KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI

EXTREME VALUE THEORY OF FLOOD LOSSES

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE MASTERS OF PHILOSOPHY IN ACTUARIAL SCIENCE

By

JASON SATEH AKORLOR

September, 2018

Declaration

I hereby declare that this submission is my own work towards the award of the Master Of Philosophy Degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgment had been made in the text.

Jason Sateh Akorlor/PG6765916	.....	.....
Student	Signature	Date

Certified by:

Mr. Asamoah Owusu Derrick	.....	.....
Supervisor	Signature	Date

Certified by:

Professor A.O Adebanji	.....	.....
Head Of Department	Signature	Date

Dedication

I dedicate this work to my dad and my late mom Patience Adu Abra for having been very supportive, motivating and encouraging for most part of the time. I also dedicate this to the rest of my family and my friends also for having been there and supportive.

Abstract

Extreme value theory is used in recent times in most risk management fields for large risks and predictions so as to prepare sufficiently for these losses. This thesis provides an overview of the theory as a method for modelling and measuring extreme risks. There are two main models of extreme value theory. The older Block Maxima method and the more recent Peaks Over Threshold. The Peaks Over Threshold method is favoured over the older Block Maxima method because it provides a simpler tool for estimating tail risks and uses data more efficiently considering the rarity of extreme data. Two types of data sets were used. One is a simulated flood insurance claim data and the other being losses due to flooding in the Kumasi Metropolitan Assembly. The method of maximum likelihood is used to estimate the shape and scale parameters of the generalized pareto distribution, which is a natural model for the excess distribution over a high threshold. The mean residual life plot is employed and used along with the quantile-quantile plots to aid in the selection of threshold. Two risk measures required from the objectives of the study were estimated as well as the return levels and return periods of the extremes. For the simulated insurance claim data, the value at risk results obtained at the 95th and 99th quatiles were GHC58,33.40 and GHC88,270.7 respectively, while results were also obtained for the flood losses data. The expected shortfalls for the corresponding value at risks were also obtained. The return levels obtained from the data included m-observation return levels and n-year return levels. Only m-observation was obtained for the simulated data because it was not simulated in yearly sets. For the next 500, 600 and 700, the return levels were GHC159,477.02, GHC172,149.37 and GHC183,878.28 respectively. Similarly for the flood losses data which were recorded on yearly basis, return levels for the next 50,60 and 60 observations return levels were estimated and for the next 2,3 and 4 years, expected return

levels were GHC1,438,058.75 GHC1,891,37065 and GHC2,285,39.43 respectively. The theory proved to be a good model for the measurement of extremes in insurance and other areas.

Acknowledgements

The Almighty God through His infinite blessings and favour bestowed me with wisdom, strength and the ideas to complete this thesis. I am indeed forever grateful. My thanks and gratitude go to my Supervisor Dr. Asamoah Owusu Derrick and also to Dr. S.K Appiah for their unflinching support, guidance and direction toward the successful completion of this work. Not forgetting my parents for the advice, prayers, support and motivation that kept me going the entire time. God bless you abundantly.

Contents

List of Abbreviation

GPD .....Generalized Pareto Distribution **POT** .....Peaks Over Threshold

GEV ..... Generalized Extreme Value Distribution **EVT** ..... Extreme Value Theory

ARCH .....Autoregressive Conditional Heteroscedasticity **QQ** ..... Quantile-Quantile **MRL** .....Mean Residual Life

iid ..... Independent and Identically Distributed

ADF ..... Augmented Dickey Fuller **GARCH** ..... Generalised Autoregressive Conditional Heteroscedasticity

VAR .....Value At Risk

ES ..... Expected Shortfall **PDF** ..... Probability Distribution Function

CDF .....Cumulative Distribution Function

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of the study

It is of common knowledge that insurance claim data exhibit thicker than normally expected tails. In insurance risk management, the insurers are of concernn of possible potential risk of large claims due to earthquakes, floods, fire outbreaks and volcanic eruptions. There have been presence of these large claims, larger than normal in the insurance industry in Ghana especially the ones associated with floods. Talk of foods, losses caused by floods have been a problem during the rainy seasons in Ghana and modelling these extreme events and being able to predict them with some level of confidence helps in managing these huge losses.

The theory of extremes is a means to describe the occurrences and behaviour of these rare events which lead to extreme large insurance claims in this case. It has been used to develop justified models that are reliable in predicting these events. Data on these rare events, basically are just fitted by an extreme value distribution and analysed, Extreme value theory has been adopted widely in financial, hydrology, insurance and applied to environmental studies where extreme occurrences of risk is of interest. Reiss (2001) discussed the diverse ways and fields in which the theory can be applied to extremes.

In hydrology, the ability to be able to predict floods is of importance to researchers due to the magnitude with which they cause damages and hence knowing the probabiliy of the next flood event, frequency and intensity are very important and crucial for planning the future. Fabio Rossi (1984) and Daniel

Cooley and Naveau (2007) applied models in extreme value theory for flood frequency analysis, and Katz (2010) also looked into the extremities in hydrology. Sea-level analysis is also a very important topic in hydrology. Peaks or extreme sea level predictions is of importance to coastal settlements. This is to help prepare adequately for such events. Smith (1989) and A (1992) used the theory of extremes to model sea levels and peaks.

Floods in simple terms can be considered as an overflow of large amounts of water beyond its normal limits, especially over what is normally considered dry land. We can also think of floods as pouring water into an already full glass, the only place for it to go is over the edge. Practically, these may occur due to overflow from water bodies such as lakes, rivers and the sea escaping its usual boundaries or simply due to accumulation of rainwater on saturated ground. The overflow from water bodies or accumulation from rainfall are unlikely to be considered significant unless they flood property or cause domestic losses. While we usually have a fair knowledge when they will occur considering some changes in the weather conditions, we can never be certain on the magnitude of the flood and or the damage it will do. Noah from the good books, should consider himself lucky. It was rare that floods of such magnitude allowed much time to prepare for a safe escape.

