TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Eindhoven University of Technology

BACHELOR

Earthquake distribution
an elaboration on the tail

Prikken, Levi H.A.

Award date:
2017

Link to publication

Disclaimer
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science

Earthquake Distribution

An elaboration on the tail

L.H.A. Prikken

Supervisors:
prof.dr. R.W. van der Hofstad
dr. W.J.T. Hulshof

version 1.0

Eindhoven, Monday 22nd May, 2017

Abstract

In this thesis the main purpose is to analyze data of earthquakes to find out what the distribution of the seismic moment of those earthquakes would be. Some suggest that a single distribution will not be suitable and that there might be some power law cut-off at the end. Also the distribution of the maximum of an earthquake's strength is a topic of interest. A theoretical analysis and a data analysis were executed to analyze several distribution for both cases. In conclusion, there was no proper distribution found for the strength of an earthquake, and neither was the case for the maximum of an earthquake's strength. There were some reasonable fits, but there was too much deviation to really get a good estimation. Further research is advised, since the main purpose of this thesis was exploration. Other distributions, with the knowledge of this thesis, are suggested.

Contents

Chapter 1

Introduction

Nature is a beautiful thing, but can also be a source of a lot of trouble for many people. An earthquake is one of those sources and it can happen in many parts of the world, but not all earthquakes are natural. An earthquake can also be induced, meaning caused by humans. In this thesis the focus will lie on the induced earthquakes happening in Groningen, the Netherlands. These earthquakes happen due to gas extraction from 3 kilometers below ground. According to the KNMI, there have been more than 1380 earthquakes in the past thirty years. Although the strength of the earthquakes in Groningen is much lower than for instance in Asia, the Groningen earthquakes have given over more than 78.000 insurance claims throughout time. Therefore, an analysis from a mathematical perspective of how these earthquakes behave and what the maximum strength of such an earthquake could be is of great importance.

With all the recent developments on the area of data analysis, there is more data available about earthquakes. Using this data, a distribution of the earthquakes can be estimated and thus can future earthquakes be better understood.

In this thesis, there will be an examination of what distributions would fit earthquakes in Groningen best and what that distribution would mean for a better understanding of earthquakes in general.

1.1 Outline

To give an idea of the structure of this thesis, the outline will be presented. At first, the main problem tackled will be stated with the corresponding subquestions in Chapter 2: Problem Definition. To give an overview of the theory, data and software used, Chapter 3: Preliminaries will provide this information. When the basic principles have been explained, a literature study is performed in Chapter 4: Literature study, to give an idea what research already showed, but also to give a better estimation of what could be useful to theoretically analyze in the next chapter, namely Chapter 5: Theoretical Analysis. In this chapter, some different distributions are researched and the maximum distribution will also be looked at, After this theoretical part, the more practical part is up. In Chapter 6: Data Analysis most of the theoretically analyzed distribution will now be fitted on the available data to see if there is a match and the research question can be answered. This answer will be given in Chapter 7: Conclusions. In Chapter 8: Discussion, there will be an overview of the limitations in this thesis and some advice for further research on this topic. To close, the used sources are presented and the different scripts written for this thesis will be given, to give a better understanding of the underlying analysis.

Chapter 2

Problem Definition

To be able to get a better understanding of earthquakes, it is useful to know the distribution of the frequency of earthquakes with a certain strength. With the distribution of the frequency of an earthquake, it is meant the relative frequency of strength M or higher, plotted against M. This strength is best described by the seismic moment of an earthquake. The seismic moment is a way to measure the size of an earthquake. It is best explained as:

$$M = \mu AD \tag{2.1}$$

with μ the ratio of shear strain of the rocks involved in Pascal, A the area of the rupture where the earthquake occurred in m^2 and D the average displacement on A in m.

Therefore, the question is what the probability of an earthquake with a certain strength is. The Gutenberg-Richter law is a commonly used method to give a estimate of that frequency, but some say that at the tail of the distribution, this law does not hold that well. (Sornette and Sornette, 1999) This gives the topic of research.

The probability on stronger earthquakes according to the Gutenberg-Richter law seems to take larger values than data may suggest. Therefore, it can be questioned if this distribution is the most accurate. The idea pitched in an article by Sornette and Sornette (1999), is that there might be some sort of cut-off present from a certain strength M_{xg}. This cut-off can have some different forms.

The main goal of this thesis is to analyze the different distributions and see which one is the best fit for the strength of an earthquake as defined in the first paragraph. Therefore, the following research question can be composed:

What is the best fitted and statistically supported distribution of the relative frequency of a certain earthquake's strength, looking at an earthquake's seismic moment, and does this distribution have a cut-off at the end of some sort?

To support the answer to this research question the following subquestions are formed:

- What is the best fitted and statistically supported distribution of the tail probabilities of an earthquake's strength (strengths greater of equal to 2), looking at an earthquake's seismic moment?
- What is the best fitted and statistically supported distribution of the maximum of a set of earthquake strengths, looking at an earthquake's seismic moment and does this have implications for the distribution of an earthquake's strength?

Chapter 3

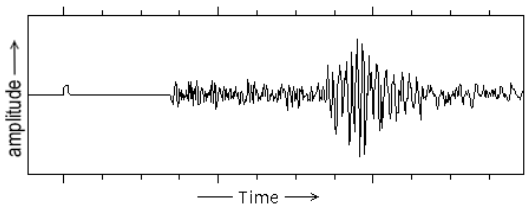
Preliminaries

When analyzing earthquakes, first the method in which the strength of an earthquake is measured should be analyzed to get an understanding of the measure. In the next paragraphs some commonly used measurement scales are addressed and elaborated on.

One of the earliest comparison methods is the Mercalli intensity scale, which was designed by Giuseppe Mercalli in 1902. This method is mainly used to measure the intensity of an earthquake. It measures the effects on the people, buildings, etcetera, that the earthquake made happen. This method is still used quite common nowadays, but is not the a good way of looking at the proposed research question. First of all, this method is very subjective to which specific class an earthquake belongs to. Second of all, a lot of different factors play a role in the earthquake's intensity. For instance, the depth of an earthquake, the soil where the earthquake took place or how densely the area is populated. All different factors that should be taken into account to know more about the real earthquake. On this scale, some strong earthquakes can be scaled beneath some weaker ones, due to the fact that the stronger one lay deeper below ground or was just not near a big city. So, the Mercalli scale does not qualify in getting a estimation of the absolute strength of an earthquake.

A more well known measurement scale is the Richter magnitude scale. The Richter scale built on the previous, more subjective Mercalli scale by offering a quantifiable measure of an earthquake's size. This newer measurement scale was made in 1935 by Charles Richter. This scale defines the magnitude of an earthquake as the logarithm of the ratio of the amplitude of the seismic waves. Seismic waves are waves of energy traveling through the earth's layers and are recorded by a seismograph. In Figure 3.1 the output is given of such a seismograph. This is called a seismogram. The definition of the x-axis is the time and of the y-axis is the amplitude of the seismic wave.

Figure 3.1: Seismogram



Unfortunately the Richter scale also has its restrictions, since it will show growing deviations

when an earthquake would be greater than 6,5 or when an earthquake is more than 500 kilometers away from a seismograph, the maximum magnitude would always be 7. (Sornette and Sornette, 1999) Therefore, a newer scale was invented. The moment magnitude scale, also called M_w which measures the size of an earthquake in terms of released energy. The moment is a product of the distance a fault moved and the force required to move it. This method is more precise and can also measure all larger earthquakes.

For the Groningen situation, these higher strengths do not occur. So both the Richter scale as well as the moment magnitude scale can be used, depending in which scale the strengths are expressed in the data. This will be further examined in Section 3.2.

3.1 Used distributions and theorems

In this thesis some different theorems and lemmas are used. To give an idea in advance of what they hold, there will be an elaboration of why they are of importance.

Besides these statements, several different distributions (and its properties) are looked at, so also in this section an overview of some of these distributions are elaborated on and some important features are stated to supply the necessary information in advance.

3.1.0.1 The Exponential distribution (Abramovich, 2013)

One of the best known power law distributions is the exponential distribution. In this thesis, the exponential distribution is used as a more basic distribution for the strength of an earthquake or functions as a cut-off for some distribution. Take X a random variable with an exponential distribution. This distribution has one parameter λ, for which holds that $\lambda > 0$. The following properties hold:

1. The cumulative distribution function (CDF):

$$F_X(x) = \mathbb{P}(X \leq x) = 1 - e^{-\lambda x} \tag{3.1}$$

2. The probability distribution function (PDF):

$$f_X(x) = \lambda \cdot e^{\lambda x} \tag{3.2}$$

3. The expected value:

$$\mathbb{E}[X] = \frac{1}{\lambda} \tag{3.3}$$

3.1.0.2 The Pareto distribution (Arnold, 2015)

The Pareto distribution is a power law probability distribution. Also a fairly common distribution for the strength of an earthquake. Take X a random variable with a Pareto distribution. This distribution has two parameters; a scale parameter x_m and a shape parameter α, for which holds that $x_m > 0$ and $\alpha > 0$. The following properties hold:

1. The cumulative distribution function (CDF):

$$F_X(x) = \mathbb{P}(X \leq x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha} \text{ with } x \geq x_m \tag{3.4}$$

2. The probability distribution function (PDF):

$$f_X(x) = \frac{\alpha \cdot x_m^\alpha}{x^{\alpha+1}} \tag{3.5}$$

3. The expected value $\mathbb{E}[X]$ can be calculated through an integral:

$$\begin{aligned}\mathbb{E}[X] &= \int_{x_m}^{\infty} x \cdot f_{X_i}(x)dx \\ &= \int_{x_m}^{\infty} x \cdot \frac{\alpha \cdot x_m^\alpha}{x^{\alpha+1}}dx\end{aligned} \tag{3.6}$$

The value of α influences the outcome, so different cases are considered. First, if $0 < \alpha < 1$. Then the next step is to rewrite the integral and calculate the primitive of the expression in the integral:

$$\begin{aligned}\mathbb{E}[X] &= \int_{x_m}^{\infty} \alpha \cdot x_m^\alpha \cdot x^{-\alpha}dx \\ &= \left[\frac{\alpha}{-\alpha+1} \cdot x_m^\alpha \cdot x^{-\alpha+1}\right]_{x_m}^{\infty}\end{aligned} \tag{3.7}$$

Since this is an improper integral, it has to be rewritten with a limit and the expected value can be calculated as the following:

$$\begin{aligned}\mathbb{E}[X] &= \lim_{R\to\infty}\left[\frac{\alpha}{-\alpha+1} \cdot x_m^\alpha \cdot x^{-\alpha+1}\right]_{x_m}^{R} \\ &= \lim_{R\to\infty} \frac{\alpha}{-\alpha+1} \cdot x_m^\alpha \cdot R^{-\alpha+1} - \frac{\alpha}{-\alpha+1} \cdot x_m^\alpha \cdot x_m^{-\alpha+1} \\ &= \lim_{R\to\infty} \frac{\alpha}{-\alpha+1} \cdot x_m^\alpha \cdot R^{-\alpha+1} - \frac{\alpha}{-\alpha+1} \cdot x_m^\alpha \cdot x_m \\ &= \infty, \text{ since } \alpha \in (0,1)\end{aligned} \tag{3.8}$$

The second case is if $\alpha = 1$. The calculations are quite similar to the previous case, so this gives:

$$\begin{aligned}\mathbb{E}[X] &= \int_{x_m}^{\infty} x_m \cdot x^{-1}dx \\ &= \lim_{R\to\infty}[x_m \cdot \ln x]_{x_m}^{R} \\ &= x_m \cdot \ln R - x_m \cdot \ln x_m \\ &= \infty, \text{ since R goes to infinity}\end{aligned} \tag{3.9}$$

180 The last case is if $\alpha > 1$. In this case,

$$\mathbb{E}[X] = \left[\frac{\alpha}{-\alpha+1} \cdot x_m^\alpha \cdot x^{-\alpha+1}\right]_{x_m}^{\infty} \tag{3.10}$$

Again we deal with an improper integral, so a limit is necessary.

$$\begin{aligned}
\mathbb{E}[X] &= \lim_{R\to\infty} \left[\frac{\alpha}{-\alpha+1} \cdot x_m^{\alpha} \cdot x^{-\alpha+1} \right]_{x_m}^{R} \\
&= \lim_{R\to\infty} \frac{\alpha}{-\alpha+1} \cdot x_m^{\alpha} \cdot R^{-\alpha+1} - \frac{\alpha}{-\alpha+1} \cdot x_m^{\alpha} \cdot x_m^{-\alpha+1} \\
&= \lim_{R\to\infty} \frac{\alpha}{-\alpha+1} \cdot x_m^{\alpha} \cdot R^{-\alpha+1} - \frac{\alpha}{-\alpha+1} \cdot x_m \\
&= 0 + \frac{\alpha}{\alpha-1} \cdot x_m \\
&= \frac{x_m \cdot \alpha}{\alpha-1}
\end{aligned} \tag{3.11}$$

In conclusion (3.8), (3.9) and (3.11) gives:

$$\mathbb{E}[X] = \begin{cases} \infty & \text{for } 0 < \alpha \leq 1 \\ \frac{x_m \alpha}{\alpha-1} & \text{for } \alpha > 1 \end{cases} \tag{3.12}$$

4. The Maximum Likelihood Estimation of the two parameters can be calculated as follows:

   - $\widehat{x}_m = \min_i X_i$.
   - $\widehat{\alpha} = \frac{n}{\sum_i (\ln X_i - \ln \widehat{x}_m)}$, with n being the total number of data points.

3.1.0.3 Gamma distribution (Stacy, 1962)

The more general case of the exponential distribution is the gamma distribution. It is also a commonly used power law. The gamma distribution will also be used in this thesis to research the strength of an earthquake. Take X a random variable with an gamma distribution. This distribution has two parameters; a shape parameter α and scale parameter β, for which holds that $\alpha > 0$ and $\beta > 0$. The following properties hold:

1. The cumulative distribution function (CDF):

$$F_X(x) = \mathbb{P}(X \leq x) = \frac{1}{\Gamma(\alpha)} \gamma\left(\alpha, \frac{x}{\beta}\right) \tag{3.13}$$

   with $\gamma(s, x)$ the incomplete Gamma function ($= \int_0^x t^{s-1}\, e^{-t}\, \mathrm{d}t$).

2. The probability distribution function (PDF):

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha\,-\,1} e^{-\frac{x}{\beta}} \tag{3.14}$$

3. The expected value:

$$\mathbb{E}[X] = \alpha\beta \tag{3.15}$$

3.1.1 Three Types Theorem (Nadarajah)

A theorem that is central in this thesis is the Three Types Theorem, also known as the Fisher – Tippett – Gnedenko theorem. This theorem will play an important part when analyzing the distribution of the maximum of random variables with a specified distribution. a more simple version is that this theorem states that the maximum of a set of random variables converges to one of the three distributions when n goes to infinity.

