An extreme value approach to modelling number of causalities in earthquakes

Henrik Steneld

Supervisor: Nader Tajvidi

January 2021

LUND
UNIVERSITY

Abstract

Earthquakes occur around the globe all the time. Most are weak enough to just pass by, some are strong enough to be felt by us humans, and some very few are completely devastating. A comprehensive database distributed by *NOAA National Centers for Environmental Information* provides a means for reviewing devastating earthquakes over the past. Extreme value theory has previously been applied to modelling earthquakes, although for the most part the modelling has been concerned with the magnitudes. In this thesis, extreme value theory has been applied to the number of casualties that are directly or indirectly the result of an earthquake.

An in-homogeneous Poisson point process is fitted to events where the death toll is at least ten or more. The events are assumed to be independent but non-stationary with respect to the magnitude of the earthquake. This leads to a Poisson point process with an intensity which is a function of magnitude. In addition, an assumption is made about the distribution of the magnitudes of earthquakes, which provides the necessary means for modelling extremes of earthquakes death toll unconditional of magnitude. With the aid of simulations and the asymptotic normality of maximum likelihood estimators, return levels and corresponding confidence intervals are calculated for three different geographical regions.

Acknowledgement

I would like to thank my supervisor Nader Tajvidi for all the help given throughout the process, it has been incredibly valuable. A special appreciation for the great patience you have had.

Contents

1 Introduction

There have been a number of papers published dealing with Extreme value theory applied to earthquakes. For example, back in 1945, John M. Nordquist published *Theory of largest values applied to earthquake magnitudes* [6]. However, while a lot of statistical research has been done on earthquakes and the effect that earthquakes have, there does not seem to be a lot of cases where extreme value theory has been applied to the resulting death toll. Therefore, the purpose of this thesis is to investigate whether univariate extreme value theory may be applied in a satisfactory manner to death tolls arising from earthquakes. If an approximate model for the number of deaths as an outcome of earthquakes can be found, return levels (quantiles) may be calculated and may perhaps so be done for different parts of the world, even those where devastating earthquakes are very rare. It is expected that the magnitude of an earthquake will play a significant role in the death toll. However, just like the death toll, the magnitude of an earthquake can be assumed random, which opens up for further investigation. Especially, if extreme earthquake events (with regards to death toll) can be modeled as a Poisson point process conditioned on the magnitude, extreme earthquake events might be modeled unconditionally and in a parametric manner given that the magnitude can be assumed to belong to some parametric family of distributions. The resulting unconditional point process model will then be a doubly stochastic Poisson process, also known as a Cox process [3, ch. 8, p. 265].

1.1 Objective

The objective is to find a suitable estimate for an approximate Poisson point process conditioned on any significant covariates, such as magnitude. With such a model, quantiles may be estimated, yielding return levels for varying values of the covariates. Furthermore, an unconditional approximate Point process is sought to be estimated. Such a point process might be more valuable in regards to explaining return levels, as those will be regardless of the values of the (perhaps random) covariates.

2 Univariate Extreme value theory

Extreme value theory might fundamentally be explained as the modelling of the maximum of some sequence of random variables. One could in some scenarios be interested in drawing conclusions from the random variable

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

where $\{X_i\}_{i=1,...,n}$ is a sequence of independent random variables, typically with an unknown but common distribution.

2.1 Extreme value distributions

Given the assumption in the previous section, that $\{X_i\}_{i=1,...,n}$ is a sequence of independent random variables with common distribution function F, it is quite straight forward to express the distribution of the maximum M_n as a function of F.

$$P(M_n \leq z) = P(X_1 \leq z, ..., X_n \leq z) = P(X_1 \leq z) \cdot ... \cdot P(X_n \leq z) = F(z)^n$$

Typical problems in practices involve F being unknown, meaning that the above expression for the distribution of M_n provides little to no help. An alternative way of gaining knowledge about the distribution of M_n could be that of investigating the approximation of F^n as $n \to \infty$. An arising problem is then that for any $z < z_+$, with z_+ being the upper endpoint of F, $F(z)^n$ will tend to zero as $n \to \infty$. Meaning that one can't find an expression for the distribution of M_n that isn't a degeneration onto z_+. It is however sometimes possible to express a function M_n^* of M_n in such a way that M_n^* tend to some non-degenerate distribution as $n \to \infty$. Let

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

with $\{a_i > 0\}_{i=1,...,n}$ and $\{b_i\}_{i=1,...,n}$ being a sequence of constants. In this setting, a central theorem in the extreme value theory follows.

Theorem 2.1 (Fisher–Tippett–Gnedenko theorem) *([1, ch. 3, p. 46]) If there exists sequences of constants $\{a_n > 0\}_{i=1,...,n}$ and $\{b_n\}_{i=1,...,n}$ such that*

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

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

$$\textit{Gumbel}: G(z) = exp\left(-exp\left(-\frac{z-b}{a}\right)\right), -\infty < z < \infty;$$

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

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

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

If one were to research the distribution of M_n in this setting, in particular carry out inference, one would have to decide upon which of the three families of distributions to assume. This could be an inconvenience as it might not always be a clear choice, especially as any subsequent inference would be under the original assumption that the correct choice of family was made.

2.2 Generalized Extreme Value Distribution

A way of conquering the problem of uncertainty in regards to which limiting distribution to decide upon could be to instead view the three distributions as one, the Generalized Extreme Value distribution (GEV).

Theorem 2.2 *([1, ch. 3, p. 48])*
If there exists sequences of constants $\{a_n > 0\}_{i=1,...,n}$ *and* $\{b_n\}_{i=1,...,n}$ *such that*

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

for a non-degenerate distribution G*, then* G *is a member of the GEV family*

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

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$*, where* $-\infty < \mu < \infty, \sigma > 0$ *and* $-\infty < \xi < \infty$*.*

If the random variable

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

is distributed according to (1) with parameters μ, σ, ξ, we say that

$$M_n^* \sim GEV(\mu, \sigma, \xi)$$

2.3 Block Maxima

In practice, it is often of interest to model the maximum over some set time frame, e.g the maximum annual event. The maximum over some decided upon time frame can then in many cases be approximated by its asymptotic distribution, that is GEV. The choice of block size becomes important in the sense of evaluating whether or not there is enough time in each block to assume asymptotic behavior, as well as if there are enough blocks to have a decent sample variance.

2.4 Peaks over threshold

Sometimes, since several extreme events might occur relatively close in time, while not in others, approaching the problem of inference with block maxima might mean that important data gets discarded. An alternative approach, that builds upon the theory of a maximums asymptotic tendency towards GEV, is peaks over threshold. A theorem, encapsulating the approach follows,

Theorem 2.3 *([1, ch. 4, p. 75])*
Let $\{X_i\}_{i=1,...,n}$ *be a sequence of independent random variables with common distribution function* F*, and let*

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

Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies Theorem 2.2, so that for large n,

$$P((M_n \leq z) \approx G(z),$$

where

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

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

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

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

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

The family of distributions defined by $H(y)$ is called the generalized Pareto family.

What theorem 2.3 says is that we can approximate the distribution of the excess $(X-u)$, for all $X > u$, and we may include data from observations where $x_i > u$ even if x_i lies close (in time) to another observation x_j that is included and satisfies $x_j > u$.

2.5 Non-stationary sequences

An assumption that was made in the introduction and in section 2.1 was that of a common distribution function F, which would imply that a stationary process laid as foundation for the maximum M_n. It is however not always the case that observations can be assumed to come from a stationary process. Non-stationarity can be present in time, or in terms of other variables. Assume that the underlying stochastic process $\{X_i\}_{i=1,\ldots,n}$ is non-stationary in the variable $t \in T$ (typically time) such that also the maximum is non-stationary in t. Then it is typically possible to model the maximum Z_n as

$$Z_n(t) \sim GEV(\mu(t), \sigma(t), \xi(t)),$$

where the parameters are for example linear functions of t:

$$\mu(t) = \beta_0 + \beta_1 t$$

or other functions, such as:

$$\sigma(t) = exp(\beta_0 + \beta_1 t),$$

which is generally suitable for σ to ensure that $\sigma > 0, \ \forall t \in T$.

2.6 Return level

Return levels z_p are quantiles associated with return periods $1/p$, in such a way that if G is the distribution function for yearly maximum, then z_p will be the level that is exceeded on average once every $1/p$ year, and is given by the equation $G(z_p) = 1 - p$.

