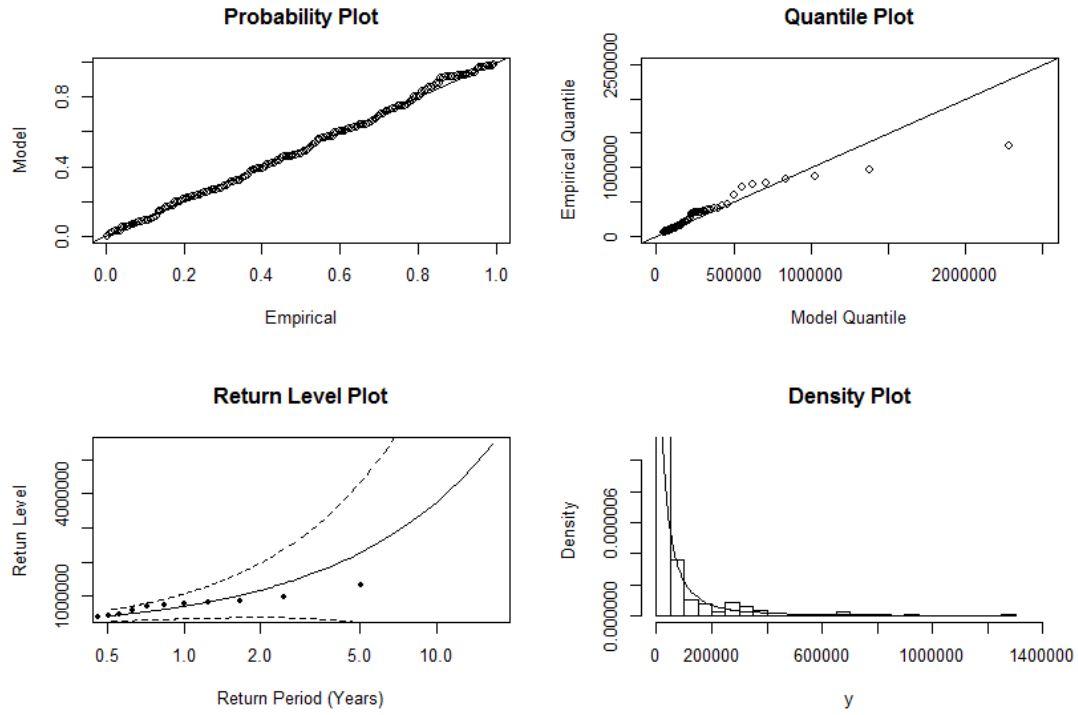

Figure 25. Goodness of fit diagnostic plots for GPD fit for the summed daily claims data.

Again, the density plot is difficult to interpret, although the fit looks basically correct. Conversely, the quantile plot and return level plots tell a clear story, where there is a fairly good fit everywhere, but in the region that is of interest, the extremes. Based on

the data available, the model seems to overestimate the claim severity. Adding to the bias, is the large uncertainty in the model and hence the estimates of the return levels, as can e.g. be seen in the profile likelihood plots. So just as in the GEV-model, the fit is less good for this data set than for the daily maxima data. The confidence intervals are also larger for e.g. the return level, even if there are more data points. Here at least the sizes of the estimates for the return levels are larger than for the daily maxima case, just as expected.