There is the need to find a way to deal with floods causing damages and in particular, ones resulting in really huge losses. In mathematics, modelling is one way to be able to predict events. Modelling the events of floods, extreme floods in this case, is of utmost importance. Actuaries will be interested in chances of these events in the future so as to adequately prepare for these losses and hence efficiently manage the risks in the insurance company. Actuaries calculate and estimate the prices of the insurance products to cover the potential risk that may be caused by floods in this case. McNeil (1997) studied and applied the Peaks

over Threshold method of extreme models to the classic Danish insurance data. Resnick (1997) pointed out that McNeils work is of great value in the application of the theory to insurance claims. Embrechts (1997) used the theory to prove as a very useful tool for managing insurance risk.

This study aims at modelling and analysing the extremes in insurance claim data and cost of damages due to flooding.. Having successfully modelled these extremities, there is the need to measure the risks using various statistical risk measuring instruments. This thesis employed two of the many risk management tools. The Value At Risk and the Expected Shortfall or Tail Value at Risk.

The Value at risk measures the quantity and the level of financial risk within a firm or an investment portfolio over a specified time period. It is used mostly by investment companies and or commercial banks to measure the extent of potential losses should they occur in their institutional portfolios. In this thesis, interests in the extent of the extremes, the level of potential loss claims due to flood, should they occur in extreme cases are considered. Insurance companies will then prepare for these potential losses by either allocating cash reserves needed to cover these losses or reinsuring.

The Expected Shortfall also known as the Conditional Value at Risk or the Tail Value at risk is the expected loss, given that the loss exceeds the value at risk. It focuses on the shape of the tail of the loss distribution. In financial portfolios, it focuses on the less profitable outcomes. It is used to reduce the probability that a portfolio will incur large losses.

1.1.1 Problem Statement

Dating back to 1972, the Mississippi River covered about 27,000 square miles of land and submerged much of that land under 30 feet of water. The rain that

supplied this flood spread the river to a width of 70 miles, claiming hundreds of lives as it poured over the land Britannica (2017). Quite recently is the 2005 and 2006 flooding in Cumbria and widespread flooding aross England in summer 2007. This claim several lives and impacted on the health and well being of people living in the affected areas. The losses amounted to more than $3.2billion in England and about $450million in Cumbria. In Ghana, there was a flooding incident in August 2007 , where roughly 350,000 people were affected with about 49 casualties in the northern sectors of the country alone, resulting in an estimated US$130million and not including long term losses. Flooding events in Ghana cannot be discussed today without mentioning the infamous June 3rd Accra-Circle flooding. The capital experienced a flash flood which led to explosions of a fuel filling station, resulting in death of 152 lives. It was argued to have been caused as a results of poor and or inadequate waste management and structural settlement, and as well as the poor hydraulic performance of the basins in Accra.(Asumadu-Sarkodie et al. (2015))

Flooding in coastal is often associated with storm surges. A surging storm occurs when a hurricane carries a large pile of water along the ocean as it moves. Essentially, this creates a giant wall of water heading towards lands and settlements. Depending on the size of the surge and how hard it hits the land, it often becomes the most dangerous part of a hurricane. The Superstorm Sandy storm surge in 2012 made history as a significant and one of the most damaging effects of the hurricane, literally washing away homes, businesses and roads across the coastal belts.

Flash floods can also be dangerous as mentioned in the Accra-Circle flood of 3rd July. These are normally quick, heavy and fast-developing floods, that occurs in a moment and mostly referred to as occurring in a 'flash'. Huge thunderstorms create temporary but big piles of water all over around and cause creeks and

smaller rivers to suddenly overflow. These sudden pools of water that come from flash floods often occur at areas of cities that are lowlands and on roadways, which pose serious danger to those driving or living in such areas.

Now, looking at hurricanes and superstorms and their total damages in properties and loss of lives, the most renowned ones include the Hurricane Catherina, which amounted to over $16billion in loss claims, The Superstorm Sandy amounted to over $8billion. The Hurricane Ike amounted to over $2billion , then there was Hurricane Irene, Hurricane Ivan and Tropical Storm Allison which amounted to over $1billion each in loss claims and the recent Hurricane Irma resulting in over 496,532 insurance claims worth an estimated $3.1billion.

It will be of great interest to insurance companies if these extremes can be predicted way ahead of time with some level of confidence so as to prepare adequately for these extremes. Extreme value theory is by far the most reliable statistical and mathematical tool for modelling these extremes. It provides some level of insight as to how frequent and intense these extremes might be. Extreme value theory provides a relatively safe method for extrapolating extreme events beyond what is normally observed (Paul Embrechts (1997)).

1.1.2 Objectives

Following flood events in Ghana, the main objective of this thesis is to model flood extreme cases by applyng the extreme value theory to claim payments simulated data on flood events. Modelling such data leads to value at risk given a time frame and then the expected shortfall. The return levels are also estimated in the process. The specific objectives of the study are as outlined below:

1. To fit simulated data based on flood insurance claims and flood losses data to the peaks over threshold method of extreme value models
2. To find the value at risk over a specified time frame
3. To calculate the conditional value at risk or the expected shortfall.
4. To calculate the return levels

1.2 Methodology

The theory of extremities generally has two methods, namely the Block Maxima method and the Peaks Over Threshold method. The block maxima method is normally preferred whenever the data set consists of maxima independent samples, mostly meteorological and hydrological data. It fits the maxima data to the the appropriate extreme value distributions, being the *Fre'chet, Gumbel* and the *weibul* distributions or it uses the generalized extreme value distribution which is a combination of the three distributions. The more recent Peaks Over Threshold method uses a threshold and fits the Generalized pareto distribution to data above the threshold. The Peaks Over Threshold method is generally more preferred since it uses data more efficiently. A simulated flood claims data is used and also data based on flood losses collected in the Kumasi metropolitan area, by the National Disaster Management Organisation. The Peaks over Threshold method is favoured in this thesis while the method of maximum likelihood estimation is used in its parameter estimation.

