university of groningen

faculty of science and engineering

mathematics and applied mathematics

Predicting Catastrophes: an Analysis of the Earthquake Data in Groningen using Extreme Value Theory

Bachelor's Project Mathematics

Author:
Aras Ali
S.No. 2291835

Supervisors:
Dr. A.E. Sterk
Dr. M.A. Grzegorczyk

July 2019

Abstract

As a result of the gas extraction in the northern part of the Netherlands many earthquakes have occurred. We argue that in order to get a better understanding of the earthquake magnitudes intensities and their occurrence rates, modelling techniques using Extreme Value Theory are required. We present estimates and confidence intervals for the expected maximum magnitudes using various EVT based techniques. Moreover, we compare the different results we obtain and argue which is the most reliable one.

Contents

1 Introduction

General model estimation in its simplest form requires a basic knowledge of statistics. These models are helpful to gain insight in the data and the process of which the data is generated from. Moreover, it helps us to extrapolate to some extent and predict future events. However, the problem with this approach is that the most important part of the data, the extremes, also widely known as outliers, are either being neglected or left out of the final analysis. These outliers could actually be considered the most important part of the data. The questions one should ask, would be; Why do the outliers not behave similarly to the rest of the data? What is the impact of these extreme values? Can we predict these extreme values? In most cases, luckily, extreme values do not have a significant effect on the process they are obtained from. They occur every so often and their effects fade away without much distress, leaving no mess behind. However, this cannot be said of every process. When predicting the financial returns in, for example, insurance companies, stock portfolios and other financial settings the extreme events tend to get significantly more attention. The reason is that the consequences of the risks may consist of going bankrupt, or even worse, as we saw in the setting of 2008, invoking a financial crisis affecting the whole economy and thus every individual. Another field in which the extreme events occur and have a non-negligible effect is in the setting of natural catastrophes. Mother nature has a tendency to not abide to our rules and be very unforgiving. Having a major effect on our surroundings and life in general, natural catastrophes have resulted in over 50.000 deaths annually on a global level since the beginning of the millennium and over a few hundred thousand annually in the preceding century [1]. These natural catastrophes include storms, floods, earthquakes, and droughts. The prediction of such extreme events could therefore benefit our whole society.

In this research our goal is **to compute a maximal expected magnitude using the available earthquake data obtained from the Groningen gas fields between 1986 and 2019** provided by the KNMI by applying Extreme Value theory based techniques and comparing the outcomes with old researches.

2 Problem Formulation

2.1 Modelling of Natural Hazards

The modelling of natural phenomenona using EVT has received much attention. To model hurricane speeds, Coles and Casson [6] have worked with a re-parameterisation of the Generalized Pareto distribution. This allowed them to compare models with different thresholds. Here they have used the method of maximum likelihood to estimate the parameters. They state that reasons to use this method of inference include its asymptotic efficiency and it allows them to specify the estimation uncertainty. However, since the amount of data used is relatively small this would result in very large confidence intervals for the parameters. Another important reason to use the method of maximum likelihood is the possibility in developing a spatial model of the variability of the hurricanes. This allows them to observe if any regions are at more risk than others.

In 'Fighting the arch-enemy with mathematics' [7] de Haan utilizes multiple different techniques to estimate the necessary height of the dutch dikes to lower the probability of a flood to approximately one in 10000. Among other techniques, he makes use of the Block Maximum method and the Peaks Over Threshold method. For both methods he uses the maximum likelihood as a method of estimation and inference. The other techniques being used are derivations from the EVT and are described in the paper, for even more detailed derivations see de Haan and Ferreira [8]. He also mentions that using the BM method information is lost, however an advantage is that no selection procedure is necessary.

2.2 Modelling of Earthquake Magnitudes

The biggest natural gas reserve of Europe is located in the most northern part of the Netherlands. The extraction of this gas began in 1963 and most of it has been extracted since then. A consequence of the extraction of this huge gas reserve are induced earthquakes [10]. These earthquakes occur frequently with mostly low magnitudes. However, from the presented data it follows that the number of occurrences of earthquakes have increased. Moreover, a number of earthquakes have occurred with a magnitude of approximately 3.5

on the Richter scale. Even though these magnitudes are not catastrophic, damage is done to surrounding civilian owned properties, damage that has to be compensated. In order to quantify the risks that are paired with the gas extraction we need a reliable model to estimate the expected intensity and frequency of these induced earthquakes.

Methods on modelling of earthquakes can usually be divided into two different main categories. These two methods consist either of a deterministic basis or a probabilistic basis. The deterministic approach most often applied is based on empirical relationships between the magnitude and multiple geological parameters. Fault parameters which are important to earthquake hazard analysis include; rupture length, downdip rupture width and rupture area [15]. These methods are developed for different seismic areas and different faults, dependent on the geological structures and positions of the regions. However, many more parameters are being used in the prediction procedure. In a significant amount of cases, it is therefore true that the result of any deterministic procedure is very uncertain. Kijko and Graham [12] claim that this uncertainty can even reach a value of up to one unit on the Richter scale.

The value of the maximum regional earthquake magnitude, can also be estimated based on historical seismological data of a certain area by the use of appropriate statistical estimation procedures. Many studies have been dedicated to the estimation of this parameter (Kijko and Graham [12]; Pisarenko et al. [12]; Raschke [13]; Beirlant et al. [3]). In order to estimate this maximum magnitude we will have to estimate the tails of a fitted distribution.

In our research we will focus on probabilistic methods. In particular, methods using main ideas from Extreme Value Theory will be applied and evaluated. The methods being used in previous researches are not always using EVT based analysis. However, in this research we have the aim to determine which EVT method is the most reliable method. This will allow us to make a comparison between our results and the results of previous researches based on EVT.

3 Tail Modelling Theory

Outliers in data modelling become important if the risks are of an extreme type, resulting in extreme effects. The extreme events occur with very low probability and are located in the most right, or respectively left, part of the tail of the probability distribution. Often, the extreme events in the right tail are the events of interest. These risks have led to the extensive research on the prediction of non-negligible extreme events or catastrophes, also called black swan events by the writer Nassim N. Taleb [14]. Through extensive research, many methods have been developed to estimate these events, with the main focus on the estimation of the tails. The general consensus regarding these tails, also called fat-tails due to their properties, is that they follow a different distribution and are thus modeled wrong initially. One of the popular methods to evaluate the tails of a distribution is by means of Extreme Value Theory (EVT). This has been developed as early as in the 1950's, but has had a huge development recently due to the introduction of fast computers, allowing for advanced numerical methods and computations.

The classical EVT in its general form focuses on the statistical behaviour of

$$M_n = \max\{X_1, X_2, \ldots, X_n\},$$

assuming that $X_1, X_2, \ldots, X_n$ consists of a sequence of independent random variables having a common distribution function F [5]. Here the X_i, $1 \leq i \leq n$, usually represent the monthly, yearly or other time based values of a certain process in which we are interested. The value M_n is the maximum value of the sequence, subject to the time interval chosen. Now, assuming the distribution function F is unknown, a distribution function F^n that fits the maximum values needs to be found. This follows the same analogy as the central limit theory, but then for extreme values. However, a degenerate limit is encountered in the asymptotic behaviour of the density function. This difficulty is avoided by a linear re-normalization of the variable M_n, given by

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

for sequences of constants $\{a_n > 0\}$ and $\{b_n\}$. This stabilizes the location and the scale of M_n^* as n increases. Therefore, instead of the limiting distribution of M_n, the EVT needs to find the limit distributions of M_n^* with the corresponding values of $\{a_n > 0\}$ and $\{b_n\}$ [5]. If the appropriate sequences for $\{a_n > 0\}$ and $\{b_n\}$ can be determined such that M_n stabilizes then the limit distributions to which M_n^* can belong must be one of the Gumbel, Fréchet or Weibull families, each with different parameters. Given by the Extremal Types theorem [5].

Theorem 1. *If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$Pr\{(M_n - b_n)/a_n \leq z\} \rightarrow G(z) \text{ as } n \rightarrow \infty,$$

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

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

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

$$III(Weibull) : G(z) = \begin{cases} \exp\big\{ - \big[- \big(\frac{z-b}{a}\big)^{\alpha}\big]\big\}, & z < b \\ 1, & z \geq b, \end{cases}$$

for parameters $a > 0$, b and, in the case of families II and III, $\alpha > 0$.

3.1 Generalized Extreme Value distribution (GEVD)

Determining which one of these families has the best fit is a tedious job with each of them having different properties and tail behaviour causing problems if the wrong one is chosen. Combining these three families into one gives rise to the **Generalized Extreme Value** family of distributions (GEVD). This distribution is given by **theorem 3.1.1** in [5].

Theorem 2. *If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$Pr\{(M_n - b_n)/a_n \leq z\} \rightarrow G(z) \text{ as } n \rightarrow \infty$$

for a non-degenerate distribution function G, then G is a member of the GEVD family

$$G(z) = \exp\Big\{ - \Big[1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\},$$

which has three parameters, namely: a location parameter μ, a scale parameter σ and a shape parameter ξ. The shape parameter ξ is also widely known as the tail index [5]. This combination makes fitting of the data easier. Instead of analyzing the data and making a choice on which of the three distributions should be chosen, the data can be fitted immediately using the GEVD. It should be noted that if the shape parameter $\xi < 0$ the distribution is bounded and a computable maximum value exists. This is not true for all other values of the shape parameter, resulting in unbounded distributions.

The GEVD makes use of a model for the distribution of block maxima. The method of block maxima cuts the data in blocks of equal length. Each block delivers its maximum value and the set of these maxima is then used in the prediction of the distribution of the extreme values. However, the size of each block is important. Small blocks may not give a good representation because of seasonality or trends in time series data resulting in high bias. Large blocks require much data and leave us with only a few block maxima and hence more variance. Therefore, the size of the blocks is a trade-off between bias and variance and one has to choose the size carefully keeping the data structures in mind.

3.2 Generalized Pareto distribution (GPD)

As discussed, the method of block maxima uses the highest value from each block. However, in the case of some blocks having multiple high values while others do not have any, important values might not be used in the analysis. This is a waste of useful information in determining the distribution of these high values. In order to circumvent this problem, one could set a certain high value and regard all of the observations exceeding this threshold as extreme values. This will make sure that all of the extremes will be used in the analysis. This gives rise to the Peaks over Thresholds method, which estimates the parameters of the tail distribution using all of the exceedances over a set threshold. Again, there is a trade-off between variance and bias in setting the threshold. Setting it too low, will result in observations being evaluated which are not at all extreme. Setting it too high, will leave us with not a sufficient amount of observations to make proper inferences about the parent distribution. The distribution of these exceedances is related to the GEV distribution, its main result is captured in the next theorem.

Theorem 3. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with common distribution function F, and let*

$$M_n = \max\{X_1, \ldots, X_n\}.$$

Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies ***theorem 3.1.1 (or Theorem 1 as stated previously)****, so that for large n,*

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

where

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

for some μ, $\sigma > 0$ and ξ. Then, for large enough u, the distribution function of $(X - u)$, conditional on $X > u$, is approximately

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

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

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

The family of distributions defined by H is called the **Generalized Pareto Family**. For more details on the derivation and properties of this distribution the reader is referred to the literature [5]. It follows again that only if the shape parameter $\xi < 0$, the distribution is bounded and thus a maximum expected value can be computed.

4 Parameter Estimation Theory

There exist multiple methods to estimate the parameters in the extreme value models. Common procedures include; estimation of the parameters through order statistics, moment-based techniques and likelihood based methods. However, the likelihood based techniques are very attractive [5]. Due to its adaptability and statistical properties.

The second method that will be used during this research is the L-moments method. The L-moments method uses linear combinations of order statistics in order to estimate the parameters. The theoretical background of this method is researched by Hosking [11]. This method has not yet been used to estimate the parameters of the GEVD and the GPD in the Groningen case. It is even claimed that estimates obtained through L-moments from small samples are sometimes more accurate than those obtained through MLE [11]. Therefore, it is of interest to compare the obtained estimates with those obtained through Maximum likelihood estimation.

4.1 Maximum Likelihood estimation: GEVD

The MLE method assumes that the observations are independent and come from the same GEVD distribution. In order to estimate the three parameters of the GEVD from theorem 2, the log-likelihood of the distribution distinguishes between the cases with shape parameter $\xi = 0$ and $\xi \neq 0$. In the first case the GEVD has the Gumbel limit which leads to the log-likelihood

$$\ell(\mu,\sigma) = -m\log\sigma - \sum_{i=1}^{m}\Big(\frac{z_i-\mu}{\sigma}\Big) - \sum_{i=1}^{m}\exp\Big\{-\Big(\frac{z_i-\mu}{\sigma}\Big)\Big\}.$$

For shape $\xi \neq 0$ the following log-likelihood is obtained:

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

if

$$1+\xi\big(\frac{z_i-\mu}{\sigma}\big) > 0,\ \text{for } i=1,\ldots,m.$$

Maximizing these functions in each case with respect to the parameters yields the estimators of the GEVD. Moreover, during this process the information matrix is easily obtained. Taking the inverse of the information matrix and evaluating it at the obtained estimates yields the variance co-variance matrix.

4.2 Maximum Likelihood estimation: GPD

Before it is possible to estimate the parameters of the GPD, a suitable threshold has to be set. Assume a suitable threshold has been determined and denote it by u. Denote by $y_1, \ldots, y_k$ the k excesses over this threshold u. Then for $\xi \neq 0$ the log-likelihood function of the GPD from theorem 3, is given by

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

if,

$$(1 + \xi y_i/\sigma) > 0 \text{ for } i = 1, \ldots, k. \text{ Otherwise } \ell(\sigma, \xi) = -\infty.$$

If our shape parameter takes on the value $\xi = 0$, then the distribution is unbounded. The log-likelihood of the GPD function for $\xi = 0$ is then given by

$$\ell(\sigma) = -k \log \sigma - \sigma^{-1} \sum_{i=1}^{k} y_i. \tag{2}$$

Maximizing these functions again yield the estimates of the parameters. Moreover, the mean and variance of these estimators are obtained through the same method as before.

