UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Application of extreme value theory to seismology

Filipa Maria Simões Mendes

Mestrado em Estatística e Investigação Operacional
Especialização em Estatística

Dissertação orientada por:
Professora Doutora Patrícia de Zea Bermudez

2023

Acknowledgements

I would like to thank the following people for helping me achieve this goal.

To my parents and to the rest of my family for all the support and inspiration and for giving me the possibility to take a masters degree.

To my advisor, Professor Patrícia de Zea Bermudez, for her dedication and help since day one on this project.

To my closest friends for being always so cheerful and supportive.

Thank you all.

Abstract

The occurrence of earthquakes is an issue with an extreme importance for the society due to the effects that they can cause.

There are specific areas in the world with great seismic activity, like the Pacific fire Ring, that includes North and South America the Kamchatka Peninsula and some islands in the western Pacific ocean.

The main goal of this study is to model high magnitude earthquakes. For this purpose the data from the ISG-GEM Catalogue was chosen to study the moment magnitude, m_W. Only severe earthquakes were considered ($m_W > 6$). From the available types of magnitude, m_W was the one selected to study due to the fact that for very large earthquakes it gives the most reliable estimate of earthquake size and it provides an estimate of the earthquake size valid over the complete range of magnitudes.

Similar studies were developed in a worldwide perspective (Pisarenko et al., 2014) and in specific regions, such as China (Ma, Bai, and Meng, 2021) and the Ecuadorian coast (García-Bustos et al., 2018).

In a first approach a worldwide study was conducted, followed by a study of a specific region with high seismic activity (Japan). Extreme Value Theory provides the appropriate methods for modeling earthquakes with high magnitudes.

In this master thesis, for both data sets, a study of some of the available variables from the ISG-GEM catalogue, being the moment magnitude the main one, was conducted.

Earthquakes with high magnitude, for both the worldwide and the Japan data sets, were modelled using the Block Maxima and the Peaks Over Threshold methods. The main purpose was to estimate the tail probabilities and extreme quantiles.

For both the worldwide and the Japan data sets, applying the Block Maxima method, the Generalized Extreme Value distribution was adjusted to the data, as well as a Gumbel model. Methods to compare and evaluate the statistical models were performed in order to choose the best one. Extreme quantiles were also calculated.

When considering the Peaks Over Threshold method, for both the worldwide and the Japan data sets, the Generalized Pareto Distribution was adjusted to the data, as well as a Exponential model. Goodness-of-fit tests were performed and quality measures of statistical model were calculated in order to choose the best model. Extreme quantiles were also calculated.

Keywords: Extreme Value Theory, Block Maxima, Peaks Over Threshold

Resumo

A Teoria de Valores Extremos é um ramo da estatística que trata da modelação de eventos extremos (muito elevados ou muito baixos) que ocorrem com frequência baixa. Face a este facto, as amostras destes tipos de acontecimentos são, na maior parte das vezes, de reduzida dimensão. Em comparação com a estatística clássica, que se foca fundamentalmente no comportamento central dos dados, a Teoria de Valores Extremos dedica-se ao estudo do comportamento das observações mais afastadas do centro da amostra (localizadas nas caudas da distribuição subjacente aos dados). No contexto da Teoria de Valores Extremos o Teorema de Fisher and Tippett, 1928 (e também Gnedenko, 1943) assume uma importância capital. De facto, segundo o teorema indicado, a amostra de máximos convenientemente normalizada tem assintoticamente uma distribuição que pertence a uma das três famílias de distribuições de extremos: a Fréchet, a Gumbel e a Weibull. Estas famílias podem ser englobadas numa só, a distribuição generalizada de valores extremos.

O Teorema de Fisher e Tippet assume na Teoria de Valores Extremos uma importância semelhante ao teorema do limite central, quando o foco da análise é a média da distribuição.

A Teoria de Valores Extremos é aplicada a diversas áreas: seguros (ver Cerchiara, 2008), meteorologia (ver Reis, Souza, and Graf, 2022), saúde (ver De Zea Bermudez and Mendes, 2012), sismologia (ver Pisarenko et al., 2014), entre outras.

A ocorrência de sismos é um fenómeno de extrema importância para a sociedade, tendo em conta os danos materiais e humanos que pode causar. Um sismo pode ser classificado pela energia libertada no hipocentro (magnitude) e pela sua intensidade (grau de vibração provocado pelo sismo).

Existem zonas específicas no globo terrestre que apresentam grande atividade sísmica, como por exemplo, o Anel de fogo do Pacífico. O Anel de fogo do Pacífico inclui a América do Norte, a América do Sul, a península da Kamchatka (Rússia) e algumas ilhas na parte ocidental do oceano Pacífico.

Ao longo dos anos têm vindo a ser desenvolvidas diferentes formas de exprimir a magnitude, que corresponde à energia libertada por um sismo no hipocentro. A medida mais conhecida é a proposta por Ritcher, m_L (ver Richter, 1935). No entanto, de entre os vários tipos de magnitudes existentes na literatura de sismologia a magnitude dos momentos, m_W, é a que fornece estimativas mais fiáveis do tamanho de um sismo. Esta medida tem a vantagem de que é válida para toda a escala da magnitude, ou seja, para todos os sismos independentemente do seu tamanho. Tendo em conta o valor de m_W, os sismos podem ser classificados em três classes: baixos (com m_W inferior a 5), moderados (com m_W superior ou igual a 5 e inferior a 6) e severos (com m_W superior ou igual a 6). Ao modelar o tamanho dos sismos recorrendo a m_W é preciso ter em conta a necessidade de se restringirem as ocorrências a sismos que tiveram lugar após o século XX, visto que apenas a partir dos anos iniciais do século XX é que é possível estimar o valor da magnitude dos momentos com precisão.

O principal objetivo da presente dissertação é modelar sismos de elevada magnitude e estimar os quantis extremos. Para tal, foram utilizados os dados do ISG-GEM Catalogue (catálogo requirido por *email*) de 1904 a 2018, considerando a variável de interesse principal a magnitude dos momentos, m_W. Dado o elevado número de sismos que ocorreram a nível mundial durante este período de tempo, e tendo em conta que este estudo se centra na modelação de eventos extremos, decidiu-se restringir a modelação de sismos severos que apresentam m_W superior a 6. Estes sismos são efetivamente os que causam perdas de vidas e danos materiais usualmente de considerável importância.

Estudos semelhantes já foram realizados a nível mundial (ver Pisarenko et al., 2014) e recentemente em áreas específicas, como é o caso da China (ver Ma, Bai, and Meng, 2021) ou da costa do Equador (ver García-Bustos et al., 2018).

Nesta dissertação foram estudados os sismos a nível mundial (dados globais) seguindo-se o estudo de uma zona específica, escolhida devido à sua elevada atividade sísmica, a área do Japão.

A teoria de valores extremos fornece as ferramentas adequadas para a modelação de sismos de elevada magnitude.

Os sismos de grande magnitude, quer a nível mundial quer na área do Japão, foram modelados usando o método dos Máximos Anuais e também o método dos Excessos acima de um limiar (*threshold*) elevado, método habitualmente conhecido pela sua designação anglosaxónica, *Peaks Over Threshold*.

O método dos Máximos Anuais ajusta a distribuição generalizada de valores extremos à amostra dos máximos.

O método dos Excessos acima de um limiar elevado ajusta uma distribuição generalizada de Pareto às excedências acima de um *threshold* suficientemente elevado (ver Balkema and Haan, 1974 e Pickands, 1975). Primeiramente é necessário escolher um *threshold*. A escolha de um limiar adequado pode ser na prática um problema extremamente complexo. Muitas têm sido as contribuições na literatura ao longo dos tempos, mas o problema continua em aberto. Na literatura são sugeridas várias metodologias para o efeito. A função de excesso médio (ver Scarrott and MacDonald, 2012) e dois metódos com abordagem bayesiana (ver Lee, Fan, and Sisson, 2015 e Northrop, Attalides, and Jonathan, 2017) foram os escolhidos nesta dissertação para abordar o tema da escolha do *threshold*. Os dois últimos métodos não forneceram resultados satisfatórios que permitissem auxiliar na escolha do limiar. Portanto a escolha do limiar fundamentou-se na utilização da função de excesso médio e na propriedade de estabilidade da distribuição de Pareto generalizada acima de um limiar elevado.

O objetivo ao aplicar estas duas metodologias é estimar probabilidades de cauda e quantis extremos.

No que se refere à probabildade de cauda (direita), pretende-se estimar a $P\{X > x\}$, sendo x um valor muito elevado. Um quantil extremo pode definir-se como o valor x_p tal que $P\{X > x_p\} = p$, para um p suficientemente pequeno.

Consideraram-se quatro conjuntos de dados: os dados globais, os dados globais restritos ao máximo anual (m_W) de 1904 a 2018, os dados relativos à área do Japão (com latitude entre 30.145 e 45.383 e com longitude entre 129.551 e 148.007) e os dados da área do Japão restritos ao máximo anual (m_W) de 1911 a 2018. Para cada um destes conjuntos de dados foi feito um estudo descritivo para algumas das variáveis existentes no catálogo (m_W, $depth$, $date$, lon e lat).

Para os dois conjuntos de dados relativos à amostra dos máximos anuais foi aplicado o método dos máximos anuais, ajustando-se a distribuição generalizada de valores extremos e a distribuição Gumbel, uma vez que a estimativa do parâmetro de forma da distribuição generalizada de valores extremos é muito próxima de zero para ambos os conjutos de dados. Os dois restantes conjuntos de dados foram modelados pelo método dos excessos acima de um limiar elevado, ajustando-se a distribuição generalizada de Pareto às excedências acima de um *threshold* suficientemente elevado e a distribuição exponencial pelo facto de, mais uma vez, as estimativas dos parâmetros de forma das distribuições generalizadas de Pareto serem, para ambos os conjuntos de dados, muito próximas de zero.

Seguidamente foram utilizados métodos (AIC e BIC; *Likelihood Ratio Test*) para comparar e avaliar os modelos ajustados de modo a escolher o melhor.

Para os quatro conjuntos de dados também foram calculados quantis extremos e correspondentes intervalos de confiança a 95%.

Keywords: Teoria de Valores Extremos, Método dos Máximos Anuais, Peaks Over Threshold, Método dos Excessos acima de um limiar elevado

Contents

List of Figures

List of Tables

1 Introduction

Earthquakes are one of the main concerns of the society since some cause substantial damages. It is desirable to avoid deaths and economic losses as much as possible so it becomes crucial the study of seismic hazard.

Earthquakes of high magnitude are considered extreme events so it is necessary to have the appropriate methods to study these events.

Extreme events are low frequency episodes of some random process. Extreme values are scarce which means that estimates are often required for levels of a process that are much greater than ever observed.

Extreme Value Theory (EVT) develops techniques and models for describing these unusual events. EVT has been widely applied to many areas such as insurance (Cerchiara, 2008), meteorology (Reis, Souza, and Graf, 2022), health (De Zea Bermudez and Mendes, 2012), earthquakes (Ma, Bai, and Meng, 2021), among others.

In EVT there are two approaches, the Block Maxima and the Peaks Over Threshold. Coles, 2001 describes both these methodologies and presents further topics of EVT.

The ISC-GEM Global Instrumental Earthquake Catalogue was the chosen database for this master thesis. Both methodologies will be applied to this data considering both the total catalog (worldwide) and a specific region (Japan). We aim to model the sample maxima and to fit a generalized Pareto distribution to model the exceedances above a sufficiently high threshold. We also intend to estimate extreme quantiles.

2 Basics of Seismology

Seismology is the scientific study of earthquakes and of the internal structure of the Earth.

When two blocks of the earth slip past one another it causes an earthquake, releasing energy in the form of seismic waves. The surface where they slip is called the fault. The location below the earth's surface where the earthquake starts is called the hypocenter and the location above it on the surface of the earth is named the epicenter.

Sometimes large earthquakes are followed by less intense earthquakes which happen in the same location. These are known as foreshocks.

An earthquake can be classified by its magnitude and intensity.

Magnitude is a measure of the size of an earthquake and intensity describes the degree of shaking caused by the earthquakes.

The logarithmic earthquake magnitude scale was first developed by Charles Richter in the 1930's, see Richter, 1935. This magnitude scale was referred to as m_L and became known as the Richter magnitude, with L standing for local.

As more seismic stations were installed around the world, it became apparent that the method developed by Richter was only valid for certain seismic waves frequencies and distance ranges.

In order to take advantage of the progress of the seismic stations, new magnitude scales that are an extension of Richter's original idea were developed (USGS, 2023): body wave magnitude (m_B) and surface wave magnitude (m_S). Likewise m_L, these two scales are valid for a particular frequency range and type of seismic signal, the same range as m_L.

Having in consideration these characteristics that were lacking in m_L, m_B, and m_S, a new magnitude scale, known as moment magnitude (m_W), was developed.

For very large earthquakes, moment magnitude gives the most reliable estimate of earthquake size and it provides an estimate (of the earthquake size) valid over the complete range of magnitudes.

A moment is a physical quantity proportional to the slip on the fault multiplied by the area of the fault surface that slips; it is related to the total energy released during an earthquake. The moment can be estimated from seismograms and geodetic measurements. The moment, m_0, is then converted into a number similar to other earthquake magnitudes by a standard formula and the result is the moment magnitude (see Felgueiras, 2012):

$$m_W = \frac{\log_{10} m_0 - c}{1.5},$$

in which $c = 16.1$ or $c = 9.1$ when m_0 is measure in the Newton-meter (Nm) scale.

As in Ma, Bai, and Meng, 2021, earthquakes can be classified in 3 classes according to the value of m_W: low ($m_W < 5$), moderate ($5 \leq m_W < 6$) and severe ($m_W \geq 6$).

Severe earthquakes are considered strong earthquakes and may cause damage to buildings.

Note that when considering m_W we must restrict our analysis to earthquakes that occurred after the XX century since it is only after this time period that m_W can be estimated with some accuracy, see Felgueiras, 2012.

3 Extreme Value Theory

Extreme value analysis aims to describe the stochastic behaviour of a process at unusually large (or small) levels. The estimation of the probability of events that are more extreme than the ones that have already occurred is one of the main purposes of the analysis, as well as extreme quantile estimation. Extreme quantile estimation is very important in EVT since it gives us information about future extreme events that can occur with a very low probability.

This chapter will review some of the fundamental concepts of extreme value theory and models. The main source for it was Coles, 2001.

It should be mentioned that the last section of this chapter (section 3.3) addresses general concepts of comparison and evaluation of statistical models. They were included in this chapter of EVT because they are part of the statistical tools that will be used in this dissertation.

3.1 Block Maxima

3.1.1 Asymptotic Models

Let us consider a sequence of independent random variables $(X_1, ..., X_n)$ with a common distribution function F. Let

$$M_n = \max\{X_1, ..., X_n\}. \tag{3.1}$$

The exact distribution of M_n can be derived as:

$$\begin{aligned} \mathrm{P}\{M_n \leq z\} &= P\{X_1 \leq z, X_2 \leq z, ..., X_n \leq z)\} \\ &= \mathrm{P}\{X_1 \leq z\} \times \mathrm{P}\{X_2 \leq z\} \times ... \times \mathrm{P}\{X_n \leq z\} \\ &= \{F(z)\}^n. \end{aligned} \tag{3.2}$$

However, the distribution F is generally unknown. However, even if F were known, F^n is a degenerate distribution as n increases. In fact F^n tends to zero if $|F(z)| < 1$ and to 1 when $|F(z)| = 1$. So we proceed to study the behaviour of F^n as $n \to \infty$, using a linear transformation of the variable M_n ($M_n^* = \frac{M_n - b_n}{a_n}$). This transformation surpasses the difficulties raised with the variable M_n. The asymptotic behaviour of M_n^* is given in the following theorem due to Fisher and Tippett, 1928; see also Gnedenko, 1943.

Extremal Types Theorem: *If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

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

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

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

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

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

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

These three families of distributions are called the extreme value distributions types I, II and III known as the Gumbel, Fréchet and Weibull families, respectively. Each family has a location and scale parameter, b and a respectively; additionally, the Fréchet and Weibull families have a shape parameter α. This theorem implies that M_n^* has a limiting distribution that must be one of the three types of extreme value distributions. The three types of extreme value distributions are the only possible limits for the distributions of M_n^*, regardless of the distribution F of the underlying population.

These distributions can be combined into a single family known as the generalized extreme value distribution (GEV).

The Generalized Extreme Values Distribution: *If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

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

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

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

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

When $\xi \to 0$ in the above expression then the Gumbel family is obtained:

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

The GEV provides a model for the distribution of block maxima, being μ, σ and ξ the location, scale and shape parameters, respectively.

The Block Maxima (BM) method consists in fitting a GEV to the sample of block maxima (or minima). In general the estimation of the GEV parameters is carried by maximum likelihood (ML), although alternative methods are available in the literature, see Hosking, Wallis, and Wood, 1985.

In many practical situations the blocking structure is naturally defined. For instance, we may have annual/monthly or daily data, data in batches, etc...

Following the parameter estimation (by maximum likelihood) the distribution function of F is found.

Provided we possess a large sample, by the Extremal Types Theorem, $G(z)$ can be approximated to one of the following distributions: Gumbel (if $\xi = 0$), Fréchet (if $\xi > 0$) or Weibull (if $\xi < 0$).

3.1.2 Parameter Estimation

When fitting the GEV model to a data set we have to be careful with the choice of the block size. This choice has to be a trade-off between bias and variance: with blocks that are to small the asymptotic basis of the model may be violated leading to bias and large blocks lead to few block maxima provoking high variance.

The existence of the ML estimators and their properties depend on the values of the shape parameter ξ of the GEV. When $\xi < -1$, the estimators of the GEV parameters do not exist. Their existence is guaranteed for ξ in the interval (-1,-0.5), although in this case the desirable asymptotic properties are not satisfied. Finally, if ξ is larger than -0.5, the ML estimators exist and have the usual ML asymptotic behaviour (see Coles, 2001).

Let $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)$ be independent variables all having the GEV distribution. The log-likelihood (note that log is the natural logarithm) for the GEV parameters when $\xi \neq 0$ is

$$\mathcal{L}(\mu, \sigma, \xi|\mathbf{z}) = -n\log\sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} \log\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{n} \log\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi} \tag{3.3}$$

where

$$1 + \xi\left(\frac{z_i - \mu}{\sigma}\right) > 0, \quad i = 1, ..., n$$

and $z = (z_1, ..., z_n)$ is the observed sample. When $\xi = 0$ the log-likelihood is

$$\mathcal{L}(\mu, \sigma, \xi|\mathbf{z}) = -n\log\,\sigma - \sum_{i=1}^{n}\left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{n} \exp\left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\}. \tag{3.4}$$

The maximum likelihood estimate for the GEV parameters is obtained maximizing (3.3) and (3.4).