Now the exact theorem will be stated, followed by an elaboration on the distributions that are used in the theorem. The theorem states as follows:

Theorem 3.1.1. *Let $M_n = \max_{i=1}^{n} X_i$ with distribution function $F_{M_n}(x)$ and $X_1, ..., X_n$ independent random variables with a known distribution independent of n. If there exist sequences a_n and $b_n > 0$ and a distribution function of H such that*

$$\frac{M_n - a_n}{b_n} \xrightarrow{d} H \tag{3.16}$$

Then the distribution of H must be of the same type as one of the three following distributions: Fréchet, Weibull or Gumbel, or the distribution of H is degenerate and H is a constant or infinity.

3.1.1.1 Fréchet distribution (Cordeiro, 2011)

The Fréchet distribution is commonly used in extreme value theory. Take X a random variable with a Fréchet distribution. This distribution has three parameters; a shape parameter α, a location parameter m and a scale parameter s, for which holds that $\alpha > 0$ and $s > 0$. The following properties hold:

1. The cumulative distribution function (CDF):

$$F_X(x) = \mathbb{P}(X \leq x) = e^{-\left(\frac{x-m}{s}\right)^{-\alpha}} \text{ if } x > m \tag{3.17}$$

2. The probability distribution function (PDF):

$$f_X(x) = \frac{\alpha}{s}\left(\frac{x-m}{s}\right)^{-1-\alpha} \cdot e^{-\left(\frac{x-m}{s}\right)^{-\alpha}} \tag{3.18}$$

3. The expected value:

$$\mathbb{E}[X] = m + s \cdot \Gamma\left(1 - \frac{1}{\alpha}\right) \text{ for } \alpha > 1 \tag{3.19}$$

with $\Gamma(x)$ the Gamma function. If $\alpha \in (0, 1]$, the expected value diverges.

3.1.1.2 Weibull distribution (Rinne, 2009)

The Weibull distribution was first identified by Maurice Fréchet, so this explains why the Fréchet and Weibull distribution closely resembles each other. Take X a random variable with a Weibull distribution. This distribution has two parameters; a shape parameter k and a scale parameter λ, for which holds that $k > 0$ and $\lambda > 0$. The following properties hold:

1. The cumulative distribution function (CDF):

$$F_X(x) = \mathbb{P}(X \leq x) = 1 - e^{-(x/\lambda)^k} \tag{3.20}$$

2. The probability distribution function (PDF):

$$f_X(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} \tag{3.21}$$

3. The expected value:

$$\mathbb{E}[X] = \lambda\,\Gamma(1+\frac{1}{k}) \tag{3.22}$$

with $\Gamma(x)$ the Gamma function.

3.1.1.3 Gumbel distribution (Nadarajah and Kotz, 2004)

The Gumbel distribution is the last of the three distributions in the Three Types Theorem. The Gumbel distribution is commonly used to model the distribution of a maximum (or the minimum) of a number of samples of various distributions, so it used a lot in extreme value theory. Take X a random variable with a Gumbel distribution. This distribution has two parameters; a location parameter μ and a scale parameter β, for which holds that $\beta > 0$. The following properties hold:

1. The cumulative distribution function (CDF):

$$F_X(x) = \mathbb{P}(X \leq x) = e^{-e^{-\frac{(x-\mu)}{\beta}}} \tag{3.23}$$

2. The probability distribution function (PDF):

$$f_X(x) = \frac{1}{\beta}\cdot e^{-(z+e^{-z})}, \text{ where } z = \frac{x-\mu}{\beta} \tag{3.24}$$

3. The expected value:

$$\mathbb{E}[X] = \mu + \beta\gamma \tag{3.25}$$

with γ meaning the Euler's constant ($\approx 0,5772$).

3.1.2 Maximum of i.i.d. random variables

This subsection focuses on the method of finding out what the convergence in distribution of a maximum of i.i.d. random variables does in general. This lemma will be very useful in this thesis, since these random variables will have different distributions in some cases within this thesis. So, each set of random variables from the different distributions will give different results, while the approach stays the same. The lemma is given as follows:

Lemma 3.1.2. *If $X_1, ..., X_n$ independent random variables with a known identical distribution, $M_n = \max_{i=1}^n X_i$ with a_n and b_n given, then*

$$F_{M_n}(b_n \cdot x + a_n) = \left(F_{X_1}\left(b_n \cdot x + a_n\right)\right)^n \tag{3.26}$$

and for n to infinity, that will converge to $F_M(x)$.

Proof. First rewrite $F_{M_n}(b_n \cdot x + a_n)$ to a more convenient form to work in, so

$$\begin{aligned} F_{M_n}(b_n \cdot x + a_n) &= \mathbb{P}\left(M_n \leq b_n \cdot x + a_n\right) \\ &= \mathbb{P}\left(\max_{i=1}^{n} X_i \leq b_n \cdot x + a_n\right) \end{aligned} \tag{3.27}$$

If the maximum of a set of random variables has to be smaller than a certain value, it holds that every random variable itself has to be smaller than that value. This gives that

$$F_{M_n}(b_n \cdot x + a_n) = \mathbb{P}(X_1 \leq b_n \cdot x + a_n, ..., X_n \leq b_n \cdot x + a_n) \tag{3.28}$$

$X_1, ..., X_n$ are alle independent random variables of each other, so the probability can be taken apart. This gives that

$$F_{M_n}(b_n \cdot x + a_n) = \mathbb{P}(X_1 \leq b_n \cdot x + a_n) \cdot ... \cdot \mathbb{P}(X_n \leq b_n \cdot x + a_n) \tag{3.29}$$

$X_1, ..., X_n$ are also identically distributed, so $X_1, ..., X_n$ are i.i.d. . This gives that the probabilities in (3.29) are all equal to each other. So using this fact

$$F_{M_n}(b_n \cdot x + a_n) = \mathbb{P}(X_1 \leq b_n \cdot x + a_n)^n \tag{3.30}$$

The CDF of X_1 is known, so the probability can be replaced with

$$F_{M_n}(b_n \cdot x + a_n) = (F_{X_1}(b_n \cdot x + a_n))^n \tag{3.31}$$

Since the rewrite of this expression is dependent on the distribution, there is no general way to approach this expression. But after these rewrites let n go to infinity. Then it is known that

$$F_{M_n}(b_n \cdot x + a_n) \overset{n\to\infty}{\longrightarrow} F_M(x) \tag{3.32}$$

□

3.2 Used data

In this thesis, there will be quite some analysis on earthquake data. It is important that the used data is trustworthy and representative. One of the bigger trustworthy organizations in the Netherlands is the KNMI, the Royal Dutch Meteorological Institute. They have a data portal so everyone can have some insight in what they measure. The data from this portal is used in this thesis (KNMI, 2016). The KNMI gathers data from 45 different measurement stations all over the Netherlands, but also has close contact with other countries, to get an overview of what is happening where. (KNMI, 2017)

The choice for this organization is that is property of the Dutch government and so a lot of checks on trustworthiness are performed on a yearly basis.

The data set has the following features:

date:	the date when the earthquakes occurred in format YYYYMMDD.
time:	the specific time on that date in format HHMMSS (hours, minutes and seconds).
location:	The location of the earthquake, where the location is selected after which city has the shortest distance to the epicenter.
latitude & longitude:	The location of the epicenter expressed in geographical coordinates.
depth:	The depth of the earthquake from the surface to the epicenter.
magnitude:	The magnitude of the earthquake expressed on a Richter magnitude scale.
evaluation:	The evaluation method how the data is checked. In this specific dataset, every row is evaluated manually.

To get an idea of how the data looks, without already starting a data analysis, some plots are made and information is gathered. Starting with a summary of the data. What has to

be taken into account is that some research facilities, like TNO, state that some of the smaller earthquakes, smaller than 1.0 - 1.5, depending on the location, are missed. (Bourne et al., 2015) This would give some false information about the quantity in which smaller earthquakes occur.

- The dataset contains 1330 data points.
- The exactness of the data is on one decimal. Which is a sufficient exactness.
- The lowest value in the data set is 0.1, since all values lower or equal to 0 were deleted, since these values do not make any sense in context. So in this dataset an earthquake must have a value of at least 0.1.
- The maximum value of the set is 3.6. This is the strongest earthquake that happened in Groningen (Rijksoverheid, 2017). It happened in Huizinge on the 16th of August 2012.
- The mean of this dataset is 1.242 and the median 1.2.
- The variance of the dataset is 0.368.

To get a visual idea of the data a density plot is made, as shown in Figure 3.2.

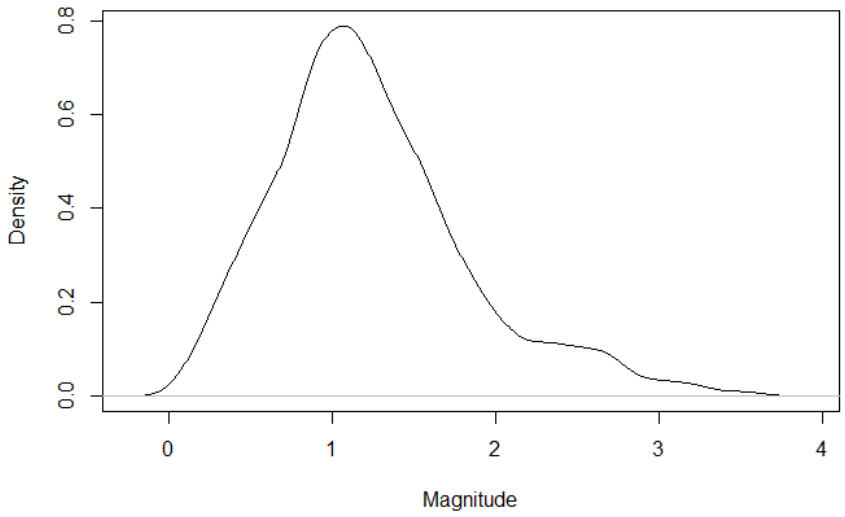


Figure 3.2: Density of the dataset

3.3 Used software and programs

In this thesis the main programming language is R. The choice for this language was because of its easy way of programming. Also the experience from the author about this language is more than other possible languages. The software used is R and R Studio. Where in R there is a package called R commander, which makes it easier to generate some more elaborate plots.

There were 5 different programs written for this thesis. The first was a basic program to generate a plot of the Gutenberg-Richter law. The code for this program can be found in Appendix A.

The second program is the data analysis of the strength of an earthquake executed in Section 6.1. Each different distribution is examined and ways to estimate the parameters are sought. Also some plots are generated. The code for this program can be found in Appendix B. In this program there is a method used from Clauset and Shalizi (2009), where they have found a good way to estimate the parameter values of a Pareto distribution on a given dataset. The code for this method can be found in Appendix C.

The fourth program is a way of generating a dataset of maximum earthquake strengths. A further elaboration will be given in Subsection 6.2.1 and the code can be found in Appendix D.

The last program has the same setup as the second program about the data analysis in Section 6.2. The only difference is that different distributions are tested on a different dataset. The generated dataset from the fourth program will be used and then a data analysis of the maximum of a set of earthquake strengths will be executed. The code for this data analysis can be found in Appendix E.

Chapter 4

Literature study

4.1 General theory of the modified Gutenberg-Richter law for large seismic moments

An article by D. Sornette, and A. Sornette.

In this article a mathematical approach is taken to analyze earthquakes. The main interest is to find out what modification has to be made to the Gutenberg-Richter law to also give a good calculation for the larger seismic moments. Since this article suggests that the current Gutenberg-Richter law does not provide such accurate answers.

The question formulated is whether at the tail of the distribution there should be a hard cut-off or a soft one, or whether the value of the b parameter (as given in equation (4.1)) should be increased upward of some magnitude.

In the introduction there is an elaboration on the Gutenberg-Richter law. This law states the relationship between the number $N(m)$ of earthquakes with magnitude $\geq m$ and the relationship with the seismic moment M.

These are

$$\log_{10} N(m) = a - bm \qquad \text{with } b \approx 1 \tag{4.1}$$

$$P(M) = \frac{\mu M_t^{\mu}}{M^{1+\mu}} \tag{4.2}$$

where a is the log of the number of $M \geq 0$ earthquakes, b a parameter that describes the relative decline in abundance of larger earthquakes relative to smaller ones, and where M_t is the lower seismic moment cut-off.

To give an illustration of the law in equation (4.1), it states that the number of earthquakes with magnitude of at least M is decreasing with a certain factor when the magnitude is increased with that same factor.

Also the relationship between seismic moment M and magnitude m is given. Namely,

$$m = \frac{1}{\beta}[\log_{10}(M) - 9] \tag{4.3}$$

with β equal to 1.5. The value of β and the number 9 in equation (4.3) are found empirically.

The reason why a lower seismic moment is chosen, is because many real-world distributions do not follow a power law over their entire range. When $M \to 0$ the function $P(M)$ diverges. In reality this is not possible, so the distribution must deviate form the power-law form below some minimum value M_t. (Newman, 2005)

The first law (4.1) states that the number of earthquakes with at least magnitude m happen 10 times more frequent than earthquakes with at least magnitude $m + 1$. To understand why this previous statement is true and to find out what the meaning of a and b is in this law, the law is rewritten to

$$N(m) = 10^{a-bm} = 10^a \cdot 10^{-bm} = 10^a \cdot (\frac{1}{10})^{bm} \qquad \text{with } b \approx 1 \tag{4.4}$$

As seen in this rewrite, $(\frac{1}{10})^{bm}$ is the factor that states the segment of the number of earthquakes with at least magnitude m. To illustrate this further, when the law is rewritten as

$$\frac{N(m)}{10^a} = (\frac{1}{10})^{bm}, \tag{4.5}$$

it can be seen that 10^a represents the total number of earthquakes that occurred in the chosen region and time, so $\frac{N(m)}{10^a}$ represents the relative frequency of an earthquake having a magnitude of at least m. So according to equation (4.5), $(\frac{1}{10})^{bm}$ represents this relative frequency too. So if m is increased by 1, the relative frequency is multiplied with a factor of $\frac{1}{10}$, so the relative frequency of having an earthquake with a magnitude of at least $m + 1$ is ten times smaller than having an earthquake with a magnitude of at least m.

When equation (4.5) is plotted, the output given is seen in Figure 4.1. It is known that this law has a power law, so if this formula is plotted in a log-log plot, it gives a straight line. Note that m already is on a logarithmic scale. The result can be seen in Figure 4.2. This result is not surprising, since the law is already known, but what is more interesting is to provide some log-log plots on data, where their distribution is not yet known. This will be further researched in the Chapter 6: Data Analysis.