4.3 L-moments method: GEVD

If X is a real-valued random variable then for a sample of size n drawn from the distribution of X the L-moments of X are given by

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} EX_{r-k,r}, \quad r = 1, 2, \ldots \tag{3}$$

Where X_{r-k}, r are the order statistics $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ and $EX_{r-k,r}$ is the expected value of $X_{r-k,r}$.

The values of the different λ_r are obtained through the `R` package `Lmoments`. The moments relate to the parameters of the GEVD through the following equations

$$\lambda_1 = \mu - \frac{\sigma}{\xi}(1 - \Gamma(1 - \xi)), \tag{4}$$

$$\lambda_2 = \frac{-\sigma}{\xi}(1 - 2^{\xi})\Gamma(1 - \xi), \tag{5}$$

$$\tau_3 = \frac{\lambda_3}{\lambda_2} = 2\frac{1 - 3^{\xi}}{1 - 2^{\xi}} - 3. \tag{6}$$

Where Γ is the Gamma function. Note that (6) is not numerically solvable, therefore using the `R` package `RootSolve` a Newton method is required to obtain the solution for the shape parameter ξ. Substituting the obtained results in (5) and subsequently in (4) yields all of the parameter estimates. Again the mean of the parameters is equal to their estimated value. However, in order to obtain the variance co-variance matrix a transformation has to be made with respect to the relations above. First, each parameter has to be written in terms of λ_i for $1 \leq i \leq 3$. Thereafter, using the delta method, the error of the estimated shape parameter is obtained by

$$\mathrm{Var}(\hat{\xi}) \approx \nabla\xi^T V \nabla\xi \tag{7}$$

where,

$$\nabla\xi^T = \left[\frac{\partial\xi}{\partial\lambda_1}, \frac{\partial\xi}{\partial\lambda_2}, \frac{\partial\xi}{\partial\lambda_3}\right] \tag{8}$$

and V the variance co-variance matrix of λ_i for $1 \leq i \leq 3$, which is obtained through the `R` package `Lmoments`. Going through the same steps for the location and scale parameters, the error estimates for the respective parameters are obtained. For a more detailed look into the theory behind the L-moments and its derivations the reader is referred to the literature [11].

4.4 L-moments method: GPD

The moments relate to the parameters of the GPD through the following equations:

$$\lambda_2 = \frac{\sigma}{(1-\xi)(2-\xi)}, \tag{9}$$

$$\tau_3 = \frac{\lambda_3}{\lambda_2} = \frac{1+\xi}{3-\xi}. \tag{10}$$

Since the values of λ_r and its variance co-variance matrix are obtained by the `R` package `Lmoments`, (10) is numerically solvable. Substituting the obtained estimate in (9) yields the required parameter estimates. Writing the parameters as functions of the L-moment values allows us to calculate the errors using the Delta Method, using the same reasoning as in the previous section. Note that the variance co-variance matrix for the L-moment estimators is readily obtained through the `Lmoments` package in `R`. Moreover, the `R` package `RootSolve` is used to calculate the required gradients for each parameter as in (8).

5 Data Exploration

The data set contains 1606 observations as of may 2019. The events are collected by the KNMI since 1986 [2]. However, between 1986 and 1992 only a few observations are available. These observations are mostly of earthquakes with magnitude above 2.0 on the Richter scale. This is in line with the literature, suggesting that many lower magnitudes are missing from older catalogues [12]. The data set contains the date of the events, the magnitude of the earthquakes and the latitude and longitude of the epicenter of the earthquakes.

Plotting the data in frequencies, figure 1 is obtained. This figure gives much insight into the distribution of the data. The first observation, is that the data does not follow a symmetric distribution. Particularly, the data is heavily left skewed indicating that it is mathematically incorrect to make inferences using the normal distribution. Thus, a different distribution is indeed needed to model the tail correctly.

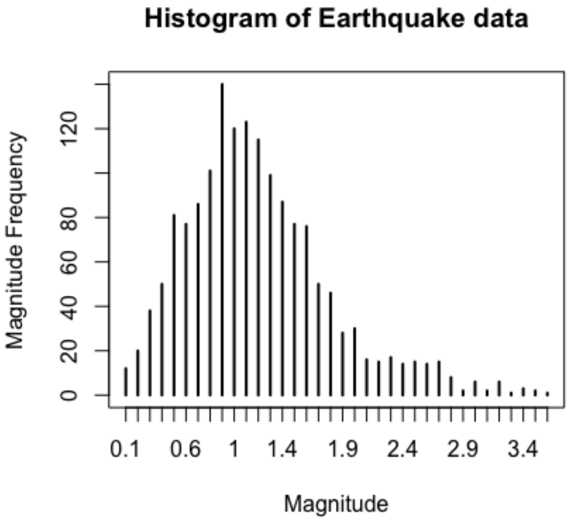


Figure 1: Histogram of the Data. Shows a left skewed distribution. Most of the magnitudes are below 2.0. However, a significant amount of occurrences remain above this level.

5.1 Stationarity of Data

The data has mean $\mu = 1.2$ and a standard deviation of $\sigma = 0.6$. This tells us that less than 5% of the earthquakes have magnitude above 2.5. Moreover, the maximum occurred earthquake had a magnitude of 3.6 which occurred on the sixteenth of august 2012. The observations with a high magnitude which occur rarely are the part of the data we are most interested in during this research. Therefore, its behaviour throughout the years needs to be examined. More particularly, it is of interest to check if the data has become more extreme and therefore if its distributional features have changed. This requires us to test for stationarity of the data. Plotting the magnitudes of the observations against the time figure 2 is obtained. This figure does not indicate any strong form of violation of stationarity of the data, nor are there any signs of trends or seasonality. The only remarkable observation is the vast increase of yearly earthquakes throughout the years as is seen by the increased density in the right. However, it does not seem that the underlying distribution has changed significantly.

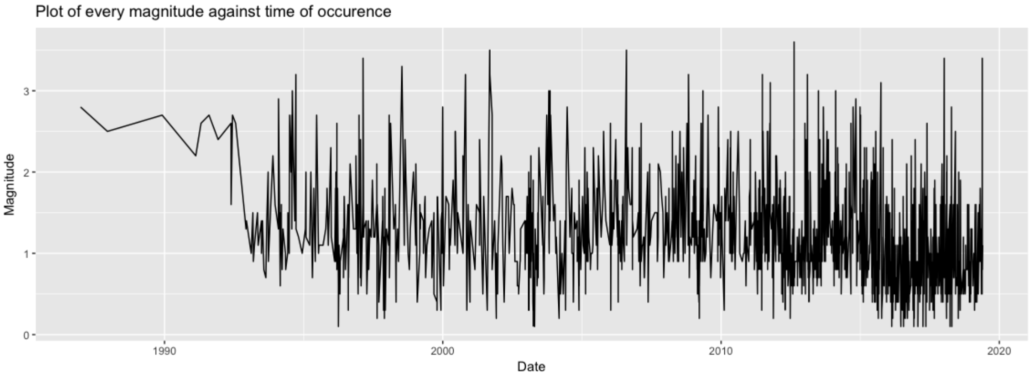
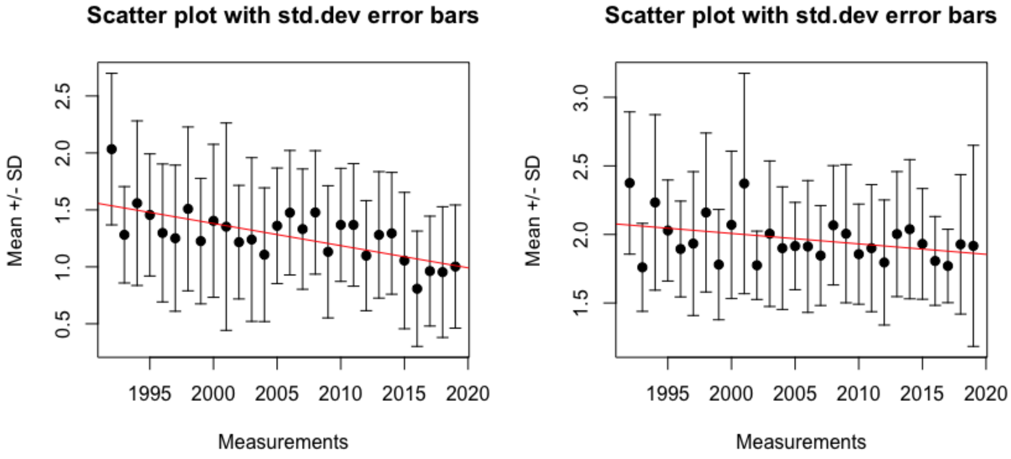


Figure 2: Every earthquake its magnitude is plotted against its time of occurrence. There are no signs of trends or seasonality. It is clear that the number of earthquakes have increased in the last 15 years.

In order to check if the data has remained similar throughout the years, the following figures are obtained. Figure 3a shows the mean and variance for each individual year. It is noticed that the variance estimates remain relatively constant. However, it seems as if the means of the subsets of data decrease constantly throughout the years. As stated earlier, in the first few years only data on earthquake magnitudes above 2.0 is available. Moreover, the equipment to detect earthquakes of lower magnitude has improved throughout the years , resulting in more observations for earthquakes of lower magnitude. Correcting for these factors, figure 3b is obtained. This figure shows all of the earthquakes with magnitude above the threshold of 1.5. From this figure it is clear that the means and variances of the larger observations have remained relatively constant, showing no significant signs of violation of stationarity of the data.

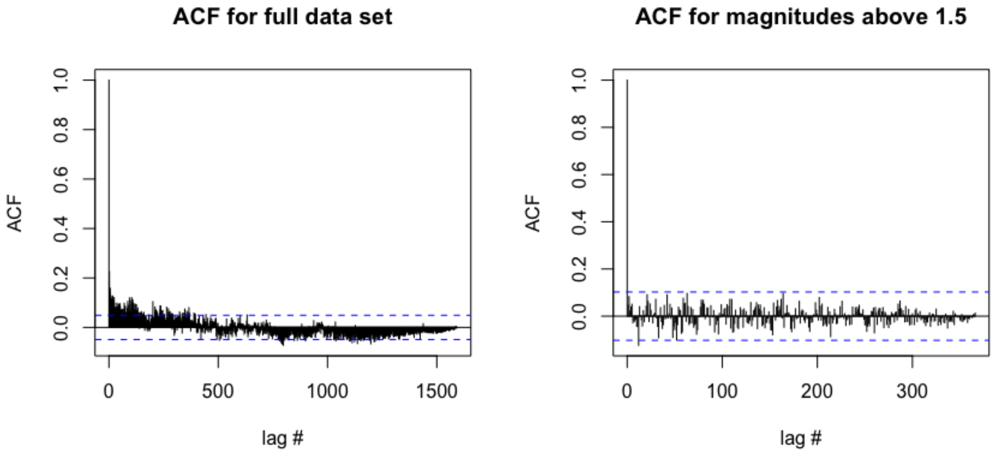


(a) Yearly means and variance of magnitudes

(b) Same as (a) for magnitudes above 1.5

Figure 3: Figure (a) shows the mean and standard errors for all of the data. Figure (b) shows the same for all of the data above a magnitude of 1.5. Both show a small downward sloping relationship, however figure (b) remains somewhat linear.

Applying an Augmented Dickey–Fuller (ADF) t-statistic test for unit root in `R` using the package `tseries` will tell if the data has remained stationary. If the series has a trend line, it will result in a large p-value. The test is done on the data set containing all observations and the data set containing only the observations with magnitude higher than 1.5. Both tests result in significant p-values well below 0.01. Therefore, it is safe to conclude that our data is indeed stationary and thus the underlying distribution has not changed over time. Note that the choice of setting the lower bound for the magnitudes at 1.5 is because of the many lower magnitudes occurring in the last 15 years.



(a) ACF with full data.

(b) ACF with select data.

Figure 4: Figure (a) shows the ACF for all of the data. Figure (b) shows the ACF for the data above a magnitude of 1.5. Both include a 95% significance region. Figure (a) shows some exceedances outside the confidence interval. Figure (b) shows that the spikes are just white noise.

By looking at the autocorrelation functions (ACF) of each signal, we can check for correlation. For a stationary signal, because we expect no dependence with time, we would expect the ACF to go to 0 for each time lag. This is done for all of the observations and the data set containing only the observations with magnitude higher than 1.5. Figure 4a shows small exceedances for the first part of the data. Note that the data set does not seem to contain all of the lower earthquake magnitudes for this model, therefore these exceedances are expected. From figure 4b it is clear that there are no autocorrelations for the magnitudes above 1.5 and the deviations are just due to white noise.

6 Parameter estimation of the GPD

The parameters of the Generalized Pareto distribution can be estimated in multiple ways. The methods used in this research are the maximum likelihood method and the L-moments method. Historically the MLE method performs very well and is the preferred choice in many works [4][5] [7]. For comparison the L-moments method will be applied as well. Its asymptotic standard errors are in comparison with the maximum likelihood estimators reasonably efficient [11]. In this chapter both of the methods will be applied to the Groningen earthquake data and the results will be displayed.

The process of estimating the maximum possible earthquakes in Groningen requires to find a distribution that fits the available data well. In this part the peaks over threshold method will be applied to the data from the KNMI containing 1606 earthquake observations. This method follows the Generalized Pareto Distribution and has two parameters to be estimated see. These parameters are the shape and scale parameters, which will be denoted respectively as ξ and σ. The main difficulty that is encountered in the process of finding these parameters is that a sufficiently large threshold has to be set for the data. As discussed earlier, this threshold is a trade-off between bias and variance.

One way to find a suitable threshold is by calculating the mean excesses for each threshold, also known as the mean residual life spans. It follows that if the exceedances after a certain threshold are linear to some extent, then the data can be considered to be of GPD type. The exceedances are calculated by

$$ME = \frac{1}{n_u}\sum_{i=1}^{n_u}(x(i) - u).$$

Where the $x(i)$ are the observations that exceed a threshold u for $1 \leq i \leq n_u$. Figure 5 is generated using the KNMI data set and the `R` package `evd`. The graph shows the mean excesses for each threshold, together with a 95% confidence interval. The data is of GPD type if the exceedances become somewhat linear after a certain threshold. Examining figure 5, this would suggest to take the threshold around $T = 2.7$, defining T as the threshold. However, if the threshold of $T = 2.7$ is taken, we are left with a mere 31 observations.