1.3 Study Organisation

The first chapter of this study begins with an introduction to the theory of extremes and its importance in insurance as well as other fields. This chapter discusses the importance and the need for the theory and outlines the specific objectives for this particular paper. The methods used to reach the required

objectives are also introduced briefly as well as the limitations that are encountered during the study.

The second chapter of the study looks at the history and the development of the theory as well as the studies that used the theory to attain solutions to problems in various fields. The theory mostly require the use of data to estimate parameters in a particular distribution and there are several methods to do this estimation. The chapter also discusses several other methods that could be used but is not used in our study.

The third chapter captures the method employed among several other methods entailed in the theory. The chapter describes the mathematics involved in the estimations and the resulting estimators in the particular method employed and then the estimations required to aid in attaining the objectives of the study. The chapter also describes briefly the other methods and approaches that could be used to attain roughly the expected results required at the end of the study.

Chapter Four describes the data used, the analysis involved in the application of our employed theory, the various tests required to validate our study, the outcome and results obtained, the validity of the results and the inferences made on the results.

The Final chapter draws conclusions, covering all that has been done in the study. The implications of results obtained in the previous chapter as it impacts on the insurance company, suggestions necessary to control such impacts if negative.

1.4 Limitations of The Study

Within the methodology of the study, choosing a suitable threshold is quite compromising and subjective since we employ a graphical method and is actually one of the best and commonly used methods of choosing a threshold.

Availability of real data was a major limitation. It was quite challenging to obtain claim data from the major insurance companies in the country. A simulated claim data were used instead since we could not get any from the insurance companies. Another data set were obtained from the National Disaster Management Organisation (NADMO), covering the amount in losses of flood events in the Kumasi Metropolitan Assembly, Ghana which contained very little data points and could have resulted in biassed parameter estimates.

Chapter 2

Literature Review

2.1 Introduction

Devloping The Extreme Value Theory

Today, the theory based on extremes is applied in several areas of interest and study. Some of the oldest fields in which EVT was and is still being applied include fields dealing with natural occurrences and phenomena including rainfalls and floods, winds, temperature, air pollution and corrosion. So engineers, hydrologists and theoretical probabilists were the first to go into developing theories involving extrem events. It is also applied in Statistics and the financial fields such as insurance and the financial market in these recent days.

The earliest works on the development of EVT dates as far back as 1709, where *N. Bernoulli* talked about the average largest distance with respect to the origin when n points randomly lie on a straight line of length t (citeJohnson1995). Bernoulli minimizes a challenge of the required lifetime with respect to the last survivor among n men to finding the expected value of the largest or maxima of n independent and identical uniform variates. Harter (1978) summarizes this and several other early studies and works that looked into extremes.

Between the years 1902-1985, Leonard Tippett was tasked by an English cotton producing industry where he aimed at making cotton threads even stronger. During his studies, he figured out that the strength of the weakest fibres accounted basically for the strength of the threads. The study showed that, intuitively, the best possible comparison among extremes is obtained by the return time. That is, after a given magnitude event occurs, the time duration taken before another event of that magnitude occurs. Enabling the statistics of the size of an event to be predicted from the variability in the size over time (citeWest2016) . Following this discovery in later years *R.A Fisher* assisted Tippet to obtain three asymptotic limits that described the distributions of extremes assuming certain variables. The theory was codified by *Emil Julius Gumbel* in 1958, in a book titled *Statistics of Extremes*, as well as the Gumbel Distribution that bear his name.

The theory first made its presence in Germany in 1922, where *Bortkiewicz*, wrote a paper which entailed the distribution of the range of random samples from the Gaussian distribution. His contribution was the proposal of the concept of *largest values distributions*. A year later, *Von Mises*, who is also German, brought about the concept and the study into the *expected value of the largest number of a sample observations* from the normal distribution.. He essentially began the research and studies into the asymptotic distribution of extreme values considering samples

from the normal distribution. Dodd (1923) studied the largest values from other distributions about the same time.

The first major step though in the theory of extremities was by Tippet in 1925. He presented a table of the largest values and thier respective probabilities for different various sample sizes from the normal distribution, and also the average of the range of these samples. The first asymptotic distributions of largest values from a class of individual distributions was introduced and written in a paper by Fre'chet in 1927. Fisher and Tippet published their first paper in 1928, which is considered as the foundation and basis of the *asymptotic theory of extreme value distributions*. They independently constructed two more distributions after they found Fre'chet's . The resulting distributions over time were found to adequately describe the extreme value distributions of all statistical distributions. Mises (1936) suggested a few quite simple and useful sufficient conditions to help deal with the weak convergence of the largest order statistic to each of the three types of limit distributions. Gnedenko (1943) meticulously came up with the foundation and basics for the Theorem and its sufficient and necessary properties for weak convergence of the extreme order statistics. The theory was first applied by Gumbel(1941) where he applied the thoery by the consideration of the largest possible duration of a life age. He then consecutively followed by implying that the statistical distribution of floods could be understood by EVT. Applications of the theory then followed in several other fields to date.

Extreme value theory normally assumeshat t observations are independent and identically distributed(iid) and one way of dealing with these issues is adopting volatility models to capture the dependence which results in the clustering of extreme values. The Autoregressive Conditional Heteroskedasticity
(ARCH) and the Generalized Autoregressive Conditional Heteroskedasticity

(GARCH) volatility models are used quite often in financial modelling. McNiel (2000) developed a two stage approach by fitting the GARCH model to the tails of the residual. Zhao (2010) at the second stage further extended the method by fitting the GPDs on both upper and lower tails of the residuals.

2.1.1 History Of Flood In Ghana

The capital of Ghana, Accra, registered a record rainfall of 5-inches on the 4th of July, 1968, prior to the preceding last nine years. On the 29th of June, 1971, buildings and structures were severely damaged, some having collapsed, following a downpour which began at night and into the day. This rendered thousands of Ghanaians in the area homeless.