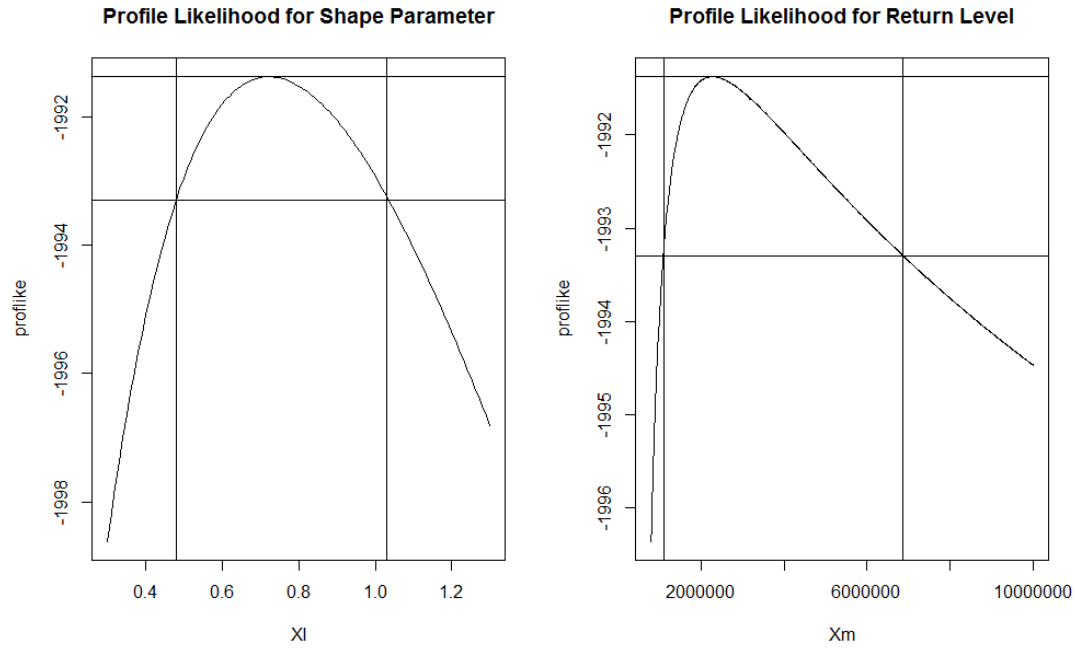

Figure 26. Profile likelihood for ξ, the shape parameter and for 5 year return level for generalized Pareto distribution fit for the summed daily claims data.

4.3.4 Poisson-GPD

The same threshold of 50 000 was used as in the GPD-model. The same large limit of 15 000 000 was also chosen for the conditional probability distribution. The probability of exceeding that limit during a period of five years was estimated to be 7 %. Also the median of the next excess over the limit was estimated to be 9 800 000 and the median of the next loss that is larger than the previously largest loss to be 2 160 000. The five year return level was estimated to be 2 110 000. These are only slightly larger than for the case where only the daily maxima were modelled. This was expected, as the data does not include very many days with several large claims, and thus the days with a large maxima and large sum of all claims that day are the same days. In other words, the difference between the largest claim on a day and the second largest day is often significant.

Table 17. Estimated parameters and 95 % confidence intervals obtained with delta method for the Poisson-GPD model

	Log Likelihood	Intensity $(\hat{\lambda})$	Scale $(\hat{\sigma})$	Shape $(\hat{\xi})$
Estimate	2543	0,0882	42 000	0,723
95 % CI		[0,0752; 0,101]	[29 600; 54 400]	[0,448; 0,998]

Table 18. Probable Maximum Losses for different time periods and confidence levels.

Confidence degree	1 year	5 years	10 years
10%	4 000 000	12 000 000	19 000 000
5%	6 000 000	20 000 000	32 000 000
1%	20 000 000	64 000 000	105 000 000

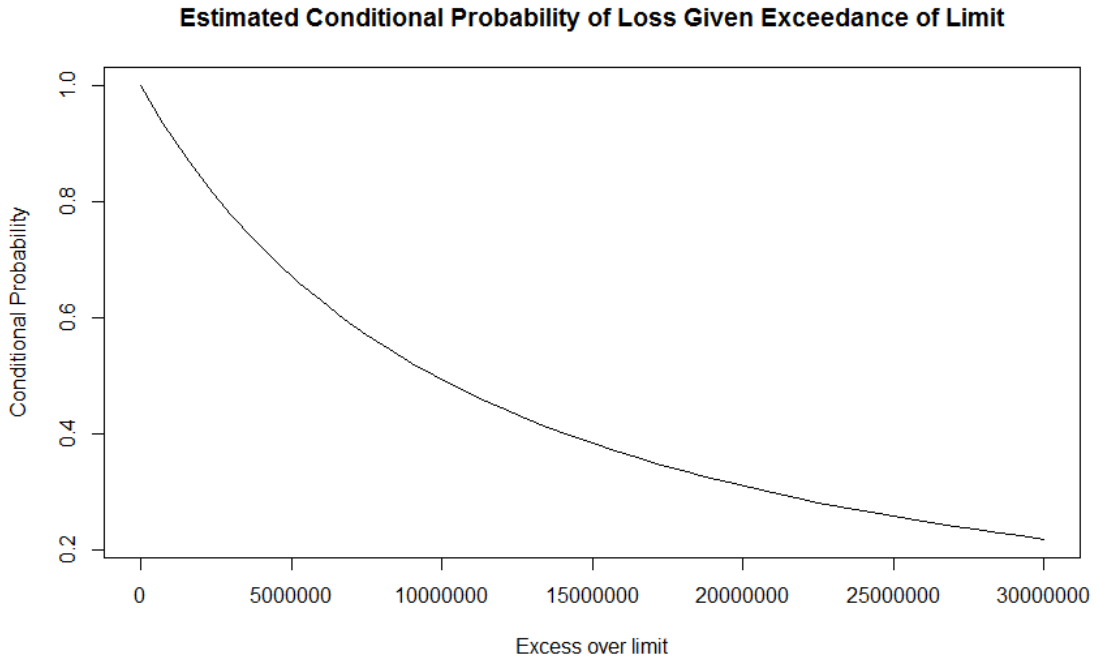

Figure 27. Conditional probability that excess loss over 15 000 000 is larger than x.