2.7 Goodness Of Fit (GOF)

Model diagnostics is an essential part of fitting data to a model, i.e. doing inference. This section focus on the creation of diagnostic plots that may visually be interpreted as a mean of determining the goodness of fit. Of special interest are qq-plots for non-stationary models. Such plots can be created after some transformations. Given that a model can be estimated as

$$Z_t \sim \text{GEV}(\mu(t), \sigma(t), \xi(t)),$$

one can make a transformation such that

$$\tilde{Z}_t = \frac{1}{\xi(t)} log\left(1 + \xi(t)\left(\frac{Z_t - \mu(t)}{\sigma(t)}\right)\right),$$

with $\tilde{Z}_t$ having the standard Gumbel distribution. Then, with known standard distribution, producing a qq-plot can be done by plotting the pairs

$$\{\tilde{z}_{(i)}, -log(-log(i/(m+1)))); \ i = 1, ..., m\},$$

given that $\tilde{z}_{(1)}, ..., \tilde{z}_{(m)}$ is the ordered values of observed $\tilde{z}_t$. If instead a model has been estimated such that $Y_t \sim \text{GP}(\sigma(t), \xi(t)$, with GP being the generalized Pareto distribution, then the transformation

$$\tilde{Y_{t_k}} = \frac{1}{\xi(t)} log\left(1 + \xi(t)\left(\frac{Y_{t_k} - u_t}{\sigma(t)}\right)\right),$$

implies that $\tilde{Y_{t_k}}$ will have the standard exponential distribution. If $\tilde{y}_{(1)}, ..., \tilde{y}_{(k)}$ are the ordered observations of $\tilde{Y_{t_k}}$, then a qq-plot can be produced by the pairs

$$\{\tilde{y}_{(i)}, -log(1 - i/(k+1))); \ i = 1, ..., k\}.$$

3 Generalized Likelihood Ratio (GLR)

A powerful tool in testing nested models is the Generalized Likelihood Ratio Test. It is encapsulated with the following two theorems.

Theorem 3.1 *([1, ch. 2, p. 35])*
Let $x_1, ..., x_n$ be independent realizations from a distribution within a parametric family $\mathcal{F}$, and let $\hat{\theta}_0$ denote the maximum likelihood estimator of the d-dimensional model parameter $\theta_0 = (\theta^{(1)}, \theta^{(2)})$, where $\theta^{(1)}$ is a k-dimensional subset of θ_0. Then, under suitable regularity conditions, for large n

$$D_p(\theta^{(1)}) = 2\{\ell(\hat{\theta}_0) - \max_{\theta^{(2)}} \ell(\theta^{(1)}, \theta^{(2)})\} \sim \chi^2_k.$$

Theorem 3.1 may be exploited in such a way that schemes for trying out nested models can be laid up,

Theorem 3.2 *([1, ch. 2, p. 35])*
Suppose $\mathcal{M}_0$ with parameter $\theta^{(2)}$ is the sub-model of $\mathcal{M}_1$ with parameter $\theta_0 = (\theta^{(1)}, \theta^{(2)})$ under the constraint that the k-dimensional sub-vector $\theta^{(1)} = \mathbf{0}$. Let $\ell_0(\mathcal{M}_0)$ and $\ell_1(\mathcal{M}_1)$ be the maximized values of the log-likelihood for models $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively. A test if the validity of model $\mathcal{M}_0$ relative to $\mathcal{M}_1$ at the α level of significance is to reject $\mathcal{M}_0$ in favor of $\mathcal{M}_1$ if $D = 2(\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)) > c_\alpha$, where c_α is the $(1-\alpha)$ quantile of the χ^2_k distribution.

4 Point Process

A point process over a set $S \subset \mathbb{R}^d$ is a rule for the occurrence and position of point events. If $S = \mathbb{R}_+$, then a point process could for example model the occurrence of significant earthquakes in time. The properties of a point process can be characterized by a set of non-negative integer-valued random variables, $N(A)$, for every $A \subset S$, such that $N(A)$ describes the number of points in A. In such a way, the point process, N, is characterized by the distributions $F_{N(A)}$ of each $N(A)$. One measure that is usually of importance is the intensity measure $\Lambda(A)$, which describes the expected number of points in A. Thus,

$$\Lambda(A) = E(N(A))$$

.

4.1 Poisson point process

Perhaps the most central of point processes are those that can be defined as a Poisson point process. A Poisson point process is such that

$$N(A) \sim \text{Poi}(\Lambda(A))$$

and if $A,\ B$ are disjoint, then $N(A)$ and $N(B)$ are independent.

4.2 Application to Extreme Value Theory

Assume that $\{X_i\}_{i=1,...,n}$ is a sequence of independent and identically distributed random variables. Then if one define

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

one have that for certain regions, N_n is approximately a Poisson process. More precisely, with the above setting one have the following theorem,

Theorem 4.1 *([1, ch. 7, p. 133])*
For sufficiently large u, on regions of the form $(0,1) \times [u, \infty)$, N_n is approximately a Poisson process, with intensity measure on $A = [t_1, t_2] \times (z, \infty)$ given by

$$\Lambda(A) = (t_1, t_2) \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi}$$

Here, u is seen as a threshold, where events that lie above u are considered extreme in an analogous way as those events above the threshold in the POT approach (see section 2.4). In the case of a Poisson process for extremes, the parameters, μ, σ and ξ are all invariant of the threshold [1, ch. 7, p. 136]. This, together with the fact that the intensity (thus also the distribution of $N(A)$) being a function constant w.r.t the threshold, makes for a more natural way of modelling non-stationary sequences, as the threshold may easily be non-constant (for example a function of time) without affecting anything other than the bias-variance trade-off.

4.2.1 Cox point process for extremes

Assume that a process N_n,

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

can be approximated as a Poisson point process with intensity measure Λ on $A = [t_1, t_2] \times (u, \infty)$, such that

$$\Lambda(A) = (t_2 - t_1) \left[1 + \xi(y) \left(\frac{u - \mu(y)}{\sigma(y)} \right) \right]^{-1/\xi(y)}.$$

for some large enough u. An interpretation of this would be that the extreme of $\{X_i(y_i)\}$ can be approximated under the GEV distribution with parameters as functions of y, i.e. the extreme is non-stationary in y. With y_i's known, $\Lambda(A)$ is deterministic and the number of points $N_n(A)$ in A can be modeled as an in-homogeneous Poisson point process, $N_n(A) \sim \text{Poi}(\Lambda(A))$. Assume now that y_i are in fact realizations of some random variable Y such that $\Lambda(A)$ is in fact a realization of a random field $\Psi(A)$. Then one might be interested in the unobserved effects of Y, i.e. when the intensity measure of the point process is considered random. Under such assumptions, we have that

$$N_n(A) \mid\ (\Psi(A) = \Lambda(A)) \sim \text{Poi}(\Lambda(A)).$$

This implies that the unobserved process N_n is in fact not a Poisson process, but an instance of a more general process named Cox process or doubly stochastic Poisson process [3, ch. 8. p. 265]. Such a process does not share all the same properties as a Poisson process. $N_n(A)$ will in general not be distributed as a Poisson random variable. For instance

$$\text{Var}(N_n(A)) \geq E(N_n(A)),$$

with equality only if $\Psi(A)$ is almost surely constant [4, ch. 3, p. 8]. However, as the conditional distribution of N_n is known, it is fairly straight forward to establish the intensity measure and the void probabilities of N_n. One have that the intensity measure $\psi(A)$ of $N_n(A)$ is given by

$$\psi(A) = E(\Psi(A))$$

[4, ch. 3, p. 8]. For the void probabilities $P(N_n(A) = 0)$, first note that by the law of the unconscious statistician one have that

$$E(e^{-\Psi(A)}) = \int_{-\infty}^{\infty} e^{(t_2-t_1)\left(1+\xi(y)\left(\frac{u-\mu(y)}{\sigma(y)}\right)\right)^{-1/\xi(y)}} f_Y(y)dy. \tag{2}$$