July 5th, 1995, midnight rains which continued to morning caused minor floods in the lower lands of the capital. This led to power cuts having affected commuters and the Achimota substation.

Many institutions were closed down following a flood caused by hours of intermittent downpours during June 13th, 1997. There were a few more events of floods through the years of 1999 and 2001. The 1999 downpours affected about three hundred thousand people across the northern regions of the country and that of 2001 submerged some areas in the capital, Madina, Achimota, Avenor, Santa Maria and Adabraka Official Town.

There have been a lot more events of flooding in the country across the years through to present day, including the already mentioned infamous June 3rd, 2015 flooding in Accra-Circle. There have been minor incidents relating to floods since

then. The recent minor floods on the 28th on June 2018 at Anloga Junction-Kumasi caused quite the turmoil among the people of the area.

Rains are very crucial to human survival, though the storms and the floods following heavy downpours can lead to loss of lives and property, they can not be halted. There will be a lot more in the future and we can only prepare adequately on how to manage their destructive force. (Asumadu-Sarkodie et al. (2015))

2.2 Extrem Value Theory

As will be discussed in more details in the next chapter, the extreme value theory is very important in several fields in recent times. The theory enables the modelling and prediction of extreme events aiding in setting up appropriate counter measures in the fields where these events could cause great losses. There are two main approaches to applying the theory, the older Block Maxima method and the newer Peaks Over Threshold(POT) approach. In recent times the latter is more preferred because it uses data more efficiently. The preferred method will be the POT approach, where parameters will be estimated given sufficient data. The statistical estimation method preferred will be The maximum likelihood method which is discussed in details in the next chapter. There are several other estimation methods such as The Pickands, Hills,adapted Hill, moment, Moments Ratio and the Peng's and W estimators.

2.2.1 The Pickands Estimator

The Pickands' estimator for the parameter γ, generally known as the extremal or extreme valued index of a distribution function F on the real line determines how

heavy the right tail of F is or will be. Pickands(1975) suggested a simple estimator for γ. It was investigated and improved upon by several authors.

Smith (1989) considered the maximum likelihood estimation which has many advantages such as efficiency, invariancy with respect to the changes in the location and scale parameters and its ability that enables it to be extended to various regression models. Pickans (1975) already did suggest a simple location and scale invariant estimator of γ, denoted by $X_{1,n} \leq X_{2,n} \leq .... \leq X_{n,n}$, which is the ascending order statistic from an independent sample of size n from F. Pickands suggested the estimator

$$\hat{\gamma}_{n,k}^{pick} = \frac{1}{log2} log[\frac{X_{n-k+1,n} - X_{n-2k+1,n}}{X_{n-2k+1,n} - X_{n-4k+1,n}}], for k = 1, ..., [n/4] \qquad (2.1)$$

Dekkers (1989) then showed consistency and asymptotic normality for all γ. The downside to the estimator is that it could be somewhat volatile if it is a function of k, and its asymptotic variance is large. Several authors in their attempt to improve the estimator realized that Pickands' estimator is a linear combination of log-spacings of order statistics. Some of these authors included Pereira (1994)), Falk (1994), Fraga Alves (1995), Drees (1995), Yun(2000), and Yun(2002).

2.2.2 Hills Estimator

Hill (1975) suggested a simple way to make conclusions and inferences about the tail behaviour of a distribution. It is unnecessary specifying the underlying distribution F, but merely the tail behaviour where it is desired to make inferences. In other terms, given a sample of size k, we condition upon the r + 1 upper or lower order statistic. Following that, the r + 1 upper order statistic is

used to obtain the conditional likelihood for the parameters that describe the distribution's tail. Hills' estimator can be constructed based on the conditional likelihood function.

There are ways for making inference about the tail bahaviour of a distribution which are well developed, However, Pickans (1975) also developed alternative methods to Hills' that are based on extreme order statistic. There are quite a significant number of modifications of the Hills' estimator, one well known is the moment estimator proposed by Dekkers (1989).

Hill proposed that, for a set ($X_t, t \geq 1$) independent and identically distributed random variables with distribution function $F \in D(H(\gamma))$, the maximum domain of attraction of the generalized extreme value distribution H, where $\gamma \in <$, the sample path is $X_t : 1 \leq t \leq n$ where n is the sample size, If $\{k(n)\}$ is an intermediate order sequence, i.e $k(n) \in \{1,....,n-1\}$, $k(n) \to \infty$ and $k(n)/n \to 0$, then the Hill tail-index estimator is

$$\hat{\gamma}^{Hill}_{(k(n),n)} = \frac{1}{k(n)} \sum_{i=n-k(n)+1}^{n} (in(X_{i,n}) - in(X_{n-k(n)+1,n}))^{-1} \tag{2.2}$$

where $X_{i,n}$ is the i-th order statistic of $X_1,...,X_n$. The probabilities of this estmator converges to γ, and is asymptotically normal, provided $k(n) \to \infty$ is restricted based on a higher order regular variation property. Consistency and asymptotic normality hold under i.d.d assumptions (Paul Embrechts (1997)).

2.2.3 The Moment Estimator

The moment estimator is another adaptation of the Hills' estimator. Dekkers (1989) proposed this estimator with the aim of obtaining consistency for all $\gamma \in <$. This estimator is given by

$$\hat{\gamma}_M = M_1 + 1 - \frac{1}{2}(1 - \frac{(M_1)^2}{M_2})^{-1} \tag{2.3}$$

where $M_j = \frac{1}{k}\sum_{i=1}^{k}(InX_{i,n} - InX_{(k+1),n})^j, forj = 1, 2.$

Dekkers (1989) proved asymptotic normality as well as strong and weak consistency for the estimator.