4.3.4.1 Simulation Study Using the Estimated Poisson-GPD Model

Again, 1000 simulations were drawn using inverse transform sampling. The maxima and sums of the threshold exceedances were plotted as histograms. The estimated GEV-distribution does not match the simulated histogram as well as in the previous section. The threshold model's fit seemed to be worse than for the block maxima, based on the goodness-of-fit plots. Therefore it is only natural that the Poisson-GPD model does not either accurately describe the data, but has a tendency to overestimate the possible claims. The results should therefore be viewed with caution.

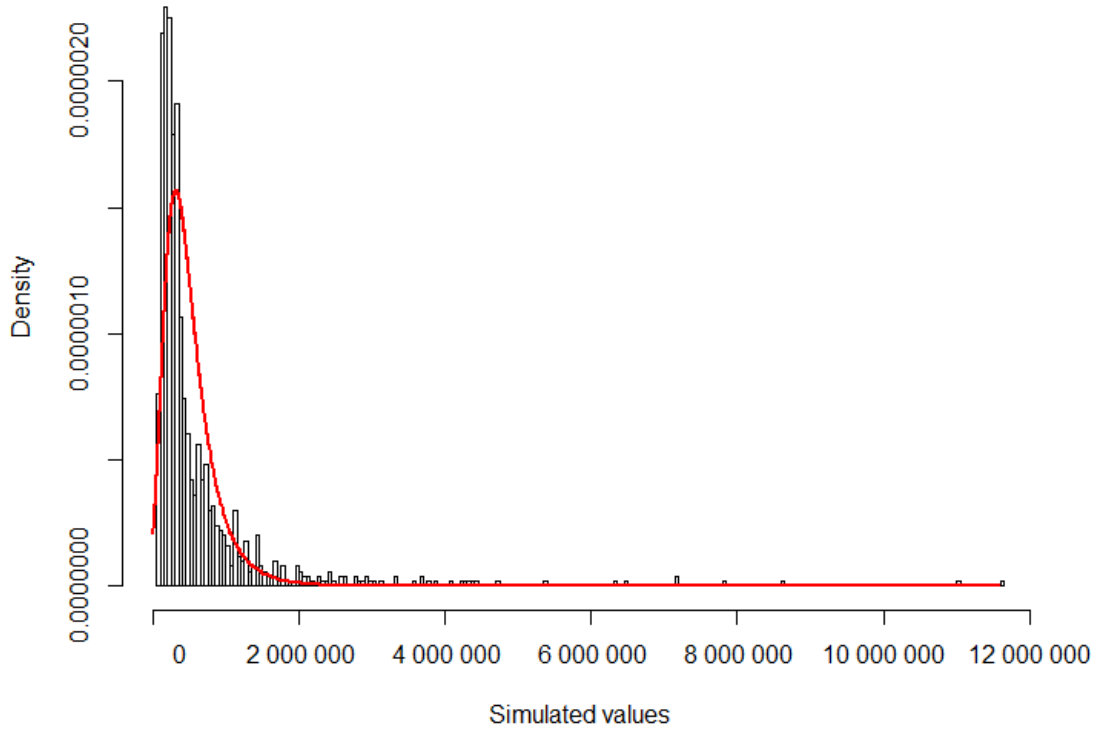

Figure 28. Histogram of simulated results for quarterly maxima of summed daily claims. The fitted GEV-pdf is overlaid in red.

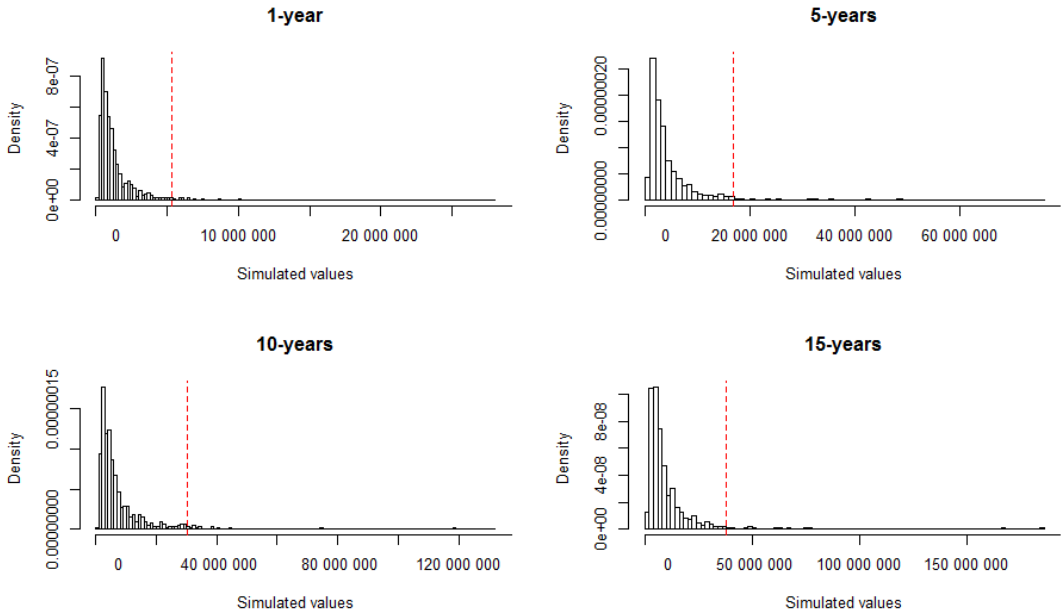

Figure 29. Histograms for 1000 simulated maxima for summed daily claims for different time periods. The dashed red line represents the 95 % quantile.