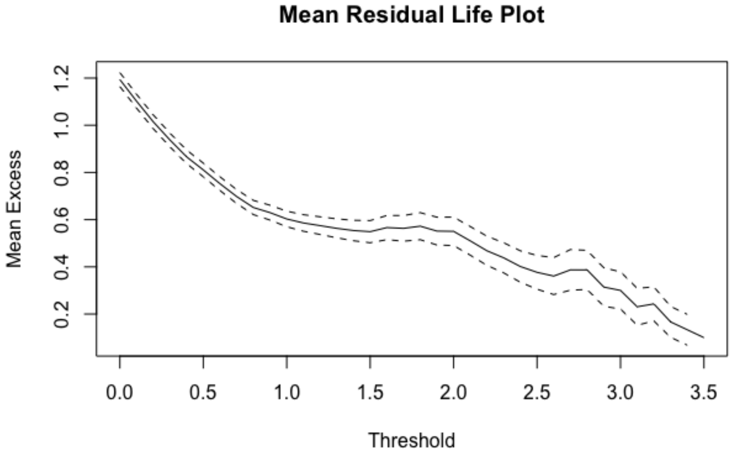


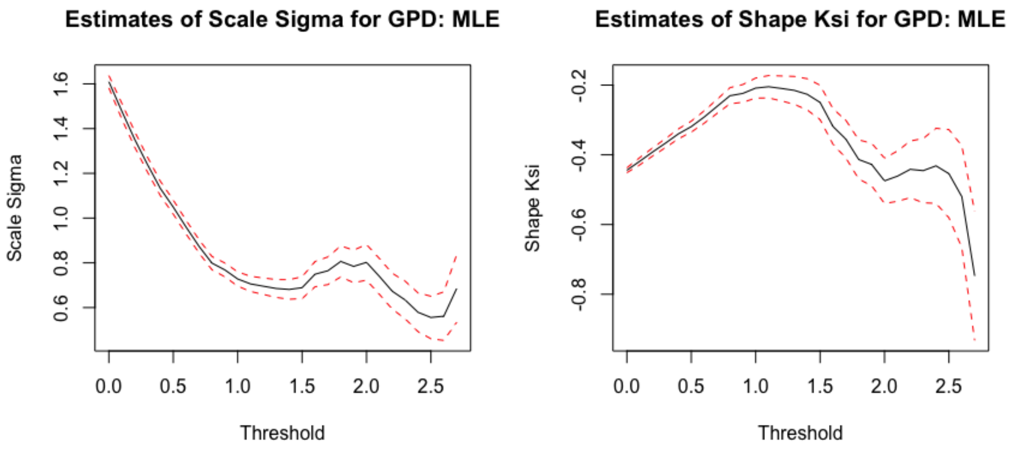
Figure 5: The Mean Excesses of the thresholds. Showing an approximately linear relationship after $u = 1.8$. Indicating that a proper threshold level should be close to $u = 1.8$.

This would result in a high bias in the generated results. Taking the confidence interval into account, there is evidence for linearity above $T = 1.5$. Accordingly, it is better to take the threshold somewhere between $T = 1.5$ and $T = 2.0$.

6.1 Parameter Estimation of the GPD using the MLE

Another way to find a suiting threshold is by fitting the GPD at different threshold levels. Using the `R` package `evd`, the parameters are estimated at different thresholds. This yields a range of shape and scale parameters for each threshold. The estimates of these parameters are shown in figure 6 plotted against the different threshold levels.

Figure 6 shows that the error estimates increase as the thresholds increase. If we find that the excesses of a threshold u follow a GPD, then it is should be true that the exceedances over higher threshold models follow a GPD as well. Moreover, the GPD parameters of these higher threshold models should be similar to the estimated parameters of the GPD with the original threshold u. Verifying if this is true for the parameters, it is observed from



(a) Scale $\hat{\sigma}$ at different threshold levels

(b) Shape $\hat{\xi}$ at different threshold levels

Figure 6: The estimates of the parameters using MLE including error estimates at different threshold levels. Figure b shows an approximately constant relationship between the thresholds $u = 1.8$ and $u = 2.5$.

figure 6b that the estimates of the shape parameter are somewhat constant between $u = 1.8$ and $u = 2.5$, if the 95% confidence intervals are taken into account. It is important to note that after the threshold of $u = 2.5$ hardly any observations remain, resulting in a significantly higher variance for the estimates. Therefore, estimates for these thresholds have been left out of the analysis.

In order to make assertions on the scale parameter it has to be noted that the parameter changes with the threshold u during the estimation. Therefore, it follows that a reparameterization of the scale parameter σ is required in the following way $\tilde{\sigma} = \sigma - \xi u$ [5]. The confidence intervals for $\tilde{\sigma}$ are then calculated by the Delta Method:

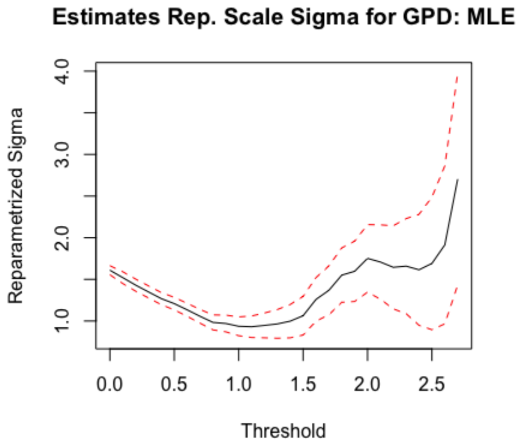
$$\mathrm{Var}(\tilde{\sigma}) \approx \nabla \tilde{\sigma}^T V \nabla \tilde{\sigma},$$

where

$$\nabla \tilde{\sigma}^T = \left[\frac{\partial \tilde{\sigma}}{\partial \sigma}, \frac{\partial \tilde{\sigma}}{\partial \xi} \right] = [1, -u].$$

Note that V is the variance-covariance matrix for the shape and scale parameters, calculated by the `evd` package in `R`.

Applying these transformations yields figure 7. Inspecting figure 7 on the same interval as the estimated shape parameter $\hat{\xi}$, we can conclude that the reparameterized scale parameter $\tilde{\sigma}$ remains approximately constant between $u = 1.8$ and $u = 2.5$. This is especially true if the 95% confidence intervals are taken into account.



(a) scale $\tilde{\sigma}$

Figure 7: Reparametrized scale $\tilde{\sigma}$ with error estimates. An approximately constant relationship is seen between the thresholds $u = 1.8$ and $u = 2.5$.

Therefore, taking the lowest of these thresholds would yield the most reliable results as this would minimize the variance and maximize the observations. That is, in the next part of the research the threshold will be set at the level $u = 1.8$. In order to light some contrast on this threshold the results for the thresholds between $u = 1.4$ and $u = 2.2$ will also be treated. This is in contrast with previous researches done on this topic, where always a threshold of $u = 1.5$ is assumed [9][4]. One reason might be that at the time of the previous researches the data set was smaller and not enough observations remained to make proper inferences. However, at the threshold of $u = 1.8$ there remain 241 observations as of June 2019.

6.2 Parameter Estimation of the GPD using L-moments

Another way to estimate the parameters of the GPD is using the L-moments method as described by Hosking [11].

$$\begin{aligned}
\lambda_1 &= \mu + \frac{\sigma}{1-\xi}, \\
\lambda_2 &= \frac{\sigma}{(1-\xi)(2-\xi)}, \\
\tau_3 &= \frac{\lambda_3}{\lambda_2} = \frac{1+\xi}{3-\xi}.
\end{aligned}$$

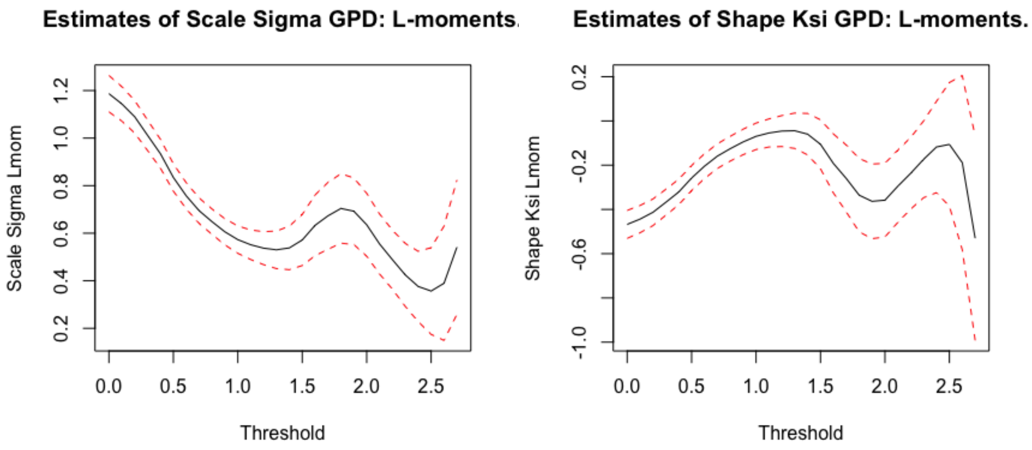
These equations have analytical solutions given by:

$$\begin{aligned}
\xi &= \frac{3\tau_3 - 1}{\tau_3 + 1}, \\
\sigma &= \lambda_2(1-\xi)(2-\xi), \\
\mu &= \lambda_1 - \frac{\sigma}{1-\xi}.
\end{aligned}$$

Using the `R` package `Lmoments` we obtain solutions for the first set of equations, evaluated at each threshold between $u = 0$ and $u = 2.7$. This includes the variance co-variance matrix for the first three moments. Using the relations in the second set of equations, estimates for the parameters evaluated at each threshold are obtained. Moreover, using the Delta method error estimates are obtained for the shape parameter ξ and scale parameter σ. This is possible due to the fact that the L-moment estimators follow an asymptotic multivariate normal distribution [11].

It is again important to find a threshold such that the parameters take on constant values for each higher threshold. Similar to the MLE method, these results have been plotted and yield figure 8.

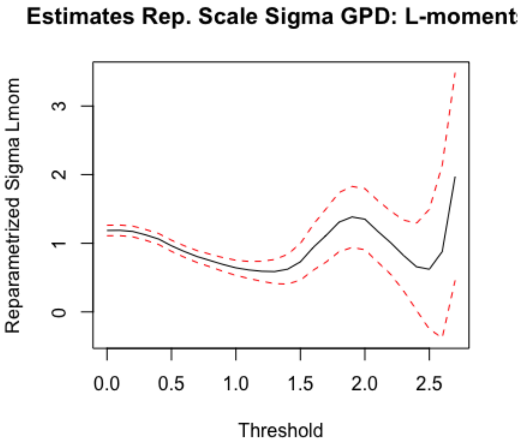
Evaluating figure 8, the estimates do not seem to get stable after any threshold. However, within the interval of thresholds $u = 1.7$ and $u = 2.3$ the same parameter estimates are possible if the error estimates are taken into account. In order to check the behaviour of the estimated scale parameters, the same reparameterization as was used previously on the scale parameter using the MLE is required.



(a) Scale $\hat{\sigma}$ at different thresholds.

(b) Shape $\hat{\xi}$ at different thresholds.

Figure 8: The estimates of the parameters using L-moments including error estimates at different threshold levels. From figure (b) a constant relationship is possible between the thresholds $u = 1.7$ and $u = 2.3$.



(a) Reparametrized scale $\tilde{\sigma}$.

Figure 9: Reparametrized scale $\tilde{\sigma}$ with error estimates obtained through L-moments. Indicating a possible constant relationship between $u = 1.7$ and $u = 2.3$.

The reparameterized scale values are plotted against the different thresholds in figure 9. Evaluating the figure, it is possible for the value of $\tilde{\sigma}$ to remain constant between the threshold values of 1.8 and 2.5. Therefore, this method complies with what was found using the MLE method. Thus, this validates the choice of setting the threshold at $u = 1.8$ even further.

7 Parameter estimation of the GEVD

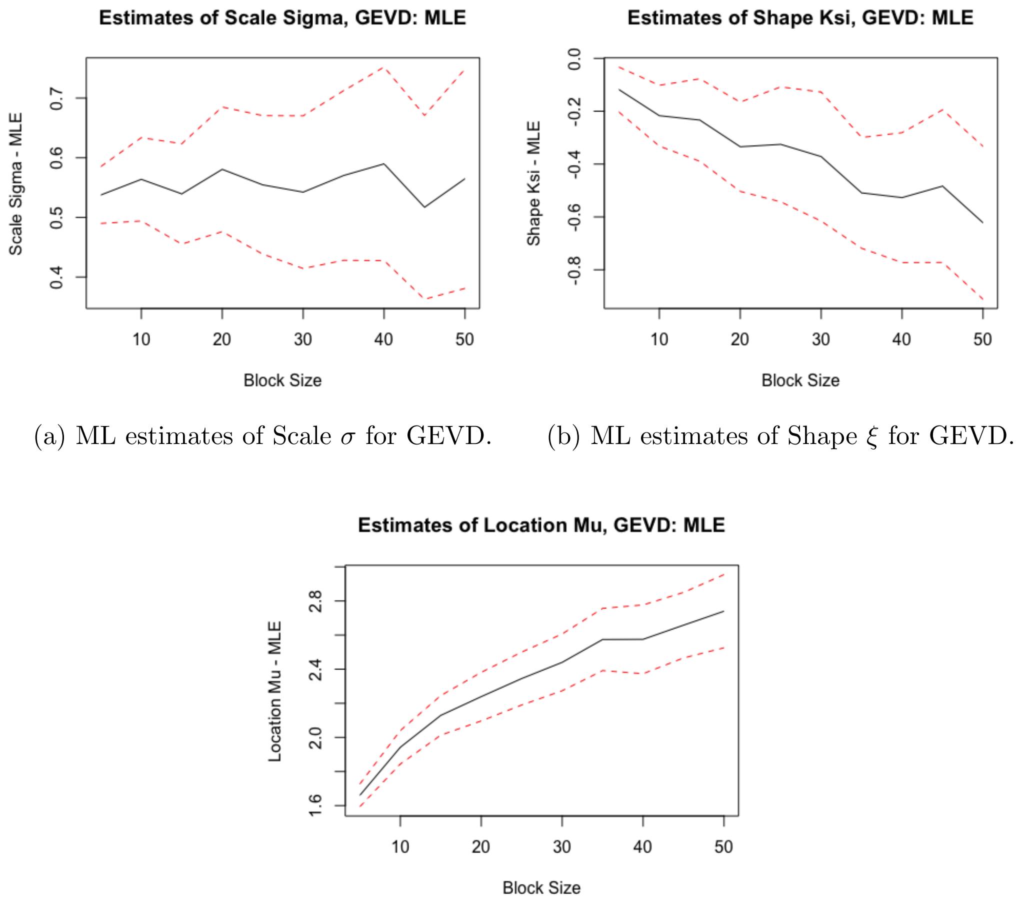
An other extreme value method is by using the method of Block Maxima instead of the Peaks over Threshold. This method follows the Generalized Extreme Value Distribution. The distribution takes on three parameters. In order to estimate these three parameters the observations are split in to blocks of even length. Each block then delivers the maximum value from within the block. Subsequently, the parameters of the GEVD are estimated using the subset consisting of all of the maxima generated by each individual block.