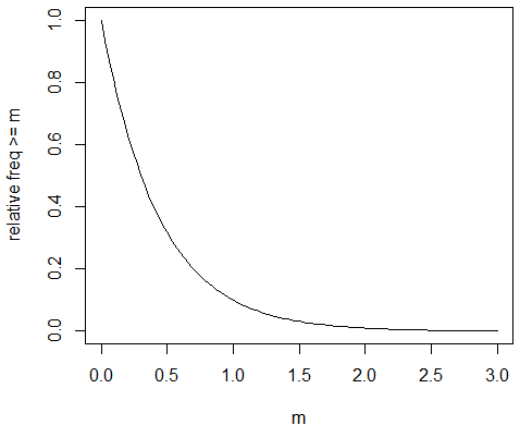


Figure 4.1: Gutenberg-Richter law

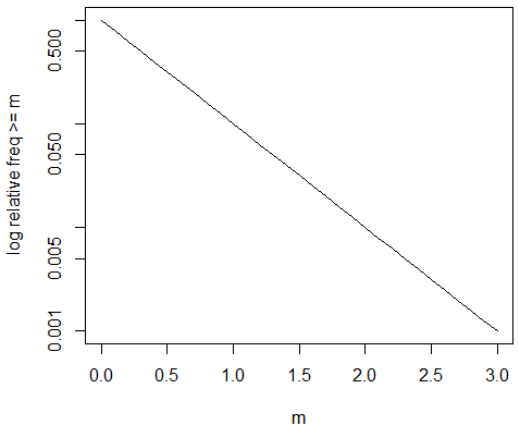


Figure 4.2: log-log Gutenberg-Richter law

From the relation stated in equation (4.3) it can also be analyzed what the relation is between an increase in scale and an increase in seismic moment. Equation (4.3) can be rewritten as

$$m = \frac{1}{\beta}[\log_{10}(M) - 9] \tag{4.6}$$

$$\Rightarrow m\beta + 9 = \log_{10}(M) \tag{4.7}$$

$$\Rightarrow M = 10^{m\beta+9} \tag{4.8}$$

$$\Rightarrow M = 10^{\beta m} \cdot 10^9 \tag{4.9}$$

Since β is generally taken equal to 1.5, this means that if m increases by 1, the seismic moment increases with a factor of $10^{\beta} = 10^{1.5} \approx 31.6$. So the total seismic moment created in an earthquake with a magnitude of $m+1$, is 31.6 higher than an earthquake with magnitude m.

4.2 A seismological model for earthquakes induced by fluid extraction from a subsurface reservoir

An article by S.J. Bourne, S.J. Oates, J. van Elk, and D. Doornhof.

This article chooses an approach with the emphasis on the physics of an earthquake in Groningen. Bourne et al analyzes the total strain that is caused by an earthquake, using the seismic moment of an earthquake.

The Gutenberg-Richter law, where $N(M)$ represents the number of earthquakes with magnitude of at least M, is stated by

$$\log_{10} N(M) = a - bM \tag{4.10}$$

where a is the log of the number of $M \geq 0$ earthquakes and b a parameter that describes the relative decline in abundance of larger earthquakes relative to smaller ones. This law is also stated in Section 4.1. They also analyze the relationship between the magnitude, M, and the seismic moment, M_0. This takes the form of

$$\log_{10} M_0 = c + dM \tag{4.11}$$

Where normally $c = 9.1$ and $d = 1.5$. These values are found empirically. Equation (4.11) is also stated in the article by Sornette and Sornette in Section 4.1.

Equation (4.11) can be rewritten to an equation expressing the magnitude M, with the seismic moment as the parameter, as follows

$$\log_{10} M_0 = 9.1 + 1.5M \tag{4.12}$$

$$1.5M = \log_{10} M_0 - 9.1 \tag{4.13}$$

$$M = \frac{1}{1.5}(\log_{10} M_0 - 9.1) \tag{4.14}$$

4.3 Similarities and Remarks

Different scientific papers, can support each other or can be opposite of each other. Some important remarks are placed here.

In the articles elaborated in Section 4.1 and Section 4.2, there can be seen that both analyze the same quantity in their models; namely the seismic moment. The reason for this is that the seismic moment is a clear and objective quantity that states what the strength of an earthquake is, since some measurement scales, as elaborated in Chapter 3: Preliminaries, are not that always that suitable. But there is also a difference. While Sornette and Sornette focuses more on the distribution of the seismic moment of one earthquake or on the maximum seismic moment of an earthquake, Bourne et al focuses more on the sum of the seismic moments. This difference is probably due to having different goals. Sornette and Sornette focuses on finding out more about an earthquake itself and what the risks are on stronger earthquakes. Therefore, an analysis on the maximum is appropriate. While Bourne et al focuses on finding out more on the total strain of induced earthquakes, so finding out more on the sum of seismic moments is more interesting then.

Chapter 5

Theoretical Analysis

There are many different ways to look at the prediction of an earthquake. One could analyze the frequency of the earthquakes on a specific location or could look at the power of such an earthquake, to prevent or to anticipate on them. It could be useful to know what the distribution of an earthquake's strength and of the maximum of such earthquakes are, to know how strong the buildings must be to withstand such powers. In this first section, there is an analysis of the distribution of an earthquake's strength itself, while the second section analyzes the distribution of the maximum of a set of earthquakes.

5.1 Approach for an earthquake distribution

5.1.1 Soft exponential cut-off

Define $X_1, ..., X_n \sim \text{Pareto}(x_m, \alpha)$ with $x_m > 0$ and $\alpha > 0$ the strengths of n earthquakes. This distribution is a model for the seismic moment of a set of earthquakes.

Since it is said that the correct approximation problem lies in the tail of the distribution, a soft cut-off is suggested in Sornette and Sornette (1999). In this thesis, an exponential cut-off is chosen. Define $Y_i \overset{d}{=} \min(X_i, E_i) = X_i \wedge E_i$, where $E_i \sim \exp\left(\frac{1}{A}\right)$, with $i = 1, ..., n$.

Lemma 5.1.1. *The random variable Y_i has $F_{Y_i}(x) = \mathbb{P}(Y_i \leq x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha} \cdot e^{-\frac{x}{A}}$ as a CDF.*

Proof. When proving this Lemma, the following is true according to the definition, so

$$\mathbb{P}(Y_i \geq x) = \mathbb{P}(X_i \wedge E_i \geq x) \tag{5.1}$$

Since the minimum of X_i and E_i must be greater than x, both variables must be greater than x, so

$$\mathbb{P}(Y_i \geq x) = \mathbb{P}(X_i \geq x) \cdot \mathbb{P}(E_i \geq x) \tag{5.2}$$

The CDF of E_i ($\exp(\frac{1}{A})$) is $F_{E_i}(x) = 1 - e^{-\frac{x}{A}}$ and the CDF of X_i (Pareto) is $F_{X_i}(x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha}$ according to respectively definition 3.1.0.1 and 3.1.0.2. These can be used to substitute the probabilities in (5.2). This gives that

$$\mathbb{P}(Y_i \geq x) = \left(\frac{x}{x_m}\right)^{-\alpha} \cdot e^{-\frac{x}{A}} \tag{5.3}$$

So $F_{Y_i}(x) = \mathbb{P}(Y_i \leq x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha} \cdot e^{-\frac{x}{A}}$. □

5.2 Analysis of the maximum of a set of earthquakes

5.2.1 Maximum of Pareto random variables

To get an estimation of how the distribution of the maximum of an earthquake set could be, first the distribution of an earthquake's strength has to be assumed. In this case, this will be a Pareto distribution. Now define $M_n^X = \max_{i=1}^n X_i$ as the maximum of the Pareto distributed seismic moments. In the following theorem, n will be taken to infinity to see what this maximum does when the data set grows larger. Then for M_n^X holds the following theorem.

Theorem 5.2.1. *Given $X_i, .., X_n \sim$ Pareto(x_m, α) and $M_n^X = \max_{i=1}^n X_i$. For each $x_m > 0$, $\alpha > 0$,*

$$n^{-\frac{1}{\alpha}} M_n^X \xrightarrow{d} M^X \tag{5.4}$$

with $F_{M^X}(x) = \mathbb{P}(M^X \leq x) = e^{-\left(\frac{x}{x_m}\right)^{-\alpha}}$. This is a Fréchet distribution with shape parameter α and scale parameter x_m.

Proof. When proving this theorem, we make use of the convergence in distribution of a maximum of i.i.d. random variables (see Lemma 3.1.2) and the fact that for a Pareto distribution (see Subsubsection 3.1.0.2) the CDF $F_{X_1}(x) = 1 - \left(\frac{x}{x_m}\right)^{-\alpha}$. This gives

$$\mathbb{P}\left(n^{-\frac{1}{\alpha}} M_n^X \leq x\right) = \left(1 - \left(\frac{x \cdot n^{\frac{1}{\alpha}}}{x_m}\right)^{-\alpha}\right)^n \tag{5.5}$$

When rewriting we get an expression which resembles the limit of e^{-x}, but now x is replaced with $\left(\frac{x_m}{x}\right)^{\alpha}$. So

$$\begin{aligned} \mathbb{P}\left(n^{-\frac{1}{\alpha}} M_n^X \leq x\right) &= \left(1 - \frac{x_m^\alpha}{n \cdot x^\alpha}\right)^n \\ &= \left(1 - \frac{\left(\frac{x_m}{x}\right)^\alpha}{n}\right)^n \\ &\xrightarrow{n\to\infty} e^{-\left(\frac{x_m}{x}\right)^\alpha} = e^{-\left(\frac{x}{x_m}\right)^{-\alpha}} \end{aligned} \tag{5.6}$$

So $F_{M^X}(x) = \mathbb{P}\left(M^X \leq x\right) = e^{-\left(\frac{x}{x_m}\right)^{-\alpha}}$. It can be seen that this is the CDF of a Fréchet distribution with shape parameter α, scale parameter x_m and location parameter 0. The Three Types Theorem (Theorem 3.1.1) supports this conclusion. □

The Fréchet distribution can be a way of interpreting the distribution of an earthquake's maximum strength. So this gives the properties of M^X according to Section 3.1.1.1:

1. The cumulative distribution function (CDF):

$$F_{M^X}(x) = \mathbb{P}(M^X \leq x) = e^{-\left(\frac{x}{x_m}\right)^{-\alpha}} \text{ if } x > x_m \tag{5.7}$$

2. The probability distribution function (PDF):

$$f_{M^X}(x) = \frac{\alpha}{x_m}\left(\frac{x}{x_m}\right)^{-1-\alpha} \cdot e^{-\left(\frac{x}{x_m}\right)^{-\alpha}} \tag{5.8}$$

3. The expected value:

$$\mathbb{E}[M^X] = x_m \cdot \Gamma\left(1 - \frac{1}{\alpha}\right) \text{ for } \alpha > 1 \tag{5.9}$$

with $\Gamma(x)$ defined as the Gamma function.

Since M^x is a more theoretical distribution, because the amount of data will never be infinite, it is also useful to look at some properties of M_n^X, since this distribution is dependent on the amount of data available. Some properties of M_n^X are:

1. The CDF:

$$F_{M_n^X}(x) = \left(1 - \left(\frac{x}{x_m}\right)^{-\alpha}\right)^n \text{ with } x > x_m. \tag{5.10}$$

since the CDF of a Pareto random variable is $1 - \left(\frac{x}{x_m}\right)^{-\alpha}$ with $x > x_m$.

2. The PDF:

$$\begin{aligned} f_{M_n^X}(x) &= \frac{d}{dx}\left(F_{M_n^X}(x)\right) \\ &= \frac{d}{dx}\left(\left(1 - \left(\frac{x}{x_m}\right)^{-\alpha}\right)^n\right) \\ &= n \cdot \left(1 - \left(\frac{x}{x_m}\right)^{-\alpha}\right)^{n-1} \cdot \alpha \cdot \left(\frac{x}{x_m}\right)^{-\alpha-1} \cdot \frac{1}{x_m} \\ &= \frac{n\alpha}{x} \cdot \left(\frac{x}{x_m}\right)^{-\alpha} \cdot \left(1 - \left(\frac{x}{x_m}\right)^{-\alpha}\right)^{n-1} \end{aligned} \tag{5.11}$$

5.2.2 Maximum with soft exponential cut-off

Since it is not certain that the exponential distribution as suggested in Subsection 5.1.1 is a good cut-off fit, another possibility is examined. What if the parameter of the exponential cut-off is dependent on n? So now take $A = \gamma \cdot n^{\frac{1}{\alpha}}$. A maximum analysis will be preformed on the Y_i variable, but with a new parameter choice for the exponential part. Unfortunately, it can be seen that it does not give an outcome as in the Three Types Theorem (Theorem 3.1.1). This is because the variable Y_i is dependent on E_i which is now dependent on n, given this choice of A.

Theorem 5.2.2. *Define $M_n^Y = \max_{i=1}^n Y_i$, with $Y_i = X_i \wedge E_i$, where $E_i \sim exp(\frac{1}{\gamma \cdot n^{\frac{1}{\alpha}}})$. For each $x_m > 0$, $\alpha > 0$:*

$$n^{-\frac{1}{\alpha}} M_n^Y \xrightarrow{d} M^Y \tag{5.12}$$

with $F_{M^Y}(x) = \mathbb{P}\left(M^Y \leq x\right) = e^{-\left(\frac{x}{x_m}\right)^{-\alpha} \cdot e^{-\frac{x}{\gamma}}}$.

Proof. When proving this theorem, we make use of the convergence in distribution of a maximum with i.i.d. random variables (Lemma 3.1.2) and the CDF of Y_i stated in Lemma 5.1.1. This gives:

$$\mathbb{P}\left(n^{-\frac{1}{\alpha}} \cdot M_n^Y \leq x\right) = \left(1 - \left(\left(\frac{x \cdot n^{\frac{1}{\alpha}}}{x_m}\right)^{-\alpha} \cdot e^{-\frac{x \cdot n^{1/\alpha}}{\gamma \cdot n^{1/\alpha}}}\right)\right)^n \tag{5.13}$$

When rewriting we get an expression which resembles the limit of e^{-x}, but now x is replaced with $\left(\frac{x}{x_m}\right)^{-\alpha} \cdot e^{-\frac{x}{\gamma}}$. So

$$\begin{aligned}\mathbb{P}\left(n^{-\frac{1}{\alpha}} \cdot M_n^Y \leq x\right) &= \left(1 - \left(\left(\frac{x}{x_m}\right)^{-\alpha} \cdot \frac{1}{n} \cdot e^{-\frac{x}{\gamma}}\right)\right)^n \\ &= \left(1 - \frac{\left(\left(\frac{x}{x_m}\right)^{-\alpha} \cdot e^{-\frac{x}{\gamma}}\right)}{n}\right)^n \\ &\overset{n\to\infty}{\longrightarrow} e^{-\left(\frac{x}{x_m}\right)^{-\alpha} \cdot e^{-\frac{x}{\gamma}}}\end{aligned} \tag{5.14}$$

So $F_{M^Y}(x) = \mathbb{P}(M^Y \leq x) = e^{-\left(\frac{x}{x_m}\right)^{-\alpha} \cdot e^{-\frac{x}{\gamma}}}$. This CDF does not hold for the Three Types Theorem, since E_i is dependent on n. □

5.2.3 Maximum of exponential random variables

Another possibility is to look if the earthquake strengths might have a exponential decay. If this is true, it could be interesting to know more about the maximum distribution with exponential random variables. Therefore, a variable G_n^A is defined, which is this maximum. The $A \cdot \log n$ is added for stability of the distribution.