Table 19. 95 % quantiles from 1000 simulated maxima of threshold exceedances for summed daily claims.

	Quarter	1 year	5 years	10 years	15 years
95 % quantile	2 143 355	5 344 190	16 798 242	30 315 343	37 995 889

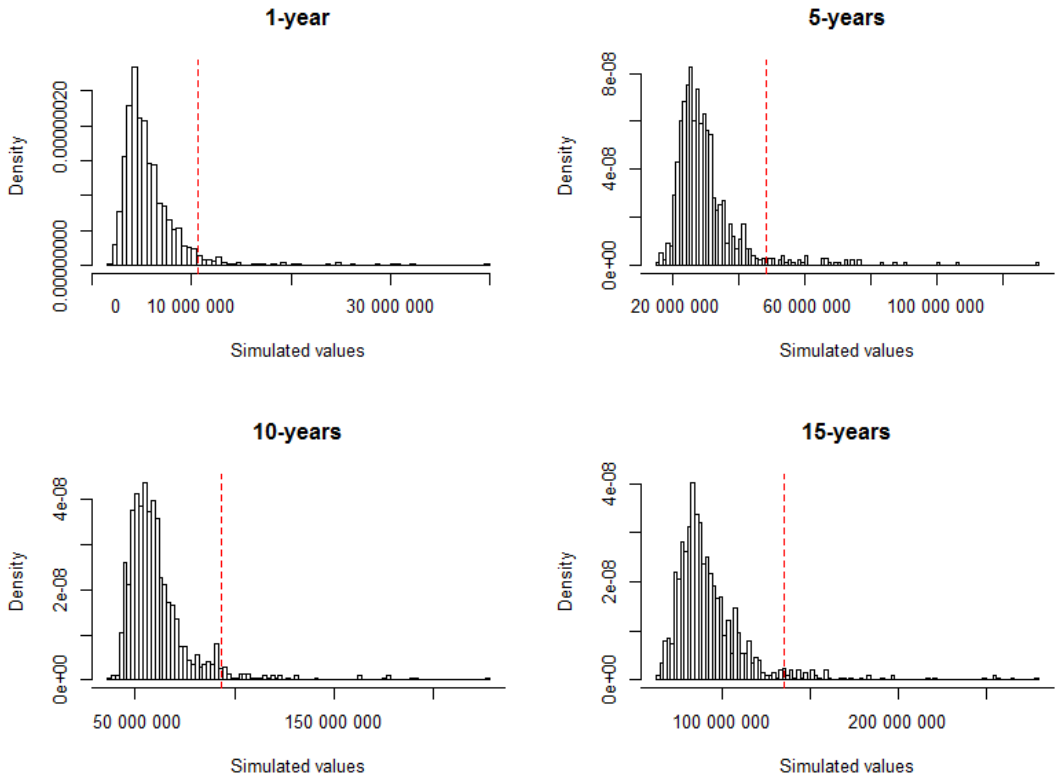

Figure 30. Histograms for 1000 simulated sums of threshold exceedances for summed daily claims for different time periods. The dashed red line represents the 95 % quantile.

Table 20. 95 % quantiles from 1000 simulated sums of threshold exceedances for summed daily claims.

	Quarter	1 year	5 years	10 years	15 years
95 % quantile	3 324 870	12 175 884	53 210 558	88 815 370	138 702 270

The sum of all summed daily claims exceeding 50 000 under five years is 95 % times under 53 210 558 in the simulation performed. The corresponding figure in the five year data is approximately 25 300 000, so the order of magnitude is reasonable.