3.1.3 Extreme Quantiles Estimation

Let's consider a GEV family with parameters μ, σ and ξ, we have

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]^{-1/\xi}\right\}. \tag{3.5}$$

Equaling (3.5) to $1 - p$ and inverting the equation, for very small probability p, the $(1 - p)^{th}$ quantile can be obtained as follows:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1 - p)\}^{-\xi}\right] & \text{for } \xi \neq 0, \\ \mu - \sigma\log\{-\log(1 - p)\} & \text{for } \xi = 0. \end{cases} \tag{3.6}$$

Confidence intervals can be calculated. By the delta method

$$var(\hat{z_p}) \approx \nabla z_p^T V \nabla z_p, \tag{3.7}$$

where V is the variance-covariance matrix of $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ and considering $y_p = -\log(1 - p)$ we have

$$\nabla z_p = \begin{bmatrix} \frac{\partial z_p}{\partial \mu} \\ \frac{\partial z_p}{\partial \sigma} \\ \frac{\partial z_p}{\partial \xi} \end{bmatrix} = \begin{bmatrix} 1 \\ -\xi^{-1}(1 - y_p^{-\xi}) \\ \sigma\xi^{-2}(1 - y_p^{-\xi}) - \sigma\xi^{-1}y_p^{-\xi}\log\, y_p \end{bmatrix}$$

computed at $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

z_p is the return level associated with the return period $\frac{1}{p}$. The quantile z_p is expected to be exceeded, on average, once every $N = \frac{1}{p}$ years.

The delta method referred here is a method that enables to approximate the asymptotic behaviour of functions of a random variable, if the random variable is itself asymptotically normal, allowing us to calculate confidence intervals (CI). See Robinson, 2022 for more information about this topic.

3.2 Peaks Over Threshold

Consider a sequence of independent and identically distributed (i.i.d) random variables, $X_1, X_2, ...,$ having marginal distribution function F. Let us define extreme events as those of the X_i that exceed some high threshold u. Being X an arbitrary term in the X_i sequence, it follows that a description of the stochastic behaviour of extreme events is given by the conditional probability

$$\mathrm{P}\{X > u + y | X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \quad y > 0. \tag{3.8}$$

If the distribution F were known then the distribution of the threshold exceedances would be known exactly. Unfortunately, this is not so. In this section we will approach the Peaks Over Threshold methodology (POT), whose goal is to fit a generalized Pareto distribution (GPD) to the values that exceed a high threshold. See Coles, 2001 for more details on this topic.

3.2.1 The Generalized Pareto Distribution

Theorem: Let $X_1, X_2, ..., X_n$ be a sequence of independent random variables with common distribution function F, and let

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

For large enough n we have

$$\mathrm{P}\{M_n \leq z\} \approx G(z), \tag{3.9}$$

where

$$G(z) = \exp\left\{ -\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \right\}, \tag{3.10}$$

for some $\mu \in \mathbb{R}, \sigma > 0$ and $\xi \in \mathbb{R}$. Then for a large enough threshold u, the distribution function of the random variable $(X - \mu)$, conditional on $X > \mu$, is approximately

$$H(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}} \tag{3.11}$$

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

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

The distribution function defined in (3.11) is known as the generalized Pareto family, being μ the location parameter, σ the scale parameter and ξ the shape parameter.

If block maxima have approximating distribution G, then the threshold excesses have a corresponding approximate distribution within the generalized Pareto family. The parameter ξ in (3.11) is equal to the corresponding GEV distribution.

If $\xi < 0$ the distribution of excesses has a finite upper bound of $u - \tilde{\sigma}/\xi$ and if $\xi > 0$ the distribution has no upper limit. For $\xi = 0$ the distribution is unbounded. When $\xi \to 0$ in (3.11) we get

$$H(z) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), \quad y > 0 \tag{3.12}$$

which is the exponential distribution with parameter $1/\tilde{\sigma}$.

3.2.2 Threshold Selection

Being the data a sequence of i.i.d measurements $x_1, x_2, ...x_n$ and u a high threshold, we can define $\{x_i : x_i > u\}$ as the n_u exceedances above u. So for those x_i, $i = 1, 2, ..., n_u$, we can denote $y_j = x_j - u$ as the j^{th} excess, $j = 1, 2, ..., n_u$. The excesses are realizations of a i.i.d. random variable, whose distribution can be approximate by a member of the generalized Pareto family. The fit of a generalized Pareto distribution to the sample of excesses or exceedances requires a previous choice of an adequate threshold u. This is generally a very difficult task.

We should not choose a too low threshold because it may violate the independence assumption of the data leading to bias, while a too high threshold is not an option either because this selection tends to increase the variance of the estimators of the GPD parameters due to the lack of observations. So the choice of the ideal threshold is based on a balance between bias and variance.

We can use several procedures for this purpose. Exploratory analysis (*qqplot*, for example) is an option namely to assess the tail weight, as well as the use of the Mean Residual Life Function to be defined in 3.2.3, fitting models across a range of potentials thresholds to evaluate the stability of parameter estimates is also a standard technique.

Several methods have bean proposed in the literature for choosing the threshold (see Scarrott and MacDonald, 2012 for a review). More recent methods were also proposed by Lee, Fan, and Sisson, 2015 and Northrop, Attalides, and Jonathan, 2017. Both of these methods were developed in a bayesian framework. In 2020, in a classical set up, a method based on L-moments was proposed (see Silva Lomba and Fraga Alves, 2020)

The next subsection will explore one of the traditional procedures.

3.2.3 Mean Residual Life Function

The Mean Residual Life Function, also known as Mean Excess Function (MEF), is based on the mean of the generalized Pareto distribution. Being Y a random variable with a generalized Pareto distribution with parameters σ and ξ then

$$\mathrm{E}(Y) = \frac{\sigma}{1-\xi} \tag{3.13}$$

when $\xi < 1$. For $\xi \geq 1$ the mean is infinite. Applying (3.13) for some threshold u_0, usually chosen from $X_1, X_2, ...X_n$ being X an arbitrary term of the n random variables, we get:

$$\mathrm{E}(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1-\xi}. \tag{3.14}$$

If the generalized Pareto distribution is valid for excesses above the threshold u_0, it is also valid for all thresholds $u > u_0$. Hence for $u > u_0$,

$$\begin{aligned} \mathrm{E}(X - u | X > u) &= \frac{\sigma_u}{1-\xi} \\ &= \frac{\sigma_{u_0} + \xi u}{1-\xi} \end{aligned} \tag{3.15}$$

which implies that $\mathrm{E}(X - u | X > u)$ is a linear function of u and is the mean of the excesses over the threshold u.

By (3.15) these estimates are expected to change linearly with u, at levels of u for which the generalized Pareto model is appropriate.

The points

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{max} \right\}, \tag{3.16}$$

where $x_{(1)}, x_{(2)}, ..., x_{(n_u)}$ are the n_u ordered observations that exceed u and x_{max} is the sample maxima are usually plotted creating a mean residual life plot. Above a threshold u_0 the mean residual life plot should be approximate linear in u.

3.2.4 Parameter Estimation

After choosing an appropriate threshold we can proceed to estimate the parameters of the generalized Pareto distribution by maximum likelihood.

Let $\mathbf{y} = (y_1, y_2, ..., y_{n_u})$ be the n_u excesses of a threshold u. For $\xi \neq 0$ the log-likelihood of the generalized Pareto distribution is

$$\mathcal{L}(\sigma, \xi | \mathbf{y}) = -n_u \log\sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n_u} \log\left(1 + \frac{\xi y_i}{\sigma}\right), \tag{3.17}$$

for $(1 + \sigma^{-1}\xi y_i) > 0$ for $i = 1, ..., n_u$. When $\xi = 0$

$$\mathcal{L}(\sigma | \mathbf{y}) = -n_u \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{n_u} y_i. \tag{3.18}$$

The analytical maximization of the log-likelihood function for $\xi \neq 0$ is not possible so numerical techniques are necessary.

3.2.5 Extreme Quantiles Estimation

Assuming that a generalized Pareto distribution with parameters σ and ξ is suitable to model the exceedances above a threshold u, we have:

$$\mathrm{P}\{X > x | X > u\} = \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}}, \quad x > u. \tag{3.19}$$

Considering that

$$\begin{aligned} \mathrm{P}\{X > x | X > u\} &= \frac{\mathrm{P}\{X > x, X > u\}}{\mathrm{P}\{X > u\}} \\ &= \frac{\mathrm{P}\{X > x\}}{\mathrm{P}\{X > u\}} \end{aligned} \tag{3.20}$$

it follows that

$$\mathrm{P}\{X > x\} = \mathrm{P}\{X > x | X > u\} \times \mathrm{P}\{X > u\} \tag{3.21}$$

An extreme quantile x_p is a number such that $\mathrm{P}\{X > x_p\} = p$ for a very small probability p and it is the solution of

$$\left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}} \times \zeta_u = p. \tag{3.22}$$

where $\zeta_u = \mathrm{P}\{X > u\}$.

Thus

$$x_p = \frac{\sigma}{\xi}\left[\left(\frac{\zeta_u}{p}\right)^{\xi} - 1\right] + u \tag{3.23}$$

where p has to be a probability close to 0 to guarantee that $x_p > u$. If $\xi = 0$ and proceeding in a similar way, we have

$$x_p = u + \sigma \log\left(\frac{\zeta_u}{p}\right). \tag{3.24}$$

In (3.23) and (3.24) the $(1-p)^{th}$ quantile is presented.

The estimation of extreme quantiles requires the substitution of the parameters by their estimates. However ζ_u is unknown, therefore we need to estimate it. An obvious estimator for ζ_u is

$$\hat{\zeta_u} = \frac{N_u}{n},$$

where N_u is the number of order statistics that exceed u and n is the sample size. The number of exceedances above u follows a binomial distribution, $\mathrm{Bin}(n, \zeta_u)$ and $\hat{\zeta_u}$ is the maximum likelihood estimate of ζ_u.

Confidence intervals for x_p can also be computed by the delta method. However the uncertainty in the estimate of ζ_u should be also taken into account. The variance of $\hat{\zeta_u}$ can be calculated applying the properties of the binomial distribution:

$$\mathrm{var}(\hat{\zeta_u}) = \frac{1}{n^2} var(N_u) = \frac{1}{n^2}[n\hat{\zeta_u}(1-\hat{\zeta_u})] = \frac{1}{n}[\hat{\zeta_u}(1-\hat{\zeta_u})].$$

So the variance-covariance matrix for $(\hat{\zeta_u}, \hat{\sigma}, \hat{\xi})$ is given by:

$$V = \begin{bmatrix} \hat{\zeta_u}(1-\hat{\zeta_u})/n & 0 & 0 \\ 0 & v_{1,1} & v_{1,2} \\ 0 & v_{2,1} & v_{2,2} \end{bmatrix}$$

where $v_{i,j}$ is the (i,j) term of the variance-covariance matrix of $\hat{\sigma}$ and $\hat{\xi}$. Thus by the delta method,

$$var(\hat{x_p}) \approx \nabla x_p^T V \nabla x_p,$$

where

$$\nabla x_p = \begin{bmatrix} \frac{\partial x_p}{\partial \zeta_u} \\ \frac{\partial x_p}{\partial \sigma} \\ \frac{\partial x_p}{\partial \xi} \end{bmatrix} = \begin{bmatrix} \sigma p^{-\xi} \zeta_u^{\xi-1} \\ \xi^{-1}\{(p^{-1}\zeta_u)^{\xi} - 1\} \\ -\sigma\xi^{-2}\{(p^{-1}\zeta_u)^{\xi} - 1\} + \sigma\xi^{-1}(p^{-1}\zeta_u)^{\xi}\log(p^{-1}\zeta_u) \end{bmatrix}$$

computed at $(\hat{\zeta_u}, \hat{\sigma}, \hat{\xi})$.

In the POT approach x_p is commonly expressed as the m-observation return level x_m instead ($m = \frac{1}{p}$). It represents the level which is surpassed, on average, once in m observations. Considering that when applying the POT method we will have a random number of exceedances each year, to calculate the N-year return level, the use of the average number of observations per year is a solution, see Gilleland and Katz, 2016. Then, provided n is the total number of observations and $nyears$ the time period (in years), thus $m = \frac{n}{nyears} \times N$.

3.2.6 Goodness-of-fit Tests for the Generalized Pareto Distribution

In this subsection we will describe two procedures to verify if the generalized Pareto distribution is suitable for the data being studied, the Cramér-von Mises test and the Anderson-Darling test.

The Anderson-Darling statistic (A^2) is a modification of the Crámer-von Mises statistic (W^2) giving more weight to observations in the tail of the distribution, which is useful in detecting outliers.

The null hypothesis is H_0: the random sample $x_1, x_2, .., x_n$ comes from a generalized Pareto distribution and the alternative hypothesis is H_1:the random sample $x_1, x_2, .., x_n$ does not come from a generalized Pareto distribution.

The test statistics are given by

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-u)[\log(z_i) + \log(1 - z_{n+1-i})], \tag{3.25}$$

and

$$W^2 = \sum_{i=1}^{n}(2i-u)\left[z_i - \frac{(2i-1)}{(2i)}\right]^2 + \frac{1}{12n}, \tag{3.26}$$

where $z_i = F(x_i)$ and F is the distribution function of the generalized Pareto distribution.

There are several cases considering the knowledge about the parameters of the distribution, see Choulakian and Stephens, 2001. We will describe the case in which both σ and ξ are unknown. In this situation the procedure is the following:

1. By maximum likelihood obtain the estimates of σ and ξ and make the transformation $z_{(i)} = F(x_{(i)})$ for $i = 1, .., n$.
2. Calculate A^2 and W^2 as described in (3.25) and (3.26).

The tables with the critical values of A^2 and W^2 for the case of σ and ξ unknown can be found in Choulakian and Stephens, 2001.

3.3 Methods to Compare and Evaluate Models

3.3.1 Likelihood Ratio Test

The Likelihood Ratio Test (LRT) is a statistical test to compare two models.

Let M_1 be a model with q parameters and M_2 a model with $q + 1$ parameters, such that M_1 and M_2 are nested models ($M_1 \subset M_2$).

The null hypothesis H_0 is the additional parameter of M_2 is equal to zero and the alternative hypothesis H_1 states that the additional parameter of M_2 is not equal to zero.

Being $\mathcal{L}_1$ and $\mathcal{L}_2$ the likelihood scores of M_1 and M_2, respectively, we can define the test statistic

$$D = 2\{\log(\mathcal{L}_1) - \log(\mathcal{L}_2)\}. \tag{3.27}$$

Under H_0, D given in (3.27) has asymptotically a chi-square distribution with 1 degree of freedom. Then using this information, we can determine the critical region of the test from the χ^2 table (Reject H_0 if $D > c_\alpha$, being c_α the $(1-\alpha)$ quantile of the χ^2 distribution, Coles, 2001) or calculate the p-value .

3.3.2 Quality Measures of Statistical Models

The Akaike Information Criteria (AIC) (see Akaike, 1998) and the Bayesian Information Criteria (BIC) (see Schwarz, 1978) are quality measures of statistical models. They provide a way to select the best model from a set of possible models fitted to a data set.

Let q be the number of parameters of a model, $\mathcal{L}$ the maximum value of the likelihood and n the sample size. The quantities AIC and BIC are defined as follows:

$$\text{AIC} = 2q - 2\log(\mathcal{L}), \tag{3.28}$$

$$\text{BIC} = -2\log(\mathcal{L}) + q\log(n). \tag{3.29}$$

The smaller the values of AIC and BIC are, the better the model is.

The BIC criteria attributes greater penalization to the models which have more parameters than the AIC criteria.

4 Literature Review

To develop the case study to be presented in chapter 5, several articles were studied. However we will specially focus on three in which we based most of our study, namely due to the similarity of the main goal.

In these articles the studies were performed in different regions, such as Mainland China (Ma, Bai, and Meng, 2021), the Ecuadorian Coast (García-Bustos et al., 2018) and even worldwide (Pisarenko et al., 2014).

In each one of them EVT was applied to model earthquakes' magnitudes.

In this chapter we will briefly describe the data analysed presented in these papers, the methods used and also highlight aspects that are considered relevant.

Note that the p-values presented in the following tables result from applying the LRT.

For Mainland China the data consisted of 907 seismic events which occurred from 1920 to 2020. The BM and the POT methods were applied and the results were the following:

Table 4.1: BM Method for Mainland China

	GEV			**Gumbel**	
	μ	σ	ξ	μ	σ
Estimated Parameters	6.28	0.70	-0.19	6.21	0.68
Standard Error Estimates	0.08	0.06	0.07	0.08	0.05
AIC	203.32			207.34	
BIC	209.65			221.69	
p-values (LRT)				0.01	

Table 4.2: POT Method with $u = 6.2$ for Mainland China

	GPD		**Exponential**
	σ	ξ	σ
Estimated Parameters	0.79	-0.28	0.62
Standard Error Estimates	0.07	0.05	0.05
AIC	177.95		189.88
BIC	189.49		193.06
p-values (LRT)			2.0×10^{-4}

When it comes to the Ecuadorian Coast the data consisted of 6099 records from 1906 to 2016. The BM and the POT methods were also applied. The results are summarized in tables 4.3 and 4.4, respectively.

Table 4.3: BM Method for Ecuadorian Coast

	GEV			**Gumbel**	
	μ	σ	ξ	μ	σ
Estimated Parameters	5.16	0.53	0.11	5.19	0.56
Standard Error Estimates	0.07	0.05	0.10	0.06	0.05
AIC	176.03			175.26	
BIC	183.33			180.12	
p-values (LRT)				0.27	

Table 4.4: POT Method with $u = 4.88$ being the chosen threshold, for Ecuadorian Coast

	GPD		**Exponential**
	σ	ξ	σ
Estimated Parameters	0.49	-0.10	0.54
Standard Error Estimates	0.05	0.09	0.04
AIC	184.32		183.72
BIC	191.25		187.18
p-values (LRT)			0.24

The last set of data considered was downloaded from the Harvard catalogue (worldwide) and the study was limited to the period 1977-2006, to the observations that satisfied the conditions $depth < 70$ km and $m_W < 5.5$. The sample resulted from imposing the previous conditions had 4193 records. The BM and the POT were applied. The results were the following:

Table 4.5: GEV and GPD results for the Harvard catalogue

	GEV			**GPD**	
	μ	σ	ξ	σ	ξ
Estimated Parameters	4.982	0.847	-0.185	0.529	-0.204

Looking at table 4.1, for the GEV model $\hat{\xi}$ is negative and the standard error estimate is small, so ξ should be different from zero. The authors also fitted a Gumbel model to perform a likelihood ratio test ($H_0 : \xi = 0 \quad vs \quad H_1 : \xi \neq 0$). The p-value for the likelihood ratio test is low and consequently for any level of significance larger than 0.01 (for instance for 5%) H_0 is rejected. Also the AIC and BIC criterias for the GEV model are lower than the ones obtained for the Gumbel model. Therefore, the authors concluded that the GEV model is a better model for this data when applying the BM method to the Mainland China data than the Gumbel.