Theorem 5.2.3. *Define $G_n^A = M_n^E - A \cdot \log n$. With $M_n^E = \max_{i=1}^n E_i$ and $E_i \sim exp(\frac{1}{A})$, with $i = 1, ..., n$. Take A fixed. Then holds:*

$$G_n^A \xrightarrow{d} G^A \tag{5.15}$$

with $F_{G^A}(x) = \mathbb{P}(G^A \leq x) = e^{-e^{-\frac{x}{A}}}$. So G^A has a Gumbel distribution with scale parameter A.

Proof. When proving this theorem, we make use of how to calculate the CDF of a maximum given in Lemma 3.1.2 and the CDF of E_i, which is $F_{E_i}(x) = 1 - e^{-\frac{x}{A}}$ as given in Definition 3.1.0.1. These combined give that

$$\mathbb{P}(G_n \leq x) = \left(1 - e^{-\frac{1}{A} \cdot (x + A \cdot \log n)}\right)^n \tag{5.16}$$

When rewriting we get an expression which resembles the limit of e^{-x}, but now x is replaced with $e^{-\frac{x}{A}}$. So

$$\begin{aligned}\mathbb{P}(G_n \leq x) &= \left(1 - e^{-\frac{x}{A}} \cdot e^{-\log n}\right)^n \\ &= \left(1 - e^{-\frac{x}{A}} \cdot \frac{1}{n}\right)^n \\ &= \left(1 - \frac{(e^{-\frac{x}{A}})}{n}\right)^n \\ &\overset{n\to\infty}{\longrightarrow} e^{-e^{-\frac{x}{A}}}\end{aligned} \tag{5.17}$$

So $F_G(x) = \mathbb{P}(G \leq x) = e^{-e^{-\frac{x}{A}}}$. Since all requirements for the Three Types Theorem hold, the distribution of G is indeed a Gumbel distribution with scale parameter A. □

Chapter 6

Data Analysis

6.1 Distribution of an earthquake's strength and its tail

The main goal here is to find the the best estimation of the distribution for an earthquake. So first, a summary of some distributions that deviate a lot from each other that were found in the previous sections will be presented:

- Pareto distribution
- Gutenberg-Richter law
- Gamma distribution

6.1.1 Pareto distribution

The most commonly used distribution for earthquake strengths, is the Pareto distribution. It is a power law distribution, with two parameters. The first being the scale parameter x_m. This is the minimum possible value of X_i for which the power law holds, while still being larger than 0. The second parameter is α, which is the shape parameter.

The first step is to find the values for the parameters that suit our dataset best. Clauset and Shalizi (2009) have written a program in R that analyses what the best parameter values are for a certain dataset.

To verify if, for instance, a Maximum Likelihood Estimation would give the same output, also the parameter values according to this method are calculated. When looking how to calculate these parameter values, the only premature decision made, is to not take x_m equal to the minimum of the dataset. The Pareto distribution is a decreasing function, so the value of x_m is chosen the same way as Clauset chooses it. This value is used in the calculations. When executing these methods, it gives the values of our parameters, which are as follows:

Clauset $x_m = 1, \quad \alpha \approx 3.588243$

Modified MLE $x_m = 1, \quad \alpha \approx 2.29607$

These models are shown in Figure 6.1. It seems that both models fit reasonably at the tail. Clauset's model fits best at the tail, but has a huge difference when looking at the magnitudes between 1 and 1.5. The MLE might fit better between 1 and 1.5, but still is not a very suitable option. So, the conclusion is reached that just the Pareto distribution is not a suitable estimator

for the whole dataset, since the lower magnitudes are not estimated well enough and the tail is also not that good fitted. But what the main purpose of this thesis is, is to analyze the tail of the set. So there will be zoomed in on the values higher than 2, since these earthquakes are normally the ones that are felt by humans. (UPSeis, 2017)

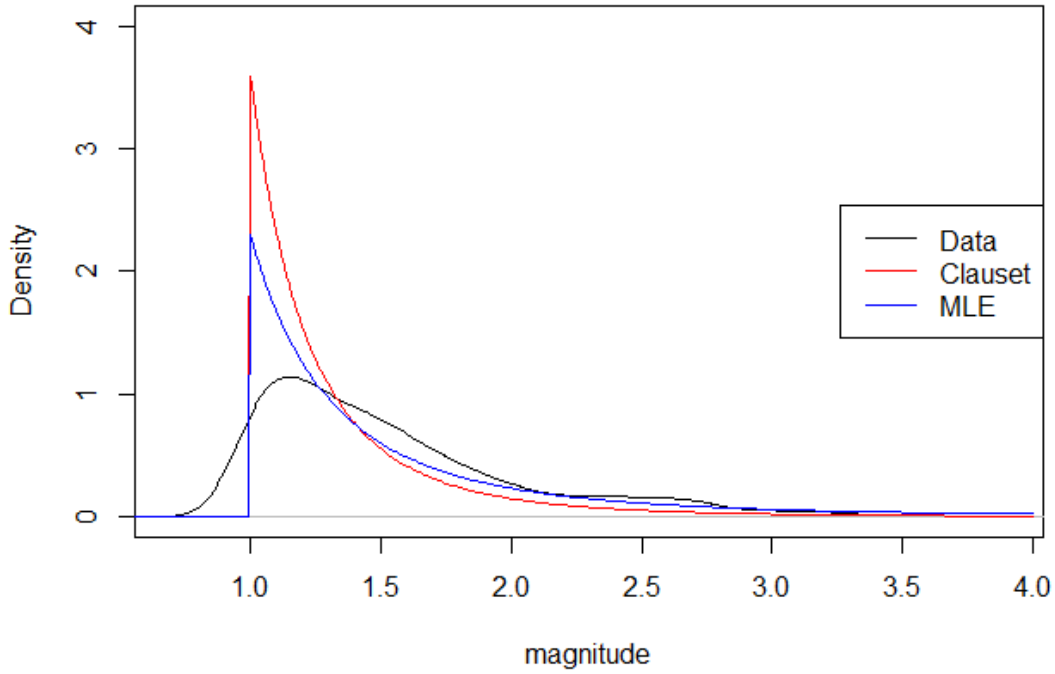


Figure 6.1: Pareto distribution fitting

When looking at Figure 6.2, it can be seen that both distributions still have some deviations. The modified MLE methods gives sort of an upper bound, while the Clauset method gives more of an lower bound. The benefit of Clauset's fit is that it does not have such a large tail while the modified MLE fit does. Research suggested that the maximum earthquake strength in Groningen would be 4.5. (NOS, 2015) This is more in alignment with Clauset.

The Q-Q plots, as seen in Figures 6.3 and 6.4, give an estimation if the Pareto distribution does fit well enough at the tail. It can be seen that there is not much difference between the two Q-Q plots and both have some trouble allocating the higher values in the data. So, in conclusion, other distributions should be researched, but until a better distribution is found, Clauset's Pareto distribution fit is preferred.

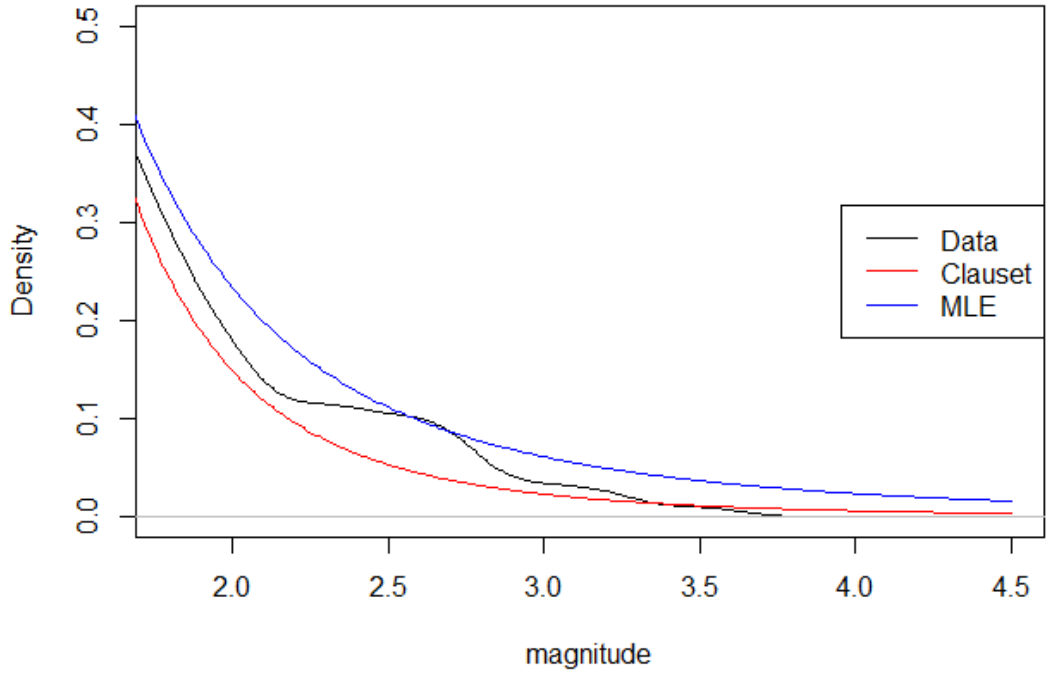


Figure 6.2: Pareto distribution fitting on the tail

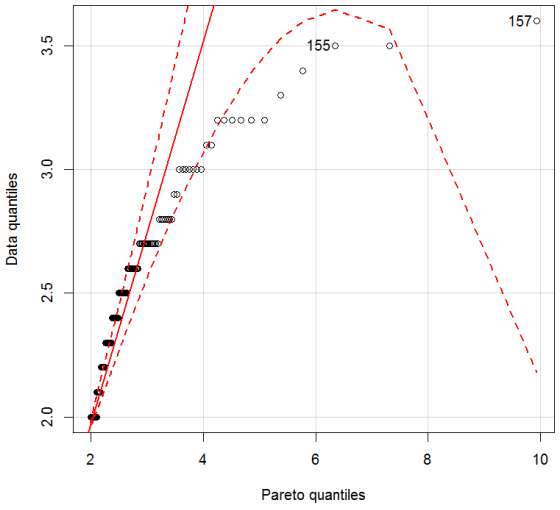


Figure 6.3: Q-Q plot with Clauset's method

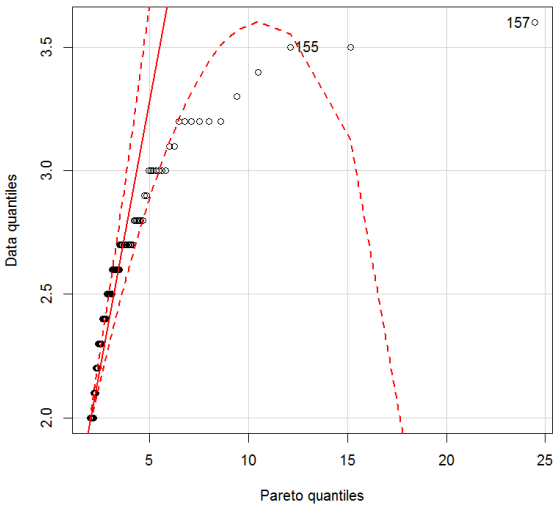


Figure 6.4: Q-Q plot with modified MLE

6.1.2 The Gutenberg-Richter law

The Gutenberg-Richter law is also one of the most commonly used distributions to fit to earthquake data, since it is an easily understood distribution. The Gutenberg-Richter law, where $N(M)$ represents the number of earthquakes with magnitude of at least M, is stated by

$$\log_{10} N(M) = a - bM \tag{6.1}$$

where a is the log of the number of $M \geq 0$ earthquakes and b a parameter that describes the relative decline in abundance of larger earthquakes relative to smaller ones.

In Section 4.1, the law was plotted to give an idea of how this distribution looks like, see Figure 4.1. When analyzing the Pareto distribution, it was already seen that a distribution with only a decreasing probability is not suitable at the beginning of the data. So when fitting the Gutenberg-Richter law, a lower cut-off is also assumed. This lower cut-off, x_m, will again take the value of 1. The result of this can be seen in Figure 6.5.

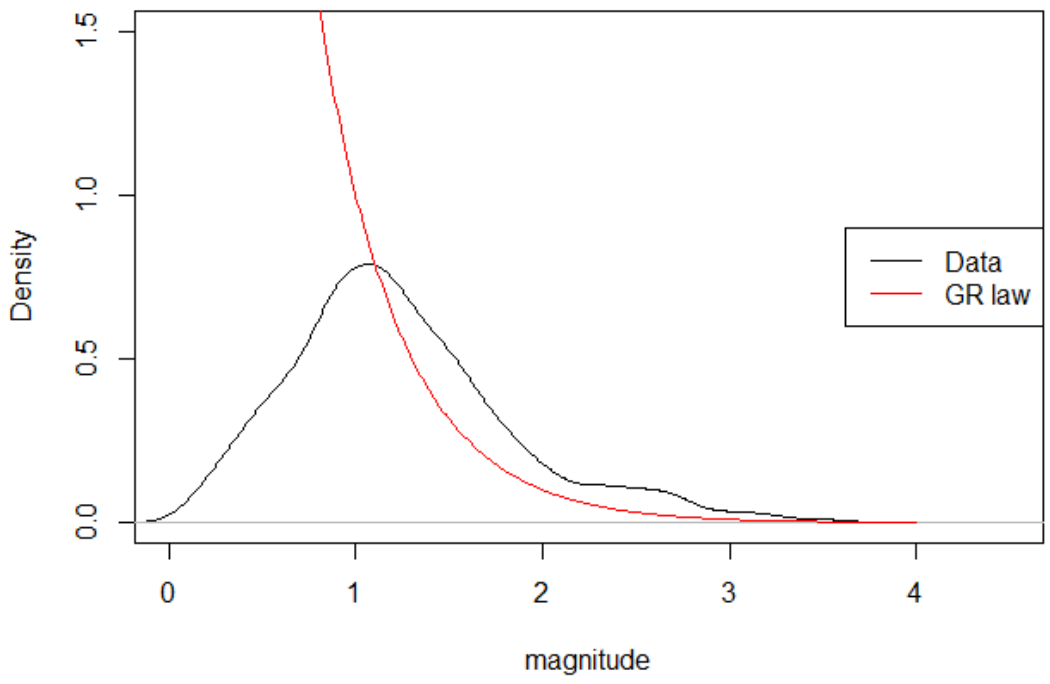


Figure 6.5: Gutenberg-Richter distribution fitting

From Figure 6.5 not much can be concluded about the tail. What can be seen is that the law does lie beneath the dataset, so will probably also give more of an lower bound. To research this in more detail, Figure 6.6 presents the tail of the dataset, also starting from 2.