Secondly, note that the *continuous version of the law of total probability* [3, ch. 3, p. 39) implies that

$$\begin{aligned} P(N_n(A) = 0) &= \int_{-\infty}^{\infty} P((N_n(A) \mid Y = y) = 0) f_Y(y)dy \\ &= \int_{-\infty}^{\infty} e^{-\Lambda(A)} f_Y(y)dy \\ &= \int_{-\infty}^{\infty} e^{(t_2-t_1)\left(1+\xi(y)\left(\frac{u-\mu(y)}{\sigma(y)}\right)\right)^{-1/\xi(y)}} f_Y(y)dy \end{aligned} \tag{3}$$

Thus, (2) and (3) together implies that the void probabilities are given by

$$E(e^{-\Psi(A)}).$$

4.2.2 Special case

A particular case that will demonstrate relevance in the subject of modelling significant earthquakes will be when one can assume that $\xi(y)$ and $\sigma(y)$ are both constant but $\mu(y) = \mu_0 + \mu_1 y$ and y is the realization of a random variable Y, $Y \sim \text{Exp}(\theta)$. Then

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

is a point process such that for regions A of the form $A = [t_1, t_2] \times (u, \infty)$ one have that

$$N(A) \mid (\Psi(A) = \Lambda(A)) \sim \text{Poi}(\Lambda(A))$$

with the intensity measure $\Lambda(A)$ given by

$$\Lambda(A) = (t_2 - t_1)\left(1 + \xi\left(\frac{u - (\mu_0 + \mu_1 y)}{\sigma}\right)\right)^{-1/\xi},$$

and for unobserved y the intensity measure $\psi(A)$ of $N_n(A)$ can be expressed (according to the previous section) as

$$E(\Psi(A)) = \int_0^{\infty} (t_2 - t_1)\left(1 + \xi\left(\frac{u - (\mu_0 + \mu_1 y)}{\sigma}\right)\right)^{-1/\xi} \theta e^{-\theta y} dy.$$

Or equivalently, with the aid of the upper incomplete gamma function $\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}dt$, in the following manner,

$$\psi(A) = (t_2 - t_1)\beta^{-\alpha}e^{\beta z}\Gamma(\alpha+1, \beta z)$$

with change of variables $\alpha = -1/\xi$, $\beta = -\theta/(\xi\mu_1/\sigma)$, and $z = 1 + (\xi/\sigma)u - \xi\mu_0/\sigma - \xi\mu_1/\sigma$. The void probabilities can be obtained for example by numerically solving

$$E(e^{-\Psi(A)}) = \int_0^\infty e^{-(t_2-t_1)\left(1+\xi\left(\frac{u-(\mu_0+\mu_1 y)}{\sigma}\right)\right)^{-1/\xi}}\theta e^{-\theta y}dy.$$

5 Data

Data from the *Significant Earthquake Database* laid as core for the analysis. Data from the *ANSS Comprehensive Earthquake Catalog* was used as a way of complementing in the analysis of unconditional probabilities. The *Significant Earthquake Database* consists of thousands of events that stretches from 2150 BC to present time. However, not all earthquakes during the time is included (earthquakes of low intensity occurs incredibly frequent). Events that meets at least one of the following criteria is included: moderate damage (approximately $1 million or more), 10 or more deaths, magnitude 7.5 or greater, Modified Mercalli Intensity X or greater, or the earthquake generated a tsunami. To analyse whether or not there is a correlation between societal aspects and the number of deaths, data from *World Development Index* was imported. This data set, which take the form as a time series contains a vast amount of information related to countries, such as population density and GDP.

5.1 Data filtering with regards to death toll

An important part of the analysis was that of choosing events from the *Significant Earthquake Database* that could be considered unbiased. Unbiased in the sense that all events included are included on the same premises. Since the primary objective of the analysis was to model the number of deaths from an earthquake, it seemed reasonable to include only events where there are 10+ deaths as those are all included, regardless of whether or not the event satisfied any other criteria. Suppose events where there are less than 10 deaths where to be included, then those event would have to meet one of the other criterion, which would mean that those particular events would be "more extreme" than other events with the same number of deaths (that are excluded). It is also noteworthy that when modelling extremes (such as yearly maximum), one can expect the number of deaths to be high and above 10, meaning that events with sub 10 deaths would have little to no impact on the statistics.

5.2 Transformation of death tolls

The analysis considers the logarithm over the number of deaths rather than just the number of deaths. The motivation for this is twofold. First, the non-transformed data is highly skewed, and outliers seems to have an unreasonable large impact on the modeling (parameter estimation). Second, the number of deaths was assumed dependent on the magnitude of the corresponding earthquake, and the magnitude is logarithmically proportional to the amplitude of the ground motion [15]. Therefore, if a linear trend seems evident in the relation between the logarithm of the number of deaths and the magnitude, then that would suggest a linear trend between the number of deaths and the ground motion.

5.3 Transformation of magnitudes

The magnitudes of the earthquakes were shifted and scaled such that $m^* = 0.4(m-6)/10$, so that the magnitudes m^* had origin at 6 and was scaled such that an earthquake with magnitude 10 was an earthquake with modified magnitude 1. Those below magnitude 6 were dropped. The shifting countered the issue of model instability for low magnitudes. As it was deemed highly unlikely (or perhaps impossible) to have any other death toll than zero for "extreme" earthquakes with magnitude close to zero, it meant that the statistical model in place would have to predict a value for the logarithm of deaths approaching negative infinity. The scaling was done such that any magnitude would fall in the interval $[0, 0.95]$, i.e. close to $[0, 1]$. (The strongest ever earthquake on record is of magnitude 9.5, it happened 1960 in Chile [16]).

5.4 Data with respect to geographic location

The *Significant Earthquake Database* includes information regarding where the earthquakes occurred, in the sense of geographical coordinates, country and region. It was reasonable to assume that earthquakes have varying effects on the number of deaths dependent on location.

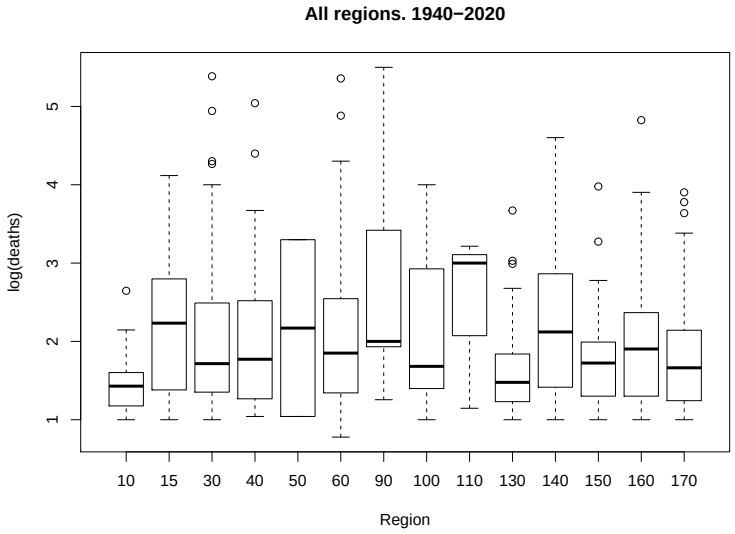

Figure 1: Death toll shown in correlation with location. Region codes are predefined in the *Significant Earthquake Database* and can can be found on the NOAA website.

Although the data set could be separated into subsets for different regions, the analysis was mostly based upon the entirety, primarily as a way of reducing variance. The result suggested that a fine model could be established nevertheless.

5.5 Data with respect to time

Another aspect in the choice of events was the year of which they occurred. It is not obvious (and probably false) to assume that the number of deaths in an event is identically distributed regardless of time. However, viewing the the annual maximum over the last 80 years, Figure 2 shows that any trend might be negligible. It was therefore decided to include data only of events that occurred between 1940-2020 (the last 80 years), thus assuming local stationarity.

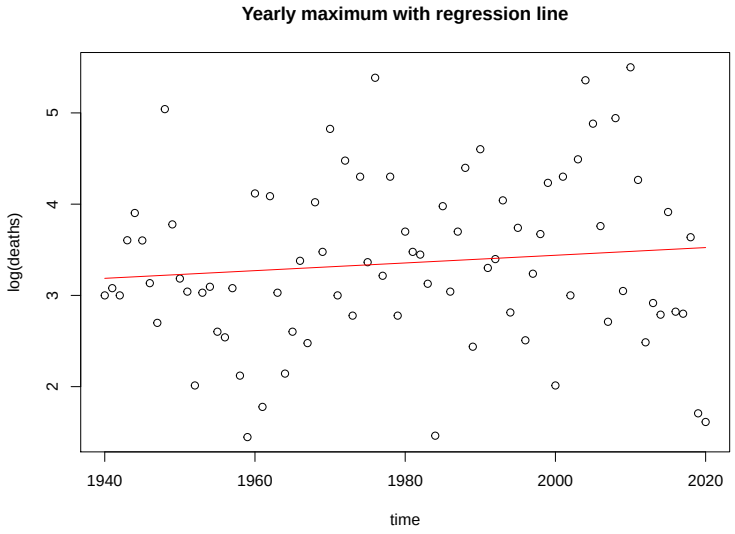

Figure 2: A weak trend can be shown with respect to time.