2.2.4 The Moments Ratio Estimator

Considering situations where $\gamma > 0$, the downside of the Hills' estimator is that it has the tendency to be highly biased, if behaviour of the second order of the underlying distribution *F* is considered. Based on an asymptotic second order expansion of the distribution function *F*, from which one gets the bias of the Hill estimator, Danielsson (1996) proposed the moments ratio estimator to be.

$$\hat{\gamma}_{MR} = \frac{M_2}{2M_1} \tag{2.4}$$

which has proven to have an asymptotic square bias which is lesser or lower as compared to the Hills' estimator if they are both evaluated at the same threshold, that is for the same *k*, though with the same convergence rates.

2.2.5 The Adapted Hill Estimator

The Hills' estimator, due to its fame, created a compelling problem to try to extend it considering its simplicity and good properties. In the general case for $\gamma \in <$, Beirlant (1996) developed the adapted Hills' estimator, applicable for any γ over a range of real numbers given by

$$\hat{\gamma}_{adH} = \frac{1}{k}\sum_{i=1}^{k}(In(U_1) - In(U_{k+1} \tag{2.5}$$

))

where $U_i = X_{(i+1):n}(\frac{1}{i}\sum_{j=1}^{i}(InX_{j:n} - InX_{(i+1):n}))$

2.2.6 The QQ-Estimator

The qq-plot is a common graphical technique used to visually asses the goodness of fit and the estimations of the location and scale parameters. This is normally applied to data from the Pareto distribution or any data generated by a heavy tailed distribution. This is one of the approaches concerning the Hills' derivation Beirlant (1996). The Hill Estimator is approximately the slope of the line fitted to the upper tail of the Pareto QQ plot. Kratz (1996) suggested a more detailed and precise approach and derived an estimator ofγ given by

$$\hat{\gamma}_{qq} = \frac{\sum_{i=1}^{k} In\frac{i}{k+1}\{\sum_{j=1}^{k} InX_{j:n} - kInX_{i:n}\}}{k\sum_{i=1}^{k}(In\frac{i}{k+1})^2 - (\sum_{i=1}^{k} In\frac{i}{k=1})} \tag{2.6}$$

The estimates proved weak consistency and asymptotic normality under conditions similar to the ones imposed by the Hills' estimator. However, the asymptotic variance of the QQ-estimator is twice the asymptotic variance of the Hills' estimator. Though similar results are obtained from simulations of small samples,tThe QQ-estimator is preferred as compared to the Hills' estimator because the residuals contain information which potentially can be utilised to confront the bias in the estimates when the approximation is not exactly valid.

2.3 Pareto Tail Behaviour

Statistics of extreme values, emphasize on the characteristics associated with distributions tails such as describing tail decay powers and indices, very high and very low quantiles, tail probabilities that may be too small or extremal depending on the indicators. γ is the shape parameter of the Generalized Extreme Value Distribution (GEV), with probability distribution function given by

$G_\gamma(x) = exp(-(1+\gamma x)^{-1/\gamma}), 1+\gamma x > 0, \gamma$ 6= 0 and $G_\gamma(x) = exp(-exp(-x)), x \in <, \gamma = 0$

(2.7)

This turns out to be the only non-degenerate limiting distribution for a sequence of appropriately normalized maximum values of pure random samples.

Considering a set of independent and identically distributed random variables $X_1,...,X_n$ according to some distribution function F, and if $X_{1,n},...,X_{n,n}$ is the ascending order statistic, then for a sequence of constants $(a_n > 0)_n$ and $(b_n)_n$

$$lim_{n\to\infty} P(\frac{X_{n,n} - b_(n)}{a_n} \leq x) = lim_{n\to\infty} F^n(b_n + a_n x) = G(x \quad) \qquad (2.8)$$

The sign of γ distinguishes three cases. for $\gamma > 0$, G_γ assumes the class of Paretotype distributions with an index $1/\gamma$. These are the heavy tailed distributions and possess infinite right endpoint. G_0 is known as the Gumble class and consists of distributions with moderate tails such as the Weibull, Gamma, Normal and Lognormal Distributions. $G_{\gamma<0}$ contains distributions with finite right endpoint. This study concentrates on classes of heavy tailed Pareto-type distributions. It can be shown in these types of distributions that the first-order condition can be expressed in an equivalent way in terms of the survival function $1 - F(X)$.

$$1 - F(x) = x^{-1/\gamma} l_F, x > 0 \qquad (2.9)$$

where l_F denotes a function slowly varying at infinity, i.e $\frac{l_F(\lambda x)}{l_F(x)\to 1}$ as $x \to \infty$ for all $\lambda > 0$ This can also be expressed in terms of the tail quantile function U, defined as

$$U(x) = inf\{y : F(y) \geq 1 - 1/x\}, x > 1 \text{ and } U(x) = x^\gamma l_U(x) \qquad (2.10)$$

where l_U again stands for a slowly varying function at infinity.

A random sample $X_1,...,X_n$, fitted to a statistical distribution can visually be assessed by inspection of a quantile-quantile or QQ plot. In the case of a strict

Pareto distribution, the log-transformed Pareto Random variables are exponentially distributied. It is natural to consider an exponential quantile plot based on the log-transformed observations, leading to the QQ plot coordinates

$(log\frac{n+1}{j}, logX_{n-j+1,n}), j = 1, ..., n$

Supposing data is taken only from the Pareto distribution, the quantile-quantile plot of the data against the pareto distribution will yield a straight line with which the slope of the straight line is given by the etremal index. Taking a Pareto type distribution, $loglUx/logx \to 0 as x \to \infty$ implying $logU(x)\ \gamma logx$ as $x \to \infty$. This is to say, for strict parentian data, the quantile quantile plot of the pareto distribution will ultimately be linear with the slope of the line being approximately γ. Many researchers and authors studied and exploited this ultimate linearity of the Pareto quantile plot to construct and derive estimators for γ, (Hill (1975),Csorgo (1985),Kratz (1996), Schultze (1996).