It can be seen quite easily that at the tail this distribution is even a more strict lower cut-off than Clauset's Pareto fit already was. To support this statement, the probability of the

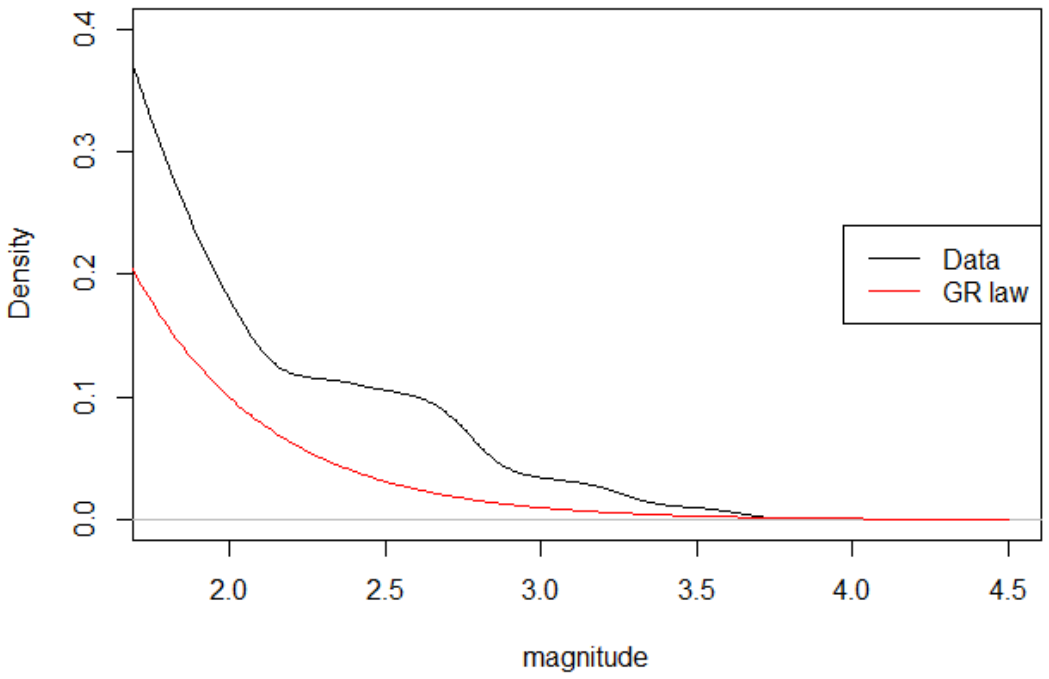


Figure 6.6: Gutenberg-Richter distribution fitting of the tail

earthquake with a strength of 4.5 or higher with the Pareto distribution is 0.003612303, while with the Gutenberg-Richter law, the probability equals 0.0003162278. This a 10 times smaller. So this supports the earlier statement of the NOS (2015), but in total the Clauset method supports the rest of the Groningen dataset more accurate. Another argument for Clauset's fit of the Pareto distribution is that the chance of an earthquake with strength 4.5 or higher is still extremely low (0.3%), so in conclusion Clauset's fit of the Pareto distribution is, so far, preferred.

6.1.3 Gamma distribution

In the article by Sornette and Sornette (1999), the main topic is that the distribution of an earthquake does not resemble the Gutenberg-Richter power law very well. As elaborated in Subsection 6.1.2, this thesis supports that statement. Sornette and Sornette advise another distribution, namely the gamma distribution. In this report the gamma distribution will be examined as a candidate distribution and the parameter values will be sought.

There will be two methods used to seek the shape and scale parameter of the gamma distribution, namely the Method of Moments and the function 'fitdistr' in R, which uses a MLE approach.

The methods give the following parameter values:

MoM $\alpha \approx 4.19325, \quad \beta \approx 3.376224$

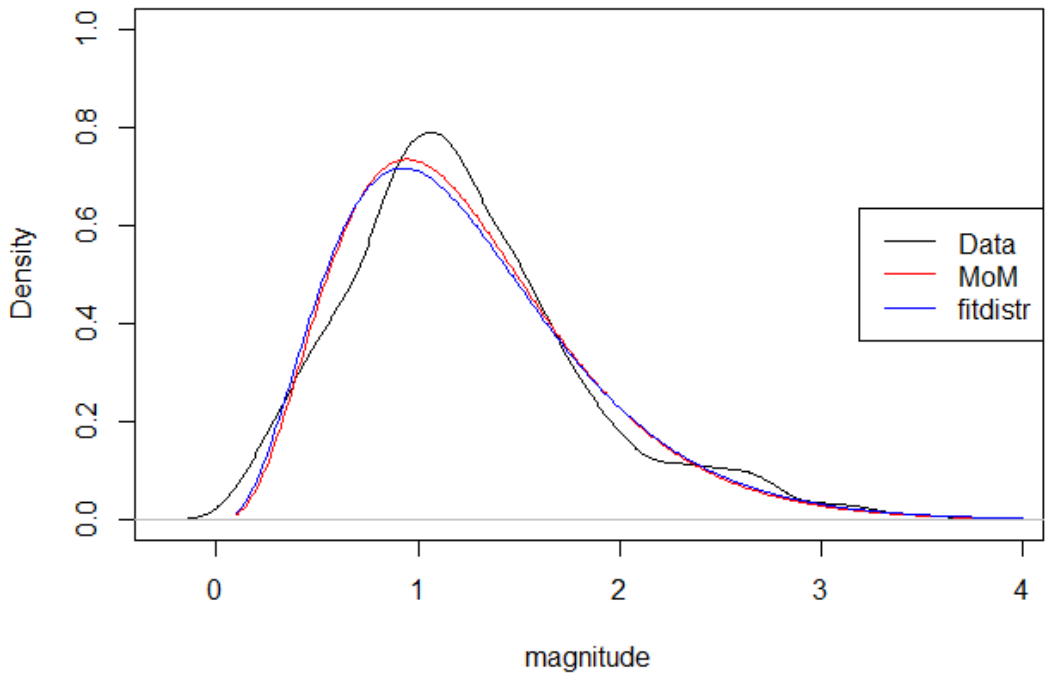


Figure 6.7: Gamma distribution fitting

fitdistr $\alpha \approx 3.92629, \quad \beta \approx 3.161281$

It can be seen that the parameter values estimates differ a bit. To get an idea of how they fit the data, the distributions are plotted in Figure 6.7. It seems that this distribution has a reasonable fit in general. But there is still quite some deviation. But before checking which one of the parameter sets is a better fit, first a Quantile Comparison plot is made to check if the distribution itself would be a good fit for the entire dataset. Figure 6.8 illustrates a Q-Q plot for the distribution. It can be seen that a lot of values do not lie within the dotted confidence lines, Which could indicate that a gamma distribution does not fit that well for the entire set.

To give a qualitative conclusion, an Anderson-Darling Goodness of Fit test is executed. This test has the following hypotheses:

H_0: The entire dataset follows a specified distribution, here a gamma distribution.
H_1: The entire dataset does not follow a specified distribution, here a gamma distribution.

If the p-value of the test is lower than 0.05, then it can be concluded that there is enough evidence to reject H_0. In this specific case, R produced a p-value of 0.005601 and 0.002983 for respectively the MoM and 'fitdistr' method. So H_0 can be rejected with enough evidence. A main effect for this is the size of our data in combination with the Anderson-Darling test. This test assumes the suggested distribution is true and then searches at each data point if there are

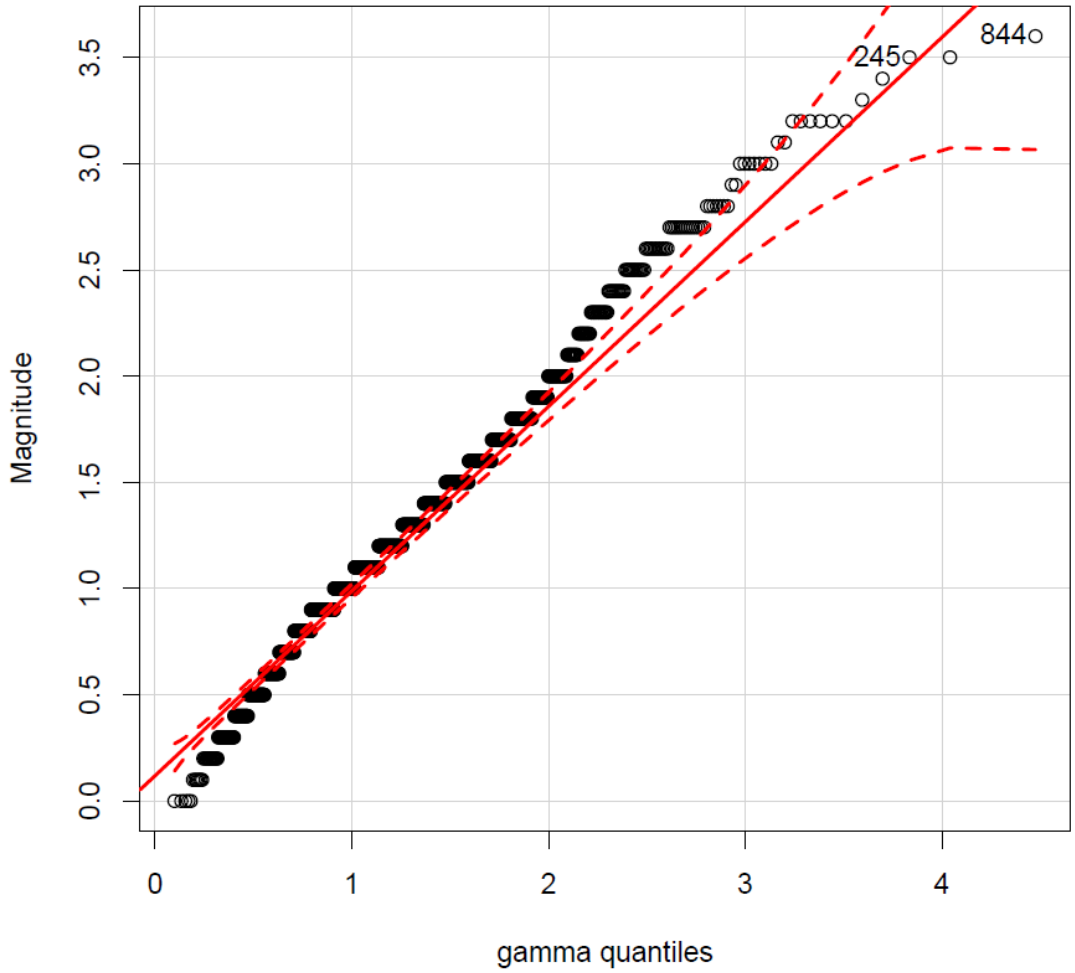


Figure 6.8: Q-Q plot with gamma distribution

reasons to conclude it is not the suggested distribution. When having such a large dataset, the test is very strict. If a sample of 50 data points would be chosen, the p-value varies between 0.3 and 0.8, which says there is not enough evidence to reject H_0, and so a gamma distribution is kept for the data.

In the current research, the conclusion is reached that the gamma distribution is not an ideal fit for the distribution of the magnitude of an earthquake, therefore, now the analysis on the tail will be performed. In Figure 6.9 it can be seen that the gamma distribution does fit quite well on 2 and higher. It even fits better then the suggested Pareto fit, since it is not a lower bound. To support this statement, a Q-Q plot of the data starting at 2, will be made. This is shown in Figure 6.10. Here it can be seen that only the data at the start (value 2) do not fit. The rest lies in the 95% confidence interval given. In the best case scenario all dots lie on the straight line presented, but this could deviate because of natural perturbation. So in conclusion, the gamma distribution might not fit the entire dataset best, but it is a relatively good fit for the distribution of the tail. It could still be that some distribution fits better, but out of the researched distributions the gamma distribution scores best.