5.6 Subset for block maxima

A subset of the data was created by selecting the events that was the annual maximum with regards to death toll. Thus, the subset contained 80 events. No smaller (half-year or quarterly) subsets was created as those would not have been complete in the sense of each frame containing at least one event.

6 Software

The data review and subsequent extreme value analysis was implemented via *RStudio* in the programming language *R*. The package *in2extRemes* was used for most of the extreme value analysis, including parameter estimation, creation of GOF plots, GLR testing and data simulation. The package *rcomcat* was used to search information from the *ANSS Comprehensive Earthquake Catalog*. The *wdi* package was used to search and download content from the *World Development Index* database.

7 Formulation of the objective

The objective was to establish a reliable point process that could be used to model the number of deaths that is the outcome of earthquakes with extremely high death tolls. An expected correlation between the number of deaths from

the worst earthquakes and the earthquakes magnitude could at first be strengthened by a visual inspection of the annual maximums, see Figure 3. This suggested that the worst earthquakes should perhaps be modeled with distribution dependent on magnitude. Define,

$$\begin{aligned}\tilde{X}_i(m_i) &:= \text{Logarithm of death toll from a single} \\ &\quad\ \text{earthquake with known magnitude } m_i \in [0,1]. \\ X_i &:= \text{Logarithm of death toll from a single} \\ &\quad\ \text{earthquake with random magnitude.} \\ Z_n(m) &:= max\{\tilde{X}_i(m)\}_{i=1,..,n}.\end{aligned}$$

As a mean of finding a model for the extremes of n X_i's, a first goal was set to establish a point process for the extremes of n $\tilde{X}_i(m_i)$'s. This point process would be of the form suggested in section 4.2, i.e.

$$N_n = \{(i/(n+1),\ \tilde{X}_i(m_i) : i = 1, ..., n\}$$

such that for regions A of the form $A = [t_1, t_2] \times (u, \infty)$, the number of points in A, $N_n(A)$, is Poisson distributed $N_n(A) \sim \text{Poi}(\Lambda(A))$ with

$$\Lambda(A) = (t_2 - t_1)\left(1 + \xi(m)\left(\frac{u - \mu(m)}{\sigma(m)}\right)\right)^{-1/\xi(m)}.$$

. Thereafter, with aid from the theory established in section 4.2.1, a more general point process was hoped to be found for the extremes of X_i. An assumption was made that there occurs n earthquakes annually with magnitude larger than six, i.e. a fixed amount of earthquakes.

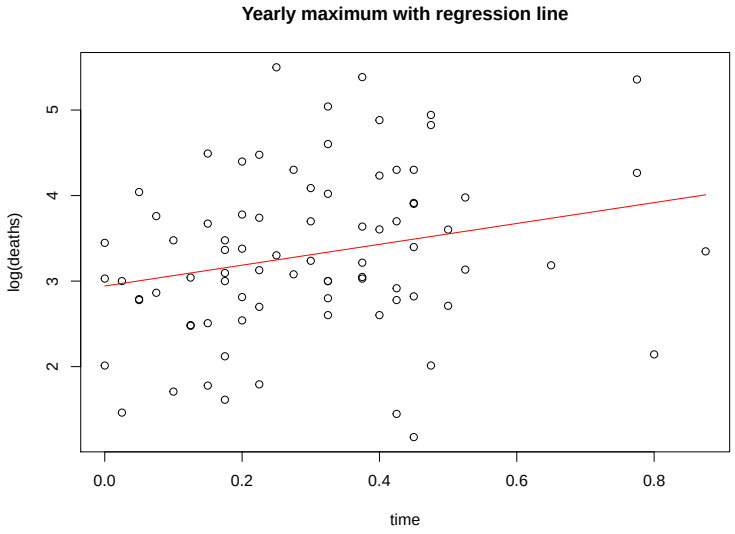

Figure 3: Yearly maximum with respect to magnitude.

8 Parameter functions for the process $\{\tilde{X}_i(m_i)\}_n$

Let $\boldsymbol{\theta}(m) = (\mu(m), \sigma(m), \xi(m))$. Due to the link between the parameters given by modeling $Z_n(m)$ as

$$Z_n(m) \sim \text{GEV}(\boldsymbol{\theta}(m))$$

and the parameters in the intensity measure of a point process (such as suggested in section 4.2), it made sense to start with the simpler model, block maxima, to determine the basic construction of the function $\boldsymbol{\theta}(m)$. The approach for finding suitable functions was to try nested models. Especially, performing formal generalized likelihood ratio (GLR) tests in order to determine significant parameters. Goodness of fit (GoF) plots were reviewed for each model. Evaluation of each model was then based on both the GLR test and the GoF plots with the principle of parsimony in mind. Five models were evaluated,

$$\begin{aligned}
\mathcal{M}_1 :\ \boldsymbol{\theta}(m) &= (\mu, \sigma, \xi) \\
\mathcal{M}_2 :\ \boldsymbol{\theta}(m) &= (\mu_0 + \mu_1 m, \sigma, \xi) \\
\mathcal{M}_3 :\ \boldsymbol{\theta}(m) &= (\mu_0 + \mu_1 m + \mu_2 m^2, \sigma, \xi) \\
\mathcal{M}_4 :\ \boldsymbol{\theta}(m) &= (\mu_0 + \mu_1 m + \mu_2 e^m, \sigma, \xi) \\
\mathcal{M}_5 :\ \boldsymbol{\theta}(m) &= (\mu_0 + \mu_1 m, exp(\sigma_0 + \sigma_1 m), \xi)
\end{aligned}$$

The evaluation for each model was made by fitting a GEV distribution to

the annual maximum. The corresponding Log-Likelihood values were,

$$\begin{aligned}\mathcal{M}_1 &: -106.906\\ \mathcal{M}_2 &: -104.028\\ \mathcal{M}_3 &: -103.603\\ \mathcal{M}_4 &: -103.590\\ \mathcal{M}_5 &: -102.950\end{aligned}$$

Thus, the deviance statistic $D = 2(\ell(\mathcal{M}_2) - \ell(\mathcal{M}_1)$ was realized to $d = 5.756$, which implied that the hypothesis that $\mathcal{M}_1$ is a plausible reduction of $\mathcal{M}_2$ could be rejected with p-value $p = 0.0164$, i.e significance level lower than 0.05. Rejection on level 0.05 could not be done for any of the more complex models $\mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5$ against $\mathcal{M}_2$. GoF plots were reviewed for $\mathcal{M}_1$ and $\mathcal{M}_2$ (see Figure 4 and 5) and they did not seem to speak in any great disfavor for $\mathcal{M}_2$ even though it could be noted that two observations were a bit off line for $\mathcal{M}_2$. Thus it was said that, out of the five models considered, model $\mathcal{M}_2$ provided the best description of $Z_n(m)$ and would be the one used for further analysis.

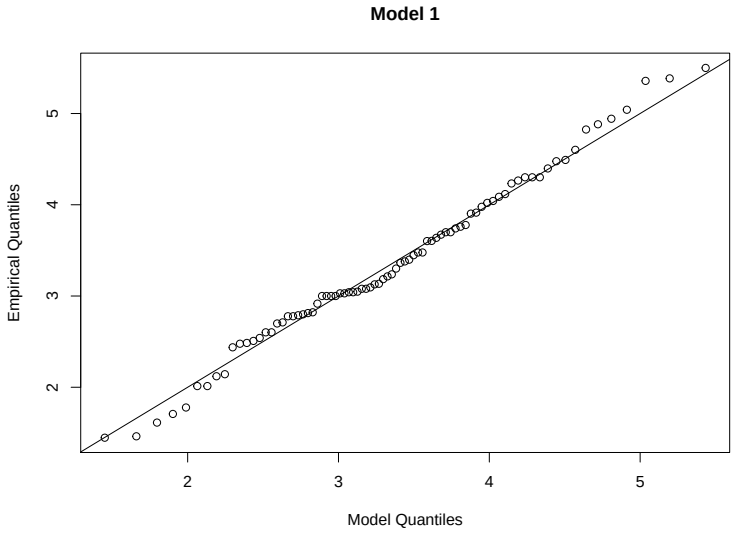

Figure 4: qq-plot for model 1.