4.3.5 Point Process

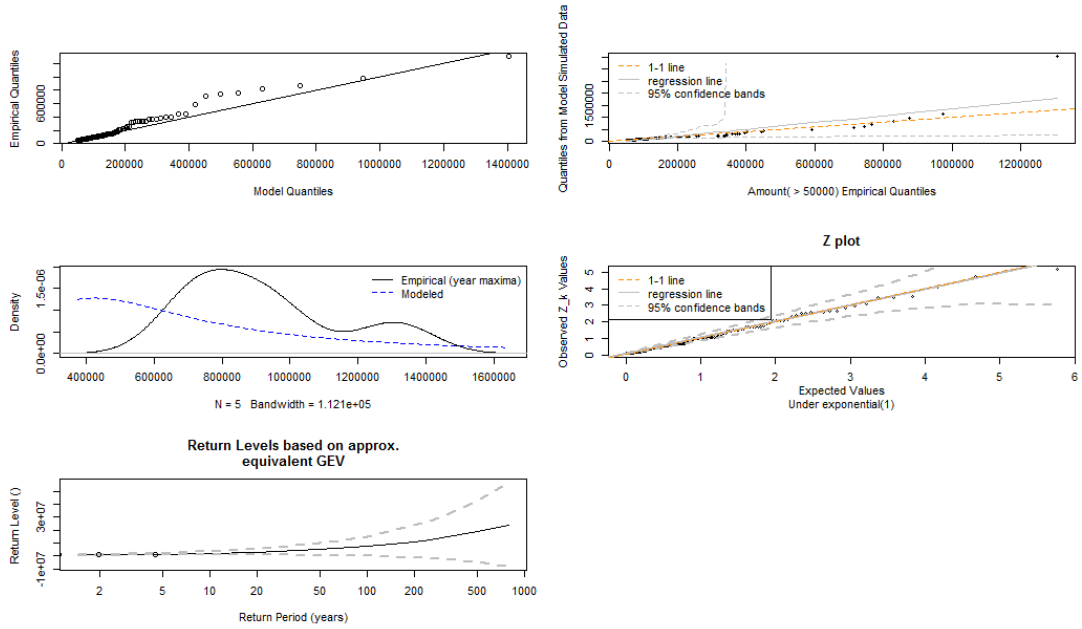

Figure 31. Goodness of fit diagnostic plots from package in2extRemes for point process fit for the summed daily claims data.

Table 21. Estimated parameters and return levels for the point process model, with 95 % confidence intervals obtained using the delta method

	Log Likelihood	Location ($\hat{\mu}$)	Scale ($\hat{\sigma}$)	Shape ($\hat{\xi}$)	5-year Return Level
Estimate	1594	555 000	328 000	0,556	1 320 000
95 % CI		[380 000; 730 000]	[159 000; 497 000]	[0.377; 0,733]	[658 000; 1 990 000]

The results look similar to the daily maxima case; that is the goodness-of-fit plots give a dual message; the occurrences of the events seem to be well modelled according to the z-plot, but the modelled and empirical densities do not coincide at all. The quantile plots hint of a certain bias in the model. Furthermore, the confidence intervals of the return level are considerably smaller than for the other models, especially when comparing to the confidence intervals received through profile likelihoods. This is of course on its own desirable, but coupled with the ill fit, it casts further doubt over the validity of the model.

5 DISCUSSION AND CONCLUSIONS

5.1 Summary of the Results

The results for the two data sets are, perhaps surprisingly, rather different. It seems that the first part of the analysis, that is for the data where only the daily maxima were used, the data is in general fairly well modelled by the different models proposed. This holds for both the block maxima and the threshold approaches. Conversely the data where all the claims per day where summed for each day, does not seem to be modelled very well, i.e. the results forecasted by the fitted distributions do not seem to describe the data well enough to be of practical use. The worse fit is of course unfortunate, especially since the summed data incorporates all the claims during the time period for the data and thus gives a more holistic picture of the claims process.

For the first part of the analysis and the block maxima approach, where quarterly maxima were modelled, the GEV-distribution resulted in a better model fit than the Gumbel distribution. The five year return level was estimated to be around 1,4 million, the order of magnitude being well in line with the data. The same figures for the Generalized Pareto fit and the GP-Poisson model were approximately 1,7 and 1,6 millions respectively. The thresholds used in the study were 45 000 and 50 000. These are comparable to the ones used by the company when defining large claims according to Mr. Nygård.

15 million was chosen as an example of a large single loss and potential retention level in reinsurance terms. The probability of exceeding this within the next five year was estimated to be 3 % and the estimated conditional distribution of exceedances over this limit is shown in Figure 13. This distribution could be used to price the net premium for an excess-of-loss reinsurance contract with a retention level of 15 million and suitable upper limit, as presented on page 2.

The simulation study showed a consistency between the GP-Poisson model and the fitted GEV-distribution. The simulated quantiles and calculated probable maximum losses also coincided well. The 95 % quantiles for the sum of all threshold exceedances obtained from the simulation and shown in Table 10, represent the closest working estimate of the aggregated risks the company faces presented in this study. For example, the sum of the daily maxima that exceed 45 000 has an estimated 5 % chance of exceeding 36 million during a period of five years. The sum of the threshold exceedances in the data is around 20 million which is located around the middle of the histogram for five years in Figure 16. All in all the figures obtained correspond fairly well to the data. The confidence interval for the intensity parameter is very small (since there are sufficient observations). The intervals for the GP-distribution's parameters are larger. Thus, the frequency of threshold exceedances is well modelled, but there is

greater uncertainty in their sizes. Also, as can be seen in Figure 11, the model seems to overestimate the sizes of the largest claims.