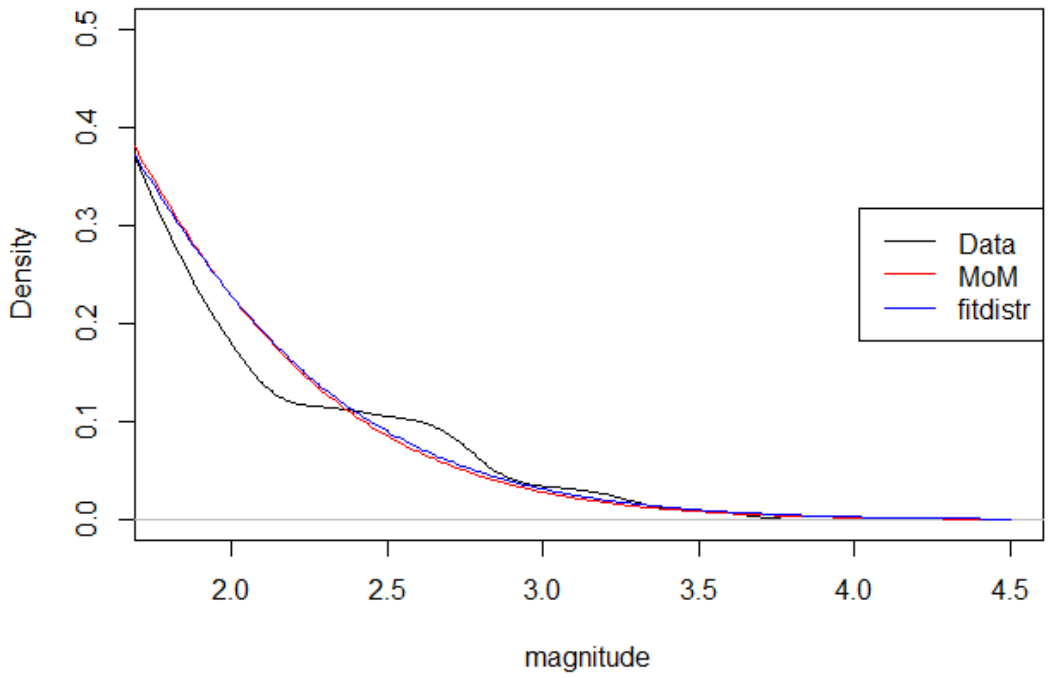


Figure 6.9: Gamma distribution fitting at the tail

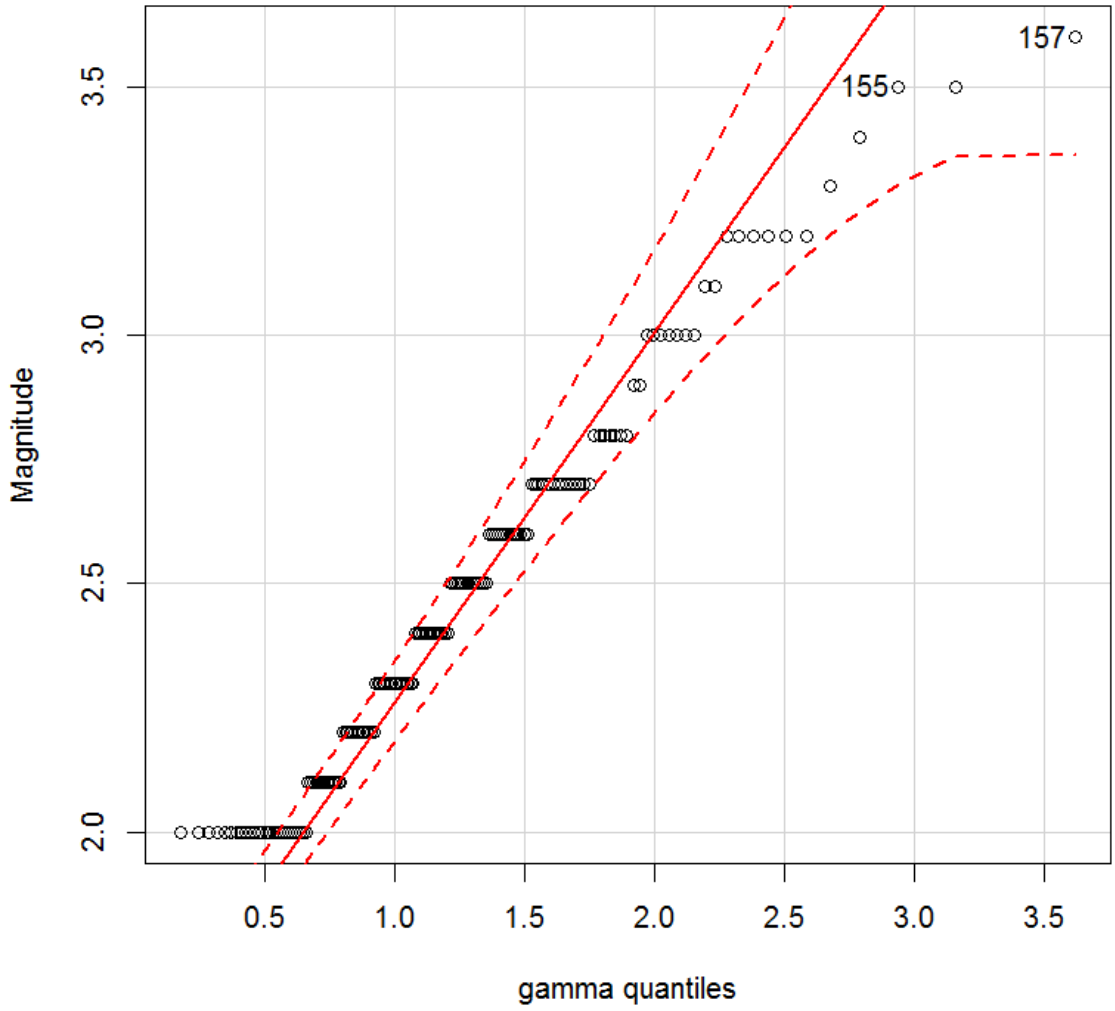


Figure 6.10: QQ plot with Gamma distribution at the tail

6.2 Distribution of the maximum of an earthquake

6.2.1 Data production

Since in this thesis only one dataset of earthquake strengths is available, it is necessary to make a set of the maximums of sets. The approach to this dataset generation is as following:

Take the dataset of n values and shuffle the order of the values, divide it in y approximately equal parts. Each part consists out of approximately a rounded $\frac{n}{y}$ points. Now take the maximum value of each of the y parts and store that value in a new dataset. This new dataset is now a dataset of maximum earthquake strengths and contains y points. Each point in the original dataset will only be used once, so no value is used twice or more. Also now there are no dependencies between the values in the maximum dataset. In this thesis the dataset contains approximately 1300 values and the value of y is chosen to be the square root of n, so in the written program, the dataset is split in approximately 36 parts containing 36 points. The R

code for this method can be found in Appendix D. A plot for this generated data can be seen in Figure 6.11.

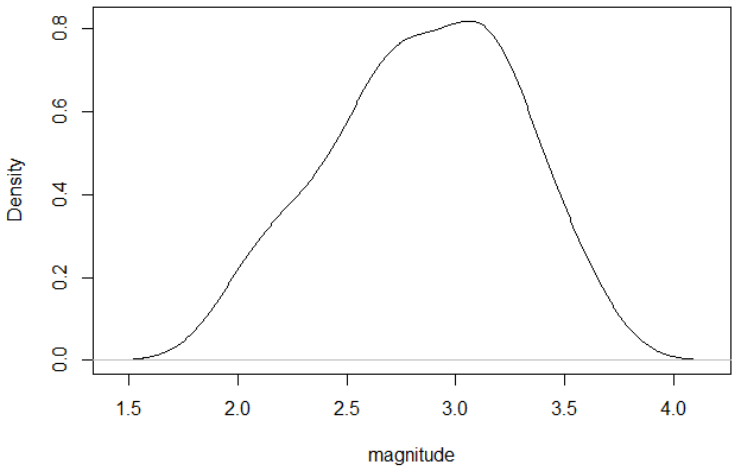


Figure 6.11: Plot of generated data

Now the dataset is generated, a summary of some interesting candidate distributions is presented. The choice for these distributions is based on Theorem 3.1.1 (Three Types Theorem):

- Fréchet distribution
- Gumbel distribution
- Weibull distribution

6.2.2 Fréchet distribution

The Fréchet distribution can function, like mentioned in Theorem 5.2.1, as a maximum of i.i.d. random variables with a Pareto distribution. So if this distribution gives a good fit, then it could suggest that the distribution of the strength of an earthquake is following a power law. First, the parameter values of the distribution need to be found. Two methods have been chosen. One will take the parameter values of the earlier Pareto distribution, because of Theorem 5.2.1. The other method is simply trial and error, finding the parameter values which resemble the best by plotting some options and looking with a bare eye which one resembles the most. The result is shown in Figure 6.12.

The conclusion from this figure is simple. The Pareto parameters do not fit well. The trial and error approximation does fit better, but the important deviation that can be seen is that the distribution of the actual data is left-skewed (with a skewness coefficient of -0.24), while the Fréchet distribution is right skewed. In conclusion, the Fréchet distribution is not a good fit for the maximum, and therefore, the suggestion that the distribution of an earthquake's strength follows a power law can also be rejected.

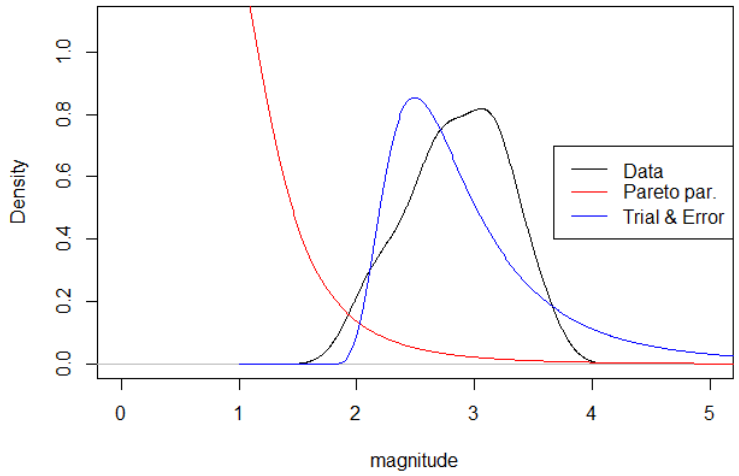


Figure 6.12: Fréchet fitting

6.2.3 Gumbel distribution

The next distribution that will be examined will be the Gumbel distribution. If a Gumbel distribution fits well, according to Theorem 5.2.3, this could suggest that the distribution of an earthquake's strength follows an exponential trend. To start, it is wise to analyzing some standard properties of the Gumbel distribution. One of the main properties is that the skewness of the distribution equals 1.14 for any set of parameters. (Nadarajah and Kotz, 2004) This means the distribution is right-skewed, just like the Fréchet distribution. This would mean the Gumbel distribution would not be a good fit for the data. To confirm this thought, a plot is given in Figure 6.13. In this plot an estimation of the Gumbel distribution parameters, using an R generated MLE parameter estimating function, is also given in red.

The hypothesis that the distribution would not fit well can now be confirmed. The skewness of the distributions does not lean to the same side, so the Gumbel distribution is not a good fit for the maximum, and therefore, the suggestion that the distribution of an earthquake's strength follows an exponential trend can also be rejected.

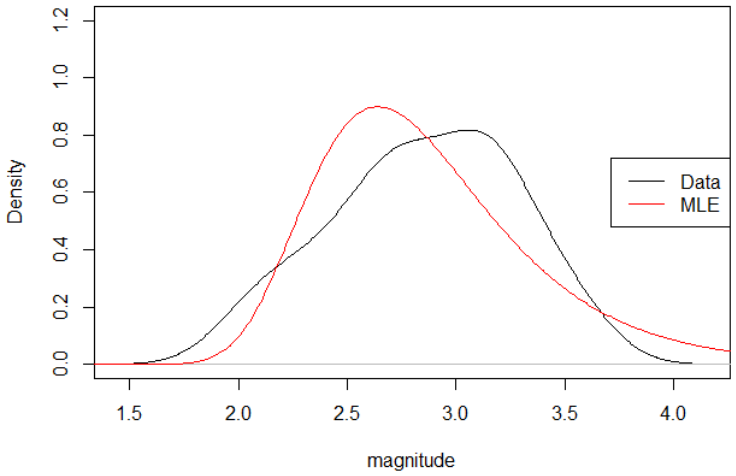


Figure 6.13: Gumbel fitting

6.2.4 Weibull distribution

The last distribution that will be examined will be the Weibull distribution. There is one main method used, namely the Maximum Likelihood Estimation. These give the two parameter value estimations for the data. This estimation together with the data is seen in Figure 6.14.

The conclusion can be drawn from this figure that it does fit better than the previous two distributions, since the skewness is on the same side. A better method to see how well this distribution fits is to make a Weibull probability plot for the data and the estimation. If the data is on or close to the estimated distribution, the dataset plotting will be linear. The result is given in Figure 6.15.

It can be seen that most points lie on the estimated line, but also that there are some deviating values. This indicates that although it does look quite promising, there is still too much deviation to conclude that the distribution of the maximum of a set of earthquake strengths follows a Weibull distribution. However, it does fit best of the three different distributions.

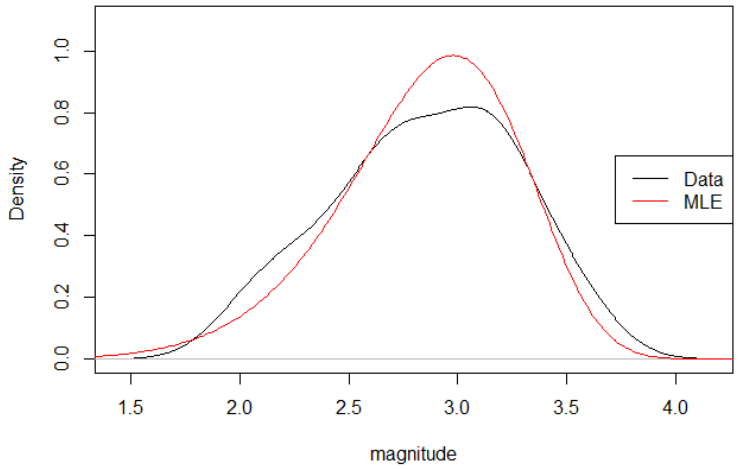


Figure 6.14: Weibull fitting

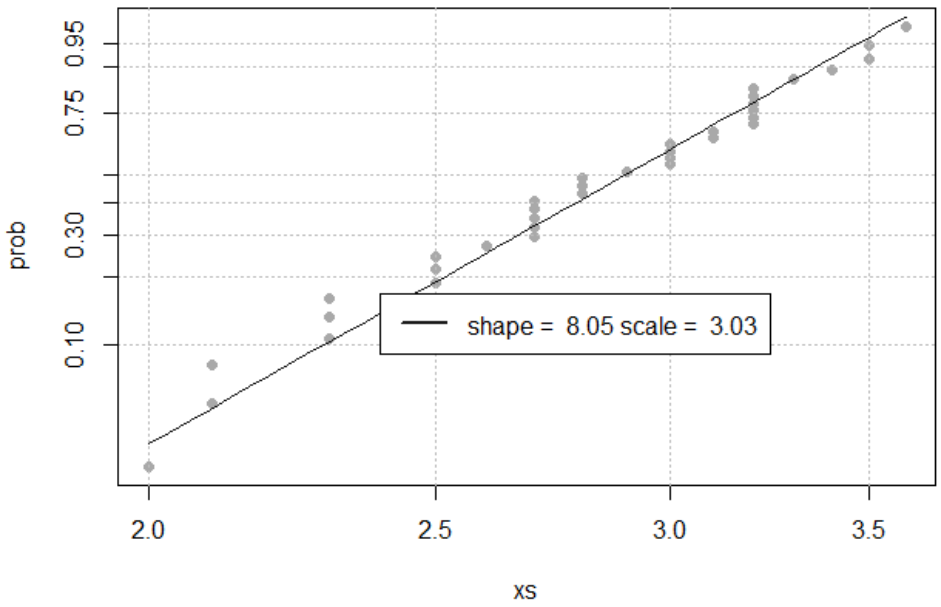


Figure 6.15: Weibull probability plot

Chapter 7

Conclusions

After some theoretical analysis and following some data analysis, the main research question of this thesis can be answered. The question was:

"What is the best fitted and statistically supported distribution of the relative frequency of a certain earthquake's strength, looking at an earthquake's seismic moment, and does this distribution have a cut-off at the end of some sort?"

And the answer to this question is none. In this thesis there is no distribution that gives a good fit for the entire range of earthquake strengths. The distribution that scored best on a overall fit was the gamma distribution, but there was still to much deviation to conclude this would be a good fit.