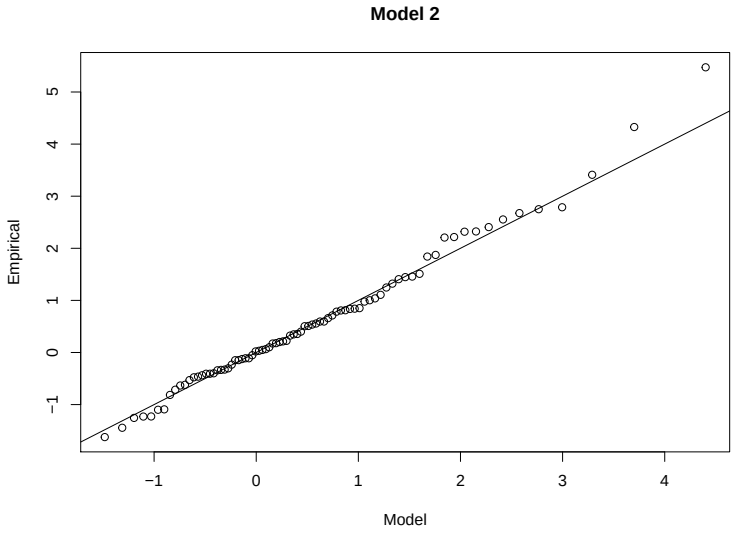

Figure 5: qq-plot for model 2.

8.1 Other possible covariates

Several functions for the parameter was tried with generalized likelihood ratio tests against $\mathcal{M}_1$ in addition to those already suggested. Especially with regards to year of occurrence linear in μ ($p = 0.922$), population density in the country of occurrence linear in μ ($p = 0.132$), and GDP per capita in current \$ linear in μ ($p = 0.264$). Assumptions that any of these could be significant was based on information from World Health Organization (WHO) [12] and Our World in Data [13]. However, none were determined to be an acceptable extension of the constant model $\mathcal{M}_1$.

9 Model Construction for the process $\{\tilde{X}_i(m_i)\}_n$

This section concerns the method for estimating the parameters in the intensity measure of the Poisson point process. Although parameters for the Poisson point process could be estimated using block maxima or POT, it was decided to first and foremost fit a point processes. The strategy to fit a point process (as well as would have been with POT) is beneficial over fitting a block maxima in the sense that some data which could very well contribute to parameter estimation might be spared. Consider that $X_i(m)$ depends on magnitude m, then what might be considered an unusual high value for $X_i(m)$ will depend on m. This suggests a non-constant threshold over which observations are considered into the parameter estimation.

As the function for $\boldsymbol{\theta}(m)$ was decided (section 8) to be linear in μ and constant for $\sigma,\ \xi$, i.e. $\boldsymbol{\theta}(m) = (\mu_0 + \mu_1 m,\ \sigma,\ \xi)$, the aim for the parameter estimation was to estimate $\mu_0,\ \mu_1,\ \sigma$ and ξ.

9.1 Threshold Selection

As the parameterization of a point process is invariant of the threshold, the selection of a threshold function would only affect the bias-variance trade-off. Earthquakes with high magnitude can be expected to cause a significantly higher death toll compared to a low magnitude earthquake, it was therefore sensible to have a threshold that increased with magnitude. One approach that was tried was to view the empirical quantiles: $q_{0.9}(m)$, and base the threshold function on those, as to approximately keep an uniform rate of exceedances. However, since only data from events with 10 or more deaths are included, the empirical quantiles produced could be assumed higher than the ones that would have been obtained if data from all earthquakes was included in the data set. In essence, Figure 6 depicts what could be assumed to be some estimated upper bound for $q_{0.9}(m)$. The threshold function was then being selected by trial-and-error, with the quantiles as reference and with GoF plots used for comparing thresholds of the same shape. A function that seemed sensible was found and is depicted in Figure 7.

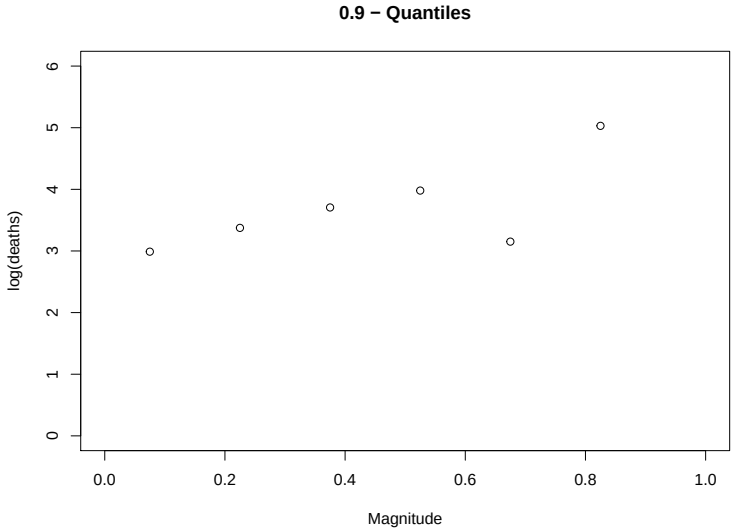

Figure 6: 90% quantiles for the magnitudes of earthquakes.

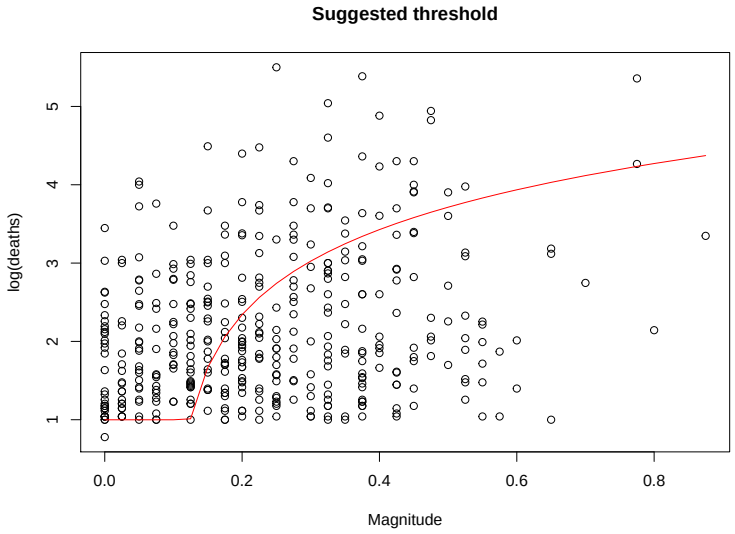

Figure 7: A function for the threshold.

9.2 Parameter Estimation

With a threshold selected, the parameters μ_0, μ_1, σ and ξ was numerically estimated as the ML-estimators $\hat{\mu_0}$, $\hat{\mu_1}$, $\hat{\sigma}$ and $\hat{\xi}$. Logistically, this was made by first estimating the parameters of the corresponding generalized Pareto distribution, which was then transformed to those approximately equivalent for a GEV distribution [17]. Confident intervals was also established numerically from the observed information matrix [17].

$$\hat{\mu}_0 = 3.13,\ \hat{\mu}_{=}2.45,\ \hat{\sigma} = 0.63,\ \hat{\xi} = -0.23$$

and 95% confidence intervals obtained by normal approximation,

$$I_{\mu_0} = (2.83, 3.42),\ I_{\mu_1} = (1.50, 3.39)$$
$$I_{\sigma} = (0.54, 0.72),\ I_{\xi} = (-0.32, -0.14)$$

9.3 Model Diagnostics

Diagnostics for the model was primarily made by inspecting plots. Especially three plots. First, a qq-plot with modified excesses over the threshold that was compared to the standard exponential distribution through the means of first transforming the parameters to those approximately equivalent of generalized Pareto. The data was sought to line up somewhat neatly.