7.1 Parameter Estimation of the GEVD using MLE

In this section the MLE method is used in order to estimate the values of the parameters of the Generalized Extreme Value distribution. To do so, the dataset is subsetted in to blocks of the same size. This is done for blocks of size 5 up to 50 by increments of 5. Taking blocks of bigger size would leave us with too small of a data set to get any significant results.

Since the data is generated over a span of over 30 years it would make sense to subset the data in to smaller blocks dependent on time. However, as most of the data is from recent years this would cause biased results. This is seen from the fact that more than 75% of the earthquake data is collected during the last 15 years. Therefore, the blocks have been subsetted without regard to the time interval. Since the blocks differ in size with the smallest consisting out of 5 data points and the largest consisting out of 50 data points a total of 10 estimates are obtained for each of the three parameters. Using the `R` package `evd` the parameters of the GEVD are estimated using a Maximum Likelihood method. Moreover, the standard errors for these estimates are obtained through the variance covariance matrix of the parameters. The results obtained from this analysis are plotted against the different block sizes in figure 10.

Examining figure 10 it is clear that the shape parameter ξ is estimated to be somewhere between -0.5 and 0. Therefore, it follows that the maximum likelihood estimators are regular and they have the usual asymptotic properties [5]. Using the asymptotic normality of these parameters, the standard



(a) ML estimates of Scale σ for GEVD.

(b) ML estimates of Shape ξ for GEVD.

(c) ML estimates of Location μ for GEVD.

Figure 10: Parameter estimates and std. errors for different block sizes of the GEVD. Estimated using Maximum Likelihood estimation.

errors are calculated. This is done with a function from the `evd` package in `R`. Moreover, focusing on the scale parameter σ its estimates remain significantly constant throughout the calculations using different block sizes. In contrast to the estimates of the shape ξ and location μ parameters, which seem to decrease and increase respectively.

7.2 Parameter Estimation of the GEVD using L-moments

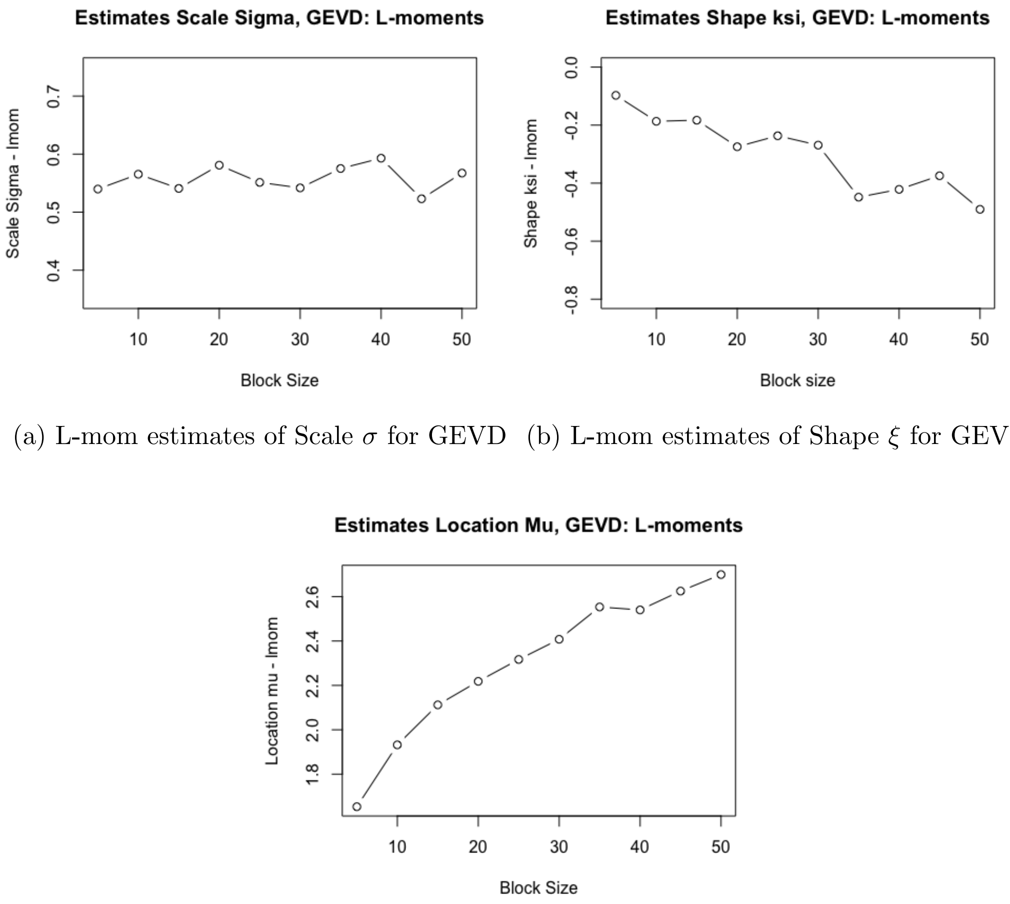
In this section the same methodology of the previous section is used to obtain different blocks of data. However, the L-moments methodology is used in the parameter estimation process. The moments relate to the parameters of the GEVD through the following equations:

$$\lambda_1 = \mu - \frac{\sigma}{\xi}(1 - \Gamma(1-\xi)), \tag{11}$$

$$\lambda_2 = \frac{-\sigma}{\xi}(1 - 2^{\xi})\Gamma(1-\xi), \tag{12}$$

$$\tau_3 = \frac{\lambda_3}{\lambda_2} = 2\frac{1-3^{\xi}}{1-2^{\xi}} - 3. \tag{13}$$

In order to solve these equations, first (13) has to be solved for the shape parameter ξ. This is done by using a Newton method in `R` from the package `rootSolve`. Afterwards, the scale and location parameters σ and μ respectively, are solved by substitution of the values. Using the relations between the parameters, estimates for the parameters for each block size are obtained. These results are plotted in figure 11. It is noteworthy how constant the estimates of the scale parameter σ remain for each block size. Moreover, it is noticeable how similar the parameter estimates are in comparison to the estimates using the MLE in the previous section. Which further strengthens the validity of the estimated parameters.



(a) L-mom estimates of Scale σ for GEVD (b) L-mom estimates of Shape ξ for GEVD

(c) L-mom estimates of Location μ for GEVD

Figure 11: Parameter estimates for different block sizes of the GEVD. Estimated through the method of L-moments. Figure (a) shows how constant the scale parameter remains using different blocksizes.

8 Results

Using the parameters obtained from the different estimation procedures the return levels have been calculated. The return levels are the expected maximum earthquake magnitudes for different numbers of earthquakes. Table 1 found below summarizes the results from the sections of this chapter. Every procedure and how its values have been determined will be discussed in this chapter, except for the GEVD model with blocksize 20. The results of the GEVD for blocksize 20 have been added as a tool for comparison. These estimates had a significantly higher variance compared to those of blocksize 35. However, their parameter estimates had a much lower variance compared to those estimated with a blocksize of 35. Note that for the estimates of the GEVD using the L-moments method, the error estimates are missing.

Furthermore, we conclude that the GPD estimates work better than those generated by the GEVD. Comparing our data, consisting of 1606 observations, the estimates generated by the GPD have a much better fit. This is seen through the fact that our data set has 21 observations with magnitude 3.0 or higher, and only three observations with magnitude 3.5 or higher. These would be modelled correctly only for the GPD models. While the GEVD estimates tend to overestimate the maximum expected earthquake magnitude for any verifiable time interval.

Expected Earthquake Magnitudes and 95%-cf upper bound for each method				
Return level:	100	1000	10000	Lim $\to \infty$
GPD: MLE	3.06 (3.20)	3.48 (3.62)	3.65 (3.83)	3.75 (4.01)
GPD: L-mom	2.99 (3.16)	3.48 (3.74)	3.70 (4.12)	3.90 (4.59)
GEVD (35): MLE	3.59 (3.73)	3.66 (3.86)	3.68 (3.88)	3.69 (3.89)
GEVD (35): L-mom	3.67	3.78	3.82	3.84
GEVD (20): MLE	3.60 (4.01)	3.80 (4.21)	3.90 (4.30)	4.04 (4.57)
GEVD (20): L-mom	3.74	4.02	4.17	4.33

Table 1: Table with estimates for GPD and GEVD methods. For the GPD the threshold is set at $u = 1.8$. For the GEVD method the blocksize is between parenthesis.

8.1 Return levels and Estimates: GPD

In the previous sections of this research different estimators for the parameters of the GPD and GEVD have been obtained. However, these parameters do not tell us much about the probabilities of the extreme events itself. Therefore, it is more convenient to calculate the probabilities of certain earthquake magnitudes in order to get a better understanding on what can be expected. For the GPD it follows by the literature [5] that for an event X,

$$Pr\{X > x\} = \zeta_u \Big[1 + \xi\Big(\frac{x - u}{\sigma}\Big)\Big]^{-1/\xi}, \tag{14}$$

where $\zeta_u = Pr\{X > u\}$ and u a set threshold. Rewriting (14) assigns a probability to how often an earthquake with magnitude bigger than x_m is exceeded on average once every m observations, given by

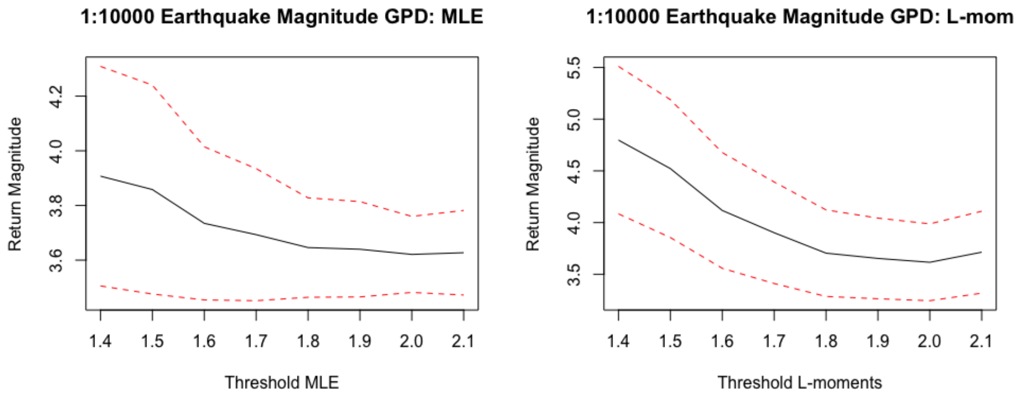
$$\frac{1}{m} = \zeta_u \Big[1 + \xi\Big(\frac{x_m - u}{\sigma}\Big)\Big]^{-1/\xi}. \tag{15}$$

Moreover, rewriting (15) gives the possibility to calculate the maximum expected magnitude x_m given a certain number of earthquakes. Since it is known how many earthquakes occur on average it gives the possibility to calculate what magnitude could be expected in a certain time interval. For any $x_m > u$,

$$x_m = u + \frac{\sigma}{\xi}\big[(m\zeta_u)^{\xi} - 1\big]. \tag{16}$$

In figure 12 the expected maximum magnitudes of 1:10000 events have been plotted using different thresholds. The choice of 1:10000 is based on the data set provided. The data set shows approximately between 80 and 120 earthquakes annually for the last 10 years. Therefore, 10000 earthquakes is approximately equal to an event with an expected probability of occurring of 10% in 10-years, or a probability of 1% in 1-year. For the estimation of x_m from (16) the obtained parameter estimates for σ and ξ need to be used. In figure 12a the estimators of the parameters obtained through MLE have been used. In figure 12b the estimators of the parameters obtained through L-moment method have been used. The estimated probability of exceedance denoted by ζ is simply the proportion of exceedances of a threshold to the total number of observations, that is

$$\zeta_u = \frac{k}{n}, \tag{17}$$



(a) MLE estimates at different thresholds. (b) L-mom estimates at different thresholds.

Figure 12: Figure (a) shows the expected magnitude of a 1:10000 earthquake using ML estimators with different thresholds for a fitted GPD model. Figure (b) shows the same using L-moment estimators.

where k the number of exceedances over a threshold u and n is the total number of available observations.