When applying the POT method, (table 4.2), with 6.2 being the chosen threshold, $\hat{\xi}$ is lower than zero and the standard error estimate associated is again small. Therefore ξ should be different than zero. An Exponential model was also fitted to the data. Performing the LRT ($H_0 : \xi = 0 \quad vs \quad H_1 : \xi \neq 0$) the authors concluded that the p-value is very low and consequently H_0 is rejected at any usual significance level. The AIC and BIC criterias for the GPD model are lower than the ones associated with the Exponential model. Thus, the GPD model is more suitable for the Mainland China data than the Exponential.

The results obtained by applying the BM and the POT methods are completely in accordance.

Considering the Ecuadorian Coast data and applying the BM method (table 4.3), the authors concluded, for the GEV model, that $\hat{\xi}$ is positive. However $\hat{\xi}$ is also close to 0 and its standard error is relatively high and so a Gumbel model was fitted to the data and the LRT was performed. The authors came to the conclusion that H_0 should be not rejected, since the p-value is much higher than the usual significance levels considered. Therefore there is not statistical evidence that ξ is not 0; moreover the AIC and BIC criterias associated with the Gumbel model are lower than the ones corresponding to the GEV model. Consequently the authors concluded that the Gumbel model is a better model using the BM method.

For the POT method (table 4.4), a threshold of 4.88 was considered and $\hat{\xi} = -0.10$ (note that the parametrization of the GPD model used in this paper is not the same as the one presented in subsection 3.2.1. Thus, when in table 4.4 $\hat{\xi} = -0.10$ it means that the authors indicated $\hat{\xi} = 0.10$) for the GPD model. Since $\hat{\xi} < 0$ and its standard error is high, an Exponential model was adjusted. The p-value associated with the LRT is very high and consequently H_0 is not rejected. Furthermore, the values of AIC and BIC criterias obtained for the Exponential model are lower than the ones associated with the GPD model. Having this in consideration, there is no evidence that ξ is not equal to zero, thus the Exponential model is more suitable for the Ecuadorian Coast data than the GPD.

Likewise the China data, the results obtained by applying the BM method and the POT method to the Ecuadorian data are in agreement as would be expected.

In table 4.5 we see that the estimated parameters are similar with the ones presented in tables 4.1 and 4.2 in the sense that they are all negative.

It should be referred that some authors argue that earthquakes cannot have an infinite upper bound. They claim that natural upper bounds exist and consequently they right truncate the distribution (see e.g. Beirlant, Fraga Alves, and Gomes, 2016 and Ma, Bai, and Meng, 2021).

5 Extreme Value Modeling

5.1 Exploratory Analysis for the Global data set

The data analysed in this master thesis were downloaded from the ISC-GEM Global Instrumental Earthquake Catalogue.

The Supplement Catalogue was ignored due to the poor data availability that prevented the authors of the catalogue to determinate with accuracy the epicentre or/and the magnitude parameters of the earthquakes.

The ISC-GEM catalogue that was requested for this thesis is the version 9.1 released on 27-06-2022. It contains information about 48606 earthquakes that occurred between 04-04-1904 and 31-12-2018.

For the case study only a few variables were selected: *lat* and *long* (latitude and longitude, respectively, of the epicenter), *depth* (depth of the epicenter, in km), m_W (moment magnitude) and *date* (date of the earthquake origin). The last variable was used to create a new variable called *year*; we also imposed the restriction that m_W should be greater than 6 in order to study only severe earthquakes.

By applying the previous restriction to the data we were left with 12046 observations.

In this section we will explore and analyse the data.

Table 5.1: m_W by intervals of amplitude 0.5

(6,6.5)	**[6.5,7)**	**[7,7.5)**	**[7.5,8)**	**[8,8.5)**	**[8.5,9)**	**[9,9.5)**	**[9.5,10]**
8068	2661	884	344	73	12	3	1
66.98%	22.09%	7.34%	2.86%	0.61%	0.10%	0.02%	0.01%

As we can see in table 5.1, the great majority of the earthquakes has moment magnitude between 6 and 7. Fewer observations belong to the interval [8,10].

Considering now the years of the occurrence of the earthquakes, we can summarize the information as follows:

Table 5.2: Number of earthquakes by intervals of 19 years

[1904,1923)	**[1923,1942)**	**[1942,1961)**	**[1961,1980)**	**[1980,1999)**	**[1999,2018]**
886	1950	2077	2221	2299	2613

By table 5.2, it can be observed that in the earlier years not as many earthquakes occurred comparing to the most recent years.

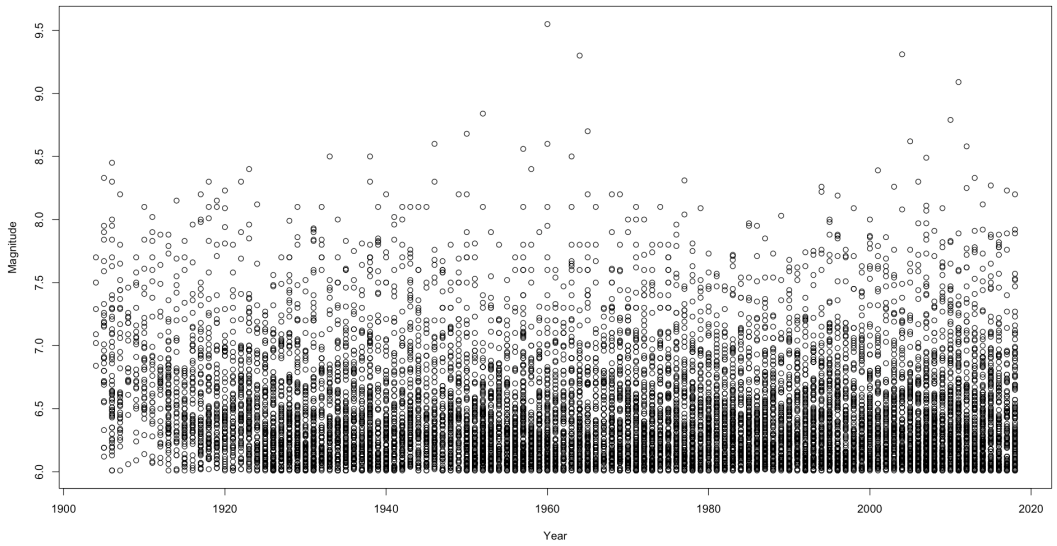

Figure 5.1: m_W through the years

Table 5.3: Summary statistics of m_W by intervals of 19 years

$year$	**Min**	1^{st} **Qu.**	**Median**	**Mean**	3^{rd} **Qu.**	**Max**
[1904,1923)	6.01	6.31	6.56	6.67	6.92	8.45
[1923,1942)	6.01	6.13	6.31	6.44	6.62	8.50
[1942,1961)	6.01	6.13	6.30	6.42	6.55	9.55
[1961,1980)	6.01	6.12	6.30	6.43	6.60	9.30
[1980,1999)	6.01	6.13	6.31	6.42	6.58	8.26
[1999,2018]	6.01	6.13	6.31	6.44	6.61	9.31

However, figure 5.1 and table 5.3 show that the values of m_W do not follow any trend as the years go by.

The 8 intervals have approximately the same behaviour in terms of m_W. The earthquake with the highest moment magnitude occurred in the [1942,1961) interval, more precisely in 1960.

More recently, several earthquakes with high m_W were also recorded.

Table 5.4: Summary statistics of m_W and *depth*

Variable	**Min**	1^{st} **Qu.**	**Median**	**Mean**	3^{rd} **Qu.**	**Max**
m_W	6.01	6.14	6.32	6.45	6.62	9.55
$depth$	0.00	15.00	25.00	56.59	35.00	690.40

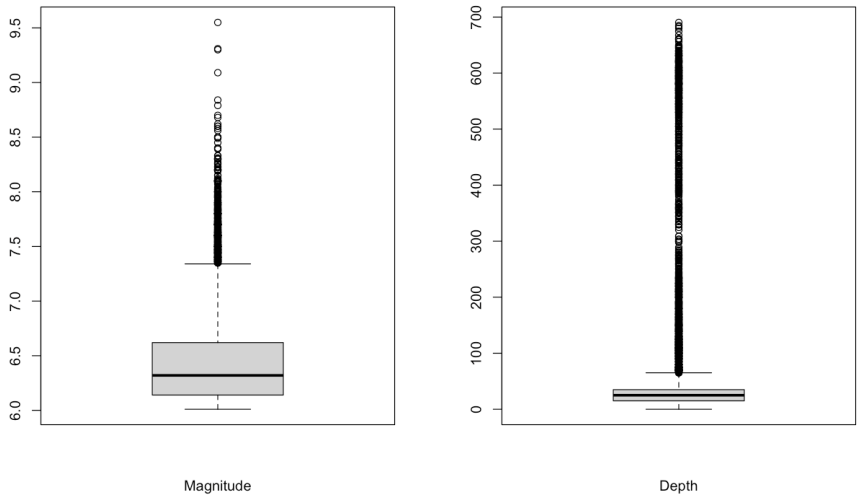

Figure 5.2: Boxplots of m_W and *depth*

Table 5.5: Summary statistics of m_W by *depth* intervals

$depth$	**Min**	1^{st} **Qu.**	**Median**	**Mean**	3^{rd} **Qu.**	**Max**	**n**
[0,100)	6.01	6.13	6.31	6.44	6.60	9.55	10757
[100,200)	6.01	6.16	6.35	6.50	6.70	8.30	577
[200,300)	6.01	6.15	6.34	6.49	6.64	8.10	181
[300,400)	6.01	6.20	6.46	6.57	6.82	7.72	71
[400,500)	6.03	6.16	6.46	6.55	6.79	7.90	102
[500,600)	6.01	6.25	6.48	6.58	6.85	8.20	230
[600,700]	6.01	6.16	6.52	6.65	6.92	8.33	128

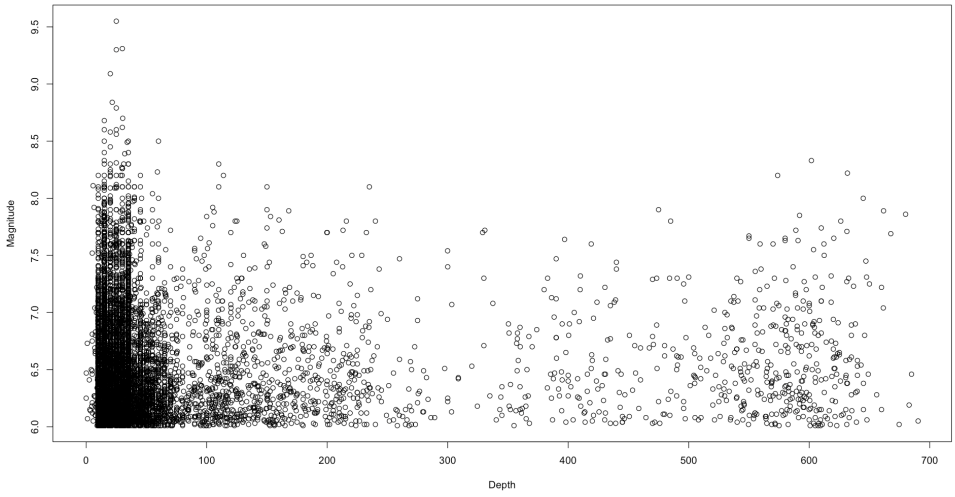

Figure 5.3: m_W vs *depth*

By tables 5.4, 5.5 and figures 5.2,5.3, we can observe the behaviour of the variables m_W and $depth$.

As seen before, most of the earthquakes have m_W between 6 and 7. The depth of the epicenter for approximately 89% of the earthquakes is between 0 and 100 km.

Figure 5.3 illustrates the relation between the moment magnitude and the depth of the epicenter. Again, most of the points are concentrated in the interval [0,100]; note that this interval also contains a large number of earthquakes with high moment magnitude.

As referred in chapter 2, earthquakes mainly occur in faults. In figure 5.4 a map of the tectonic plates (ESC, 2023) is presented, in which the faults are identified by black bold lines.

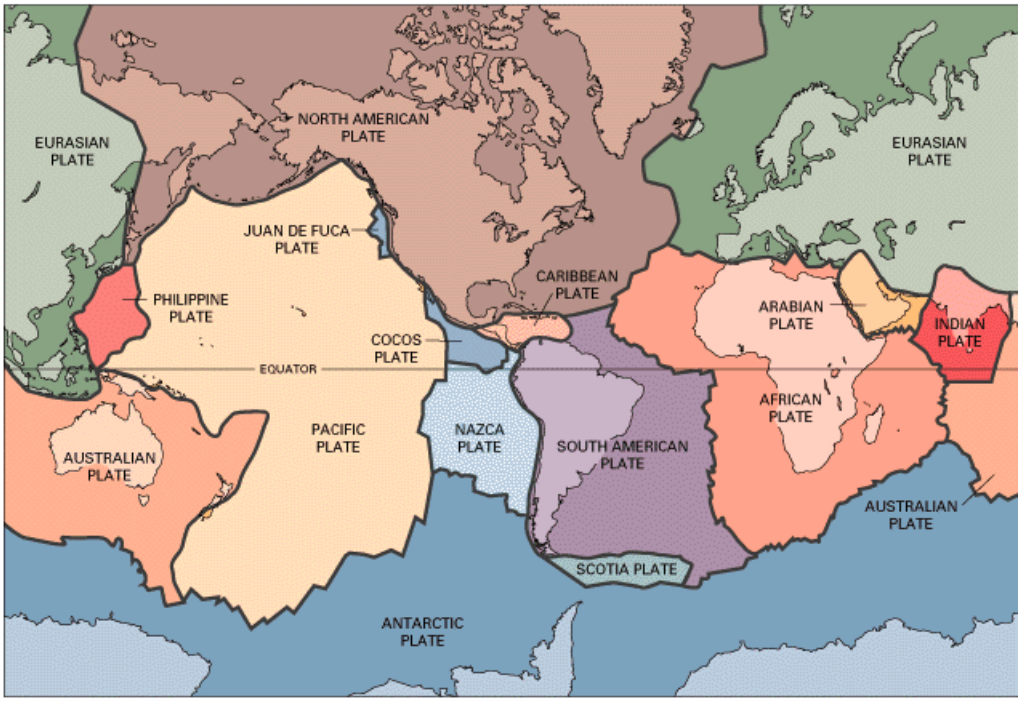

Figure 5.4: Tectonic Plates Map

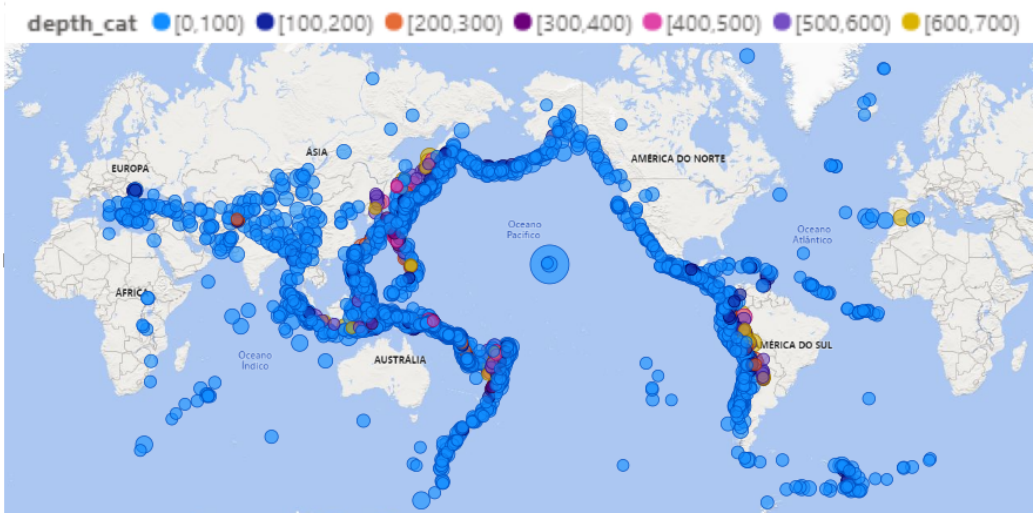

Figure 5.5: Earthquakes represented by m_W and $depth$

The coordinates of the earthquakes epicenters enable us to represent them on a map (Figure 5.5). So when comparing the two figures (5.4 and 5.5), we can see that the locations where the earthquakes occurred are mostly overlapping the limits of the tectonic plates (faults).

The goal of this master thesis is to model extreme values of m_W obtained in the ISC-GEM catalogue. The following subsections will describe the models created for this purpose.

5.2 BM Method for the Global data set

To apply the BM method the appropriate data has to be put together first. For doing so, the 12406 observations referred to in section 5.1 will be considered.

For each year that we have record, the earthquake with the highest value of m_W was selected; our new data set consists in 115 observations, one for each year, from 1904 to 2018. Let's call this data the BM data set.