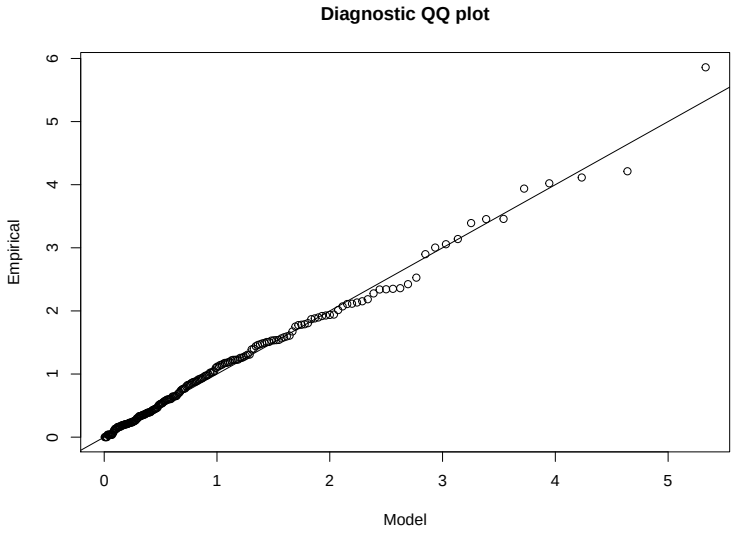

Figure 8: qq-plot with modified excesses compared to standard exponential distribution.

Secondly, a qq-plot with model simulated data compared to excesses. Transformations of the parameters once again made to those that are approximately equivalent of generalized Pareto. Similar to the first plot, data was sought to line up, although it could not be expected to behave just as well [17].

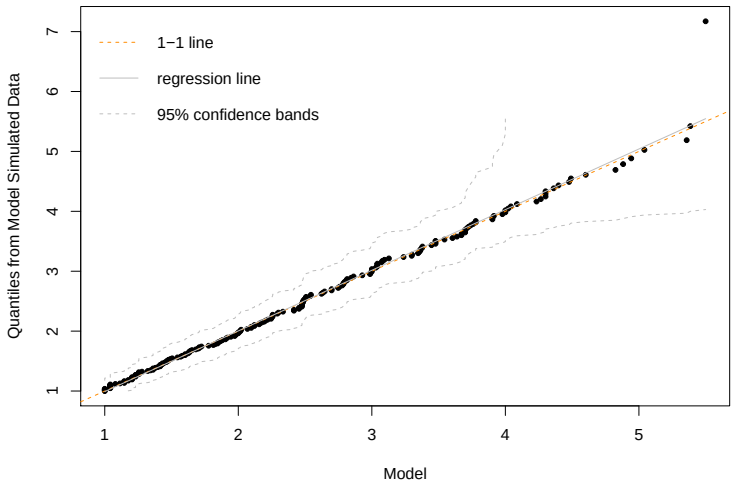

Figure 9: Simulated data compared to excesses. Data sough to line up.

Third, a return level plot. A plot depicting lines for some estimated return levels r_y, where r_y is the y-year return level, as well as data points. An indication of an ill-behaved model would be for example if there would be a tendency for data to come out above all or most return levels r_y more than $80/y$ times (or analogously, below).

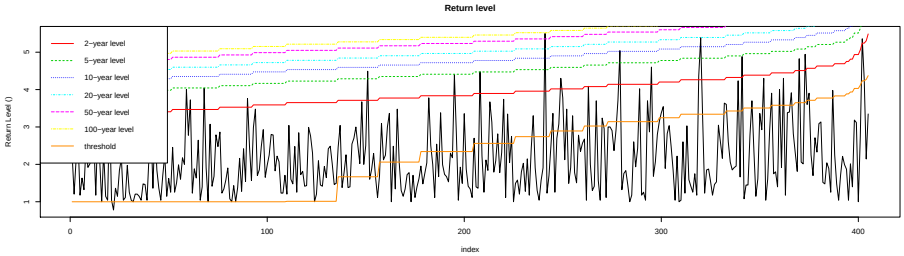

Figure 10: Return level plot. The black line crossing any of the coloured lines depicts an event above corresponding return level.

10 Further Justification

The data was fitted using block maxima and peaks over threshold strategy as well. This was primarily done for the purpose of inspecting GoF plots. If those were reasonable, they could possibly strengthen the arguments for that

the model assumptions that had been made were reasonable.

10.1 Block Maxima

Parameters in an intensity measure for a point process of extremes are closely linked to those of an approximate distribution for maximums obtained by a block maxima model. In fact, by default, the software used constructs the estimates from a point process fitting to those that in theory are the same as fitted with block maxima [17]. Therefore, it was sensible to compare the estimated parameters from the point process for $\tilde{X}_i(m_i)$ with thus obtained by modelling $Z_n(m)$ as $Z_n(m) \sim \text{GEV}(\mu(m), \sigma(m), \xi(m))$. The estimates for the block maxima were

$$\hat{\mu}_0^{BM} = 0.90,\ \hat{\mu}_1^{BM} = 2.89,\ \hat{\sigma}^{BM} = 0.93,\ \hat{\xi}^{BM} = -0.28$$

and for the point process

$$\hat{\mu}_0^{PP} = 3.13,\ \hat{\mu}_1^{PP} = 2.45,\ \hat{\sigma}^{PP} = 0.63,\ \hat{\xi}^{PP} = -0.23$$

Something noteworthy and sought for, is that the shape parameter is negative in both cases. A qq-plot was also produced which did not indicate any direct issues, see Figure 11.

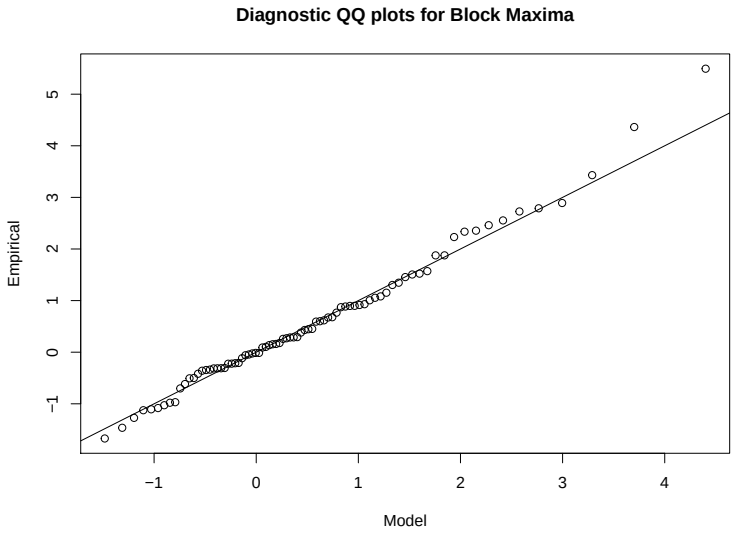

Figure 11: qq-plot for a model that has been fitted using the block maxima approach on yearly maximums.

10.2 Peaks Over Threshold

The data was fitted to a General Pareto distribution using the POT strategy. The same threshold that was decided upon in section 9.1 was used for this model. A linear trend in magnitude was added to σ with a log-link to ensure positivity. The qq-plot could be deemed a bit off in one section, see Figure 12.

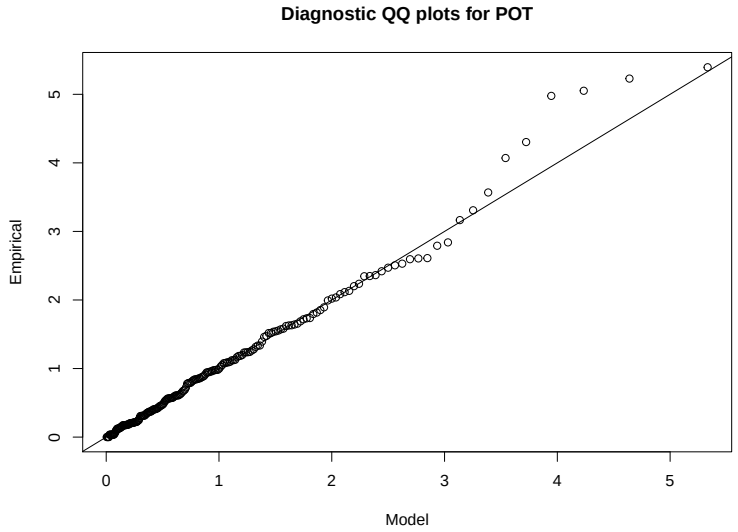

Figure 12: qq-plot for a model that has been fitted using the peaks over threshold approach with the same threshold as the one used for the point process.