Since the main question is not satisfyingly answered, the next step was to research the first subquestion, which looks at the tail (strengths greater or equal to 2). The approach was to see if there was a distribution that would score well here, but also to see what kind of cut-off distribution would give a good estimation for the tail. The same distributions as before were tested, but also the same conclusion were reached. No distribution follows the data very well, but the one that did score well enough out of the researched distributions is the gamma distribution.

The next subquestion was to analyze the maximum distribution of a set of earthquake maximums, to see if this would give some implications for the distribution of an earthquake's strength. Three different distributions were tested, but none fitted the data that good. The Fréchet and Gumbel distribution did not fit well, since the skewness of the distribution was to the wrong side. The Weibull distribution did have its skewness to the correct side, but still had quite some deviation comparing to the data.

In conclusion, what do all these answers to the (sub)questions mean? According to this thesis, there is no distribution that gives a very well fit for the strength of an earthquake, or even its tail or maximum. Only some distributions that have an average fit, but where there is still much deviation. This could be due to natural perturbation or measurement errors, or that those distributions are just not a suitable model for the strength of an earthquake.

Chapter 8

Discussion

8.1 Advice

The advice that can be given according to this thesis is that there needs to be more research on other possible distributions and it is wise to look further into the way earthquakes are measured and when the data can be trustworthy. Also some more features can be linked to the earthquakes, for instance, induced or natural. In this dataset of Groningen most earthquakes will probably be induced, but it might be that the perturbation is due to the earthquakes that are natural. This could give deviating results. There are many more different features and factors that can be linked to the earthquakes, and this might give a better understanding of how each type of earthquake behaves.

To elaborate on the first advice, namely the research in other distributions. In this thesis several distribution are tested for the strength of an earthquake, as well as the maximum of a set of earthquake strengths.

For the distribution of the strength of an earthquake, the following suggestions are made for further research, but due to that some results were found in this thesis in a relative late state and due to limited time, they could not be researched in this thesis:

exponential cut-off Due to limited time, the case of a Pareto distribution with a exponential cut-off could only be performed in the theoretical analysis, while it could also be interesting to see this implemented in the data analysis. This could maybe give a good estimation for the strength of an earthquake.

different cut-off In this thesis the Pareto distribution was used with an exponential cut-off in the theoretical analysis, but there might be a better cut-off than an exponential. There could be a further analysis if this is indeed true.

gamma with cut-off The gamma distribution provided the best fit in the data analysis, but maybe this distribution could use a cut-off at the end of some sort, which would provide a better fit.

second distribution Like mentioned in the first paragraph of the discussion, it could be that the data needs to be split in different groups, since those groups have different distributions. If this is the case, than the current thesis would not give any good estimations, since it tries to find a distribution for the entire set, instead of for the different groups. It could be due to these underlying distributions that there is nu distribution found for

the strength of an earthquake. Basically, this comes down to the fact that the data is not identically distributed. So the assumption that the data is i.i.d. could possibly not correct. If this is the case, a theoretical analysis will be harder to perform.

For the maximum analysis, also some suggestions are made:

exponential cut-off As explained in the paragraph about the suggestions for the analysis of the strength of an earthquake, it could be interesting to not only analyze the Pareto distribution with a exponential cut-off in the theoretical analysis, but also in the data analysis. In this case the maximum of those Pareto random variables with an exponential cut-off.

different cut-off In the suggestions for the strength of an earthquake, it was stated that a different cut-off could be a possible research direction. If this is performed, it can be wise to also analyze the maximum of these random variables with a new cut-off. One of the possible different cut-off distributions is the gamma distribution, since it scored reasonably well on the tail.

maximum of gamma Since the gamma distribution gave the best comparison to the data, it could be interesting to research the maximum of gamma distributed random variables. The educated guess made is that the gamma distribution and exponential distribution are so closely related and the fact that the case in question also qualifies for the Three Types Theorem (Theorem 3.1.1), that the maximum of the gamma distributed random variables will also give a Gumbel distribution. But this is just speculation and it is smart to actually perform this analysis to make sure this is the case.

8.2 Limitations

There were some limitations in this thesis that will be elaborated on here.

small maximum set Currently, there were around 1300 data points, which is quite a good size for the analysis of the earthquake size, but when it comes to the analysis of the maximum, the square root is taken, which then only gives 36 data points. This is quite a low amount of data. It also has as a consequence that, when making a new dataset with the written program, it could look very differently relative to the dataset used now. This makes it hard to really get an idea of how the data of the maximum really looks like. so if in another thesis, the maximum is analyzed, the advice is to use a large enough dataset, so enough data points are generated. Also, it is advised to do some research on how to get a representative dataset.

statistical analysis The current thesis focused mainly on finding some new possibilities for the distribution of an earthquake's strength. There has been some statistical analysis using a few testing methods, but there is more that can be done in this area. Due to limited time and the amount of possible distributions that were tested, this more elaborate statistical analysis stayed behind.

lower strength values Research from TNO showed that there is quite some uncertainty about the lower values of earthquake strengths (lower than 1.0 - 1.5, depending on the location). Some are missed, which makes it difficult to really have a good data set of all the values.

Bibliography

Didier Sornette and Anne Sornette. General theory of the modified Gutenberg-Richter law for large seismic moments. *Bulletin of the Seismological Society of America*, 89(4):1121–1130, 1999. ISSN 00371106. 2, 4, 16, 24

Felix Abramovich. *Statistical Theory: A Concise Introduction*. 2013. ISBN 9781482211849. 4

Barry C Arnold. Pareto Distribution. pages 1–10, 2015. doi: 10.1002/9781118445112.stat01100.pub2. 4

E. W. Stacy. A Generalization of the Gamma Distribution. 33(3):1187–1192, 1962. URL http://www.jstor.org/stable/pdf/2237889.pdf?refreqid=excelsior{%}3Abe3d382baf648be0d83bee0ea3385300. 6

Saralees Nadarajah. The Extremal types theorem. URL http://www.maths.manchester.ac.uk/{~}saralees/ettproof.pdf. 7

Gauss M Cordeiro. The generalized inverse Weibull distribution. pages 591–619, 2011. doi: 10.1007/s00362-009-0271-3. 7

Horst Rinne. *The Weibull Distribution: A Handbook*. 2009. ISBN 978-1-4200-8743-7. 7

Saralees Nadarajah and Samuel Kotz. The beta gumbel distribution. 4(March):323–332, 2004. 8, 30

KNMI. No Title, 2016. URL http://www.knmi.nl/kennis-en-datacentrum/dataset/aardbevingscatalogus. 9

KNMI. Over het KNMI, 2017. URL http://www.knmi.nl/over-het-knmi/over. 9

S. J. Bourne, S. J. Oates, J. Van Elk, and D. Doornhof. A seismological model for earthquakes induced by fluid extraction from a subsurface reservoir. *Journal of Geophysical Research B: Solid Earth*, 119(12):8991–9015, 2015. ISSN 21699356. doi: 10.1002/2014JB011663. 10

Rijksoverheid. Aardbevingen door gaswinning in Groningen, 2017. URL https://www.rijksoverheid.nl/onderwerpen/gaswinning-in-groningen/inhoud/aardbevingen-door-gaswinning-in-groningen. 10

Aaron Clauset and Cosma Rohilla Shalizi. Power-Law Distributions in Empirical Data $*$. 51(4):661–703, 2009. 11, 20, 43

M E J Newman. Power laws , Pareto distributions and Zipf ' s law. 7514, 2005. doi: 10.1080/00107510500052444. 13

UPSeis. Earthquake Magnitude Scale, 2017. URL http://www.geo.mtu.edu/UPSeis/magnitude.html. 21

NOS. 'Aardbevingen Groningen maximaal 4,5', 2015. URL http://nos.nl/artikel/2042760-aardbevingen-groningen-maximaal-4-5.html. 21, 24

List of Figures

Appendix A

Generation of a Gutenberg-Richter law plot

```
#define the Gutenberg-Richter law
y <- function(x) {
  a <- (1/10)^{x}
  return(a)
}

#plot the GR law on a regular plot and on a Q-Q plot
x <- seq(0,3,0.1)
curve(y(x), from=0, to=3, type="l", main = "", xlab="m", ylab="relative freq >= m"
    ) #GR law
plot(x,y(x), log="y", type="l", main="", xlab="m", ylab="log relative freq >= m")
    #Q-Q plot
```

Appendix B

Data analysis on the strength of an earthquake

```
820 library(readxl)
    library(actuar)
    library(MASS)
    library(ADGofTest)
825 library(sdcMicro)

    #-------------------------histogram plot of data-------------------------------

    #read the data from the excel
830 KNMI_Groningen <- read_excel("~/Vakken/BEP/Data/KNMI Groningen.xlsx",col_types = c
        ("numeric", "numeric", "text", "numeric", "numeric", "numeric", "numeric", "
        text"))
    mag <- KNMI_Groningen$MAG

835 #remove all non positive values
    magpos <- mag[-which(mag <= 0)]


    #plot some information about the data
840 plot(density(magpos), main="", xlab="Magnitude")

    #------------------------ Pareto implementation ------------------------------

    #edit the dataset starting from 1
845 magpos1 <- magpos[-which(magpos <= 1)]

    #find the parameter values through modified MLE
    xmin1 <- 1
    alpha1 <- length(magpos1)/(sum(log(magpos1)-log(xmin1)))
850


    #make a plot of the data with the fitted parameter values
    x <- seq(0.1,4,0.01)
    plot(density(magpos[which(magpos >= 1)]), ylim=c(0,4), xlab="magnitude", main="")
855 lines(x, dpareto1(x, plfit(magpos)$alpha, plfit(magpos)$xmin), col='red') #MoM
    lines(x, dpareto1(x, alpha1, xmin1), col='blue')
    legend("right", legend=c(expression("Data", "Clauset", "MLE")), lty=1, col=c("
        black", "red", "blue"))

860 #Tail analysis
```

```
#edit the dataset starting from 2
magpos2 <-magpos[which(magpos >= 2)]

#make a plot of the data with the fitted parameter values
x2 <-seq(1.6,4.5,0.01)
plot(density(magpos), xlim=c(1.8,4.5),ylim=c(0,0.5), xlab="magnitude", main="")
lines(x2, dpareto1(x2, plfit(magpos)$alpha, plfit(magpos)$xmin), col='red') #MoM
lines(x2, dpareto1(x2, alpha1, xmin1), col='blue')
legend("right", legend=c(expression("Data", "Clauset", "MLE")), lty=1, col=c("
    black", "red", "blue"))

#chance of maximum
dpareto1(4.5, plfit(magpos)$alpha, plfit(magpos)$xmin)

#---------------------- Gutenberg-Richter law ----------------------------------

#define the law itself
GRlaw <- function(a,M) {
  output <- (10^{a-M}) / length(magpos)
  return(output)
}

#make a plot of the data with the fitted parameter values
plot(density(magpos), xlim=c(0,4.5),ylim=c(0,1.5), xlab="magnitude", main="")
lines(x, GRlaw(log10(1324)+1, x), col='red') #MoM
legend("right", legend=c(expression("Data", "GR law")), lty=1, col=c("black", "red
    "))

#fitting the tail
x2 <-seq(1.6,4.5,0.01)
plot(density(magpos), xlim=c(1.8,4.5),ylim=c(0,0.4), xlab="magnitude", main="")
lines(x2, GRlaw2(log10(1324)+1, x2), col='red') #MoM
legend("right", legend=c(expression("Data", "GR law")), lty=1, col=c("black", "red
    "))

#---------------------- Gamma distribution ----------------------------------

#finding parameter values

#fitdist
fitdistr(magpos, "gamma")
fita <- 3.9262919 #shape value
fitb <- 3.1612810 #rate value

#MoM
m1 <- mean(magpos)
m2 <- var(magpos) + mean(magpos)^2

gam_a<- m1^2 / (m2 - m1^2)
gam_b <- m1 / (m2 - m1^2)

#make a plot of the data with the fitted parameter values
plot(density(magpos), ylim=c(0,1), xlab="magnitude", main="")
lines(x, dgamma(x, gam_a, gam_b), col='red') #MoM
lines(x, dgamma(x, fita, fitb), col='blue') #fitdist
legend("right", legend=c(expression("Data", "MoM", "fitdistr")), lty=1, col=c("
    black", "red", "blue"))

#plot of tail
  x2 <-seq(1.6,4.5,0.01)
  plot(density(magpos), xlim=c(1.8,4.5),ylim=c(0,0.5), xlab="magnitude", main="")
```

```
  lines(x2, dgamma(x2, gam_a, gam_b), col='red') #MoM
  lines(x2, dgamma(x2, fita, fitb), col='blue')  #fitdist
925 legend("right", legend=c(expression("Data", "MoM", "fitdistr")), lty=1, col=c("
      black", "red", "blue"))

#Anderson-Darling test on data
ad.test(magpos, pgamma, fita, fitb)
930 ad.test(magpos, pgamma, gam_a, gam_b)

#Anderson-Darling test on smaller data set
sample <- sample(magpos, 100)
935 ad.test(sample, pgamma, fita, fitb)
```