2.4 Fields Of Application Of Extreme Value Theory

Extreme value theory has gained lots of recognition over time in several fields. In the meteorological field, it is applied by using it to predict the probability distribution of extreme storms hence extreme floods for that matter. It is used to model the maximum sizes of ecological populations, mutations in evolution, environmental natural load effects on structures, it is applied in the sports field to determine by measuring in time how fast a person is capable of taking the 100 meters dash and performance in several other fields of athletics. It is applied in the medical field to model the side effects of drugs such as *Ximelagatran*. It can be applied in the distribution of income maximas, like some of the surveys done in virtually all the national Offices of statistics. In Insurance, it is used to model and predict the frequency and amount of large insurance losses, it is also applied in Equity risks; Day to day market risk.

2.4.1 EVT and Geological Anomaly

When predicting mineral deposition, searching for mineral deposits involves identification of a geological anomaly indicating the economic value is high Darehshiri (2015). A geological anomaly is a geological body or complex of bodies with obvious different compositions, structures, or orders of genesis as compared with the surrounding circumstances,(Zhao et al. (1998)). Data acquired from Cu and Au originate from 26 exploration lines of the Jiguanzui Cu-Au mining area in Hubei, China, was used to conduct the study. An EVT model was proposed for the anomalies in the geological make up of the area where it identifies the anomalies in the Jiguanzui Cu-Au mining area. Several steps were adopted while modelling the anomalies. The first step was conditioned, it demanded before the EVT model of the geological anomaly was used, the stationary and post-tail of the sample data needed to be tested. The common method of the conditional test adopted the use of probability and the quantile-quantile(Q-Q) plots, augmented Dickey-Fuller(ADF) test and other required tests. They proceeded by estimating the parameters in scale and shape using the moment method. A good threshold was then chosen and the distribution of the excesses was determined. At the end of the study, the results showed that the model can effectively differentiate between the anomalies of the geological components of the region of Cu and Au. Inference showed the anomalies in the a'reas of Cu and Au is consistent with the size in ranges of ore bodies of actual engineering exploration. They further concluded that the EVT model of the anomalies effectively identifies and differentiates anomalies with high indicating function with regards to the ore prospecting.

2.4.2 EVT and Health

The prediction likelihood of an extreme event occuring soon or in the not so distant future is a major concern for resource planning in Public Health. These

events could be for instance, an unusual community epidermic, a major heat wave or an accidental toxic exposure.

Maud Thomas, Magali Lemaitre, Mark L. Wilson, Cecile Viboud, Youri Yardanov, Hans Wackernagel and Fabrice Carrat published a paper on July 15 2016 Thomas M (2016) about the application of EVT on public health. Their objective was to show how EVT could be employed in public health as a means of predicting and aid in making intelligent guesses of future extreme events. The theory was applied to rates of Pneumonia and influenza deaths sorted on weekly basis over the years 1979-2011. The attendance at the emergency department in a network of 37 hospitals over 2004-2014 was observed and recorded on daily basis. The maximum or maxima were taken from the consecutively grouped samples, these maxima was then fitted to the Generalized Extreme Value Distribution. Having fitted, the probabilities of the occurances of the maximas was then estimated over specific periods.

The outcomes of the study showed an annual Pneumonia and Influenza death rate of 12per 100,000(largest maximum recorded) is expected to exceed once in next 30, it is also expected that there should be an estimated 3% risk that the diseases' (Pneumonia and influenza) rate of death will exceed this value. The maximum observed rise in counts of daily attendance from the same weekday within two consecutive weeks over the past 10years was 1133. An estimated 0.37% chance of exceeding a daily increase of 1000 on each month.

Inferences made suggested EVT models can be and should be applied to various areas and topics in epidermiology thus contributing to public health planning for extreme cases.

2.4.3 EVT and Meteorology

Situations involving phenomenological cases of extreme events have been studied over long periods of time. Some of these Phenomenological cases include

earthquakes, floods, water levels, hot and cold spells etc. For at least 3000 years, there have been records of earthquakes in forms of newspaper articles and texts all over the world. Another example worth mentioning is the water levels records of the Nile river, there are records of its highest water levels and lowest water levels for over 5000 years in order to analyze too high or too low levels that could possibly lead to famine and other disasters.

Sheng (2012) conducted a research with the goal measuring and being able to predict the extremes of hot and cold spells with extreme value theory.

Hot spells and heat waves are extreme meteorological phenomena, the only difference between heat waves and hot spells is their duration, heat waves are longer than hot spells. Records in history has it that, the average temperature over a specified period can be obtained considering a particular location. Temperatures are abnormally higher than the averages of the location during periods of hot spells or heat waves. Heat waves tend to be very fatal compared to other natural phenomena such as floods, hurricanes etc, they cause higher toll on victims than any other natural hazard.

Sheng Gongs' paper mainly focused in details on the analysis of hot spells and properties such as frequency of spells, duration of spells, the average maxima/minima and finally testing for trend of these properties analyzed.

The empirical analysis of hot spell in the paper used the temperatures recorded between 1900 and 2001 in Uppsala in Sweden. Both Block Maxima and Peaks over threshold methods were applied to the data. The mean residual life plot was employed to obtain the suitable threshold for the Peak over threshold method. Results were obtained for length of cold spells as well as hot spells.

2.4.4 EVT and Insurance

Loss severity estimation from historical data plays an important role in actuarial work hence in insurance. There are times in insurance where it is required to choose or price a high-excess layer, estimating the tails of the loss distribution is of interest in actuarial work. It is necessary to find a good statistical model for the largest observed historical losses; it is less important if the model describes smaller losses. Fact is if the model fit is for all historical losses then it would not be a good fit for large losses, and such a model will not be suitable for pricing a high-excess layer. EVT models only the large losses, using either the block Maxima method and the Generalized Extreme Valued distribution or the Peak of Threshold method and the Generalized Pareto Distribution.