11 A point process describing $\{X_i\}_n$

If an assumption could be made regarding the distribution of the magnitudes $m_1, ..., m_n$ such that they are i.i.d and realizations of some random variable M with known density $f_m(m)$, then it would be possible to describe a point process $N_n = \{i/(n+1), X_i : i = 1, ..., n\}$ as a Cox point process. As such, it would be possible to obtain estimates for the corresponding intensity measure $\psi(A)$ and for the void probabilities $\nu(A) = P(N_n(A) = 0)$.

11.1 Distribution of earthquakes magnitudes

A histogram (see Figure 13) of the modified magnitudes worldwide over a ten-year period was produced, visually indicating that an exponential distribution might had been to assume. Although a corresponding qq-plot did not line up perfect (see Figure 14), it did line up somewhat fine, and for that reason it was decided that an assumption from thereon would be made that magnitudes

of earthquakes (above magnitude 6) follows an exponential distribution. An estimated rate value for earthquakes worldwide was numerically calculated to be 10.2 with standard deviation 0.26. It was expected that the rate value would vary a lot when looking at geographical regions other than worldwide.

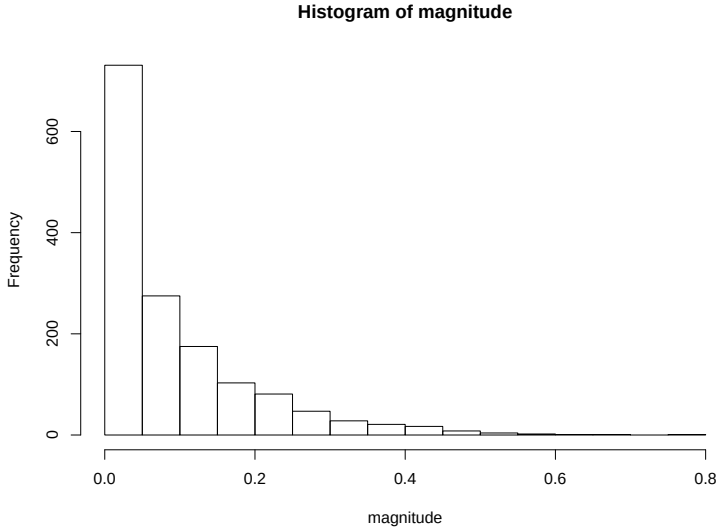

Figure 13: Histogram produced to aid in the search for a parametric family of distributions that the magnitudes can be fitted to.

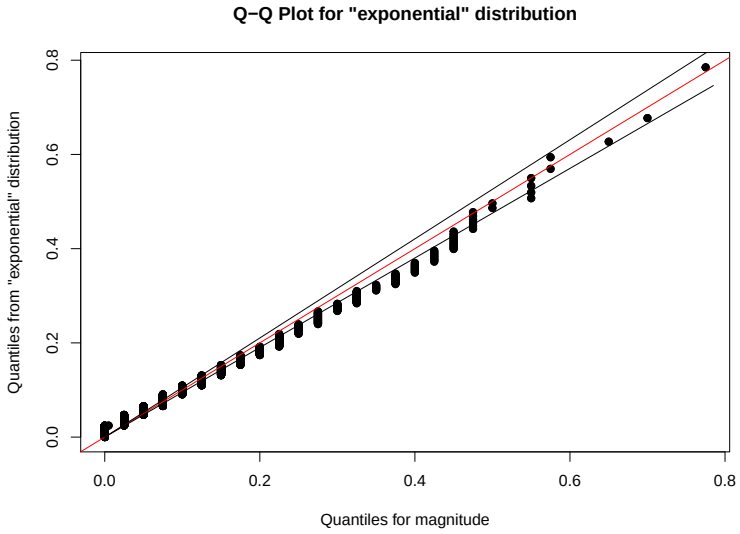

Figure 14: Diagnostics for magnitudes fitted to an exponential distribution.

11.2 Return Levels

With $\nu(A)$ being the void probability over $A = [t_1, t_2] \times (u, \infty)$. The probability of the maximum X_i over the time frame $[t_1, t_2]$ being less than u, is

$$P(max\{\{X_i \ : (i/(n+1)) \in [t_1, t_2]\}\} \leq u) = \nu(A),$$

since if the maximum is not above u, then no points can be in A. This implies of course that

$$P(max\{\{X_i \ : (i/(n+1)) \in [t_1, t_2]\}\} > u) = 1 - \nu(A).$$

By recognizing that an assumption had been made stating that n earthquakes occurs per year, letting $t_1 = 0,\ t_2 = 1$, means that one can determine the y-year return level z_y from the equation $1 - \nu([0, 1] \times (z_y, \infty)) = 1/y$.

With the density for magnitudes assumed exponential, i.e. $f_m(m) = \theta e^{\theta m}$, estimated confidence intervals for the return levels could be obtained with the aid of simulations. The method used was to, for each z_y return level of interest, draw a large number of realizations of the estimate

$$(\{(\hat{\mu_0})_i, (\hat{\mu_1})_i, (\hat{\sigma})_i, (\hat{\xi})_i, (\hat{\theta})_i\}_{i=1,\ldots,N})_y.$$

Then, y could be estimated N times and the upper and lower .05 empirical quantiles could make up the 95% confidence interval for y. This was possible since the distribution of the MLEs used for $(\mu_0, \mu_1, \sigma, \xi)$ could be approximated as

$$Normal_4((\hat{\mu_0}, \hat{\mu_1}, \hat{\sigma}, \hat{\xi}), \Sigma)$$

with Σ being the covariance given by the inverse of the observed information matrix. Similarly, the distribution of $\hat{\theta}$, which was assumed independent of the other estimates, was approximated as $N((\hat{\theta}, Var(\hat{\theta}))$ due to asymptotic normality, with $Var(\hat{\theta})$ numerically estimated.

Since the distribution for the relevant magnitudes of the earthquakes were assumed exponential, the theory in section 4.2.2 could be applied directly, and the void probabilities could be estimated by numerically approximating

$$\nu([0,1] \times (z_y, \infty)) = \int_0^\infty e^{-(t_2-t_1)\left(1+\xi\left(\frac{z_y-(\mu_0+\mu_1 m)}{\sigma}\right)\right)^{-1/\xi}} \theta e^{\theta m} dm.$$

Return level plots based on the methodology above were created for three different geographical regions as mean of model validation. The reasoning for looking at different regions was to review the adaptability with respect to the estimated rate parameter. The rate parameter for the magnitudes varied by region, while estimates for the GEV-parameters were based upon earthquakes all over the world, but scaled in μ and σ due to the varying number of earthquakes per year. If the model could be assumed good enough of an approximation, then it would perhaps be possible to apply the model for regions were large earthquakes very rarely occurs.