Both figures 12a and 12b have a 95%-confidence interval. These are obtained through the calculation of the variance of the return parameter assuming approximate normality. The variance is calculated by applying the Delta method to the variance covariance matrix of the parameters used in (16). The variance covariance matrix of the shape and scale parameters are obtained earlier during the calculation of the estimates from figures 6 and 8. Using the fact that ζ is binomial distributed and independent of the other variables [5], the variance co-variance matrix is given by:

$$V = \begin{bmatrix} \hat{\zeta}_u(1-\hat{\zeta}_u)/n & 0 & 0 \\ 0 & \mathrm{var}(\hat{\sigma}) & \mathrm{cov}(\hat{\sigma},\hat{\xi}) \\ 0 & \mathrm{cov}(\hat{\xi},\hat{\sigma}) & \mathrm{var}(\hat{\xi}) \end{bmatrix}. \tag{18}$$

Therefore, by the Delta Method, the variance of $\hat{x}_m$ is given by

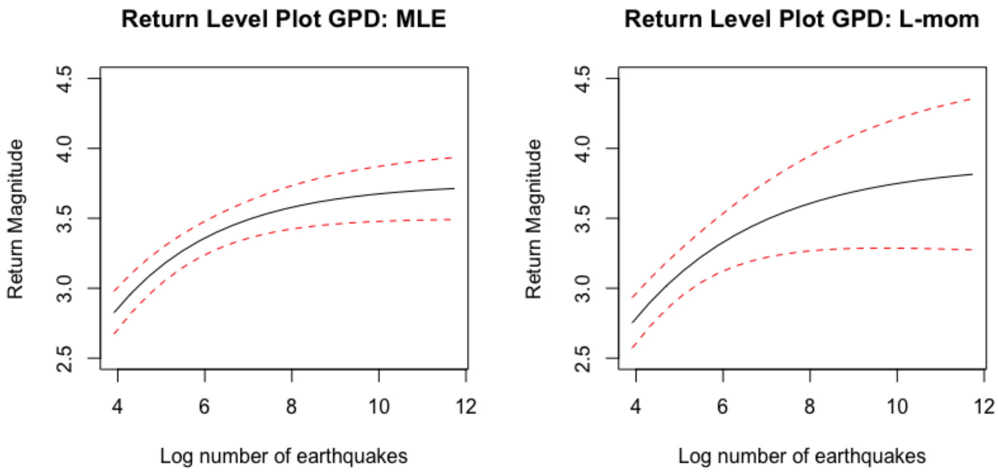
$$\mathrm{var}(\hat{x}_m) \approx \nabla x_m^T V \nabla x_m \tag{19}$$

where,

$$\nabla x_m^T = \left[\frac{\partial x_m}{\partial \zeta_u}, \frac{\partial x_m}{\partial \sigma}, \frac{\partial x_m}{\partial \xi}\right], \tag{20}$$

which is evaluated at $(\zeta_u, \sigma, \xi) = (\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})$ for x_m as in (16). From figure 12a it is clear that the expected maximum magnitude has a value of approximately $x_m = 3.5$ with varying 95%-confidence intervals. Note that at the threshold of $u = 1.8$ the errors seems to stabilize. This is in accordance with the earlier determined optimum threshold of $u = 1.8$ in section 5. For figure 12b the expected maximum magnitude changes significantly with every threshold. Setting the threshold at $u = 1.8$ and analyzing the return levels at this specific level for both the MLE and the L-moments method yields figure 13.

Figure 13 shows the expected maximum earthquake magnitudes for occurrences of earthquakes between 50 and 125.000. The highest value on the x-axis could be interpreted as the maximum expected earthquake magnitude with an approximate 1%-probability of occurring within 10 years, using the same reasoning as before. The number of earthquakes is depicted on the x-axis on a logarithmic scale. The main difference between the two methods of estimation lies in how they behave for long return levels. The estimates using the MLE method remain somewhat stable in figure 13a, even for high return levels. While the estimates using the L-moments method in figure 13b have a higher error. Having an error of more than 1.0 on the scale of Richter for high return levels, resulting in much uncertainty. We therefore conclude that the MLE method is the preferred choice in estimating the parameters of the GPD, since estimates with more certainty are preferred.



(a) ML estimates of magnitudes.

(b) L-mom estimates of magnitudes.

Figure 13: Both figures show the expected maximum magnitudes for a different number of earthquakes. The x-axis follows a logarithmic scale. It is clear that the estimates obtained through the L-moments method result in higher variance and uncertainty.

8.2 Return levels and Estimates: GEVD

For the GEVD it follows that,

$$x_m = \mu - \frac{\sigma}{\xi}\Big[1 - \{-\log(1-p)\}^{-\xi}\Big]. \tag{21}$$

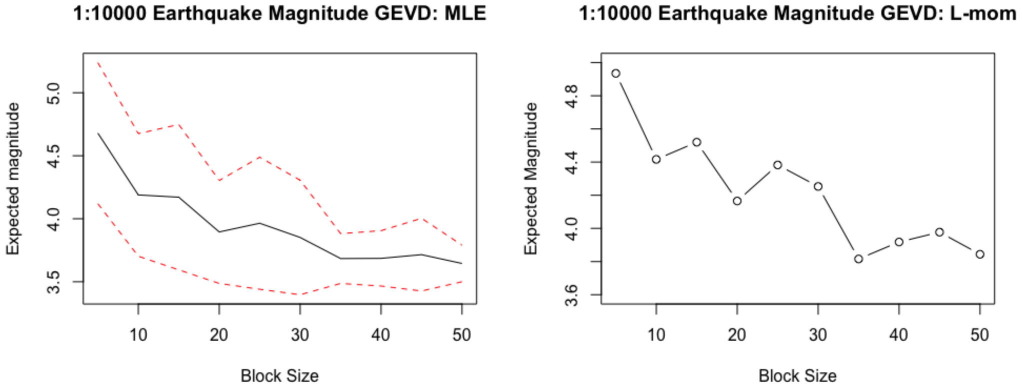
For $\xi \neq 0$ and p the probability of occurrence is equal to $1/m$ for a certain number of earthquakes m. For more details see Coles pg. 49 [5]. In order to obtain the estimated $\hat{x}_m$ the estimates for the location, shape and scale parameters are plugged in to (21). In figure 14 the estimates of x_m are shown for all the different block-sizes and their respective parameter estimations. From these plots we conclude that the blocks of size 35 have the preferred choice, due to its low error estimates. The errors of $\hat{x}_m$ are obtained through the Delta method,

$$\text{Var}(\hat{x}_m) \approx \nabla x_m^T V \nabla x_m \tag{22}$$

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

$$\nabla x_m^T = \left[\frac{\partial x_m}{\partial \mu}, \frac{\partial x_m}{\partial \sigma}, \frac{\partial x_m}{\partial \xi} \right] \tag{23}$$

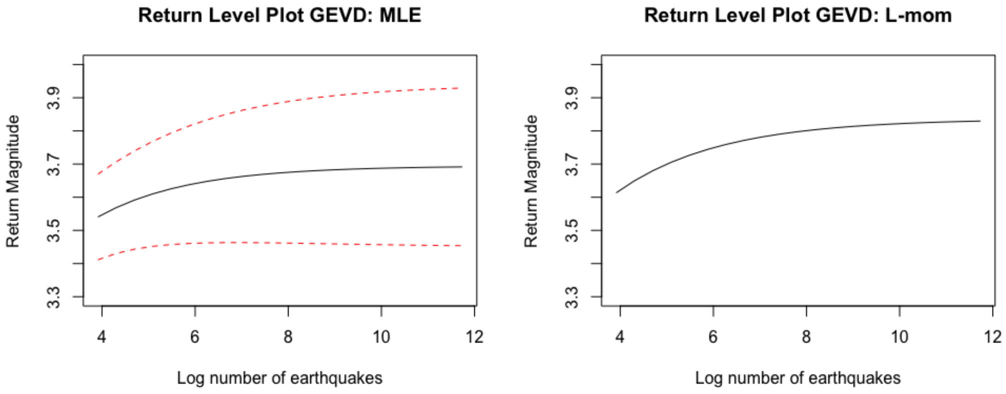
which is evaluated at $(\mu, \sigma, \xi) = (\hat{\mu}, \hat{\sigma}, \hat{\xi})$ for x_m as in (21).



(a) MLE estimates at different thresholds (b) GEVD estimates at different thresholds

Figure 14: Figure (a) shows the expected magnitude of a 1:10000 earthquake using ML estimators with different thresholds for a fitted GEVD model. Figure (b) shows the same using L-moment estimators.

It is clear from figure 14a that the lowest variance occurs at a blocksize of 35. For figure 14b the errors could not be estimated. Setting the block size at 35 and computing the maximum expected earthquake magnitudes using the MLE and L-moments methods yields figure 15. The x-axis follows a logarithmic scale of the number of earthquakes. The estimates seem to stabilize early in the predictions. However, this is very logical upon inspection of (21) for x_m. Since our shape parameter estimate is negative it follows from (21) that as $p \to 0$, or similarly $m \to \infty$, $x_m = \mu - \sigma/\xi$. Hence we have a bounded distribution. The difference in the estimated x_m is due to the different estimates of the parameters obtained through the MLE and L-moments methods. Where the first one has a maximum value of $x_m = 3.7$ and the latter has a maximum value of $x_m = 3.8$, rounded of to decimals.



(a) ML estimates of magnitudes.

(b) L-mom estimates of magnitudes.

Figure 15: Both figures show the expected maximum magnitudes for a different number of earthquakes. The x-axis follows a logarithmic scale. It is clear that the expected maximum magnitude of figure (b) is higher than that of figure (a). Moreover, the error estimates of figure figure (b) are missing.

8.3 Model fit GPD: MLE

In order to assess the quality of the estimated model we will make use of the probability and quantile plots. It is expected that both will show a diagonal line if the model is any good. Moreover, the probability plot should have a line approximately on the unit diagonal. For a threshold set at u and excesses the $y_{(1)} \leq \ldots \leq y_{(k)}$ the probability plot consists of the pairs

$$\{(i/(k+1), \hat{H}(y_{(i)}));\ i = 1, \ldots, k\} \tag{24}$$

where $\hat{H}$ is the estimated model evaluated at $(\sigma, \xi) = (\hat{\sigma}, \hat{\xi})$. Given for $\xi \neq 0$ by,

$$\hat{H}(y) = 1 - \left(1 + \frac{\hat{\xi} y}{\hat{\sigma}}\right)^{-1/\hat{\xi}}. \tag{25}$$

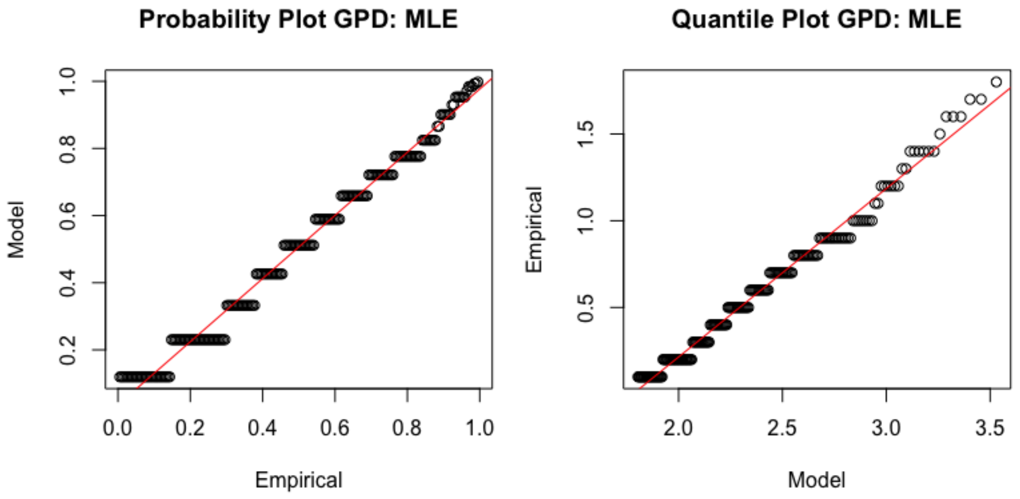
The quantile plot consists of the pairs

$$\{(H^{-1}(i/(k+1)), y_{(i)});\ i = 1, \ldots, k\} \tag{26}$$

where,

$$H^{-1}(y) = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[y^{-\hat{\xi}}\right]. \tag{27}$$

Both of these plots should be approximately diagonal if the GPD is a reasonable model to plot the excesses over the threshold u. The probability plot and quantile plot are depicted in figure 16 for the GPD model using maximum likelihood as a method of estimation. The probability plot is close to the unit diagonal as required. The quantile plot shows a good linear relationship. Both plots do not give any reason to suspect model failure.



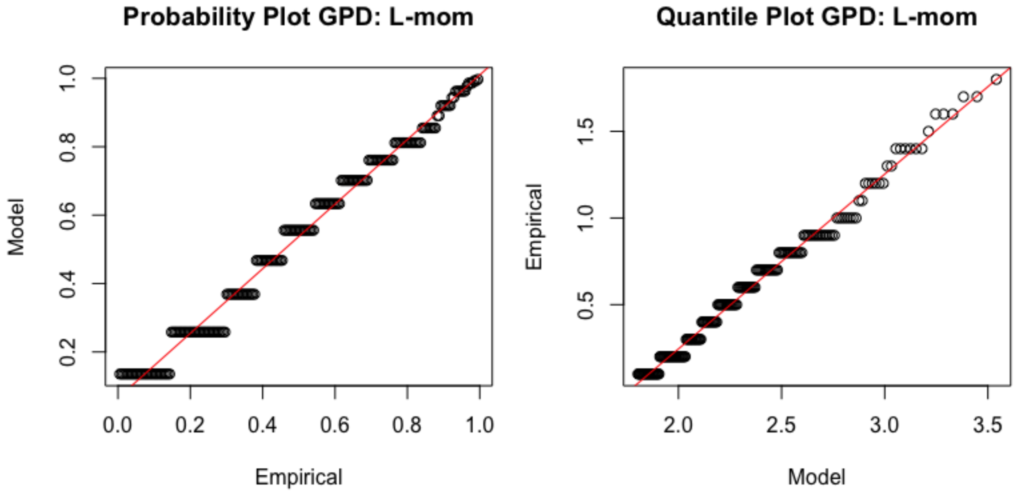
(a) Probability plot of MLE for GPD.

(b) Quantile plot of MLE for GPD.

Figure 16: Probability and quantile plot of the GPD using MLE as an estimation method. It is clear that both plots are diagonal and figure (a) is on the unit diagonal indicating no signs of model failure.

8.4 Model fit GPD: L-moments

The probability plot and quantile plots are constructed in the same way using L-moments estimation as it is using MLE. The plots are shown in figure 17. Both of them have the similar desirable form as the plots obtained using MLE in the previous section. Using the same reasoning as before there are therefore no indications of model failure.