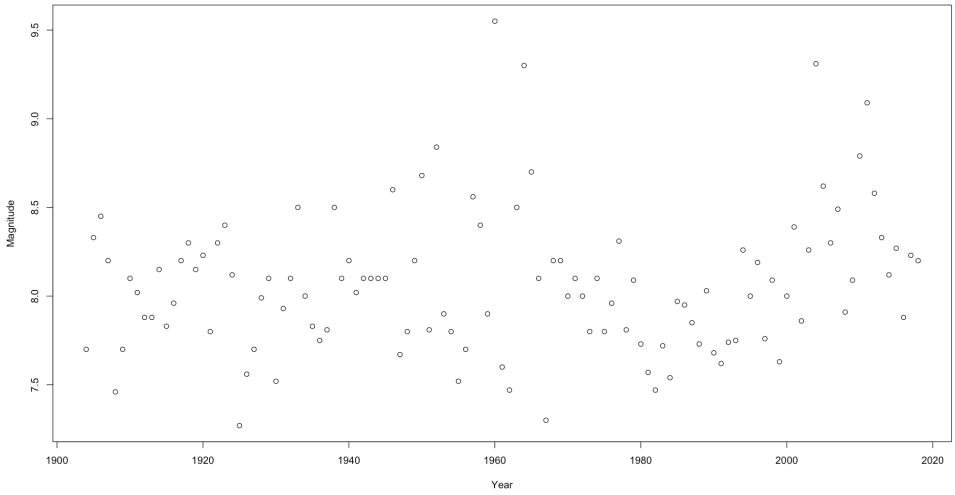

Figure 5.6: Earthquakes represented by m_W through the years for the BM data set

Considering the same intervals of m_W as in chapter 5.1, the data set can be arranged in a table as follows:

Table 5.6: m_w by intervals of amplitude equal to 0.5 for the BM data set

(6,6.5)	**[6.5,7)**	**[7,7.5)**	**[7.5,8)**	**[8,8.5)**	**[8.5,9)**	**[9,9.5)**	**[9.5,10]**
0	0	5	45	50	11	3	1
0%	0%	4%	39%	43%	10%	3%	1%

By figure 5.6 and table 5.6, we can see that the observations of the BM data set correspond to earthquakes with m_W above 7. Moreover, 82% of the data belong to the [7.5,8.5) interval.

In the following table we have the summary statistics for m_W and $depth$.

Table 5.7: Summary statistics of m_W and $depth$ for the BM data set

Variable	**Min**	1^{st} **Qu.**	**Median**	**Mean**	3^{rd} **Qu.**	**Max**
m_W	7.27	7.80	8.03	8.07	8.25	9.55
$depth$	10.00	15.00	25.00	62.97	35.00	644.80

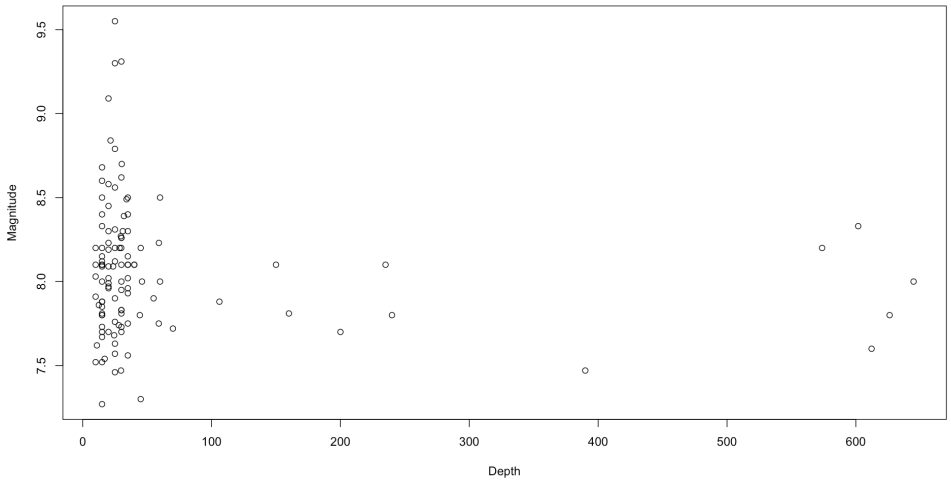

Figure 5.7: Earthquakes represented by m_W and $depth$ for the BM data set

Table 5.8: Summary statistics of m_w by $depth$ for the BM data set

$depth$	**Min**	1^{st} **Qu.**	**Median**	**Mean**	3^{rd} **Qu.**	**Max**	**n**
[0,100)	7.27	7.81	8.09	8.09	8.27	9.55	103
[100,200)	7.81	7.85	7.88	7.93	7.99	8.10	3
[200,300)	7.70	7.75	7.80	7.87	7.95	8.10	3
[300,400)	7.47	7.47	7.47	7.47	7.47	7.47	1
[500,600)	8.20	8.20	8.20	8.20	8.20	8.20	1
[600,700]	7.60	7.75	7.90	7.93	8.08	8.33	4

The BM data set contains 115 observations of earthquakes with $depth$ situated between 1 and 644.80 km (table 5.7); the depth of 90% of them belong to the interval [0,100), as we can see by table 5.8.

By figure 5.7 the depth of the epicenter of the earthquake with the highest m_W is situated between 40 and 50 km.

In figure 5.8 the epicenter coordinates of the earthquakes for the BM data set are represented by m_W and by $depth$. Most of the earthquakes occurred in the limits of the Eurasian Plate, the North America plate and the Pacific plate. The limits of the South American Plate and the Nazca Plates are also areas with a large number of earthquakes.

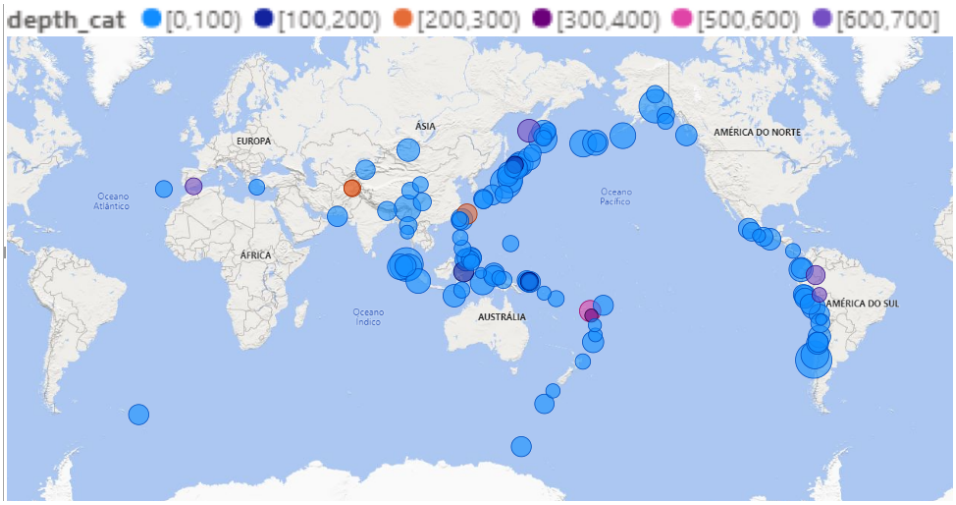

Figure 5.8: Location of the earthquakes represented by m_W and $depth$ for the BM data set

Applying the BM method we obtained the results presented in table 5.9. The confidence intervals presented are asymptotic.

Table 5.9: GEV model for the BM data set

	GEV		
	μ	σ	ξ
Estimated Parameters	7.90	0.33	-0.06
Standard Error Estimates	0.03	0.02	0.05
95% Confidence Intervals	(7.83,7.96)	(0.28,0.38)	(-0.17,0.05)

The estimated shape parameter ($\hat{\xi}$) is negative (-0.06), thus the GEV model is in fact a Weibull distribution. However $\hat{\xi}$ is very close to zero so maybe a Gumbel model would be more appropriate.

Adjusting a Gumbel model we obtained the results given in table 5.10.

Table 5.10: Gumbel model for the BM data set

	Gumbel	
	μ	σ
Estimated Parameters	7.89	0.33
Standard Error Estimates	0.03	0.02
95% Confidence Intervals	(7.82,7.95)	(0.28,0.37)

To compare the two models the AIC and BIC criterias were used and the LRT was performed.

The LRT hypotheses for this case are

$$H_0 : \xi = 0 \quad vs \quad H_1 : \xi \neq 0$$

The following table summarizes the results

Table 5.11: AIC, BIC for the adjusted models and p-value for the LRT for the BM data set

	GEV	**Gumbel**
AIC	105.75	104.74
BIC	113.99	110.23
p-value		0.32

By table 5.11 we conclude that H_0 is not rejected at any usual significance level since the p-value of the test is very high.

Moreover, the AIC and BIC values associated with the Gumbel model are lower than the ones associated with the GEV model.

A density plot is a representation of the distribution of a numeric variable, it uses a kernel density estimate to show the probability density function of the variable.

Figure 5.9 shows the kernel density estimate plot for the BM global data as well as the model estimated density. There is a good agreement between the two curves, although the empirical right tail is slightly heavier than the one given by the fitted Gumbel model.

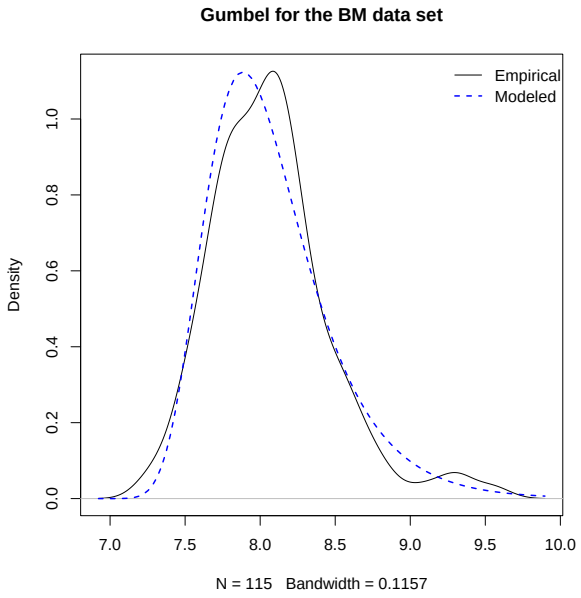

Figure 5.9: Kernel Density Estimate plot for the BM method for the global data set

In figure 5.10 the qq-plot of the model vs the data is represented, we can see that the Gumbel model is well adjusted.

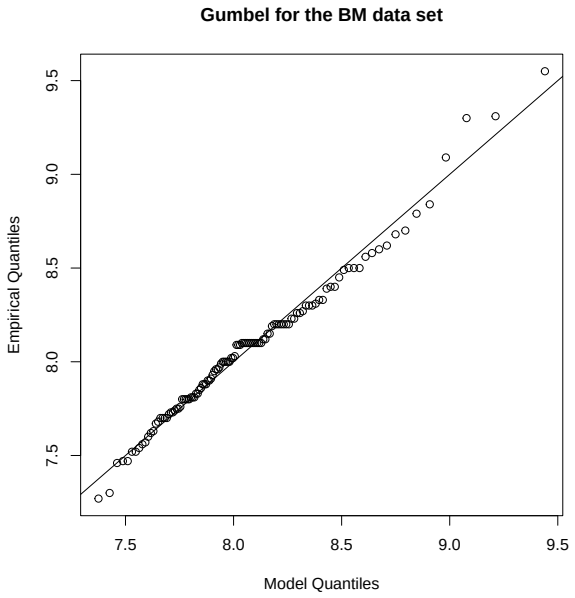

Figure 5.10: qq-plot for the BM data set

So by all the reasons reported the Gumbel model is more suitable to the data than the GEV. These results are in accordance with the ones referred in chapter 4 for the Ecuadorian coast.

5.2.1 Return Levels

In table 5.12 the return levels calculated are presented and in figure 5.11 we have the return level plot.

Table 5.12: Return levels for the BM data set

Probability	**Return Period (Year)**	**N-Year Return Level (Quantile)**	**95% CI**
0.2	5	8.38	(8.27,8.48)
0.1	10	8.62	(8.49,8.76)
0.04	25	8.93	(8.76,9.11)
0.02	50	9.16	(8.96,9.37)
0.01	100	9.39	(9.16,9.63)
0.005	500	9.92	(9.61,10.23)
0.001	1000	10.15	(9.81,10.48)

For example, the 100-year return level is 9.39, which is the level to be exceeded, in average, once in every 100 years. It has a probability of 0.01 of being exceeded in a particular year.

Figure 5.11 shows that the 95% CI is quite narrow even for large return periods which shows that the return level are estimated with small uncertainty.

However, there seems to be some tendency of the return levels of m_W to be located on the lower part of the CI. This tendency changes as the return periods increase.

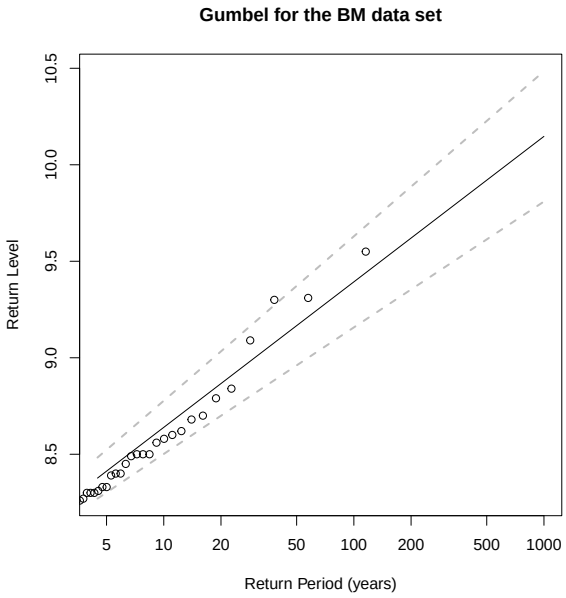

Figure 5.11: Return level plot for the BM data set

5.3 POT Method for the Global data set

In this section the POT method was applied.

The data set considered was the one initially described in section 5.1.

This method, as described before in section 3.2, fits a GPD to the exceedances above some high threshold u. To fit a GPD and estimate its parameters by ML we need to choose a threshold u.

5.3.1 Threshold Choice and Model Fitting

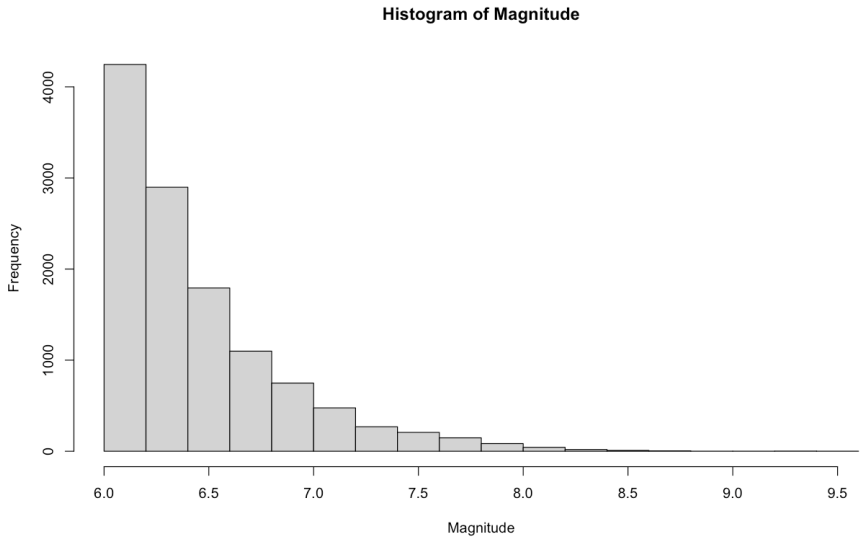

Figure 5.12: Histogram of m_W for the global data set

A histogram shows the distribution of a data set. In figure 5.12 the histogram of m_W is presented. It clearly resembles an exponential distribution.

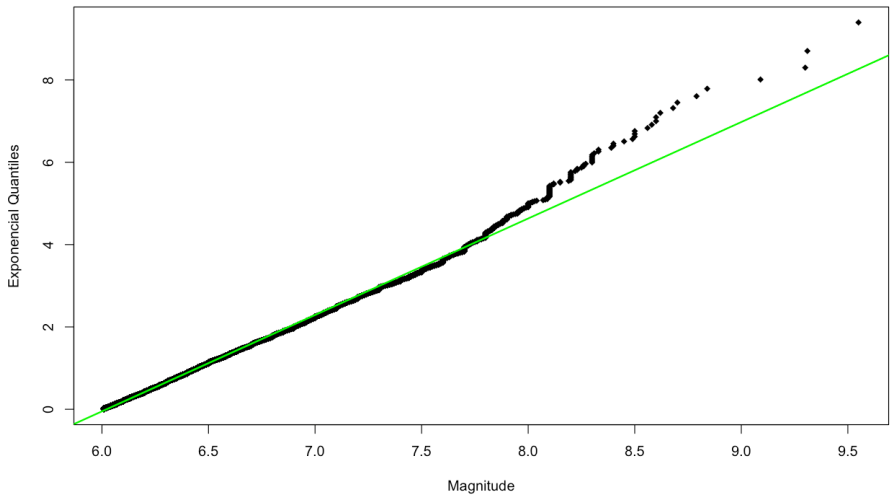

Figure 5.13: Exponential qq-plot for m_W for the global data set

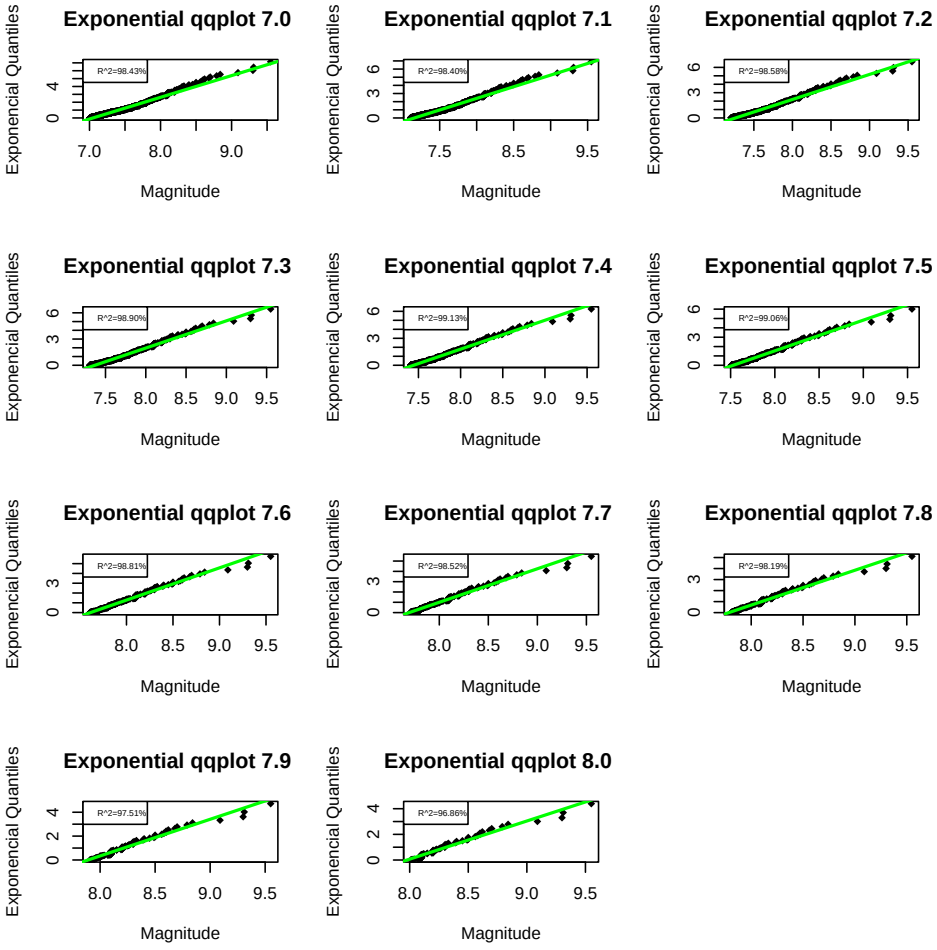

Figure 5.14: Exponential qq-plots for $u = (7.0, 7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7, 7.8, 7.9, 8.0)$

In an exponential qq-plot the exponential theoretical quantiles are plotted *versus* the empirical quantiles. The exponential qq-plot in figure 5.13 suggests that the distribution above some high u might have an

exponential tail. Moreover, in figure 5.14 we can see the exponential *qq-plots* considering a range of thresholds $u = (7.0, 7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7, 7.8, 7.9, 8.0)$.