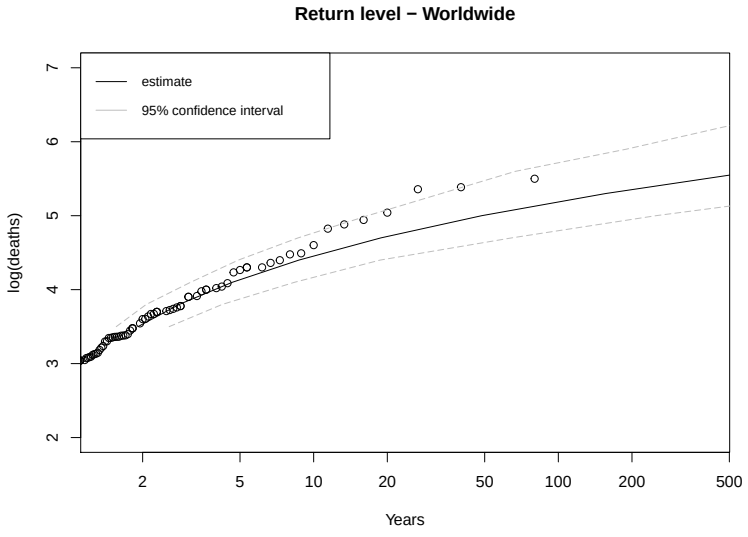

Figure 15: Rate parameter estimated to 12.5

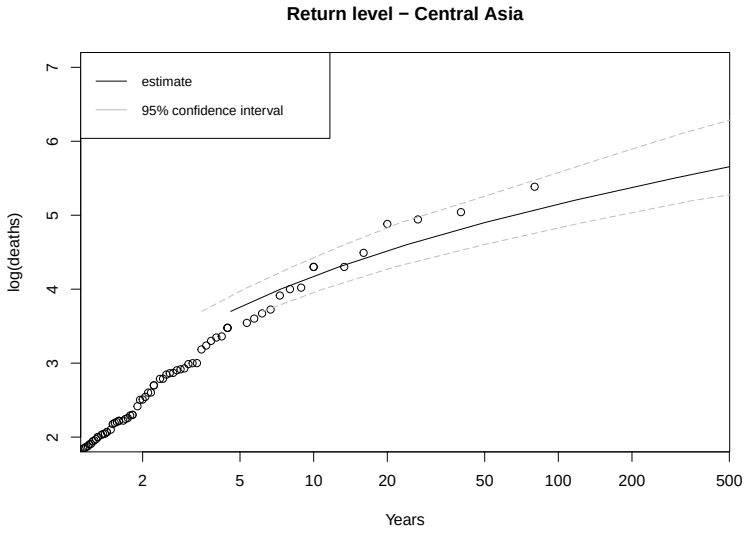

Figure 16: Rate parameter estimated to 7.88

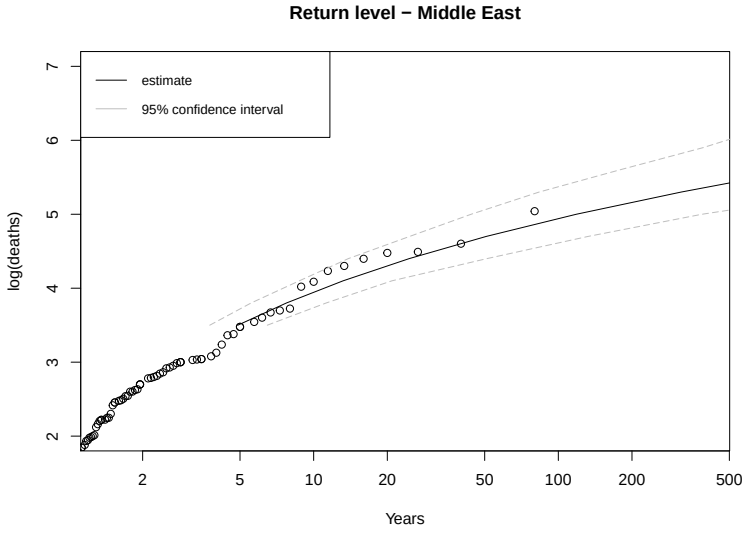

Figure 17: Rate parameter estimated to 9.33

11.3 Intensity of extreme cases

With an established assumption of earthquakes magnitude following an exponential distribution with parameter θ, what is theorized in section 4.2.2 may be applied. Therefore, with the following change of variables,

$$\alpha = -1/\xi$$

$$\beta = -\theta/(\xi\mu_1/\sigma)$$

$$z = 1 + (\xi/\sigma)u - \xi\mu_0/\sigma - \xi\mu_1/\sigma$$

an approximate intensity measure of earthquakes with death tolls above u, over the number of days t, can be given by

$$\psi(A) = (t/365)\beta^{-\alpha}e^{\beta z}\Gamma(\alpha + 1, \beta z)$$

with $\Gamma(a, x)$ being the upper incomplete gamma function.

12 Conclusions

The objective of the the thesis was first and foremost to find a suitable estimate for an approximate Poisson point process conditioned on any significant covariates. It was determined, as suspected, that the magnitude would be a relevant covariate that should be conditioned upon. Other covariates were tried out, but none were found to significantly impact the statistical model. The diagnostics performed on the estimated model that was a result of fitting a point process to selected data suggested that extreme value theory might very well be applicable under reasonable assumptions. The fact that fitting a GEV model over the yearly maximum made for a reasonable qq-plot and estimated parameters (that were in line with the point process approach), was definitely sought after and contributed to the overall reasonability of the estimated point process. It was not straight forward how to validate the general point process that relied on known magnitude distribution, however, return level plots that was produced were indeed looking good.

13 Future research

Although some covariates other than magnitude were considered, it may be of interest to investigate others. One possible covariate that was not particularly looked into, but that could probably influence the behavior of the model was geographical regions. As stated in section 5.4, it is reasonable to assume that earthquakes have varying effects on the number of deaths dependent on location and therefore one might find it interesting to look further into the topic.

Another aspect that would be interesting to look further into would be the effect of a random number of observations. An assumption was made was that there occurs n earthquakes annually with magnitude larger than six and with

death toll at least ten. This assumption is obviously a hefty one, and not true. It seems likely that n could be assumed random with Poisson distribution, however validation of this and the effect it would have on both parameter estimation and corresponding confidence intervals was not dealt with. ξ is invariant of n, but μ and σ is variant in such a way that $\sigma - \xi\mu$ will remain constant w.r.t n [1, ch. 4, p. 75-76]. For block maxima the randomness of n would not pose a direct issue as it is incredibly unlikely that no event would qualify for an entire year.

Figure 14 indicates that perhaps it is an over simplification to assume exponential distribution for earthquakes with magnitude higher than six. Although qq-plots was inspected for smaller regions, indicating a better match for the exponential distribution, it could be of interest to find a more suitable distribution, or perhaps look into non-parametric assumptions.

In addition to the number of deaths from an earthquake, the data set provided by *NOAA National Centers for Environmental Information* also contains other interesting information, such as the number of houses destroyed, economical damage, and injured people. Those may be of interest to model for similar reasons as the death toll is interesting to model.

References

[1] Coles S. (2004). An Introduction to Statistical Modeling of Extreme values.

[2] Kotz S. (2000). Extreme Value Distributions. Theory and Applications.

[3] Gut A. (2009). An Intermediate Course in Probability (second edition).

[4] Møller, J. and Waagepetersen, R. (2002) Statistical inference for Cox processes, in Spatial cluster modelling, Denison, D. and Lawson, A. B. (eds), Chapman and Hall/CRC

[5] Gilleland, E., & Katz, R. (2016). in2extRemes: Into the R package extRemes. Extreme value analysis for weather and climate applications (No. NCAR/TN-523+STR). doi:10.5065/D65T3HP2

[6] Nordquist J. M. (1945). Theory of largest values applied to earthquake magnitudes.

[7] National Geophysical Data Center / World Data Service (NGDC/WDS): NCEI/WDS Global Significant Earthquake Database. NOAA National Centers for Environmental Information. doi:10.7289/V5TD9V7K [16/10-2020]

[8] USGS Earthquake Hazard Program. ANSS Comprehensive Earthquake Catalog (ComCat) [16/10-2020]
`https://earthquake.usgs.gov/data/comcat/`

[9] The World Bank. World Development Indicators

[10] Vincent Arel-Bundock (2020). WDI: World Development Indicators and Other World Bank Data. R package version 2.7.1.
`https://CRAN.R-project.org/package=WDI`

[11] Thomas Roth (2016). qualityTools: Statistics in Quality Science. R package version 1.55
`http://www.r-qualitytools.org`

[12] World Health Organization. Earthquakes. [online] 12/01-2021
`https://www.who.int/health-topics/earthquakes#tab=tab_1`

[13] Hannah Ritchie (2014) - "Natural Disasters". Published online at OurWorldInData.org.
`https://ourworldindata.org/natural-disasters`

[14] United States Geological Survey [online] 12/01-2021
`https://www.usgs.gov/faqs/why-do-earthquakes-other-countries-seem-cause-more-damage-and-casualties-earthquakes-us?qt-news_science_products=3#qt-news_science_products`

[15] Michigan Tech. How Are Earthquake Magnitudes Measured? [online] 12/01-2021
`http://www.geo.mtu.edu/UPSeis/intensity.html)`

[16] National Geographic. May 22, 1960 CE: Valdivia Earthquake Strikes Chile [online] 12/01-2021
https://www.nationalgeographic.org/thisday/may22/valdivia-earthquake-strikes-chile/

[17] Gilleland E. fevd: Fit An Extreme Value Distribution (EVD) to Data. [online] 12/01-2021
https://rdrr.io/cran/extRemes/man/fevd.html