(a) Probability plot of L-moments for GPD. (b) Quantile plot of L-moments for GPD.

Figure 17: Probability and quantile plot of the GPD using L-moments as an estimation method. It is clear that both plots are diagonal and figure (a) is on the unit diagonal. Indicating no signs of model failure.

8.5 Model fit GEVD: MLE

In order to check the model fit of the GEVD model using MLE and L-moments the probability and quantile plots have to be assessed [5]. The probability plot compares the empirical distribution to the fitted distribution function. The empirical distribution function evaluated at $y_{(i)}$ is given by

$$\hat{F}(y_{(i)} = i/(m+1), \tag{28}$$

where $y_{(1)} \leq \ldots \leq y_{(m)}$ are the ordered block maxima.
The model estimates are given by substituting the obtained parameters in the GEVD. These are given by

$$\hat{G}(y_{(i)}) = \exp\left\{ -\left[1 + \hat{\xi}\left(\frac{y_{(i)} - \hat{\mu}}{\hat{\sigma}}\right)\right]^{-1/\hat{\xi}}\right\}. \tag{29}$$

If the estimated model is working well, it should follow that

$$\hat{F}(y_{(i)}) \approx \hat{G}(y_{(i)}), \tag{30}$$

for each $1 \leq i \leq m$.

Hence, the probability plot consists of the points

$$\left\{\left(\hat{F}(y_{(i)}), \hat{G}(y_{(i)})\right),\ i = 1, \ldots, m\right\}. \tag{31}$$

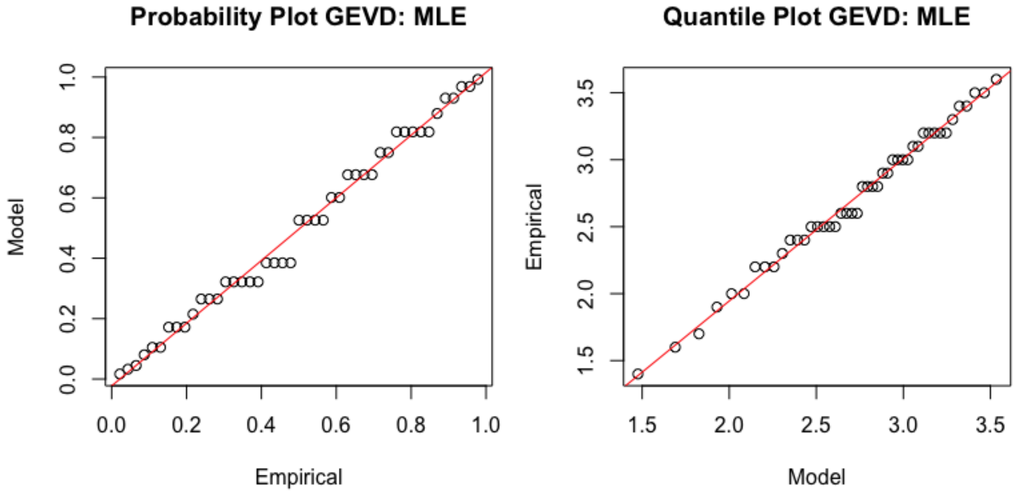
The quantile plot assesses the accuracy of large values of $y_{(i)}$ in a better way than the probability plot. Since both $\hat{F}(y_{(i)})$ and $\hat{G}(y_{(i)})$ approach 1 for higher values of $y_{(i)}$. The quantile plot is given by the points,

$$\left\{\left(\hat{G}^{-1}(i/(m+1), y_{(i)}\right),\ i = 1, \ldots, m\right\} \tag{32}$$

where we have,

$$\hat{G}^{-1}\left(\frac{i}{m+1}\right) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - \left\{-\log\left(\frac{i}{m+1}\right)\right\}^{-\xi}\right]. \tag{33}$$

The two plots should be approximately diagonal, with the points of the probability plot located approximately on the unit diagonal. The resulting plots are shown in figure 18. Both plots seem to fit well and give no indications of model failure.



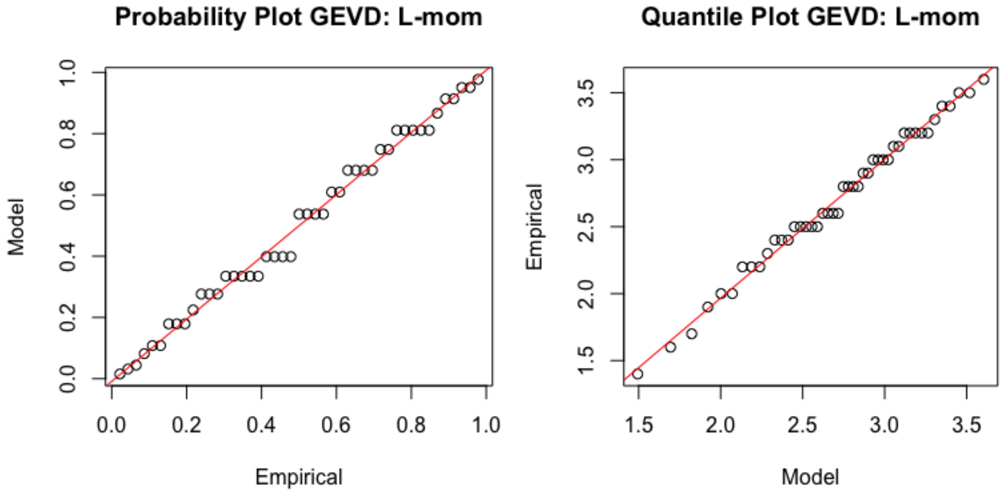
(a) Probability plot of MLE for GEVD.

(b) Quantile plot of MLE for GEVD.

Figure 18: Probability and quantile plot of the GEVD using MLE as an estimation method. It is clear that both plots are diagonal and figure (a) is on the unit diagonal, with no significant deviations. Indicating no signs of model failure.

8.6 Model fit GEVD: L-moments

The probability plot and quantile plots are constructed in the same way as in the previous section, using the parameters obtained through L-moments. The plots are shown in figure 19. Both of them have the similar desirable form as the plots obtained using MLE. Using the same reasoning as before, we conclude that there are therefore no indications of model failure.



(a) Probability plot L-moments for GEVD. (b) Quantile plot L-moments for GEVD.

Figure 19: Probability and quantile plot of the GEVD using L-moments as an estimation method. It is clear that both plots are diagonal and figure (a) is on the unit diagonal. Figure (b) seems to show some deviations but nothing worrying. There are no indications of model failure.

9 Discussion and Conclusion

In this chapter the results obtained in the previous chapter will be compared to results obtained in other works. Next to the results, the differences in research methodology will be discussed. Furthermore, we will discuss how our results should be interpreted and which assumptions have been made.

9.1 Comparison with older works

Comparing the results of this research with the research done by Beirlant et al. [4] there are some differences. First of all, it is important to note that in their research a modified form of the GPD was used. Another important difference in the setup of research is that a threshold level of $u = 1.5$ was used, while we found a threshold level of 1.8 to be more appropriate. The choice of setting the threshold at $u = 1.5$ is justified by the fact that an older research [9] used the same threshold. However, this choice should have been reevaluated as the number of observations has grown significantly for the last 10 years. Our results indicated that at a threshold of $u = 1.5$, the estimations have to high of a bias and do not reflect the behaviour of the tail well.

Another difference is that our shape parameter ξ never has the value of zero, this is even true for the lower bounds of the 95%-confidence intervals. In the research done by Beirlant et al. [4] they find that the shape parameter ξ is equal to zero. Therefore they need to use a truncated GPD, as for $\xi = 0$ the distribution would be unbounded and unreasonable high estimates would be obtained. Since our research does not show any signs of the shape parameter being zero (seen from the 95%-confidence intervals), the approach to truncate the GPD did not turn out to be useful. On the contrary, we find that the distributions are bounded and the endpoints are readily obtained.

However, the obtained results using the GPD with ML estimation ended up being very similar to the results obtained using the truncated EVT based techniques in the aforementioned research [4]. As is seen in table 1 in the results section, the 95%-confidence intervals are very similar if compared to table 2 in Beirlant et al. on page 18 [4]. Furthermore, the results are also in line with another research done by the NAM [9] in 2013.

The results are only based on data from the gas extraction in Groningen. In order to gain more insight into the underlying distributions, data collected from other induced earthquakes can be examined. These could include earthquakes induced by water extraction, oil extraction or other gas extraction sites. Comparing these data and the distributional features could give more insight in to how well our estimation procedure works.

9.2 Conclusions

In our research we computed and compared the expected magnitudes of the earthquakes in Groningen induced by gas extraction. To this end the accumulated data by the KNMI was analyzed, consisting out of 1606 observations collected between december 1986 and may 2019. Our analysis compared the predictions of the expected earthquakes for different return levels estimated by the GEVD and GPD distributions. Both the GEVD and GPD parameters are estimated using Maximum Likelihood estimation and L-moments estimation. The four estimates and their confidence intervals are compared and explained. Based on these estimates table 1 is constructed containing all of the estimates for different return levels. The last estimates are too be interpreted as the maximum possible expected earthquake magnitudes. These maximum expected magnitudes lay in the range of 3.7-4.3 and the upper 95%-confidence intervals have values between 3.9-4.6. The lower endpoints are estimated by the GEVD model with blocksize 35, having high errors. This model tends to overestimate the magnitudes for small return levels and underestimate for high return levels. The results generated by the GEVD should therefore be interpreted carefully. Moreover, as the block sizes have been set at 35 the calculations are based on a mere 45 observations. The obtained estimates are therefore not reliable.

Furthermore, the error estimates of the GPD using MLE as a method of parameter estimation resulted in smaller variances compared to the L-moments method. Therefore, we conclude that the preferred model in our research, which yields the most reliable results is the GPD model using maximum likelihood estimation for the parameters, setting the threshold at $u = 1.8$. This results in a maximum expected earthquake magnitude of 3.75 with a maximum value of 4.01 by the upper 95%-confidence interval.

These estimates are purely based on the structure of the data. Therefore, the biggest assumption made is that all other variables remain constant in our model. However, it should not be forgotten that nature is very unforgiving and surprising. Therefore, if the underlying distribution changes significantly, it would result in new unpredictable events. This could be possible due to the change in the deterministic variables of which the earthquake magnitudes are dependent.

References

[1] International disaster data - our world in data. `https://ourworldindata.org/ofdacred-international-disaster-data`.

[2] Knmi - aardbevingscatalogus. `https://www.knmi.nl/kennis-en-datacentrum/dataset/aardbevingscatalogus`.

[3] Jan Beirlant, Isabel Fraga Alves, and Ivette Gomes. Tail fitting for truncated and non-truncated pareto-type distributions. *Extremes*, 19(3):429–462, 2016.

[4] Jan Beirlant, Isabel Fraga Alves, Tom Reynkens, et al. Fitting tails affected by truncation. *Electronic Journal of Statistics*, 11(1):2026–2065, 2017.

[5] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values.* Springer, 1 edition, 2001.

[6] Stuart Coles and Edward Casson. Extreme value modelling of hurricane wind speeds. *Structural Safety*, 20(3):283–296, 1998.

[7] Laurens De Haan. Fighting the arch–enemy with mathematics ‘. *Statistica neerlandica*, 44(2):45–68, 1990.

[8] Ferreira A. De Haan L. *Extreme Value Theory: An Introduction.* Springer, 1st edition edition, 2006.

[9] Bernard Dost, Mauro Caccavale, Torild van Eck, and Dirk Kraaijpoel. Report on the expected pgv and pga values for induced earthquakes in the groningen area. *KNMI report. Royal Netherlands Meteorological Institute (De Bilt)*, 2013.

[10] Nicolas Duran, J Paul Elhorst, et al. A spatio-temporal-similarity and common factor approach of individual housing prices. Technical report, University of Groningen, Research Institute SOM, 2017.

[11] Jonathan Richard Morley Hosking. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1):105–124, 1990.

[12] Andrzej Kijko and Gerhard Graham. Parametric-historic procedure for probabilistic seismic hazard analysis part i: estimation of maximum regional magnitude mmax. *Pure and Applied Geophysics*, 152(3):413–442, 1998.

[13] Mathias Raschke. Inference for the truncated exponential distribution. *Stochastic environmental research and risk assessment*, 26(1):127–138, 2012.

[14] Nassim Nicholas Taleb. Black swans and the domains of statistics. *The American Statistician*, 61(3):198–200, 2007.

[15] Haiyun Wang and Xiaxin Tao. Relationships between moment magnitude and fault parameters: theoretical and semi-empirical relationships. *Earthquake Engineering and Engineering Vibration*, 2(2):201–211, 2003.