As pointed out earlier, the second part of the analysis was less successful and the goodness of fit was insufficient. This is especially the case for the threshold models. Based on the various goodness of fit tools, the block maxima fit is better than for the threshold models, but worse than for the similar analysis performed on the daily maxima data. Also, a difference to the first part is that it was not possible to decide whether the GEV or Gumbel models were better. The main evidence for the lack of fit can be seen in the quantile plot in Figure 25 where there is a clear deviation from the unit line. The return level tells the same story; the model adds a good measure of conservatism into the estimation by noticeably overestimating the sizes of the largest claims.

The same shortcoming is then naturally found in the GPD-Poisson model and all the results derived from it. This is unfortunate since the resulting quantiles of the sum of daily claims would have been of greater use than the quantiles for the daily maxima. The 95 % quantile for a five year period of the sum of all exceedances is estimated through the simulation to be around 53 million, which is around twice as large as the corresponding figure in the data. However, as pointed out, this is very likely to be an overestimate. The estimated distribution pictured in Figure 27 could have been used to price a stop-loss reinsurance contract, (like the XL-contract above) where it not for the insufficient fit.

Again, the point process fit obtained from the package in2extRemes is not very convincing when comparing the empirical and modelled densities in Figure 31. Also the return level plot is of no use as it extends to 1000 years and in2extRemes does not yet include a built in option to decrease the time span for the plot.

The idea behind the threshold models is to use more of the data and thus decrease the variance in the model. When comparing the results in the first part of the analysis, this seems to have worked. The confidence intervals for the block maxima model are larger than for the GP-model; using the figures obtained through profile likelihoods, the confidence interval of the shape parameter is around three times larger than the estimate and the ratio is about four times for the return level for the GEV-estimate. The same ratios are 1 and 2,5 for the GP-model, still representing large confidence bands, but a considerable improvement to the block maxima case.

5.2 Possible Sources of Uncertainty and Bias

The very large confidence intervals are clearly the biggest weakness of the study. These result from the relatively short period, five years, when the raw data was collected from.

A new database system was implemented in the company recently, so a longer data series could not be obtained for this study. Although the underlying insurance portfolio seemed to be fairly similar in character during the observation period, a longer time series would have brought that assumption into more doubt. The elapse of time is the biggest risk to the i.i.d. assumption underlying the analysis, as the behaviour of the insurance takers, the strategy of insurance companies, regulation and insurance lines change. All these factors can have an effect on the claim process, both the severity and frequency. So ironically, longer time series reduce the variance, but potentially induce a whole new bias into the simple models, where no trends or changes in the underlying process are assumed. Suggestions for signals that the claims process might be changing include changing volatility or autocorrelation in the process, changes in the regional composition or amount of legal persons in the insurance portfolio, changing average risk premium and changes in the distribution between different insurance types.

The second biggest shortcoming of the study is the insufficient fit of the second part of the analysis, performed on the sums of the daily claims. One potential reason for this is the possibility of (weakly) autocorrelated data, for which there is evidence in the plot of the autocorrelation function in Figure 2. This autocorrelation is not visible in the daily maxima data, where the fit is also much better. One solution for dealing with dependent data is the process of declustering. That means that if there are several large observations which result from the same event, e.g. a storm, only the largest one is recorded and the rest are removed from the data. However, as mentioned earlier, such clusters are not very frequent in the data. Out of the 161 exceedances of 50 000, 16 were preceded by an exceedance and only one came after two consecutive exceedances. An attempt at modifying the data by removing the related observations in these 16 consecutive exceedances could be made and perhaps the autocorrelation would diminish. There are three recurring reasons for larger claims. Many are due to fires, which are probably not related. Flooding is also a reoccurring reason for a claim. These might be manmade and hence probably not correlated or they might stem from the same weather event. Storms are the third typical reason, and these are clearly related in most cases and declustering might help reduce the correlation between consecutive daily observations.

The estimates for ξ were all positive, which would correspond to the Fréchet distribution in the original nomenclature. This distribution does not have an upper limit for the possible values (but rather a fat tail). Some of the confidence intervals for the shape parameter contained negative values, but that might rather stem from the large confidence intervals than the possibility that the data would be better described by a negative shape parameter in the model at question. In a sense, this is unfortunate as a GEV-distribution with a negative shape parameter (corresponding to the Weibull distribution in the original notation form) has a finite upper bound, $z_{sup} = \mu - {}^{\sigma}/_{\xi}$. This

would provide a more correct description for the process at hands, as the insurance policy holder's claims are in reality limited to the underwritten amount.