Applying this theory to Insurance is discussed by several authors some of which include Beirlant (1996), Mikosch (1997),McNeil (1997) and Saladin(1997), most of which applied the theory to large fire insurance losses. Tajvidi (1997) also applied EVT to windstorm insurance claims. Musah (2010) from the Kwame Nkrumah University Of Science and Technology applied EVT to crude oil prices, the highly volatile and risky crude oil prices created the possibility for spikes in crude oil prices posing a problem that needed the use of EVT to model these spikes. The Thesis presented how the theory could be and was applied to daily returns of Brent Crude Oil Prices between 1987 and 2009 in the spot market.

Chapter 3

METHODOLOGY

3.1 INTRODUCTION

The chapter presents the concepts of extreme value theory. These include the two main methods of modelling extreme events and their associated validity tests

3.2 GENERAL THEORY

From Coles (2001), Let $X_1,X_2,.,X_n$, denote a sequence of independent and identically distributed random variable with cumulative distribution function $F(F(x) = P(X \leq x))$ and M_n be the maxima or minima order statistic over the block of size n. To derive the distribution of M_n, we first define the maximum order statistic mathematically as;

$$M_n = max(X_1,X_2,....,X_n)$$

Now, to find the distribution of M_n, defined from first principles,

$$Pr(M_n \leq x) = Pr(X_1 \leq x,...,X_n \leq x) = Pr(X_1 \leq x) \times ... \times Pr(X_n \leq x) = F(x)^n$$

Knowing the population distribution F, the distribution of M_n can be exactly determined. However if we dont know the F, then we can approximate M_n, by finding the limit of F_n, through asymptotic theory of modelling. Doing so, as $n \rightarrow \infty$, the distribution of M_n degenerates to a point mass at the upper end point of F. This downside can be corrected by introducing normalizing series

$(a_n)_{n\leq 1}$ and $(b_n)_{n\leq 1}$, so that the law for $M_n* = (M_n - b_n)/a_n$, can be obtained. Having chosen an adequate a_n and b_n, the distribution of M_n can be stabilized leading to extremal type distributions.

Theorem 3.1 Assuming there exists suitable constants $a_n > 0$ and $b_n > 0$, as $n \to \infty$, so that, for all x. Fisher (1928) and Gnedenko (1943)

$$\lim_{n\to\infty} Pr\{\frac{M_n - b_n}{a_n} \leq x\} = H_\epsilon(x)$$

if $\epsilon > 0$

$$H_\epsilon(x) = \begin{cases} 0 & \text{if } x \leq u \\ exp\{-(\frac{x-b}{a})^{-1/\epsilon}\} & \text{if } x \geq u \end{cases}$$

if < 0,

$$H_\epsilon(x) = \begin{cases} exp\{-[-(\frac{x-b}{a})]^{1/\epsilon}\} & \text{if } x < u \\ 1 & \text{if } x \geq u \end{cases}$$

if $= 0$,

$$H_0(x) = exp\{-exp[-\frac{x-b}{a}]\} \text{ for all } x \in <$$

where u is a suitable threshold and $H_{epsilon}(X)$ is an extreme value Distribution.

Leadbetter (1983) proofed this theorem. Theorem (3.1) tells us that if M_n^* is stabilized with normalizing constants a_n and b_n, then the limiting distribution of M_n^* will be one of the three extreme distributions above. These distributions are normally indexed by a shape parameter (epsilon), and three domains is

established and distinguished based on the sign of (epsilon). *F* is in the Frechet domain if > 0, the Gumbel domain if = 0 and in the Weibull domain if$\epsilon < 0.$

If the population distribution is unknown, it is not appropriate to use one of the three limiting distributions. A recommended approach is using the generalized extreme value distribution, a universal extreme value distribution which compensates for all three limiting distributions.

3.2.1 The Block Maxima(Minima)Method

The older Block Maxima (or minima) method, older in the sense it was the first method developed for modelling extremes, especially in meteorological and hydrological fields. It uses the largest observations obtained from large samples of independent and identically distributed observations. Assuming we recorded monthly or daily events in sales of a supermarket or supermarkets, the block Maxima method will be appropriate to model the quarterly or yearly maximums or minimums of the observations. This method is normally used to analyse data with seasonality, like hydrological data. From the example assumed, the quarterly or yearly are the blocks. The maximum or minimum in each block is then collected and fitted to one of the extreme value distributions or the generalized extreme value distribution. The downside of the Block Maxima method is that, it is wasteful of data. Extreme observations are rare, and having to pick only the maximum in each block as our data points might and in most cases will result in very few data points. And modelling with very fewer than needed data points will result in obtaining biased and insufficient parameters.

3.2.2 Peaks Over Threshold (POT) Method

The second type of extreme value modelling is a more modern method known as the peaks-over-threshold (POT) method. This method models all high observations over a chosen high threshold. This method uses data which is often limited in EVT more efficiently and is mostly considered in most practical applications.

The Peaks Over Threshold Method has two styles of modelling, the semiparametric methods which is built around the Hill Estimator and its relatives, Beirlant (1996), Danielsson (1998) and then theres the fully parametric method based on the Generalized Pareto Distribution (GPD) Embrechts (1998). Both methods, semi-parametric and fully parametric have been justified theoretically and empirically so theres little to select between the two. However the fully parametric is preferred due to its simplicity in exposition and implementation, thus, we can easily obtain simple parametric formulae for extreme risk measures for which relatively it is quite easy to obtain estimates of statistical error using the maximum likelihood method and inference.