As discussed in section 3.2.3 the empirical mean residual life function is given by

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{\max} \right\}. \tag{5.1}$$

By this method, a threshold choice should be made by locating an area in the plot where the empirical mean residual life plot is approximately linear. The interpretation of this plot can be difficult. In figure 5.15 the mean residual life plot is presented. We can see that in the vicinity of 7.2 we have an approximately linear behaviour, as well as around the value 7.6. Above 8 we have clearly a change of pattern, so this values are not as appropriate as the ones before to be a threshold.

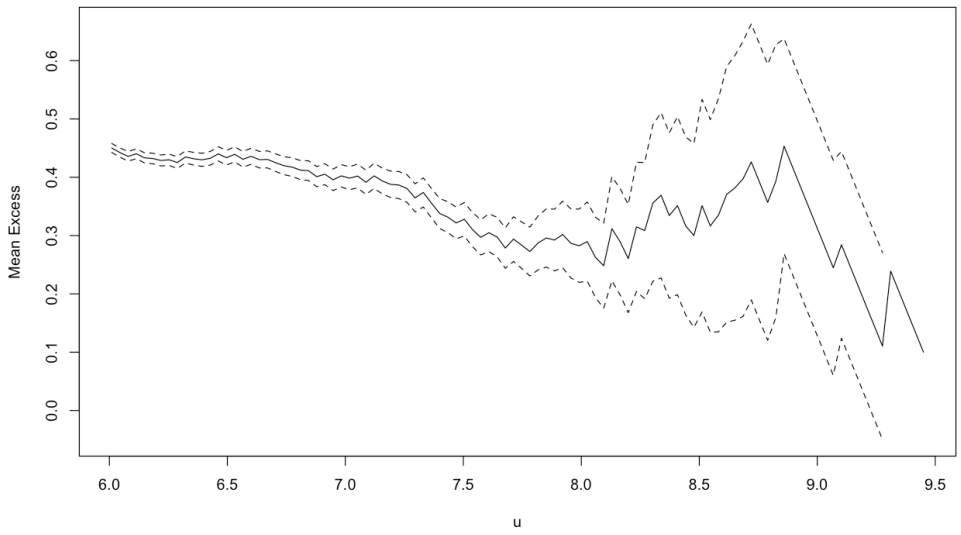

Figure 5.15: Estimated mean residual life function for the global data set

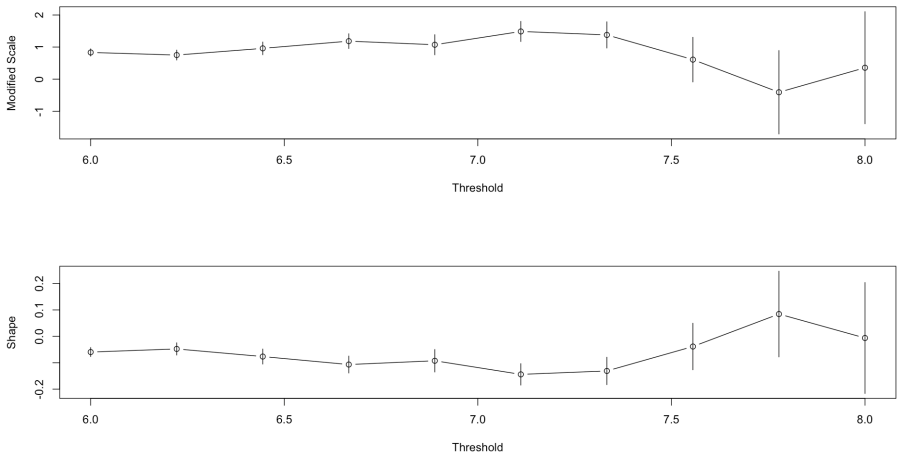

Figure 5.16: Parameter Estimates *vs* threshold for the global data set

Using the R function *gpd.fitrange*, we can calculate the parameter estimates against possible thresholds as shown in figure 5.16.

The range of threshold chosen was from 6 to 8. The height of the vertical line represents the length of the 95% confidence interval.

Figure 5.16 is in accordance with figure 5.15, above 7 we have reasonable choices for a threshold.

The methods referred in subsection 3.2.2 were applied, but the results were inconclusive due to the restriction that we applied (only considering sever earthquakes, $m_W > 6$) and to lack of variability of the global data.

Due to the difficulty to chose a threshold, for a sequence of possible threshold candidates, it was decided to fit a GPD to the exceendances above each threshold candidate considered to access stability of the GPD shape parameter. The goodness-of-fit tests for the GPD were also applied. The exponential model was fitted to perform the LRT.

Table 5.13: Results for a sequence of thresholds from 7.0 to 8.0 for the global data set

		GPD				**framed p-values (pv)**		**Exp**	**LRT**
u	$nexc$	$\hat{\sigma}$	$\hat{\xi}$	w^2	a^2	w^2	a^2	$\hat{\sigma}$	**p-value**
7.0	1263	0.46	-0.13	0.22	1.98	0.01<pv<0.025	0.001<pv<0.005	0.41	~0
7.1	989	0.46	-0.14	0.34	2.26	0.001<pv<0.005	pv<0.001	0.41	~0
7.2	787	0.45	-0.15	0.55	3.37	pv<0.001	0.001<pv<0.005	0.40	~0
7.3	620	0.44	-0.15	0.97	6.04	pv<0.001	pv<0.001	0.39	~0
7.4	518	0.39	-0.11	0.53	3.60	pv<0.001	pv<0.001	0.35	~0
7.5	406	0.36	-0.09	0.50	3.37	pv<0.001	pv<0.001	0.33	0.04
7.6	311	0.33	-0.06	0.34	2.34	0.001<pv<0.005	pv<0.001	0.31	0.25
7.7	233	0.30	-0.01	0.15	1.01	0.05<pv<0.1	0.025<pv<0.005	0.30	0.83
7.8	164	0.29	0.00	0.08	0.55	0.25<pv<0.5	0.25<pv<0.5	0.30	0.97
7.9	111	0.32	-0.03	0.20	1.19	0.01<pv<0.025	0.01<pv<0.025	0.31	0.73
8.0	80	0.31	-0.01	0.27	1.75	0.001<pv<0.005	0.001<pv<0.005	0.30	1

In table 5.13, for each u (chosen threshold) we have the number of exceedances ($nexc$), the estimation of the parameters for the GPD (GPD - $\hat{\sigma}$ and $\hat{\xi}$) and for the Exponential (Exp - $\hat{\sigma}$), the test statistics for the goodness-of-fit tests for the GPD (Crámer-von Mises and Anderson-Darling, w^2 and a^2, respectively) and the framed p-values associated, as well as the p-value for the LRT.

When performing the Crámer-von Mises and the Anderson-Darling tests the null hypothesis we are testing is

$$H_0 : x_{(1)}, x_{(2)}, ..., x_{(nexc)} \text{ comes from a generalized Pareto distribution}$$

and the alternative hypothesis is

$$H_1 : x_{(1)}, x_{(2)}, ..., x_{(nexc)} \text{ does not come from a generalized Pareto distribution}$$

The test statistics, w^2 and a^2 were calculated as in (3.26) and (3.25), respectively. To frame the p-values (pv) associated we consulted the table for the critical values of w^2 and a^2 for the case of σ and ξ both unknown (Choulakian and Stephens, 2001). Note that the GPD in this article has the following distribution function $F(x) = 1-(1-\frac{kx}{a})^{\frac{1}{k}}$, in which a is the scale parameter and k is the shape parameter. This parameterization is not the same as the one referred in subsection 3.2.1, so when looking at the table with the critical values we will consider $k = -\xi$.

For the LRT we are testing

$$H_0 : \xi = 0 \quad vs \quad H_1 : \xi \neq 0.$$

Below $u = 7.5$ we can see that the GPD does not fit the data. According to the p-values associated with the Cramér-von Mises and the Anderson-Darling tests we reject H_0, thus our data does not come from a GPD, also when adjusting an Exponential model and using the LRT we have that $\xi \neq 0$ as the p-values are approximately 0.

When considering $u = 7.6$ as a threshold, by the Cramér-von Mises and the Anderson-Darling tests, we can say that our data does not come from a GPD. However $\hat{\xi}$ is close to zero and when adjusting an Exponential model and performing the LRT we do not reject H_0 due to the high p-value. Thus there is no evidence that ξ is not equal to 0. By this we selected 7.6 as the final threshold. We were left with 311 exceedances and the Exponential model was the chosen model.

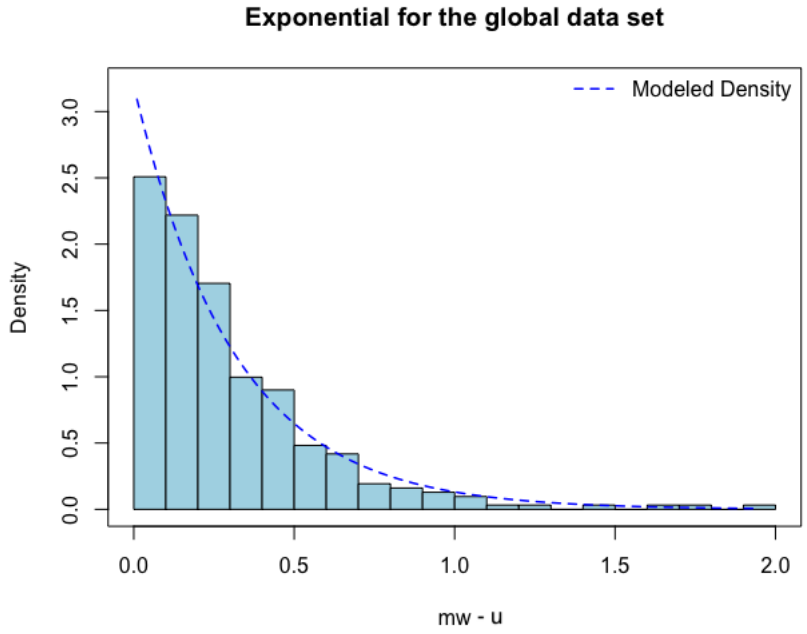

Figure 5.17: Histogram for the POT method for the excesses above the threshold u (global data set)

Figure 5.17 shows the histogram of the global data set and the probability density function of the estimated model. In this situation both (sample and fitted) right tails are very similar.

In figure 5.18 the qq-plot of the model vs the data is represented, we can see that the Exponential model is well adjusted. In table 5.14, the results for $u = 7.6$ are summarized.

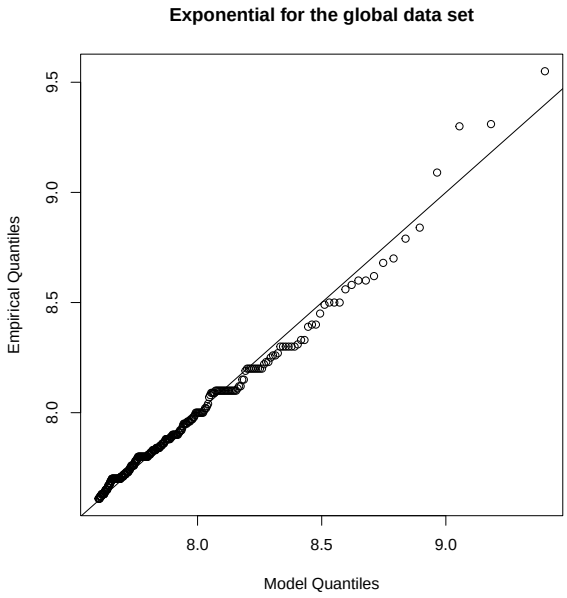

Figure 5.18: qq-plot for the POT method for the global data set

Table 5.14: Exponential Model for $u = 7.6$ for the global data set

	Exponential
	σ
Estimated Parameter	0.31
Standard Error Estimate	0.02
95% Confidence Interval	(0.28,0.35)

An Exponential distribution fits well to our data. Our conclusion is in order with the one obtained in chapter 4 for the Ecuadorian coast.

In figure 5.19 the location of the earthquakes with $m_W > 7.6$ classified by m_W and $depth$ is represented. Most of the earthquakes occurred in the limits of the Pacific Plate, the North American Plate and the South American Plate.

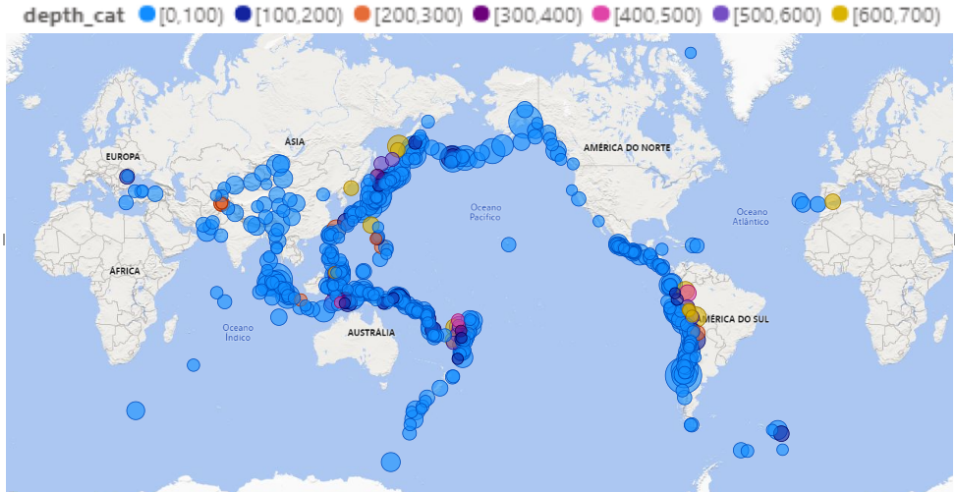

Figure 5.19: Location of the earthquakes with $m_W > 7.6$ represented by m_W and $depth$ for the global data set

5.3.2 Return Levels

With the model adjusted and tested, some return levels were calculated (table 5.15).

Table 5.15: Return levels for the POT method for the global data

m	**Return Period (Year)**	**N-Year Return Level (Quantile)**	**95% CI**
524	5	8.41	(8.33,8.50)
1048	10	8.63	(8.52,8.63)
2619	25	8.91	(8.77,9.07)
5237	50	9.14	(8.97,9.31)
10475	100	9.35	(9.16,9.55)
52374	500	9.86	(9.61,10.11)
104748	1000	10.08	(9.80,10.35)

In this case, due to the fact that the sample of excesses generally does not have the same number of observations per year, the value of m is calculated as $\frac{n}{nyears} \times N$ (see subsection 3.2.5). For our data the average number of earthquakes per year is $\frac{12046}{115} = 104.75$.

So, for instance, in a 5 year period, 1 in 524 severe earthquakes is expected to surpass the magnitude 8.41. An earthquake with $m_W = 9.14$ can occur, on average, once in 5237 earthquakes or approximately once in every 50 years.

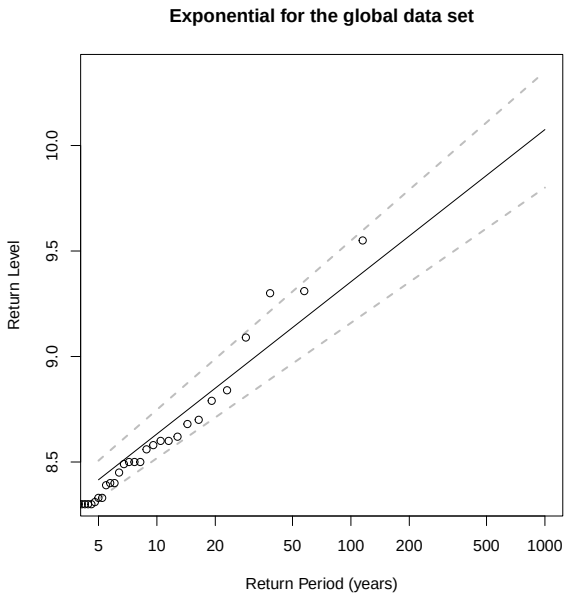

Figure 5.20: Return Level plot for the POT method for the global data set

In figure 5.20 the return level plot is presented. The figure shows that the 95% CI is quite narrow even for large return periods which shows that the return levels are estimated with small uncertainty.

5.4 Exploratory Analysis for the Japan data set

Considering the data referred in section 5.1, we will study a subset, the Japan area, due to its high seismic activity.

We will contemplate the indicated area in figure 5.21 (USGS-Catalog, 2023), restricting the latitude and the longitude to the intervals $]30.145; 45.383[$ and $]129.551; 148.007[$, respectively. We were left with 956 earthquakes. We will name this set the Japan data set.

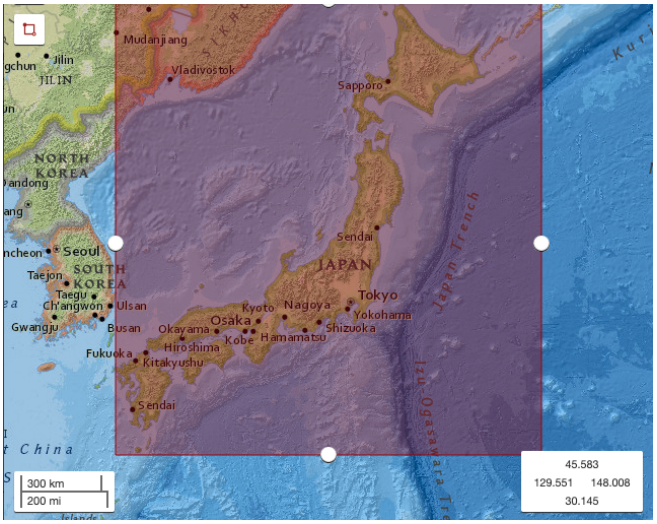

Figure 5.21: Area being considered - Japan Area

Proceeding in the same way as in section 5.1 we obtain the following table.