5.3 Suggestions for Improvements and Future Research

In relation to the possible effect of whether the day is a business day or not more could be done. The difference between the number of claims of any size depending on the day type was, for example, not studied at all. A more thorough and systematic study on this effect in general could be conducted. If it turns out that there is a statistically significant difference, a seasonal model could be used. That would, for example, entail the assumption that the claim process consists of two different sub processes, one for business days and another for none business days, and then based on that estimating a different set of parameters for both and investigating whether this model is an enhancement to the obtained ones. Also the change in the percentage of the total insurance portfolio that consists of insurances to legal persons could be monitored to see if this effect changes over time.

The declustering mentioned on page 48 could also be tried, hopefully leading to better results for the modelling of the summed daily data. Additionally, the usefulness of covariates could be studied. Possible candidates might include wind speeds, levels of rainfall or the water level in critical rivers and lakes, all potentially signalling larger property and home insurance claims. Another variable which deserves more analysis is the type of claim. It is possible that the different claim types show varying characteristics and should hence be modelled separately at first. The separate claim processes could then be aggregated into a more exact holistic model.

The threshold models used in the thesis are basically equal in the most rudimentary case where the underlying process is assumed to be constant. The Poisson-GDP and point process characterizations become more useful if it assumed that the process is in fact non-stationary. For example the intensity parameter can be assumed to be time dependent $\lambda(t)$ to account for seasonality or trends can be included in parameters if the underlying insurance portfolio is growing. So although the point process characterization is a little superfluous to the analysis presented in this paper, it is of use if the analysis were continued and expanded.

Finally, instead of improving the analysis by making the current models more flexible, completely different distributions could be tried. Reiss and Thomas (1997) mention alternative distributions employed by actuaries such as Benktander II and truncated Weibull. Right truncated distributions seem like an obvious improvement, as there is a real, finite upper-end point to the possible values the claims can reach.

BIBLIOGRAPHY

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

Coles, S., & Davidson, A. (2008, January). Statistical Modelling of Extreme Values. http://stat.epfl.ch.

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance.* Springer.

Finlands officiella statistik. (2014, September). *Konsumentprisindex, tabeller.* Retrieved November 10, 2014, from Statistikcentralen.fi: http://www.stat.fi/til/khi/2014/09/khi_2014_09_2014-10-14_tau_005_sv.html.

Fisher, R., & Tippett, L. (1928, April). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society , 24* (02), pp. 180-190.

Gesmann, M., Rayees, R., & Clapham, E. (2013). A known unknown. *The Actuary* .

Gilleland, E., & Katz, R. (2014). extRemes 2.0: An Extreme Value Analysis Package in R. *Journal of Statistical Software* , 32.

Gilleland, E., & Katz, R. (2005, 7). Extremes Toolkit (extRemes):Weather and Climate Applications of Extreme Value Statistics. Boulder, Colorado, USA.

Gilleland, E., & Katz, R. (2011, 1 11). New Software to Analyze How Extremes Change Over Time. *EOS 92(2)* , pp. 13-14.

Gulati, N. (2009). *Principle of Insurance Management.* Excel Books India.

Katz, R. (2008). *Background on Extreme Value Theory with Emphasis on Climate Applications.* Boulder: Institute for Study of Society and Environment National Center for Atmospheric Research.

Leppisaari, M. (2013). *Modeling catastrophic deaths using EVT with a microsimulation approach to reinsurance pricing.* Helsinki: Model IT & Deparment of Mathematics and Systems Analysis, Aalto University.

Lundberg, F. (1903). *Approximerad Framställning av Sannolikehetsfunktionen, Återförsäkering av Kollektivrisker.* Uppsala: Almqvist & Wiksell.

Munich Reinsurance America, Inc. (2010). *Reinsurance: A Basic Guide to Facultative and Treaty Reinsurance.* Princeton: Munich Reinsurance America, Inc.

Om oss: Folksam Skadeförsäkring Ab. (2014, 11 23). Retrieved 11 23, 2014, from Folksam Skadeförsäkring Ab: https://www.folksam.fi/sv/folksam/om-oss

Reiss, R.-D., & Thomas, M. (1997). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields.* Basel: Birkhäuser Verlag.

Rootzén, H., & Tajvidi, N. (1997). Extreme Value Statistics and Wind Storm Losses: A Case Study. *Scandinavian Actuarial* , 70-94.

Ross, S. (2010). *Introductory Statistics.* Academic Press.

Rytgaard, M. (2006). Stop-Loss Reinsurance. In *Encyclopedia of Actuarial Science. 3.* John Wiley & Sons, Ltd.

Sheaf, S., Brickman, S., & Forster, W. (2005). Claims Inflation Uses and Abuses. *GIRO conference 2005*, (pp. 1-29).

Shumway, R. H., & Stoer, D. S. (2011). *Time Series Analysis and Its Applications With R Examples.* Springer.

Smith, R. (1985, April). Maximum Likelihood Estimation in a Class of Nonregular Cases. *Biometrika , Vol. 72* (No. 1), pp. 67-90.