10 Appendix

Contents

R Code Estimating GPD Parameters using MLE

```r
getwd()
setwd("/Users/aras/Desktop")
myData <- read.csv("all_induced.csv", header = T, sep = ",")

###############################################################################
############ Calculating the GPD: MLE parameters ##############################
###############################################################################
rm(list=ls())
library("evd")
library("rootSolve")
library("Lmoments")
myData <- read.csv("all_induced.csv", header = T, sep = ",")

r <- myData$MAG
mle_estimates <- matrix(0,28,7)
colnames(mle_estimates) <- c("Thresh", "Shape xi", "Scale sig", "Shape se", "Scale se",
                             "new-sig se", "varcov sig/xi")
for(k in 0:27){

  # mle-method
  fpot_k <- fpot(r, threshold = k/10, model = c("gpd", "pp"), npp = length(r))
  mle_estimates[k+1,1] <- k/10 #threshold
  mle_estimates[k+1,2] <- fpot_k$estimate[2] #shape ksi
  mle_estimates[k+1,3] <- fpot_k$estimate[1] #scale sigma
  mle_estimates[k+1,4] <- fpot_k$std.err[2] # shape se
  mle_estimates[k+1,5] <- fpot_k$std.err[1] # scale se
  mle_estimates[k+1,6] <- (c(1,-k/10)%*%fpot_k$var.cov%*%c(1,-k/10))^0.5 # error new-sig
  mle_estimates[k+1,7] <- fpot_k$var.cov[2,1]
}

#standard errors for shape-ksi MLE- method
matplot(mle_estimates[,1],
        mle_estimates[,2] + outer(mle_estimates[,4], c(0,1,-1)),
        type="l", lty=c(1,2,2), col=c(1,2,2),
        xlab="Threshold", ylab="Shape Ksi",
        main = ’Estimates of Shape Ksi for GPD: MLE’
)
#standard errors for scale-sigma MLE- method
matplot( mle_estimates[,1],
         mle_estimates[,3] + outer(mle_estimates[,5], c(0,1,-1)),
         type="l", lty=c(1,2,2), col=c(1,2,2),
         xlab="Threshold", ylab="Scale Sigma",
         main = 'Estimates of Scale Sigma for GPD: MLE'
)
#reparametrized sigma + errors for mle
sigma_MLEnew <- mle_estimates[,3] - mle_estimates[,2]* mle_estimates[,1]
matplot( mle_estimates[,1],
         sigma_MLEnew + outer(mle_estimates[,6],c(0,1.96,-1.96)),
         type="l",lty=c(1,2,2), col=c(1,2,2),
         xlab="Threshold", ylab="Reparametrized Sigma",
         main = 'Estimates Rep. Scale Sigma for GPD: MLE'
)



#############################################################################
### Calculating the return levels, the level x_m that is exceeded on avg ####
### once every m observations for GPD- MlE:##################################
#############################################################################
x_m <- numeric(0)
var_x <- numeric(0)
covar <- matrix(0,3,3)
m <- 10000
for(u in 15:22){
  sig <- mle_estimates[u,3]
  xi <- mle_estimates[u,2]
  zeta <- length(myData$MAG[myData$MAG>(u/9.9  -0.1)])/length(myData$MAG)
  x_m[u-14] <- (u-1)/10 + sig/xi * ((m*zeta)^(xi)- 1)

  var_zeta <-  zeta*(1-zeta)/length(myData$MAG[myData$MAG>(u/9.9  -0.1)])
  covar[1,1] <- var_zeta
  covar[2,2] <- mle_estimates[u,5]^2
  covar[3,3] <- mle_estimates[u,4]^2
  covar[3,2] <- mle_estimates[u,7]
  covar[2,3] <- covar[3,2]

  delta_x <- c(sig*m^(xi)*zeta^(xi-1),
               xi^(-1)*((m*zeta)^(xi)-1),
               -sig*xi^(-2)*((m*zeta)^(xi)-1) + sig*xi^(-1)*(m*zeta)^(xi)*log(m*zeta))

  var_x[u-14] <-   (delta_x%*%covar%*%(delta_x))^0.5

}

matplot(seq(1.4,2.1,0.1),
        x_m + outer(var_x, c(0,1.96,-1.96)),
        type='l', lty=c(1,2,2), col=c(1,2,2),
        xlab="Threshold MLE", ylab="Return Magnitude",
        main = "1:10000 Earthquake Magnitude GPD: MLE"
)


#############################################################################
################### Model Validation GPD: MLE ###############################
#############################################################################

#Probability plots GPD - MLE at threshold u= 1.8 so everything higher
rm(list=ls())
myData <- read.csv("all_induced.csv", header = T, sep = ",")
```

```r
u=1.8
y <- sort(myData$MAG[myData$MAG >u] -u)
xi <- -0.413570646 #for u =1.8
sig <- 0.806181891 #for u=1.8
h <- 1 - (1 + xi*y/sig)^(-1/xi)
r <- 1:length(y)/(length(y)+1)

plot(r,h, main= "Probability Plot GPD: MLE",
     xlab = "Empirical", ylab= "Model")
abline(lm(h~ r), col = "red")

#Quantile plots GPD - MLE: Using inverse h(y)
r <- 1:length(y)/(length(y)+1)
h_1 = sort((u + (sig/xi)*(r^(-xi)-1)))
plot(h_1, y, main = "Quantile Plot GPD: MLE", xlab = "Model", ylab= "Empirical")
abline(lm( y~h_1), col = "red")

#Return level plots (m,x_m), x_m= estimated m-obs return level:
m <- numeric(0)
m[1:21] <- ceiling(50^(seq(1,3,0.1)))
zeta <- length(myData$MAG[myData$MAG>1.8])/length(myData$MAG)
x_m <- u + sig/xi*((m*zeta)^xi -1 )

covar <- matrix(0,3,3)
covar[1,1] <- zeta2*(1-zeta2)/length(myData$MAG)
covar[2,2] <- 0.004809707
covar[3,3] <- 0.003150140
covar[3,2] <- -0.003541867
covar[2,3] <- covar[3,2]

delta_x <- numeric(0)
err_mx <- numeric(0)
for(i in 1:21){
delta_x <- c(sig*m[i]^(xi)*zeta^(xi-1),
             xi^(-1)*((m[i]*zeta)^(xi)-1),
             -sig*xi^(-2)*((m[i]*zeta)^(xi)-1) +
               sig*xi^(-1)*(m[i]*zeta)^(xi)*log(m[i]*zeta))

err_mx[i] <-   (delta_x%*%covar%*%(delta_x))^0.5
}

matplot(log(m), x_m + outer(err_mx, c(0,1.96,-1.96)),
        type='l', lty=c(1,2,2), col=c(1,2,2),
        xlab="Log number of earthquakes", ylab="Return Magnitude",
        main= "Return Level Plot GPD: MLE", ylim = c(2.5,4.5)
)
```

R-code 1: R-code GPD-MLE

R Code Estimating GPD Parameters using L-moments

```r
getwd()
setwd("/Users/aras/Desktop")
myData <- read.csv("all_induced.csv", header = T, sep = ",")

rm(list=ls())
library("evd")
library("rootSolve")
library("Lmoments")
myData <- read.csv("all_induced.csv", header = T, sep = ",")

r <- myData$MAG
lmom_estimates <- matrix(0,28,8)
colnames(lmom_estimates) <- c("Thresh", "Shape xi", "Scale sig", "loc mu", "Shape se",
                              "scale se", "new-sig se", "varcov sig/xi")
var.cov_lmom <- matrix(0,2,2)
cov_ksi_sig <- numeric(0)
for(k in 0:27){
  # L-moments method
  Lcoefs <- Lcoefs(data = r[r>k/10], rmax = 3, na.rm = FALSE, trim = c(0, 0))
  shape_ksi <- (Lcoefs[1,3]*3 - 1)/(Lcoefs[1,3] +1)
  scale_s <- Lcoefs[1,2]*(1- shape_ksi)*(2-shape_ksi)
  loc_m <- Lcoefs[1,1] - scale_s/(1- shape_ksi)

  lmom_estimates[k+1,1] <- k/10
  lmom_estimates[k+1,2] <- shape_ksi
  lmom_estimates[k+1,3] <- scale_s
  lmom_estimates[k+1,4] <- loc_m

  #calculating standard error in shape ksi- Lmom
  vcov <- Lmomcov(r[r>k/10],3)

  myfunc <- function(lst){
    (3*lst[3] -lst[2])/(lst[3] +lst[2])
  }
  grad <- gradient(myfunc, c(Lcoefs[1,1],Lcoefs[1,2],Lcoefs[1,3]*Lcoefs[1,2]))
  lmom_estimates[k+1,5]<- (grad%*%vcov%*%t(grad))^0.5

  #calculating standard error in scale sigma- Lmom
  myfunc2 <- function(lst){
    lst[2]* (1-(3*lst[3] -lst[2])/(lst[3] +lst[2]))*(2-(3*lst[3] -lst[2])/(lst[3] +lst[2]))
  }
  grad2 <- gradient(myfunc2, c(Lcoefs[1,1],Lcoefs[1,2],Lcoefs[1,3]*Lcoefs[1,2]))
  lmom_estimates[k+1,6]<- (grad2%*%vcov%*%t(grad2))^0.5

  #covariance reparametrized scale sigma (building the var.cov matrx)
  cov_ksi_sig[k+1] <- (grad%*%vcov%*%t(grad2))
  var.cov_lmom[1,1] <- lmom_estimates[k+1,6]^2
  var.cov_lmom[2,2] <- lmom_estimates[k+1,5]^2
  var.cov_lmom[1,2] <- cov_ksi_sig[k+1]
  var.cov_lmom[2,1] <- var.cov_lmom[1,2]

  lmom_estimates[k+1,7] <- (c(1,-k/10)%*%var.cov_lmom%*%c(1,-k/10))^0.5 # error new-sig
  lmom_estimates[k+1,8] <- cov_ksi_sig[k+1]
```

```

}

#standard errors for shape-ksi lmom- method
matplot(lmom_estimates[,1],
        lmom_estimates[,2] + outer(lmom_estimates[,5], c(0,1.96,-1.96)),
        type="l", lty=c(1,2,2), col=c(1,2,2),
        xlab="Threshold", ylab="Shape Ksi Lmom",
        main = 'Estimates of Shape Ksi GPD: L-moments.'
)

#standard errors for scale-sigma lmom- method
matplot(lmom_estimates[,1],
        lmom_estimates[,3] + outer(lmom_estimates[,6], c(0,1.96,-1.96)),
        type="l", lty=c(1,2,2), col=c(1,2,2),
        xlab="Threshold", ylab="Scale Sigma Lmom",
        main = 'Estimates of Scale Sigma GPD: L-moments.'
)

#reparametrized sigma + errors for lmom
sigma_LMOMnew <- lmom_estimates[,3] - lmom_estimates[,2]* lmom_estimates[,1]
matplot( lmom_estimates[,1],
         sigma_LMOMnew + outer(lmom_estimates[,7],c(0,1.96,-1.96)),
         type="l",lty=c(1,2,2), col=c(1,2,2),
         xlab="Threshold", ylab="Reparametrized Sigma Lmom",
         main = 'Estimates Rep. Scale Sigma GPD: L-moments.'
)

# Calculating the return levels, the level x_m that is exceeded on avg
# once every m observations for GPD- L-moms:
y_m <- numeric(0)
var_y <- numeric(0)
covar2 <- matrix(0,3,3)
m <- 10000
for(u in 15:22){
  sigma <- lmom_estimates[u,3]
  xi <- lmom_estimates[u,2]
  zeta2 <- length(myData$MAG[myData$MAG>(u/9.99-0.1)])/length(myData$MAG)
  y_m[u-14] <- (u-1)/10 + sigma/xi * ((m*zeta2)^(xi)- 1)

  var_zeta2 <-  zeta2*(1-zeta2)/length(myData$MAG[myData$MAG>(u/9.99 -0.1)])
  covar2[1,1] <- var_zeta2
  covar2[2,2] <- lmom_estimates[u,6]^2
  covar2[3,3] <- lmom_estimates[u,5]^2
  covar2[3,2] <- lmom_estimates[u,8]
  covar2[2,3] <- covar2[3,2]

  delta_y <- c(sigma*m^(xi)*zeta2^(xi-1),
               xi^(-1)*((m*zeta2)^(xi)-1),
               -sigma*xi^(-2)*((m*zeta2)^(xi)-1) +
                 sigma*xi^(-1)*(m*zeta2)^(xi)*log(m*zeta2))

  var_y[u-14] <-   (delta_y%*%covar2%*%(delta_y))^0.5

}
matplot(seq(1.4,2.1,0.1),
        y_m + outer(var_y, c(0,1.96,-1.96)),
        type='l', lty=c(1,2,2), col=c(1,2,2),
        xlab="Threshold L-moments", ylab="Return Magnitude",
```

```r
        main = "1:10000 Earthquake Magnitude GPD: L-mom"

)

#Probability plots GPD - Lmom at threshold u= 1.8 so everything higher
rm(list=ls())
myData <- read.csv("all_induced.csv", header = T, sep = ",")
u=1.8
y <- sort(myData$MAG[myData$MAG >u] -u)
xi <- -0.335629022 #for u =1.8
sig <- 0.704043262 #for u=1.8
h <- 1 - (1 + xi*y/sig)^(-1/xi)
r <- 1:length(y)/(length(y)+1)

plot(r,h, main= "Probability Plot GPD: L-mom",
     xlab = "Empirical", ylab= "Model")
abline(lm(h~ r), col = "red")


#Quantile plots GPD - MLE: Using inverse h(y)
r <- 1:length(y)/(length(y)+1)
h_1 = sort(u + (sig/xi)*(r^(-xi)-1))
plot(h_1, y, main = "Quantile Plot GPD: L-mom", xlab = "Model", ylab= "Empirical")
abline(lm(y~h_1), col = "red")


#Return level plots (m,x_m), x_m= estimated m-obs return level:
m <- numeric(0)
m[1:21] <- ceiling(50^(seq(1,3,0.1)))
zeta2 <- length(myData$MAG[myData$MAG>1.8])/length(myData$MAG)

y_m <- u + sig/xi*((m*zeta2)^xi -1 )
covar <- matrix(0,3,3)
covar[1,1] <- zeta2*(1-zeta2)/length(myData$MAG)
covar[2,2] <- 0.005547118
covar[3,3] <- 0.007152318
covar[3,2] <- -0.005477988
covar[2,3] <- covar[3,2]

delta_y <- numeric(0)
err_my <- numeric(0)
for(i in 1:21){
  delta_y <- c(sig*m[i]^(xi)*zeta2^(xi-1),
               xi^(-1)*((m[i]*zeta2)^(xi)-1),
               -sig*xi^(-2)*((m[i]*zeta2)^(xi)-1) +
                 sig*xi^(-1)*(m[i]*zeta2)^(xi)*log(m[i]*zeta2))

  err_my[i] <-   (delta_y%*%covar%*%(delta_y))^0.5
}

matplot(log(m), y_m + outer(err_my, c(0,1.96,-1.96)),
        type='l', lty=c(1,2,2), col=c(1,2,2),
        xlab="Log number of earthquakes", ylab="Return Magnitude",
        main= "Return Level Plot GPD: L-mom", ylim = c(2.5,4.5)
)
```