Appendix C

Program from Clauset and Shalizi (2009)

```
940 ##############################################################################
    #
    # library
    #
    library(VGAM)    # zeta function
945 library(R.matlab) # read matlab file just for test purpose
    #
    ##############################################################################
    #
    # test zone
950 #
    ##############################################################################
    runtest <- function(){
      #read a matlab file
      #x generated by x = randht(10000,'xmin',15,'powerlaw',2.5) with matlab (
955       continuous law)
      x<-readMat(file("x_2.5_15_10000.mat","rb"))$x[,1]
      #plfit matlab:alpha=2.4975,xmin=18.5752,D=0.0062
       plfit(x)
      #plfit R:alpha=2.497485,xmin=18.57524,D=0.006174977
960   #x0 generated by x0 =floor(x) with matlab (big discrete approximation)
      x0<-readMat(file("x0_2.5_15_10000.mat","rb"))$x[,1]
      #plfit matlab:alpha=2.4700,xmin=15,D=0.0056
       plfit(x0)
      #plfit r:alpha=2.47,xmin=15,D=0.00564463
965


    }
    ##############################################################################
    #
970 # PLFIT fits a power-law distributional model to data.
    #
    #    PLFIT(x) estimates x_min and alpha according to the goodness-of-fit
    #    based method described in Clauset, Shalizi, Newman (2007). x is a
    #    vector of observations of some quantity to which we wish to fit the
975 #    power-law distribution p(x) ~ x^-alpha for x >= xmin.
    #    PLFIT automatically detects whether x is composed of real or integer
    #    values, and applies the appropriate method. For discrete data, if
    #    min(x) > 1000, PLFIT uses the continuous approximation, which is
```

```
#    a reliable in this regime.
#
#    The fitting procedure works as follows:
#    1) For each possible choice of x_min, we estimate alpha via the
#       method of maximum likelihood, and calculate the Kolmogorov-Smirnov
#       goodness-of-fit statistic D.
#    2) We then select as our estimate of x_min, the value that gives the
#       minimum value D over all values of x_min.
#
#    Note that this procedure gives no estimate of the uncertainty of the
#    fitted parameters, nor of the validity of the fit.
#
#    Example:
#       x <- (1-runif(10000))^(-1/(2.5-1))
#       plfit(x)
#
#
# Version 1.0   (2008 February)
# Version 1.1   (2008 February)
#    - correction : division by zero if limit >= max(x) because the unique R
    function do no sort
#                   and the matlab function do...
# Version 1.1   (minor correction 2009 August)
#    - correction : lines 230 zdiff calcul was wrong when xmin=0 (thanks to Naoki
    Masuda)
#    - gpl version updated to v3.0 (asked by Felipe Ortega)
# Version 1.2 (2011 August)
#    - correction for method "limit" thanks to David R. Pugh
#      xmins <- xmins[xmins<=limit] is now xmins <- xmins[xmins>=limit]
#    - "fixed" method added for xmins from David R. Pugh
#    - modifications by Alan Di Vittorio:
#       - correction : zdiff calculation was wrong when xmin==1
#       - the previous zdiff correction was incorrect
#       - correction : x has to have at least two unique values
#       - additional discrete x input test : discrete x cannot contain the value 0
#       - added option to truncate continuous xmin search when # of obs gets small
#
# Copyright (C) 2008,2011 Laurent Dubroca laurent.dubroca_at_gmail.com
# (Stazione Zoologica Anton Dohrn, Napoli, Italy)
# Distributed under GPL 3.0
# http://www.gnu.org/copyleft/gpl.html
# PLFIT comes with ABSOLUTELY NO WARRANTY
# Matlab to R translation based on the original code of Aaron Clauset (Santa Fe
    Institute)
# Source: http://www.santafe.edu/~aaronc/powerlaws/
#
# Notes:
#
# 1. In order to implement the integer-based methods in Matlab, the numeric
#    maximization of the log-likelihood function was used. This requires
#    that we specify the range of scaling parameters considered. We set
#    this range to be seq(1.5,3.5,0.01) by default. This vector can be
#    set by the user like so,
#
#       a <- plfit(x,"range",seq(1.001,5,0.001))
#
# 2. PLFIT can be told to limit the range of values considered as estimates
#    for xmin in two ways. First, it can be instructed to sample these
#    possible values like so,
#
#       a <- plfit(x,"sample",100)
#
```

```
#     which uses 100 uniformly distributed values on the sorted list of
#     unique values in the data set. Alternatively, it can simply omit all
#     candidates below a hard limit, like so
#
#        a <- plfit(x,"limit",3.4)
#
#     In the case of discrete data, it rounds the limit to the nearest
#     integer.
#
#      Finally, if you wish to force the threshold parameter to take a specific
    value
#      (useful for bootstrapping), simply call plfit() like so
#
#        a <- plfit(x,"fixed",3.5)
#
# 3. When the input sample size is small (e.g., < 100), the estimator is
#    known to be slightly biased (toward larger values of alpha). To
#    explicitly use an experimental finite-size correction, call PLFIT like
#    so
#
#        a <- plfit(x,finite=TRUE)
#
# 4. For continuous data, PLFIT can return erroneously large estimates of
#    alpha when xmin is so large that the number of obs x >= xmin is very
#    small. To prevent this, we can truncate the search over xmin values
#    before the finite-size bias becomes significant by calling PLFIT as
#
#        a = plfit(x,nosmall=TRUE);
#
#    which skips values xmin with finite size bias > 0.1.
#
##########################################################################
plfit<-function(x=rpareto(1000,10,2.5),method="limit",value=c(),finite=FALSE,
    nowarn=FALSE,nosmall=FALSE){
  #init method value to NULL
  vec <- c() ; sampl <- c() ; limit <- c(); fixed <- c()
  ##########################################################################
  #
  #  test and trap for bad input
  #
  switch(method,
         range = vec <- value,
         sample = sampl <- value,
         limit = limit <- value,
         fixed = fixed <- value,
         argok <- 0)

  if(exists("argok")){stop("(plfit) Unrecognized method")}

  if( !is.null(vec) && (!is.vector(vec) || min(vec)<=1 || length(vec)<=1) ){
    print(paste("(plfit) Error: ''range'' argument must contain a vector > 1;
        using default."))
    vec <- c()
  }
  if( !is.null(sampl) && ( !(sampl==floor(sampl)) ||  length(sampl)>1 || sampl<2 )
      ){
    print(paste("(plfit) Error: ''sample'' argument must be a positive integer >
        2; using default."))
    sample <- c()
  }
  if( !is.null(limit) && (length(limit)>1 || limit<1) ){
```

```
    print(paste("(plfit) Error: ''limit'' argument must be a positive >=1; using
        default."))
    limit <- c()
  }
  if( !is.null(fixed) && (length(fixed)>1 || fixed<=0) ){
    print(paste("(plfit) Error: ''fixed'' argument must be a positive >0; using
        default."))
    fixed <- c()
  }

  #  select method (discrete or continuous) for fitting and test if x is a vector
  fdattype<-"unknow"
  if( is.vector(x,"numeric") ){ fdattype<-"real" }
  if( all(x==floor(x)) && is.vector(x) ){ fdattype<-"integer" }
  if( all(x==floor(x)) && min(x) > 1000 && length(x) > 100 ){ fdattype <- "real" }
  if( fdattype=="unknow" ){ stop("(plfit) Error: x must contain only reals or only
      integers.") }

  #
  #  end test and trap for bad input
  #
  ###########################################################################

  ###########################################################################
  #
  #  estimate xmin and alpha in the continuous case
  #
  if( fdattype=="real" ){

    xmins <- sort(unique(x))
    xmins <- xmins[-length(xmins)]

    if( !is.null(limit) ){
      xmins <- xmins[xmins>=limit]
    }
    if( !is.null(fixed) ){
      xmins <- fixed
    }
    if( !is.null(sampl) ){
      xmins <- xmins[unique(round(seq(1,length(xmins),length.out=sampl)))]
    }

    dat <- rep(0,length(xmins))
    z   <- sort(x)

    for( xm in 1:length(xmins) ){
      xmin <- xmins[xm]
      z    <- z[z>=xmin]
      n    <- length(z)
      # estimate alpha using direct MLE
      a    <- n/sum(log(z/xmin))
      # truncate search if nosmall is selected
      if( nosmall ){
        if((a-1)/sqrt(n) > 0.1){
          dat <- dat[1:(xm-1)]
          print(paste("(plfit) Warning : xmin search truncated beyond",xmins[xm
              -1]))
          break
        }
      }
      # compute KS statistic
      cx   <- c(0:(n-1))/n #L CDF for the observations with value at least xmin
```

```
      cf   <- 1-(xmin/z)^a #L CDF for the power law model that best fits the data
          in region x >= xmin
      dat[xm] <- max(abs(cf-cx))
    }

    D     <- min(dat)
    xmin  <- xmins[min(which(dat<=D))]
    z     <- x[x>=xmin]
    n     <- length(z)
    alpha <- 1 + n/sum(log(z/xmin))

    if( finite ){
      alpha <- alpha*(n-1)/n+1/n # finite-size correction
    }
    if( n<50 && !finite && !nowarn){
      print("(plfit) Warning : finite-size bias may be present")
    }

  }
  #
  #  end continuous case
  #
  ###########################################################################

  ###########################################################################
  #
  #  estimate xmin and alpha in the discrete case
  #
#   if( fdattype=="integer" ){
#
#     if( is.null(vec) ){ vec<-seq(1.5,3.5,.01) } # covers range of most practical
     scaling parameters
#     zvec <- zeta(vec)
#
#     xmins <- sort(unique(x))
#     xmins <- xmins[-length(xmins)]
#
#     if( !is.null(limit) ){
#       limit <- round(limit)
#       xmins <- xmins[xmins>=limit]
#     }
#
#     if( !is.null(fixed) ){
#       xmins <- fixed
#     }
#
#     if( !is.null(sampl) ){
#       xmins <- xmins[unique(round(seq(1,length(xmins),length.out=sampl)))]
#     }
#
#     if( is.null(xmins) || length(xmins) < 2){
#       stop("(plfit) error: x must contain at least two unique values.")
#     }
#
#     if(length(which(xmins==0) > 0)){
#       stop("(plfit) error: x must not contain the value 0.")
#     }
#
#     xmax <- max(x)
#     dat <- matrix(0,nrow=length(xmins),ncol=2)
#     z <- x
#     for( xm in 1:length(xmins) ){
```

```
#       xmin <- xmins[xm]
#       z    <- z[z>=xmin]
#       n    <- length(z)
#       # estimate alpha via direct maximization of likelihood function
#       #  vectorized version of numerical calculation
#       # matlab: zdiff = sum( repmat((1:xmin-1)',1,length(vec)).^-repmat(vec,xmin
    -1,1) ,1);
#       if(xmin==1){
#         zdiff <- rep(0,length(vec))
#       }else{
#         zdiff <- apply(rep(t(1:(xmin-1)),length(vec))^-t(kronecker(t(array(1,
    xmin-1)),vec)),2,sum)
#       }
#       # matlab: L = -vec.*sum(log(z)) - n.*log(zvec - zdiff);
#       L <- -vec*sum(log(z)) - n*log(zvec - zdiff);
#       I <- which.max(L)
#       # compute KS statistic
#       fit <- cumsum((((xmin:xmax)^-vec[I])) / (zvec[I] - sum((1:(xmin-1))^-vec[I
    ])))
#       cdi <- cumsum(hist(z,c(min(z)-1,(xmin+.5):xmax,max(z)+1),plot=FALSE)$
    counts/n)
#       dat[xm,] <- c(max(abs( fit - cdi )),vec[I])
#     }
#     D     <- min(dat[,1])
#     I     <- which.min(dat[,1])
#     xmin  <- xmins[I]
#     n     <- sum(x>=xmin)
#     alpha <- dat[I,2]
#
#     if( finite ){
#       alpha <- alpha*(n-1)/n+1/n # finite-size correction
#     }
#     if( n<50 && !finite && !nowarn){
#       print("(plfit) Warning : finite-size bias may be present")
#     }
#
#   }
  #
  #  end discrete case
  #
  ###########################################################################

  #  return xmin, alpha and D in a list
  return(list(xmin=xmin,alpha=alpha,D=D))
}

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

Appendix D

Generating dataset of maximum earthquake strengths

```
#read the excel data
library(readxl)
library(actuar)
KNMI_Groningen <- read_excel("~/Vakken/BEP/Data/KNMI Groningen.xlsx",col_types = c
    ("numeric", "numeric", "text", "numeric", "numeric", "numeric", "numeric", "
    text"))

mag <- KNMI_Groningen$MAG

#remove all non positive values
magpos <- mag[-which(mag <= 0)]

#generate a maximum value set, based on dividing the data in parts and taking the
    maximum of them.
maxgen <- function(data, nPart) {
  shuffleddata <- sample(data)
  n <- length(data)

  #define the borders of the parts
  setcount <- round((0:floor(nPart))*n/floor(nPart))
  magmaxset <- c()
  setcountmin <- setcount[-c(length(setcount))]

  #check each part for its maximum and store that value
  for(i in 1:length(setcountmin)){
    maximum <- max(shuffleddata[(setcount[i]+1):(setcount[i+1])])
    magmaxset <- c(magmaxset, maximum)
  }

  return(magset)
}
```

Appendix E

Data analysis on the maximum of a set of earthquake strengths

```
#read all used libraries
library(readxl)
library(actuar)
library(e1071)
library(Renext)
library(MASS)
library(gumbel)
library(evir)
library(QRM)
library(evd)

#read the excel data
KNMI_Groningen <- read_excel("~/Vakken/BEP/Data/KNMI Groningen.xlsx",col_types = c
    ("numeric", "numeric", "text", "numeric", "numeric", "numeric", "numeric", "
    text"))
mag <- KNMI_Groningen$MAG

#remove all non positive values
magpos <- mag[-which(mag <= 0)]

#generate a maximum value set, based on dividing the data in parts and taking the
    maximum of them.
maxgen <- function(data, nPart) {
  shuffleddata <- sample(data)
  n <- length(data)

  #define the borders of the parts
  setcount <- round((0:floor(nPart))*n/floor(nPart))
  magmaxset <- c()
  setcountmin <- setcount[-c(length(setcount))]

  #check each part for its maximum and store that value
  for(i in 1:length(setcountmin)){
    maximum <- max(shuffleddata[(setcount[i]+1):(setcount[i+1])])
    magmaxset <- c(magmaxset, maximum)
  }

  return(magset)
}
```

```
#make the output
maxmag <- maxgen(magpos, sqrt(n))
maxmag2 <- maxgen(magpos, sqrt(n))

#plot of the generated data
plot(density(maxmag), main="", xlab="magnitude")

#---------------------------- Frechet distribution ----------------------------------

#calculation of the skewness of the dataset
skewness(maxmag)

#make a plot of the data with the fitted parameter values
plot(density(maxmag), xlim=c(0, 5), ylim=c(0, 1.1), main="", xlab="magnitude")
lines(y, dfrechet(y, 0, 1, 3.58), col='red') # Pareto parameters
lines(y, dfrechet(y, 1, 1.6, 3.58), col='blue') # Trial&Error
legend("right", legend=c(expression("Data", "Pareto par.", "Trial & Error")), lty
    =1, col=c("black", "red", "blue"))

#---------------------------- Gumbel distribution ----------------------------------

#finding parameter values
gumplace <- gumbel(maxmag)$par.ests["mu"]
gumscale <- gumbel(maxmag)$par.ests["sigma"]

#make a plot of the data with the fitted parameter values
plot(density(maxmag), ylim=c(0, 1.2), main="", xlab="magnitude")
lines(y, dGumbel(y, gumplace, gumscale), col='red') # MLE
legend("right", legend=c(expression("Data", "MLE")), lty=1, col=c("black", "red"))

#---------------------- Weibull distribution ----------------------------------

#Parameter fitting though Weibull MLE fitting
weishape <- fweibull(maxmag)$estimate["shape"]
weiscale <- fweibull(maxmag)$estimate["scale"]

#make a plot of the data with the fitted parameter values
y <- seq(1,6, 0.01)
plot(density(maxmag), ylim=c(0, 1.1), main="", xlab="magnitude")
lines(y, dweibull(y, weishape, weiscale), col='red') #fitdistr
legend("right", legend=c(expression("Data", "MLE")), lty=1, col=c("black", "red"))

#Weibull plot
weibplot(maxmag, shape = weishape, scale = weiscale)
```