Table 5.16: m_w by intervals of 0.5 for the Japan data set

(6,6.5)	**[6.5,7)**	**[7,7.5)**	**[7.5,8)**	**[8,8.5)**	**[8.5,9)**	**[9,9.5)**	**[9.5,10]**
643	199	74	30	8	1	1	0
67%	21%	8%	3%	0.8%	0.1%	0.1%	0%

By table 5.16, we can see that 88.08% of the earthquakes have m_W between 6 and ≈ 7 and only 10 observations belong to the interval [8,9.5). There is no record for an earthquake with m_W above 9.5.

Taking in to account the years that the earthquakes occurred, the data can be summarized as follows:

Table 5.17: Number of earthquakes for the Japan data set by intervals of 19 years

[1904,1923)	**[1923,1942)**	**[1942,1961)**	**[1961,1980)**	**[1980,1999)**	**[1999,2018]**
67	195	170	201	130	193

In table 5.17, in almost all the intervals considered we see that the number of earthquakes occurred is very similar, except for the interval [1904,1923) where it is approximately half.

Table 5.18: Summary statistics of m_w by intervals of 19 years for the Japan data set

$year$	**Min**	1^{st} **Qu.**	**Median**	**Mean**	3^{rd} **Qu.**	**Max**
[1904,1923)	6.01	6.29	6.46	6.59	6.83	7.71
[1923,1942)	6.01	6.15	6.31	6.49	6.67	8.50
[1942,1961)	6.01	6.13	6.29	6.42	6.56	8.30
[1961,1980)	6.01	6.13	6.30	6.46	6.65	8.20
[1980,1999)	6.01	6.14	6.29	6.43	6.62	8.26
[1999,2018]	6.01	6.12	6.27	6.42	6.63	9.09

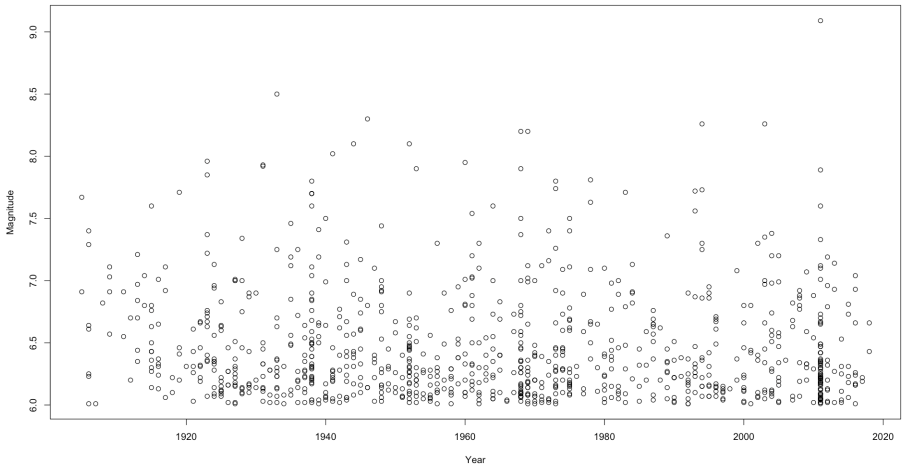

Figure 5.22: m_W through the years for the Japan data set

The moment magnitude, m_W, has the same behaviour in each interval of table 5.18. However in the first interval ([1904,1923)) the summary statistics are slightly different when compared to the remaining ones. In figure 5.22, we can not identify any trend for the value of m_W as the time evolves.

Table 5.19: Summary statistics of m_w and *depth* for the Japan data set

Variable	**Min**	1^{st} **Qu.**	**Median**	**Mean**	3^{rd} **Qu.**	**Max**
m_W	6.010	6.140	6.310	6.458	6.650	9.090
$depth$	7.60	15.00	30.00	55.74	40.00	610.00

Table 5.20: Summary statistics of m_w by *depth* intervals for the Japan data set

$depth$	**Min**	1^{st} **Qu.**	**Median**	**Mean**	3^{rd} **Qu.**	**Max**	**n**
[0,100)	6.01	6.14	6.30	6.45	6.63	9.09	875
[100,200)	6.02	6.14	6.29	6.45	6.69	7.81	25
[200,300)	6.02	6.13	6.35	6.38	6.48	6.99	6
[300,400)	6.03	6.26	6.45	6.56	6.83	7.40	20
[400,500)	6.04	6.23	6.76	6.70	7.09	7.30	12
[500,600)	6.08	6.26	6.66	6.62	6.89	7.30	16
[600,700]	6.8	7.035	7.27	7.27	7.51	7.74	2

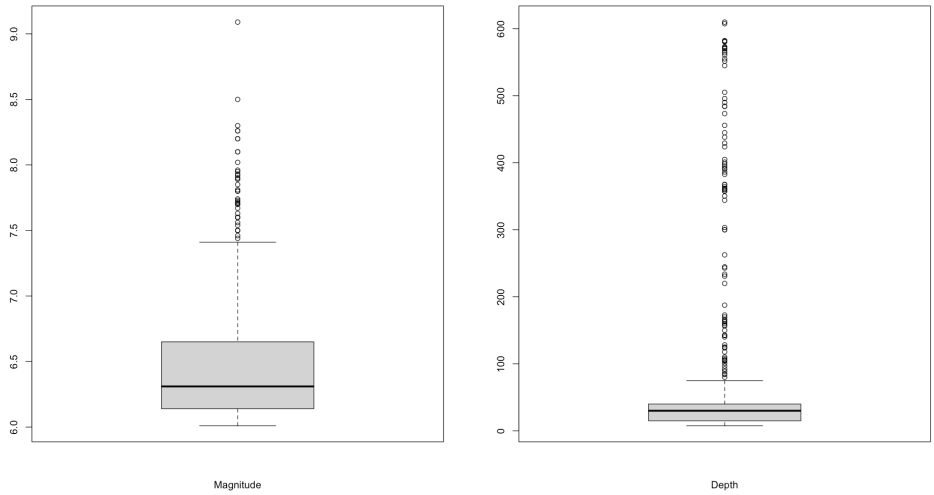

Figure 5.23: Boxplots of m_W and $depth$ for the Japan data set

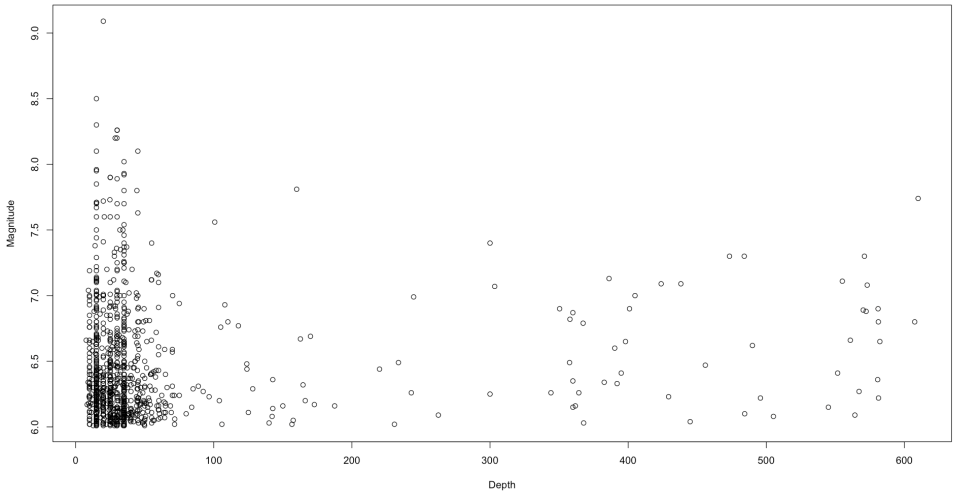

Figure 5.24: m_W vs $depth$ for the Japan data set

Observing tables 5.19, 5.20 and figures 5.23, 5.24 we can see the behaviour of the variables m_W and $depth$.

Again, most of the earthquakes have m_W between 6 and 7. Approximately 92% of the earthquakes' epicenters occurred at depths between 0 and 100 km.

In figure 5.26 the Japan data set is represented by its coordinates. Doing a parallel with the map of the tectonic plates (5.25), we conclude again that most of the earthquakes occur in the fault zone. This area is situated in the limits of the Eurasian Plate, the North American Plate, the Philippine Plate and the Pacific Plate.

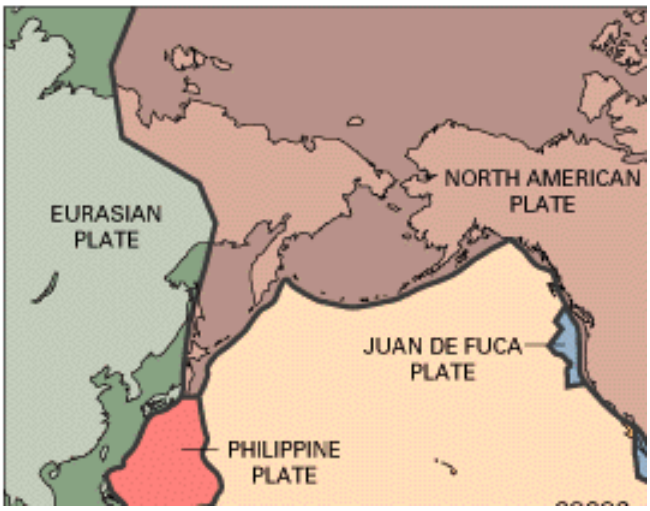

Figure 5.25: Tectonic Plates Map for the Japan Area

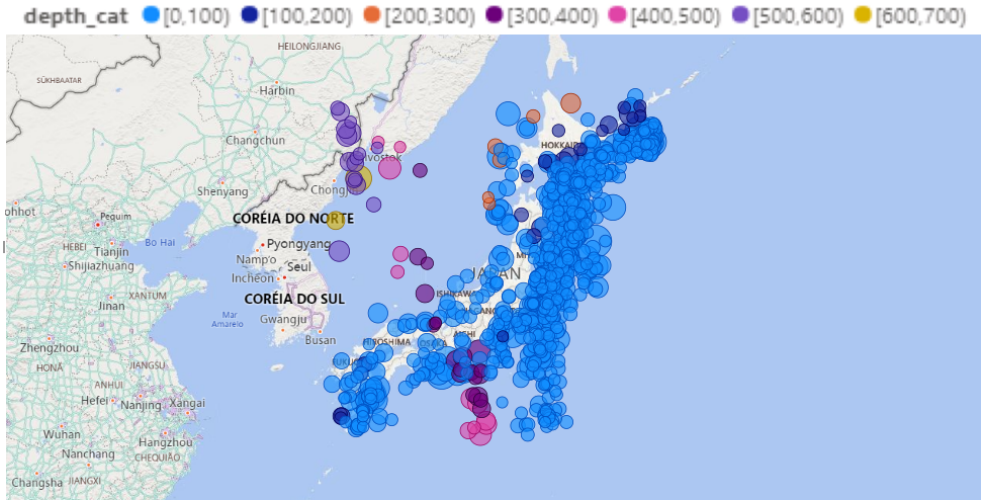

Figure 5.26: Location of the earthquakes for the Japan data set

5.5 BM Method for the Japan data set

For the Japan data set we selected, for each year, the earthquake with the maximum value of m_W. Our study contemplated the period between 1911 and 2018. Note that in 1932 there were two earthquakes candidates to the maxima of 1932 with the same value of m_W, however we considered the one that occurred first ignoring the one that occurred after. This new data set was called the BM Japan data set.

Dividing m_W in intervals of amplitude 0.5 we have

Table 5.21: m_w by intervals of amplitude 0.5 for the BM Japan data set

(6,6.5)	**[6.5,7)**	**[7,7.5)**	**[7.5,8)**	**[8,8.5)**	**[8.5,9)**	**[9,9.5)**	**[9.5,10]**
16	35	32	15	8	1	1	0
15%	32%	30%	14%	7%	1%	1%	0%

By table 5.21, 77% of the the data are situated in the interval (6,7.5).

In figure 5.27 a pattern for the values of m_W can not be seen. The highest value of m_W is approximately

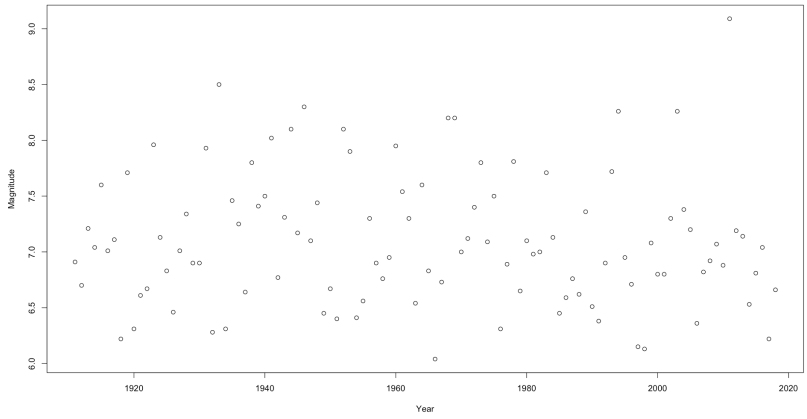

Figure 5.27: Earthquakes represented by m_W through the years for the BM Japan data set

9 and the earthquake associated with this value occurred earlier in the 2000s.

In table 5.22 the summary statistics for m_W and $depth$ are presented and in table 5.23 we have the moment magnitude summary for each $depth$ interval.

Table 5.22: Summary statistics of m_w and $depth$ for the BM Japan data set

Variable	**Min**	1^{st} **Qu.**	**Median**	**Mean**	3^{rd} **Qu.**	**Max**
m_W	6.04	6.67	7.01	7.10	7.42	9.09
$depth$	9.20	15.07	30.00	91.33	49.52	582.30

Table 5.23: Summary statistics of m_w by *depth* intervals for the BM Japan data set

$depth$	**Min**	1^{st} **Qu.**	**Median**	**Mean**	3^{rd} **Qu.**	**Max**	**n**
[0,100)	6.04	6.67	7.04	7.13	7.50	9.09	91
[100,200)	6.36	6.56	6.76	6.98	7.29	7.81	3
[300,400)	6.82	6.88	6.99	6.98	7.09	7.13	4
[400,500)	6.62	6.86	7.09	7.00	7.20	7.30	3
[500,600)	6.41	6.77	6.89	6.87	6.99	7.30	7

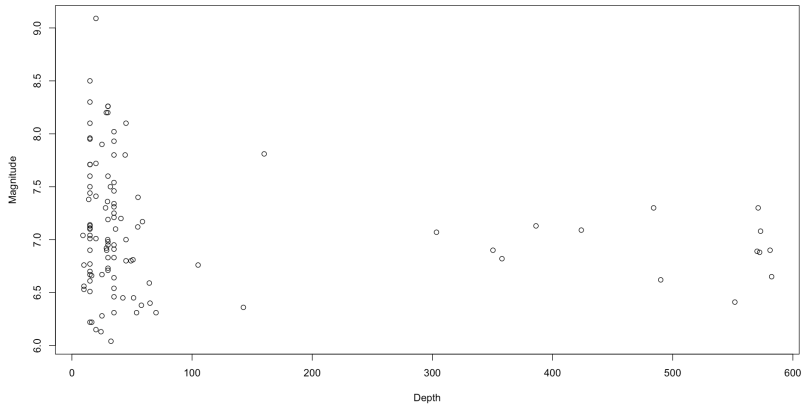

Figure 5.28: Earthquakes represented by m_W and $depth$ for the BM Japan data set

By tables 5.22 and 5.23 and figure 5.28 we can conclude that, for most of the observations of the BM Japan data set, the epicenter's depth was situated between 1 and 100 km, including the earthquake with the highest value of m_W.

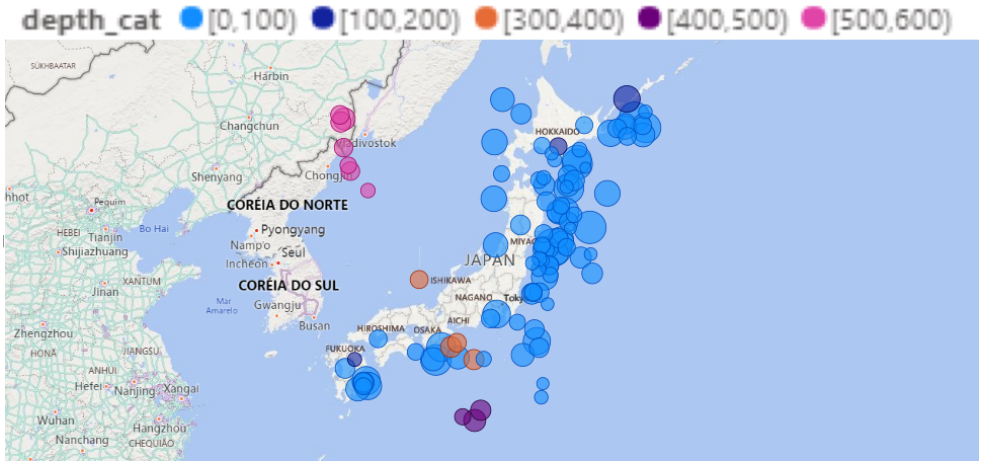

Figure 5.29: Location of the earthquakes for the BM Japan data set

For the BM Japan data set we were left with most of the locations of the earthquakes between the limit of the North American Plate and the Pacific Plate (figure 5.29).

Applying the BM method we obtained the results presented in table 5.24.

Table 5.24: GEV model for the BM Japan data set

	GEV		
	μ	σ	ξ
Estimated Parameters	6.84	0.49	-0.06
Standard Error Estimates	0.05	0.04	0.07
95% Confidence Intervals	(6.73,6.95)	(0.42,0.57)	(-0.21,0.09)

Being $\hat{\xi}$ negative (-0.06), the GEV model would be a Weibull distribution, but $\hat{\xi}$ is very close to zero so maybe a Gumbel model would be more appropriate.

A Gumbel model was then adjusted to the data and the results are summarized in table 5.25.

Table 5.25: Gumbel model for the BM Japan data set

	Gumbel	
	μ	σ
Estimated Parameters	6.82	0.49
Standard Error Estimates	0.05	0.04
95% Confidence Intervals	(6.73,6.92)	(0.41,0.49)

Comparing the two models with the AIC and BIC criterias and performing the LRT (the hypotheses are: $H_0 : \xi = 0 \quad vs \quad H_1 : \xi \neq 0$) we have

The p-value presented in table 5.26 is very high, so H_0 is not rejected at any usual significance level.