R-code 2: R-code GPD-Lmoments

R Code Estimating GEVD Parameters using MLE

```r

getwd()
setwd("/Users/aras/Desktop")
myData <- read.csv("all_induced.csv", header = T, sep = ",")

###################################################
###  check all block sizes for MLE and plot    ###
###################################################
rm(list=ls())
myData <- read.csv("all_induced.csv", header = T, sep = ",")
mag <- myData$MAG

mle_est <- matrix(0,10,8)
colnames(mle_est) <- c('bl size', 'location', 'error loc',
                       ' shape', 'err shape', 'scale', 'err scale', 'return err')
for(i in seq(5,50,5)){
  x <- vector('numeric')
  for(n in 0:floor(length(mag)/i)){
    k <- i*n
    s <- i*(n+1)
    x[n+1] <- max(mag[k:s])
  }

  gevd_mle <- fgev(x, nsloc = NULL, prob = NULL, std.err = TRUE,
                   corr = FALSE, method = "BFGS", warn.inf = TRUE)
  mle_est[i/5,1] <- i
  mle_est[i/5,2] <- round(gevd_mle$estimate[1],3) #location
  mle_est[i/5,3] <- gevd_mle$std.err[1] #std error location
  mle_est[i/5,4] <- gevd_mle$estimate[3] #shape xi
  mle_est[i/5,5] <- gevd_mle$std.err[3] #std error x
  mle_est[i/5,6] <- gevd_mle$estimate[2] #scale sigma
  mle_est[i/5,7] <- gevd_mle$std.err[2] #std error sigma

  #calculating the error of the return level z which is 1/p; coles page 56
  varcov <- gevd_mle$var.cov
  p <- 0.001 #m=1000
  y <- -1*(log(1-p))
  delta_z <- c(1, -(mle_est[i/5,4]^(-1)*(1-y^(-mle_est[i/5,4]))),
               mle_est[i/5,6]*mle_est[i/5,4]^(-2)*(1-y^(-mle_est[i/5,4]))
               -mle_est[i/5,6]*mle_est[i/5,4]^(-1)*y^(-mle_est[i/5,4])*log(y))
  mle_est[i/5,8] <- (delta_z%*%varcov%*%(delta_z))^0.5

}


# 95% confidence interval plots for location
matplot(mle_est[,1],
        mle_est[,2] + outer(mle_est[,3], c(0,1.96,-1.96)),
        type='l', lty=c(1,2,2), col=c(1,2,2),
        xlab="Block Size", ylab="Location Mu - MLE",
        main = 'Estimates of Location Mu, GEVD: MLE'
)

# 95% confidence interval plots for shape
matplot(mle_est[,1],
        mle_est[,4] + outer(mle_est[,5], c(0,1.96,-1.96)),
        type="l", lty=c(1,2,2), col=c(1,2,2),
        xlab="Block Size", ylab="Shape Ksi - MLE",
        main = 'Estimates of Shape Ksi, GEVD: MLE'
)
# 95% confidence interval plots for scale
matplot(mle_est[,1],
        mle_est[,6] + outer(mle_est[,7], c(0,1.96,-1.96)),
        type="l", lty=c(1,2,2), col=c(1,2,2),
        xlab="Block Size", ylab="Scale Sigma - MLE",
        main = 'Estimates of Scale Sigma, GEVD: MLE'
)
################################################################
################################################################
# Calculating the retun levels, the level x_m that is exceeded on avg
# once every m observations for GEVD- MlE:

r_m <- numeric(0)
for(u in 1:10){
  sigma <- mle_est[u,6]
  xi <- mle_est[u,4]
  loc <- mle_est[u,2]
  m <- 10000
  y <- -1*log(1-1/m)
  r_m[u] <- loc - sigma/xi * (1 - y^(-xi) )

}
matplot(mle_est[,1],
        r_m + outer(mle_est[,8], c(0,1.96,-1.96)),
        type='l', lty=c(1,2,2), col=c(1,2,2),
        xlab="Block Size", ylab="Expected magnitude",
        main = "1:10000 Earthquake Magnitude GEVD: MLE"
)

################################################################
################################################################
#Probability plots GEVD - MLE at Block size = 35
rm(list=ls())
yData <- read.csv("all_induced.csv", header = T, sep = ",")
data_x <-  c(1.4, 1.6, 1.7, 1.9, 2.0, 2.0, 2.2, 2.2, 2.2, 2.3, 2.4, 2.4, 2.4, 2.5,
        2.5, 2.5, 2.5, 2.5, 2.6, 2.6, 2.6, 2.6, 2.8, 2.8, 2.8, 2.8, 2.9, 2.9,
        3.0, 3.0, 3.0, 3.0, 3.1, 3.1, 3.2, 3.2, 3.2, 3.2, 3.2, 3.3, 3.4, 3.4,
        3.5, 3.5, 3.6)
xi <- -0.50899645
sig <- 0.57022056
loc <- 2.57400000

z <- (1:length(data_x))/(length(data_x)+1)
h <- exp(- (1 + xi*((data_x-loc)/sig))^(-1/xi))

plot(z,h, main= "Probability Plot GEVD: MLE",
     xlab = "Empirical", ylab= "Model")
abline(lm(h~ z), col = "red")

#Quantile plot
g <- loc - (sig/xi)*(1 - (-log(z))^(-xi))
plot(g, data_x, main = "Quantile Plot GEVD: MLE", xlab = "Model", ylab= "Empirical")
abline(lm(data_x~g), col = "red")

```

```r
#Return level plots (m,x_m), x_m= estimated m-obs return level:
m <- numeric(0)
m[1:21] <- ceiling(50^(seq(1,3,0.1)))
p <- 1/m #m=1000
y <- -1*(log(1-p))
x_m <- loc - (sig/xi)*(1- (y^(-xi)))

varcov <- matrix(c(0.008658555,-0.001619208,-0.003958419,
                    -0.001619208, 0.005261144,-0.005217842,
                    -0.003958419, -0.005217842,0.011447961),
                  nrow=3, ncol=3)

delta_z <- numeric(0)
err_rm <- numeric(0)
for(i in 1:21 ){
delta_z <- c(1, -((xi^(-1))*(1-y[i]^(-xi))),
             (sig*xi^(-2))*(1-y[i]^(-xi))-(sig*xi^(-1))*y[i]^(-xi)*log(y[i]))

err_rm[i] <- (delta_z%*%varcov%*%(delta_z))^0.5
}

matplot(log(m), x_m + outer(err_rm, c(0,1.96,-1.96)),
        type='l', lty=c(1,2,2), col=c(1,2,2),
        xlab="Log number of earthquakes", ylab="Return Magnitude",
        main= "Return Level Plot GEVD: MLE", ylim = c(3.3,4)
)
```

R-code 3: R-code GEVD-MLE

R Code Estimating GEVD Parameters using L-moments

```

getwd()
setwd("/Users/aras/Desktop")
myData <- read.csv("all_induced.csv", header = T, sep = ",")

################################################################
###  check all block sizes for L-moments and plot            ###
################################################################
# L-moments estimation of GEVD parameters
rm(list=ls())
myData <- read.csv("all_induced.csv", header = T, sep = ",")
mag <- myData$MAG

lmom_est <- matrix(0,10,7)
colnames(lmom_est) <- c(’bl size’, ’location’, ’error loc’,
                        ’ shape’, ’err shape’, ’scale’, ’err scale’)
for(i in seq(5,50,5)){
  x <- vector(’numeric’)
  for(n in 0:floor(length(mag)/i)){
    k <- i*n
    s <- i*(n+1)
    x[n+1] <- max(mag[k:s])
    x <- x[!is.na(x)]
  }
  Lcoefs <- Lcoefs(data = x, rmax = 3, na.rm = F, trim = c(0, 0))
  func_ksi <- function(b){
    2*(1-3^b)/(1-2^b) - 3 - Lcoefs[1,3]
  }
  shape_ksi <- uniroot(func_ksi,c(-4,4))$root

  scale_s <- -1*(Lcoefs[1,2]*shape_ksi)/(gamma(1-shape_ksi)*(1-2^(shape_ksi)))
  loc_m <- Lcoefs[1,1] + scale_s/shape_ksi*(1- gamma(1-shape_ksi))

  lmom_est[i/5,1] <- i
  lmom_est[i/5,2] <- loc_m
  lmom_est[i/5,4] <- shape_ksi
  lmom_est[i/5,6] <- scale_s

}

#######################################################################
#plotting the parameters of GEVD estimated by L-moments including 95% cfi
#######################################################################
plot(lmom_est[,1], lmom_est[,4], type = ’b’, ylim = c(-0.8,0),
     ylab = ’Shape ksi - lmom’, xlab = ’Block size’,
     main = ’Estimates Shape ksi, GEVD: L-moments’)

plot(lmom_est[,1], lmom_est[,6], type = ’b’,ylim = c(0.35,0.75),
     ylab = ’Scale Sigma - lmom’, xlab = ’Block Size’,
     main = ’Estimates Scale Sigma, GEVD: L-moments’)

plot(lmom_est[,1], lmom_est[,2], type = ’b’,
     ylab = ’Location mu - lmom’, xlab = ’Block Size’,
     main = ’Estimates Location Mu, GEVD: L-moments’)

#######################################################################
```

```
######################################################################
# Calculating the retun levels, the level x_m that is exceeded on avg
# once every m observations for GEVD- L-moms:
s_m <- numeric(0)
for(u in 1:10){
  sigma <- lmom_est[u,6]
  xi <- lmom_est[u,4]
  loc <- lmom_est[u,2]
  m <- 10000
  y <- -1*log(1-1/m)

  s_m[u] <- loc - sigma/xi * (1 - y^(-xi) )

}

plot(x= lmom_est[,1], y= s_m, type = "b",  xlab="Block Size",
     ylab="Expected Magnitude", ylim = c(3.6,5),
     main = "1:10000 Earthquake Magnitude GEVD: L-mom")



######################################################################
######################################################################

rm(list=ls())
yData <- read.csv("all_induced.csv", header = T, sep = ",")
data_x <-  c(1.4, 1.6, 1.7, 1.9, 2.0, 2.0, 2.2, 2.2, 2.2, 2.3, 2.4, 2.4, 2.4, 2.5,
             2.5, 2.5, 2.5, 2.5, 2.6, 2.6, 2.6, 2.6, 2.8, 2.8, 2.8, 2.8, 2.9, 2.9,
             3.0, 3.0, 3.0, 3.0, 3.1, 3.1, 3.2, 3.2, 3.2, 3.2, 3.2, 3.3, 3.4, 3.4,
             3.5, 3.5, 3.6)
xi <- -0.4483400 #for u =1.8
sig <- 0.5752321 #for u=1.8
loc <- 2.5533234

z <- (1:length(data_x))/(length(data_x)+1)
h <- exp(- (1 + xi*((data_x-loc)/sig))^(-1/xi))

plot(z,h, main= "Probability Plot GEVD: L-mom",
     xlab = "Empirical", ylab= "Model")
abline(lm(h~ z), col = "red")

#however prob plot gives least info in region of most interest, that is
#at the endpoints. Since both estimates are to approach 1 as y increases
#therefore we will use the quantile plots.

#Quantile plot
g <- loc - (sig/xi)*(1 - (-log(z))^(-xi))
plot(g, data_x, main = "Quantile Plot GEVD: L-mom", xlab = "Model", ylab= "Empirical")
abline(lm(data_x~g), col = "red")

#Return level plots (m,x_m), x_m= estimated m-obs return level:
m <- numeric(0)
m[1:21] <- ceiling(50^(seq(1,3,0.1)))
p <- 1/m #m=1000
y <- -1*(log(1-p))
x_m <- loc - (sig/xi)*(1- (y^(-xi)))

varcov <- matrix(c(0.008658555,-0.001619208,-0.003958419,
                   -0.001619208, 0.005261144,-0.005217842,
                   -0.003958419, -0.005217842,0.011447961),
                 nrow=3, ncol=3)

delta_z <- numeric(0)
err_rm <- numeric(0)
for(i in 1:21 ){
  delta_z <- c(1, -((xi^(-1))*(1-y[i]^(-xi))),
               (sig*xi^(-2))*(1-y[i]^(-xi))-(sig*xi^(-1))*y[i]^(-xi)*log(y[i]))

  err_rm[i] <- (delta_z%*%varcov%*%(delta_z))^0.5
}

plot(x = log(m), y = x_m, type="l",
        xlab="Log number of earthquakes", ylab="Return Magnitude",
        main= "Return Level Plot GEVD: L-mom", ylim = c(3.3,4)
)
```

R-code 4: R-code GEVD-Lmoments

R Code Data Exploration chapter 5

```

rm(lis=ls())
myData <- read.csv("all_induced.csv", header = T, sep = ",")
View(myData)
#install.packages(
  c("ggfortify", "changepoint",
    "strucchange", "ggpmisc")
#)

#\install.packages("lubridate")
library("lubridate")

#changing the dates so R can interpret it as dates
myData$YYMMDD <- ymd(myData$YYMMDD)
View(myData)

newdata <- myData[which(myData$MAG >0), ]
plot(newdata$YYMMDD,newdata$MAG)

r <- table(newdata$MAG)
plot(r,type = 'h', ylab = 'Magnitude Frequency', xlab = 'Magnitude', main = 'Histogram of Earthquake data')
lines(density(newdata$MAG), col = 'red')

#plotting the means and variance for each year
library(ggplot2)
p<- ggplot(newdata, aes(YYMMDD, MAG)) + geom_line() + xlab("Date") + ylab("Magnitude")
p+ ggtitle('Plot of every magnitude against time of occurence')

library('changepoint')
library("ggfortify")
library("fractal")
Box.test(newdata$MAG[newdata$MAG>1.5], lag=25, type="Ljung-Box") # test stationary signal
#Another test we can conduct is the Augmented Dickey Fuller (ADF) t-statistic test to find
#if the series has a unit root (a series with a trend line will have a unit root and result
# in a large p-value).
adf.test(newdata$MAG[newdata$MAG>1.5])
adf.test(newdata$MAG)

t <- newdata$YYMMDD#[newdata$MAG>1.5]
y_stationary <- newdata$MAG#[newdata$MAG>1.5]
plot(t,y_stationary,
     type='l',col='red',
     xlab = "time (t)",
     ylab = "Y(t)",
     main = "Stationary signal")
acf(y_stationary,lag.max = length(y_stationary),
    xlab = "lag #", ylab = 'ACF',main='ACF for magnitudes above 1.5')
```

R-code 5: R-code Data Exploration