Smith, R. (2004). Statistics of Extremes, with Applications in Environment, Insurance and Finance. In B. Finkenstädt, & H. Rootzén, *Extreme Values in Finance, Telecommunications and the Environment* (pp. 10-17). Chapman & Hall/CRC.

Smith, R., & Goodman, D. (2000). Bayesian risk analysis. In e. b. Embrechts, *Extremes and Integrated Risk Management* (pp. 235-251). London: Risk Books.

Smith, R., & Shively, T. (1995). A Point Process Approach to Modeling Trends in Troposperic Ozone. *Atmosperic Environment , 29* (23), 3489-3499.

Stephan, V. (2013). *Managing P&C Insurance Portfolios in an Uncertain Inflation Environment.* Goldman Sachs Asset Management.

Venables, W., Smith, D., & Team, R. C. (2012, 6 22). *An Introduction to R*. Retrieved 10 27, 2012, from The Comprehensive R Archive Network: http://cran.r-project.org/doc/manuals/R-intro.pdf

6 APPENDIX

These are the values removed from the histograms showing the simulated valued. They were not included in the histograms for graphical purposes. Of course the values change with every simulation, but they are nevertheless presented below, to give the interested reader a sense of the order of magnitude that can exist in the very right end of the distribution assumed.

Table 22. Removed vales from the histogram with simulated maxima for daily maxima claims.

1 year	5 years	10 years	15 years	Quarter
9 606 766	30 693 846	46 256 912	58 023 701	5 277 479
10 025 905	31 552 276	48 954 272	63 470 931	5 608 133
10 645 191	32 315 521	52 145 455	74 527 173	5 816 092
10 741 795	36 040 884	53 608 722	80 213 604	6 651 289
11 547 842	36 267 212	55 929 275	90 993 370	7 238 493
13 868 440	49 350 942	65 202 524	95 209 137	7 316 193
15 502 460	57 649 434	89 603 951	100 950 372	7 325 105
17 171 124	59 251 081	95 155 981	111 687 390	11 750 384
18 149 335	83 457 497	95 725 280	198 400 445	20 718 158
48 338 707	379 102 410	147 578 328	324 709 100	40 298 635

Table 23. Removed values from the histogram with sums of threshold exceedances for daily maxima claims.

1 year	5 years	10 years	15 years	Quarter
14 593 776	55 294 886	89 707 041	129 052 813	6 378 039
14 964 430	56 344 735	91 558 745	136 215 079	7 081 387
15 717 197	56 539 183	98 080 786	139 522 482	7 730 310
16 388 540	59 733 872	103 923 717	148 355 568	8 065 348
18 284 183	61 441 320	104 103 215	159 131 265	8 476 287
18 493 184	67 693 173	111 066 318	164 449 590	8 902 591
20 882 664	79 122 084	128 478 409	174 815 338	9 062 011
22 726 004	80 272 693	140 516 876	181 508 518	12 152 721
23 358 965	107 710 215	157 086 634	269 030 416	21 839 522
53 663 834	390 759 266	204 937 751	383 375 490	40 923 048

Table 24. Removed vales from the histogram with simulated maxima for summed daily claims.

1 year	5 years	10 years	15 years	Quarter
20 943 950	55 116 691	117 996 786	152 246 899	7 119 671
21 653 267	60 478 550	118 162 463	166 342 820	7 777 493
22 245 961	64 357 049	125 311 097	166 523 673	8 587 986
27 082 746	73 107 944	126 961 597	185 761 535	10 972 962
27 773 808	75 034 657	131 605 004	185 814 252	11 580 713
56 469 451	106 410 933	181 245 251	188 931 836	15 275 835
64 471 460	108 673 983	189 897 861	222 500 260	20 059 903
148 308 973	113 529 031	199 120 739	253 862 980	30 428 014
170 416 876	229 442 276	281 899 087	266 378 954	37 145 370
227 629 282	461 354 607	386 069 123	485 283 089	45 367 966

Table 25. Removed values from the histogram with sums of threshold exceedances for summed daily claims.

1 year	5 years	10 years	15 years	Quarter
28 820 033	87 668 692	176 895 772	254 644 617	8 236 823
30 156 771	90 421 579	177 878 228	254 740 983	9 124 207
30 784 777	100 133 141	189 722 320	256 792 851	9 692 863
32 479 116	106 348 999	190 055 120	264 891 580	11 741 611
39 722 778	130 532 957	226 292 727	278 682 682	13 085 317
61 819 401	137 030 530	228 509 837	278 870 618	20 466 760
69 923 455	143 151 390	247 726 060	310 489 990	22 437 081
155 363 555	160 118 336	263 284 019	345 328 009	31 396 649
177 039 401	259 203 495	330 848 907	412 933 513	38 014 226
234 428 126	488 227 228	442 114 705	551 724 936	46 005 045

Master's Theses in Mathematical Sciences 2015:E8
ISSN 1404-6342

LUTFMS-3273-2015

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/