Table 5.26: AIC, BIC for the adjusted models and p-value for the LRT for the Japan BM data set

	GEV	**Gumbel**
AIC	188.11	186.72
BIC	196.16	192.08
p-value		0.435

In table 5.26 we also see that the AIC and BIC values associated with the Gumbel model are lower than the ones associated with the GEV model.

In figure 5.30 shows the kernel density estimate plot for the BM Japan data as well as the model estimated density. There is a good agreement between the two curves, although the empirical right tail is slightly lighter than the one given by the fitted Gumbel model.

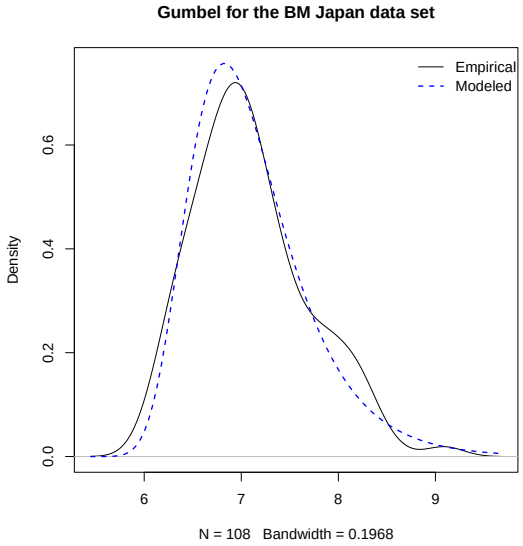

Figure 5.30: Kernel Density Estimate plot for the BM Japan data set

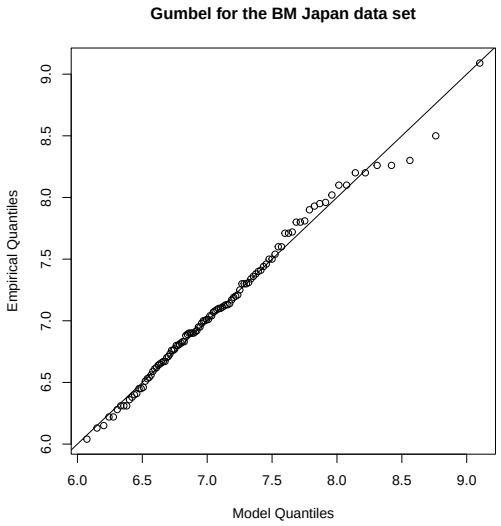

Figure 5.31: qq-plot for the BM Japan data set

In figure 5.31 the qq-plot of the model vs the data is represented, we can see that the Gumbel model fits well to the data.

Thus, by all the reasons above, we conclude that the Gumbel model fits better to the data than the GEV. These results are in agreement with the ones referred in chapter 4 for the Ecuadorian coast.

5.5.1 Return Levels

The return levels calculated are presented in table 5.27.

Table 5.27: Return levels for the BM Japan data set

Probability	**Return Period (Year)**	**N-Year Return Level (Quantile)**	**95% CI**
0.2	5	7.55	(7.39,7.72)
0.1	10	7.92	(7.71,8.13)
0.04	25	8.38	(8.10,8.65)
0.02	50	8.72	(8.40,9.04)
0.01	100	9.06	(8.69,9.43)
0.005	500	9.84	(9.36,10.32)
0.001	1000	10.18	(9.65,10.71)

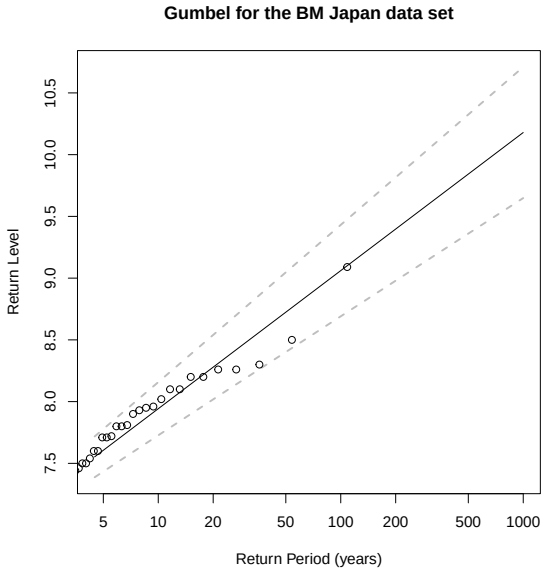

Figure 5.32: Return Level plot for the BM Japan data set

For example, the 25-year return level is 8.38, which is the level to be exceeded, in average, once in every 25 years. It has a probability of 0.04 of being exceeded in a particular year.

Figure 5.32 shows the return level plot. We can see that the 95% CI is quite narrow even for large return periods which shows that the return level are estimated with small uncertainty. However, there seems to be some tendency of the return levels of m_W to be located on the upper part of the CI. This tendency changes as the return periods increase.

5.6 POT Method for the Japan data set

In this section the POT method was applied to the Japan data set.

We want to model with a generalized Pareto distribution the exceedances above some high threshold u in this region.

5.6.1 Threshold Choice and Model Fitting

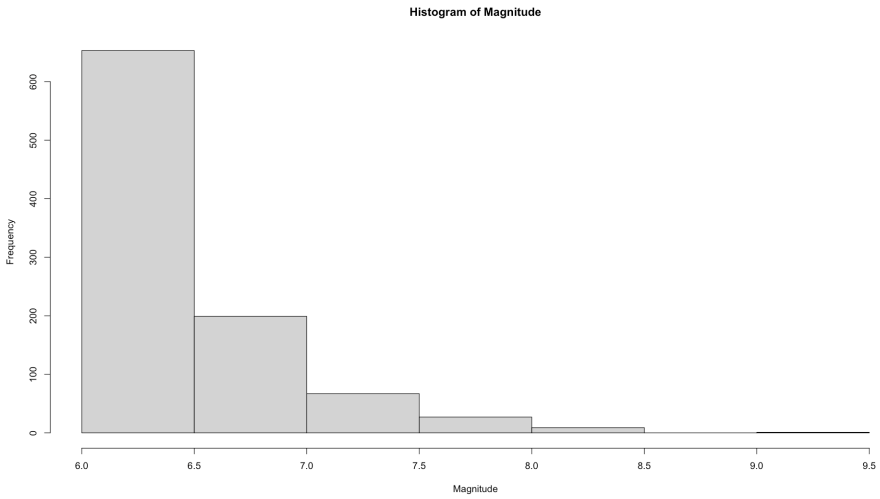

Figure 5.33: Histogram of m_W for the Japan data set

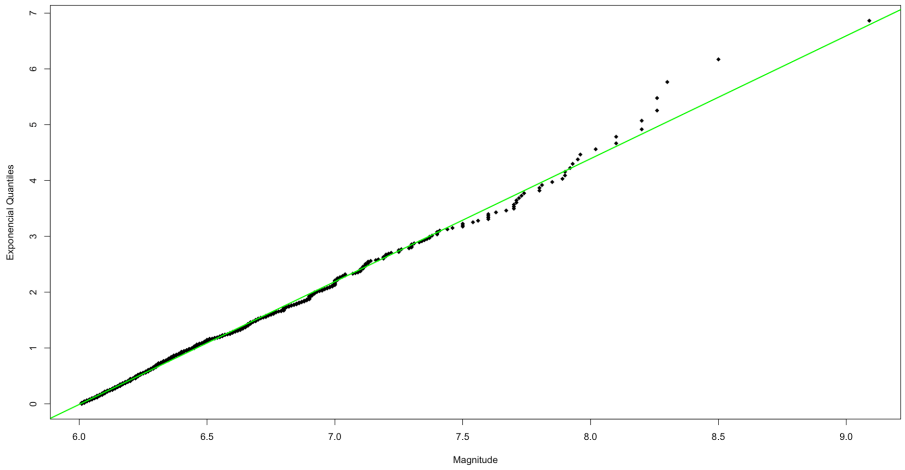

Figure 5.34: Exponential qq-plot for m_W for the Japan data set

Figure 5.33 shows the histogram of m_W which highly resembles an exponential distribution.

The exponential *qq-plot*, presented in figure 5.34, suggests that the distribution has an exponential tail.

Thus, both figures 5.33 and 5.34 are in agreement.

Moreover, in figure 5.35 we can see the exponential *qq-plots* considering a range of thresholds $u = (6.5, 6.6, 6.7, 6.8, 6.9, 7.0, 7.1, 7.2, 7.3, 7.4, 7.5)$.

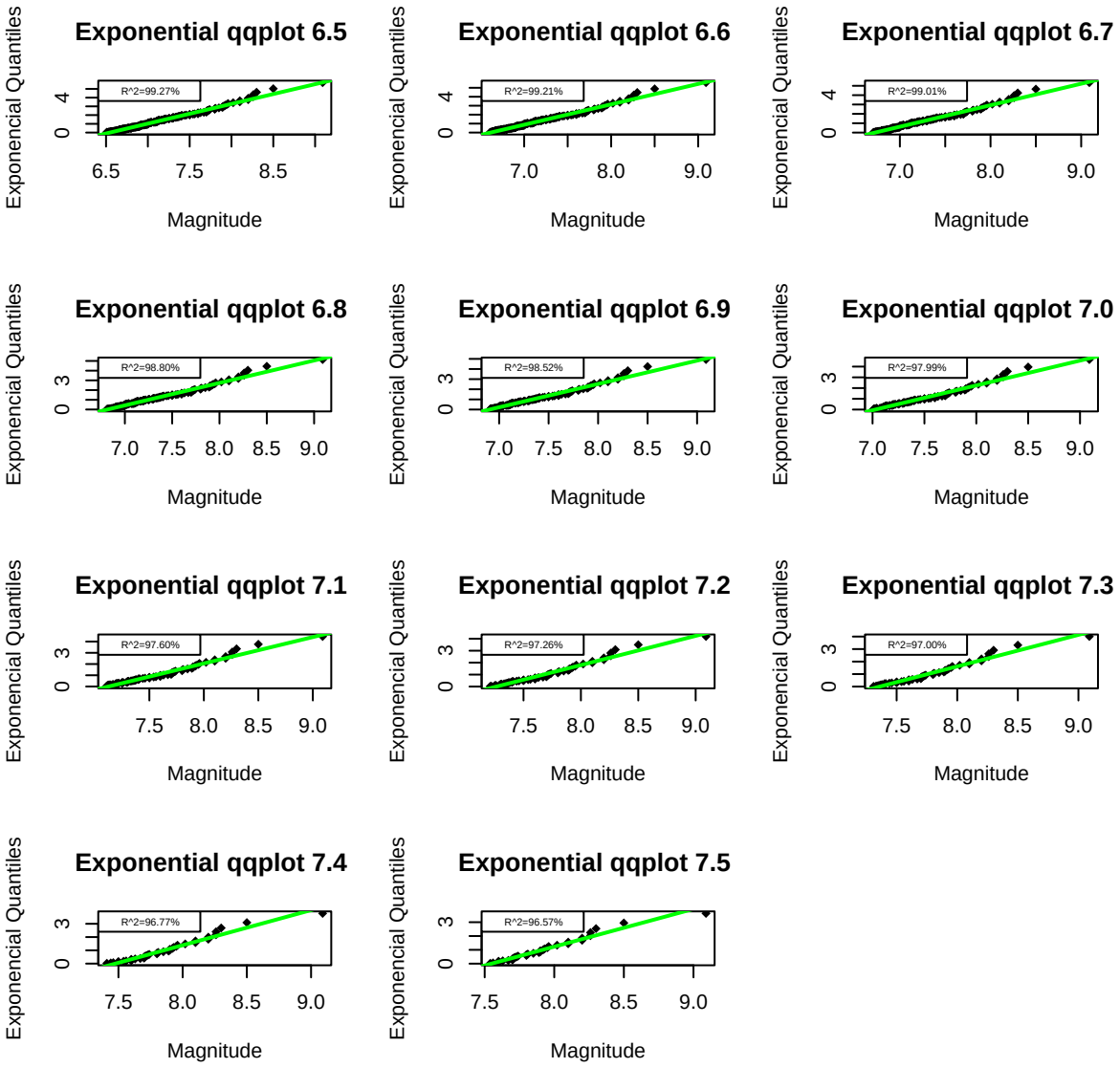

Figure 5.35: Exponential qq-plots for $u = (6.5, 6.6, 6.7, 6.8, 6.9, 7.0, 7.1, 7.2, 7.3, 7.4, 7.5)$

By figure 5.36 we can see that between 6.7 and 6.9 we have an approximately linear behaviour, as well between 7.2 and 7.3. Above 8 we see clearly a change of pattern, so this values should be not considered as potential thresholds.

With function *gpd.fitrange* from the R *package ismev*, we obtained figure 5.37. The range of thresholds considered was from 6 to 8.

Figure 5.37 is in accordance with figure 5.36, from 6.5 to 7.4 we have reasonable choices for a threshold; a choice above 7.5 will lead to high variance due to the lack of observations.

The methods referred in subsection 3.2.2 were again applied, but as before the results were inconclusive due to the lack of variability of the Japan data and to the restriction of only considering severe earthquakes with m_W larger than 6.

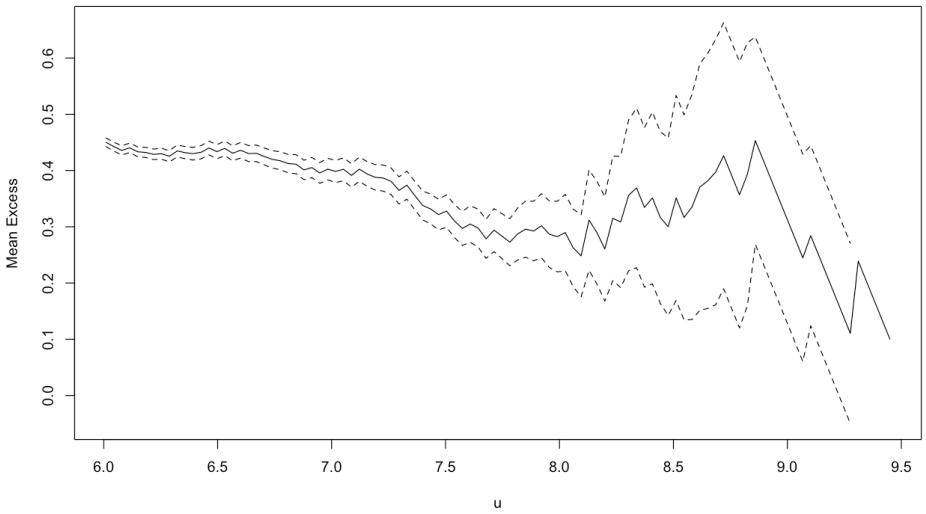

Figure 5.36: Estimated mean residual life function for the Japan data set

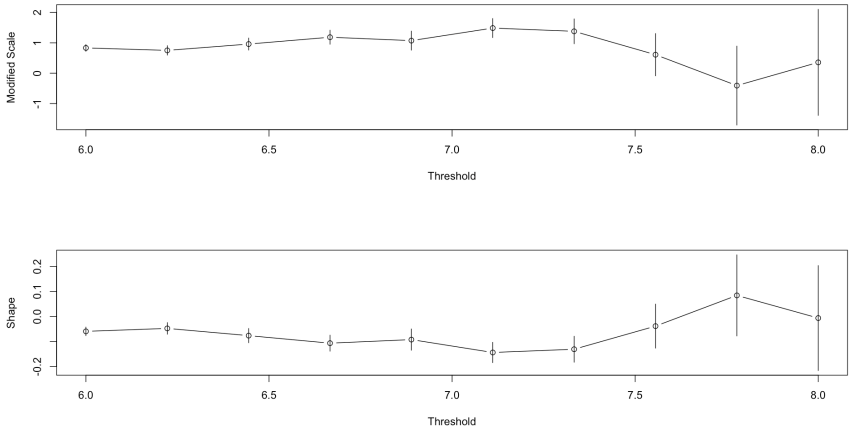

Figure 5.37: Parameter Estimates *vs* threshold for the Japan data set

Due to the difficulty to chose a threshold, for a sequence of possible threshold candidates, it was decided to fit a GPD to the exceendances above each threshold candidate considered to access stability of the GPD shape parameter. The goodness-of-fit tests for the GPD were also applied. The exponential model was fitted to perform the LRT.

In table 5.28, for each u (chosen threshold) we have the number of exceedances ($nexc$), the estimation of the parameters for the GPD (GPD - $\hat{\sigma}$ and $\hat{\xi}$) and for the Exponential (Exp - $\hat{\sigma}$), the test statistics for the goodness-of-fit tests for the GPD (Crámer-von Mises and Anderson-Darling, w^2 and a^2, respectively) and the framed p-values (pv) associated, as well as the p-value for the LRT.

When performing the Crámer-von Mises and the Anderson-Darling tests we are testing

$$H_0 : x_{(1)}, x_{(2)}, ..., x_{(nexc)} \text{ comes from a generalized Pareto distribution}$$

$$vs$$

$$H_1 : x_{(1)}, x_{(2)}, ..., x_{(nexc)} \text{ does not come from a generalized Pareto distribution}$$

Table 5.28: Results for a sequence of thresholds from 6.5 to 7.5

		GPD				**framed p-values (pv)**		**Exp**	**LRT**
u	$nexc$	$\hat{\sigma}$	$\hat{\xi}$	w^2	a^2	w^2	a^2	$\hat{\sigma}$	**p-value**
6.5	303	0.55	-0.13	0.17	1.31	0.025<pv<0.05	0.01<pv<0.025	0.49	0.02
6.6	263	0.49	-0.08	0.06	0.48	0.25<pv<0.5	0.25<pv<0.5	0.45	0.21
6.7	207	0.51	-0.11	0.11	0.74	0.1<pv<0.25	0.1<pv<0.25	0.46	0.12
6.8	170	0.49	-0.10	0.12	0.82	0.1<pv<0.25	0.1<pv<0.25	0.44	0.22
6.9	138	0.47	-0.08	0.11	0.63	0.1<pv<0.25	0.1<pv<0.25	0.43	0.38
7.0	104	0.52	-0.15	0.06	0.37	0.25<pv<0.5	$pv > 0.5$	0.45	0.14
7.1	85	0.51	-0.16	0.07	0.60	0.25<pv<0.5	0.1<pv<0.25	0.44	0.16
7.2	65	0.56	-0.22	0.06	0.39	0.25<pv<0.5	0.025<pv<0.005	0.46	0.06
7.3	54	0.53	-0.21	0.07	0.42	0.25<pv<0.5	$pv > 0.5$	0.44	0.09
7.4	43	0.54	-0.23	0.14	0.84	0.05<pv<0.1	0.1<pv<0.25	0.44	0.08
7.5	37	0.47	-0.20	0.12	0.86	0.1<pv<0.25	0.001<pv<0.005	0.30	0.18

The test statistics, w^2 and a^2 were calculated as in (3.26) and (3.25), respectively. To frame the p-values (pv) associated we consulted the table for the critical values of w^2 and a^2 for the case of σ and ξ both unknown (Choulakian and Stephens, 2001). Note again that the GPD in this article has the following distribution function $F(x) = 1-(1-\frac{kx}{a})^{\frac{1}{k}}$, in which a is the scale parameter and k is the shape parameter. This parameterization is not the same as the one referred in subsection 3.2.1, so when looking at the table with the critical values we will consider $k = -\xi$.

For the LRT we are testing

$$H_0 : \xi = 0 \quad vs \quad H_1 : \xi \neq 0.$$

In table 5.28 we see that between 6.6 and 6.9 the estimated shape parameter is between -0.08 and -0.11, it is very stable. The number of exceedances is between 138 and 263 (which represent approximately between 14.4% and 27.5% of the data. For this set of thresholds, by the GPD goodness-of-fit tests, we can say that there is no statistical evidence that our data does not come from a GPD. However, for these thresholds $\hat{\xi}$ are close to zero and by the LRT we see that we do not reject H_0 ($H_0 : \xi = 0$) for any usual significance level, since the p-values associated with the tests are high, which means that the GPD model is reduced to an exponential model.

Having the reasons above in consideration and due to the fact that it is the one with less exceedances above from the set chosen above, $u = 6.9$ was the chosen threshold. Also note that figures 5.35 and 5.37 and the fact that the 90% estimated quantile is 7 support this choice. When considering $u = 6.9$ we have 138 exceedances, as said before approximately 14.4% of the data. This may seem to much but we need to have in mind that our analysis is only for the severe earthquakes ($m_W > 6$), so a low value of u is expected.

Figure 5.38 shows the histogram of the Japan data set and the probability distribution function of the estimated model. In this situation both (sample and fitted) right tails are very similar.

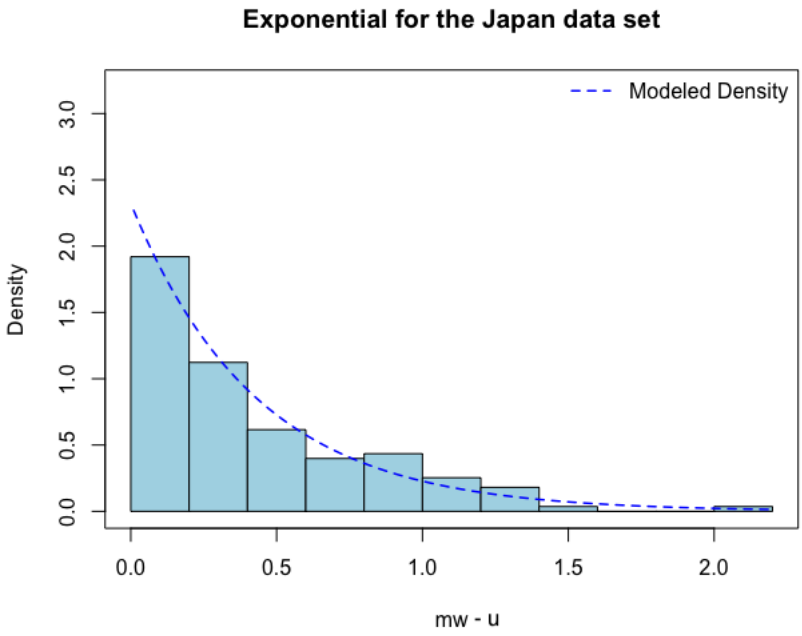

Figure 5.38: Histogram and model density function for the POT method for the excesses above the threshold u (Japan data set)

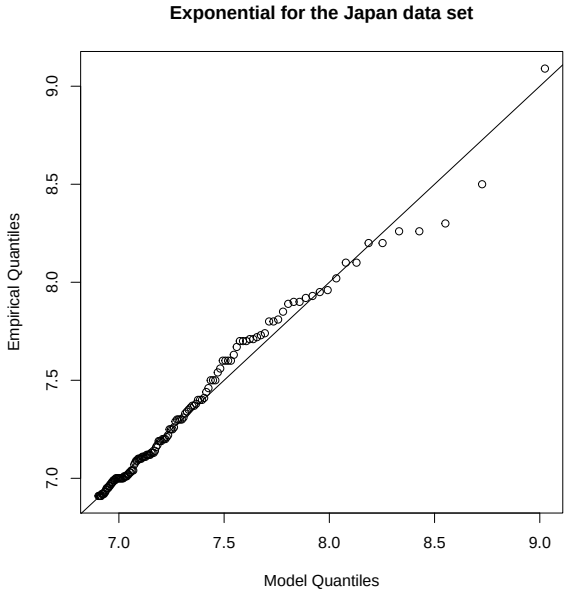

Figure 5.39: qq-plot for the POT method for the Japan data set

In figure 5.39 the qq-plot of the model vs the data is represented, we can see that the Exponential model fits well to the data.

The results for $u = 6.9$ are summarized in table 5.29.

Table 5.29: Exponential model for $u = 6.9$

	Exponential
	σ
Estimated Parameter	0.43
Standard Error Estimate	0.04
95% Confidence Interval	(0.36,0.50)

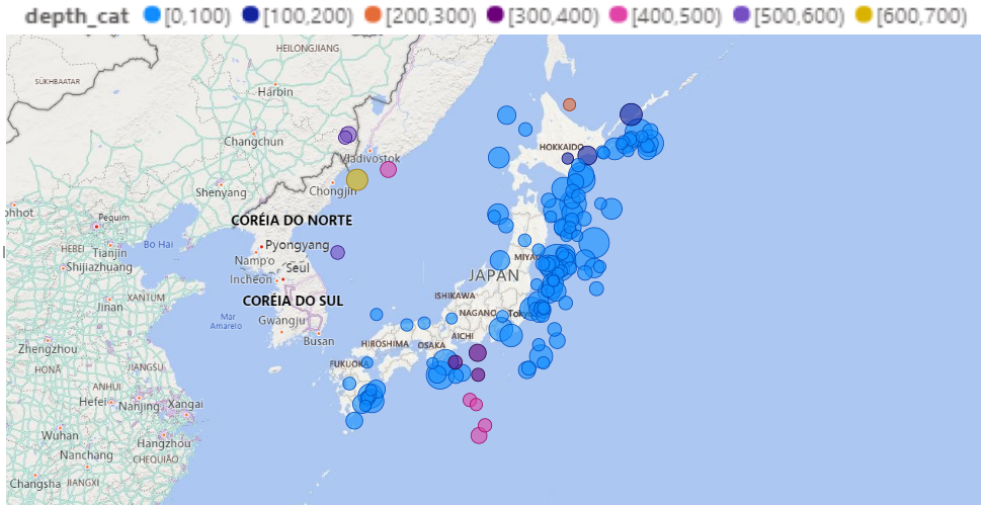

Figure 5.40: Location of the earthquakes with $m_W > 6.9$ represented by m_W and $depth$ for the Japan data set

In figure 5.40, we see that most of the earthquakes are located between the limit of the North American Plate and the Pacific Plate (see also figure 5.4).

The Exponential model is more suitable to the data. Our conclusions are the same as the ones referred in chapter 4 for the Ecuadorian Coast and are in agreement with the ones in section 5.5.

5.6.2 Return Levels

With the model adjusted and tested, some return levels were calculated.

Table 5.30: Return levels for the Japan data set for the POT method

m	**Return Period (Year)**	**N-Year Return Level (Quantile)**	**95% CI**
42	5	7.68	(7.55,7.80)
84	10	7.97	(7.79,8.15)
210	25	8.37	(8.12,8.61)
419	50	8.67	(8.37,8.96)
839	100	8.96	(8.62,9.31)
4193	500	9.66	(9.20,10.12)
8386	1000	9.96	(9.45,10.47)

For example, an earthquake with $m_W = 7.97$ can occur, on average, once in 84 earthquakes or approximately once in every 10 years.

In figure 5.41 the return level plot is presented. Again, the 95% CI is quite narrow even for large return periods which shows that the return level are estimated with small uncertainty. However, there seems to be some tendency of the return values of m_W to be located on the upper part of the CI. This tendency changes as the return periods increase.

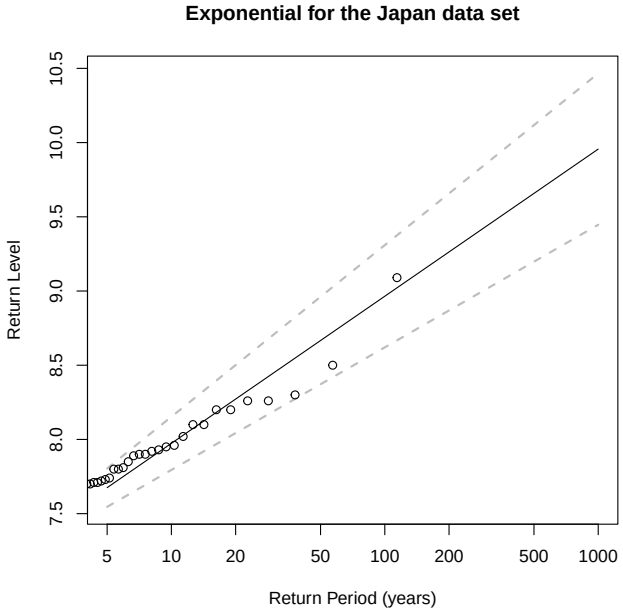

Figure 5.41: Return Level plot for the POT method for the Japan data set

6 Comments, Conclusions and Future Work

The main goal of this master thesis was to model high magnitude earthquakes. The ISC-GEM Catalogue consisted in 48606 earthquakes that occurred between 04-04-1904 and 31-12-2018. However due to the main purpose of this study we only considered the earthquakes with m_W above 6 (severe earthquakes), and so we were left with 12046 observations.

In chapter 2 some basic concepts of seismology in order to frame the topic of this master thesis are presented. This chapter seemed relevant due to the fact that most concepts are not common knowledge. It also explains the reasons why we restrict our analysis to severe earthquakes and to the time period considered (from 1900 onward).

In chapter 3 the two fundamental methodologies of the EVT are presented. Section 3.1 refers to the BM method, which consists on modelling the sample of maxima (or minima) with the generalized extreme value distribution. This distribution has three parameters: μ (location parameter), σ (scale parameter) and ξ (shape parameter) and joins together the three extreme value distributions. If $\xi > 0$ the GEV model is an element of the Fréchet family, if $\xi < 0$ we get a member of the Weibull family and finally when $\xi = 0$ we are left with the Gumbel family. Section 3.2 explains the POT approach. The main goal of this methodology is to model the exceendances above some sufficiently high threshold. The selection of an appropriate threshold can be a difficult task. The main focus when choosing a threshold should be to achieve a trade-off between bias and variance of the GPD model parameters. Two goodness-of-fit tests for the GPD are also presented, the Anderson-Darling and the Crámer-von Mises tests. Chapter 3 finalizes by addressing the likelihood ratio test and the AIC and BIC criterias (section 3.3). This section was presented due to the necessity to choose and compare models fitted to the data.

Chapter 4 presents a literature review focusing on some similar studies that were considered pertinent to this dissertation. It allows us to have results to compare with the ones that we obtained, since we have a main goal in common.

In chapter 5 we develop our data analysis. Section 5.1 introduces the data from ISC-GEM catalogue as well as an exploratory analysis. There is not any trend for value of m_W as the years go by. We also concluded that the depth of the earthquake's epicenter, for a major part of our data, is between 0 and 100 km. Moreover, and contrarily to our initial thought, there is no relation between the depth of the epicenter and the magnitude of the earthquake. When comparing the locations of the earthquakes and a tectonic plates map we conclude that most of the earthquakes occurred in the limits between tectonic plates.

Sections 5.2 and 5.3 address a first approach to extreme value modeling. For this approach the worldwide catalogue was considered. The results obtained, for both the BM and the POT methodologies, are consistent with the ones presented in chapter 4 for the Ecuadorian coast. For the BM methodology the model considered was the Gumbel and for the POT the chosen model was the Exponential. The results obtained are in perfect agreement, as it would be expected.

Sections 5.5 and 5.6 consider a subset from the worldwide catalogue, the Japan subset. Here we replicated the methods applied in sections 5.2 and 5.3. When comparing the results from the worldwide data set and the Japan data set we see that the estimated parameters of the two models are similar. For the BM method the chosen model was the Gumbel, this conclusion is in order with the one presented in chapter 4 for the Ecuadorian coast. For the POT method the model selected was the Exponential. Likewise the global data, the results obtained by applying the BM and the POT methods to the Japan data

set are in agreement as expected. This conclusions are also in accordance with the ones obtained for the Ecuadorian coast referred in chapter 4.

The ISG-GEM catalogue was selected due to a relevant aspect. The magnitude scale is a continuous one. However, most of the catalogues available only have the value of m_W with one decimal case. This catalogue has the benefit of recording the values of m_W with two decimals. That decreases the probability of having ties which definitely causes problems to the analysis.

In terms of return levels, in the future it would be interesting to have more variables associated with m_W, such as the classification of each earthquake in the Mercalli scale (scale for the intensity of an earthquake), the duration of the earthquake or even if the earthquake had a volcanic eruption and/or a tsunami associated. With these variables a more complete analysis could be made. Additionally, having information about the seismogenic crust for each region would be relevant, since the values of m_W for each area depend on it.

All the programming was carried out in R using the *packages eva*, *evir*, *ismev*, *extRemes* and *evd*. The maps presented were created in PowerBI.

References

Akaike, H. (1998). "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Selected Papers of Hirotugu Akaike*. Springer, pp. 199–213. ISBN: 978-1-4612-1694-0.

Balkema, A. A. and L. de Haan (1974). "Residual life time at great age". In: *The Annals of Probability* 2(5), pp. 792–804.

Beirlant, J., M. I. Fraga Alves, and I. Gomes (2016). "Tail fitting for truncated and non-truncated Pareto-type distributions". In: *Extremes* 19, pp. 429–462.

Cerchiara, R. R. (2008). "FFT, Extreme Value Theory and Simulation to Model Non-Life Insurance Claims Dependences". In: *Mathematical and Statistical Methods in Insurance and Finance*, pp. 61–65.

Choulakian, V. and M. A. Stephens (2001). "Goodness-of-Fit Tests for the Generalized Pareto Distributions". In: *Technometrics* 43(4), pp. 478–484.

Coles, S. (2001). *An introduction to Statistical Modeling of Extreme Values*. Springer London. ISBN: 978-1-8499-6874-4.

De Zea Bermudez, P. and Z. Mendes (2012). "Extreme value theory in medical sciences: Modeling total high cholesterol levels". In: *Journal of Statistical Theory and Practise* 6.3, pp. 468–491.

ESC (2023). *Tectonic Plates of the Earth*. URL: `https://www.usgs.gov/media/images/tectonic-plates-earth`.

Felgueiras, M. M. (2012). "Explaining the seismic moment of large earthquakes by heavy and extremely heavy tailed models". In: *Communications in Statistics - Theory and Methods* 52(3), pp. 523–542.

Fisher, R. A. and L. H. C. Tippett (1928). "On the estimation of the frequency distributions of the largest or smallest member of a sample". In: *Proceedings of the Cambridge Philosophical Society* 24, pp. 180–190.

García-Bustos, S. et al. (2018). "Statistical analysis of the largest possible earthquake magnitudes in the Ecuadorian coast for selected return periods". In: *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazars* 14, pp. 56–68.

Gilleland, E. and R. W. Katz (2016). "extRemes 2.0: An Extreme Value Analysis Package in R". In: *Journal of Statistical Software* 72, pp. 1–39.

Gnedenko, B. V. (1943). "Sur la distribution limite du terme maximum d'une série aléatoire". In: *Annals of Mathematics* 44, pp. 423–453.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985). "Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments". In: *Technometrics* 27.3, pp. 251–261.

Lee, J., Y. Fan, and S. A. Sisson (2015). "Bayesian threshold selection for extremal models using measures of surprise". In: *Computational Statistics and Data Analysis* 85, pp. 84–99.

Ma, N., Y. Bai, and S. Meng (2021). "Return Period Evaluation of the Largest Possible Earthquake Magnitudes in Mainland China Based on Extreme Value Theory". In: *Sensors* 21, p. 3519.

Northrop, P., N. Attalides, and P. Jonathan (2017). "Cross-validatory extreme value threshold selection and uncertaity with application to ocean storm severity". In: *Journal of the Royal Statistical Society Applied Statistics Series C* 66, pp. 93–120.

Pickands, J. I. (1975). "Statistical inference using extreme order statistics". In: *Annals of Statistics* 3(1), pp. 119–131.

Pisarenko, V. F. et al. (2014). "Characterization of the Tail of the Distribution of Earthquake Magnitudes by combining the GEV and GPD descriptions of Extreme Value Theory". In: *Pure Appl. Geophys.* 171, pp. 1599–1624.

Reis, C. J. dos, A. Souza, and R. Graf (2022). “Modeling of the air temperature using the Extreme Value Theory for selected biomes in Mato Grosso do Sul (Brazil)”. In: *Stoch Environ Res Risk Assess* 36, pp. 3499–3516.

Richter, C. F. (1935). “An Instrumental Earthquake Magnitude Scale”. In: *Bulletin of the Seismological Society of America* 25.1, pp. 1–32.

Robinson, T. (2022). *10 Fundamental Theorems for Econometrics*. URL: https://bookdown.org/ts_robinson1994/10EconometricTheorems/.

Scarrott, C. and A. MacDonald (2012). “A review of extreme value threshold estimation and uncertainty quantification”. In: *REVSTAT - Statistical Journal* 10.1, pp. 33–60.

Schwarz, G. (1978). “Estimating the Dimension of a Model”. In: *The Annals of Statistics* 6.2, pp. 461–464.

Silva Lomba, J. and M. I. Fraga Alves (2020). “L-moments for automatic threshold selection in extreme value analysis”. In: *Stochastic Environmental Research and Risk Assessment* 39, pp. 465–491.

USGS (2023). *Moment magnitude, Richter scale - what are the different magnitude scales, and why are there so many?* URL: https://www.usgs.gov/faqs/moment-magnitude-richter-scale-what-are-different-magnitude-scales-and-why-are-there-so-many.

USGS-Catalog (2023). *Japan Area Map*. URL: https://earthquake.usgs.gov/earthquakes/search/.