3.2.3 THE GENERALIZED EXTREME VALUE DISTRIBUTION (GEV DISTRIBUTION)

This was introduced by Mises (1936) and Jenkinson (1955), as a way of combining the three limiting types of extreme value distributions. Theorem 3.2. Following from theorem 3.1, with existing sequences of normalizing constants as defined above, *an* > 0 and *bn* > 0, such that;

$$P\{\frac{M_n - b_n}{a_n} \leq x\} \to G_\epsilon(x)$$

For *G* being a non-degenerate distribution function, then *G* is a member of the GEV family. GEV distribution is a parametrisation of the three extreme value distributions into one:

$$G(x) = exp(-[1 + \epsilon(\frac{x-\mu}{\sigma})]^{-1/\epsilon}) \quad (3.1)$$

this is defined on the set x;$1 + \epsilon(x-\mu)/\sigma > 0$, where the parameters satisfy $-\infty < \mu < \infty$ (location parameter),$\sigma > 0$ (scale parameter)and $-\infty < \epsilon < \infty$ (shape parameter)

3.2.4 The Generalized Pareto Distribution

Hosking and Wallis (1987),this is a two parameter distribution used basically for the Peaks Over Threshold Method. Its distribution function is given by;

$$G_{\epsilon,\beta}(x) = \begin{cases} 1 - (1 + \frac{\epsilon x}{\beta})^{-1/\epsilon} & \text{for} \epsilon \neq 0 \\ 1 - exp(\frac{x}{\beta}) & \text{for} \epsilon = 0 \end{cases}$$

where $\beta > 0$, $x \geq 0$ and $0 \leq x \leq -\beta/\epsilon$ when $\epsilon < 0$

It is generalized in the sense that, it assumes the basics of other distribution under a common parametric form. is relatively the most important parameter and the shape parameter, and then there is the scaling parameter β. For $\epsilon > 0$, $G_{\epsilon,\beta}$ is the ordinary pareto distribution which is used in statistics and actuarial science for modelling large loses. For < 0, it assumes the type II pareto distribution and it is the exponential distribution for $\epsilon = 0$.

Since we trying to model extremes, the first case will be the type of interest, where > 0, we get a heavy-tailed distribution appropriate for our interests. A heavytailed distribution of this sort does not have complete well defined moments. The normal distribution on the other hand has well defined moments of all orders. in the case of our heavy-tailed GPD, we notice that, the k^{th} moment $[E(x^k)]$ is infinite

for$k \geq 1/\epsilon$, and has an infinite variance distribution for $k = 1/2$, and for $k > 1/4$, the GPD has and infinite fourth moment.

The fact it allows for infinite variance means it allows for infinite losses, extremely large claims for that matter. Pickans (1975)

3.2.5 The Excess Distribution

As we stated earlier, the POT method models all observations over a particular high threshold. Every observation over this chosen threshold is considered and extreme and with this observations, an excess distribution over the threshold is defined from first principles i.e the excesses is the amount by which the observations exceed the threshold.

Theorem 3.2 McNeil (1999) Let u be a large enough threshold, then the distribution function of *excess losses* over u ,$(X - u)$ provided $X > u$, is defined from first principles as:

$$F_u(y) = P(X - u \leq y \mid X > u) \tag{3.2}$$

for $0 \leq y < x_0 - u$, where $x_0 \leq \infty$ is the *right endpoint* of F. We can see, the excess distribution produces the probability that an observation or loss exceeds the threshold u by at most the amount y, provided the loss exceeds u. We can rewrite F in equation 3.2 using principles of conditional probabilities as:

$$F_u(y) = \frac{F(y+u) - F(u)}{1 - F(u)} \tag{3.3}$$

The underlying distribution of F allows for infinite right endpoint so that it covers for extremely large losses.

Theorem 3.3 This theorem in its limiting case explains the key results and importance of the GPD in EVT.

$$lim_{u \to x_0} sup_{0 \leq y < x_0 - u} \mid F_u(y) - G_{\epsilon,\beta(u)}(y) \mid = 0 \tag{3.4}$$

This is to say, if we progressively raise the threshold u for a large class of underlying distributions, F_u converges to the generalized pareto distribution. Mathematically, this is not complete because we can not really define a large class of underlying distributions. It is sufficient though, knowing the class includes all the various common known continuous distributions in statistics and actuarial science.

The unknown excess distribution follows the GPD above sufficiently high thresholds and this is the basis on which the entire model is constructed. Simply, this is to say for X having distribution F, assumes that, for a certain high threshold u, the excesses above this threshold follows the GPD with parameters and β.

$$F_u(y) = G_{\epsilon,\beta}(y) \tag{3.5}$$

Having established this relationship, we choose a suitable high threshold u for our observations or realisations, $X_1, X_2, \ldots, X_n$. Let N_u denote all realisations exceeding the threshold u. We find the excess of these observations i.e $N_u - u$, then we statistically fit these excesses to the GPD and then we estimate the parameters and β. The Maximum likelihood estimation is idle for estimating the parameters. This is because these parameters are chosen to maximize the probability density distribution.

3.3 Test For Pareto Tail behaviour

Pareto tail behaviour was already discussed in chapter 2. There are several ways of confirming Paretian tail behaviour in a data set. This paper adopts the empirical plots or Zipf plots and the mean excess plots to test for Pareto tail behaviour the data set.

3.3.1 The Zipf Plot

When it comes to checking parentian tail behaviour in data, the Zipf plot is the most used and sometimes abused. It was first proposed by Zipf (1949) constructed on binary observations. A different one was presented in *arXiv2013* based on the empirical survival function.

Considering the standard Pareto I distribution, whose cdf is given by

$$F(x) = 1 - (\frac{x}{x_0})^{-\epsilon} \quad \text{for } 0 < x_0 \leq x$$

The corresponding survival function is then given by

$$S(x) = (\frac{x}{x_0})^{-\epsilon} = 1 - F(x) \text{for } 0 < x_0 \leq x$$

Taking the log of the resulting survival function, we arrive at a inear relationship which is negative between the survival functions' logarithm and the logarithm of x with − being the slope. This particular derivation holds for the type I pareto but also holds with ease when deriving similar results for all pareto types. Plotting the empirical survival functions' logarithm against the logs of the ordered values of x, a more or less negative linear relationship is expected if the data is Paretian type. A single straight line is naturally observed in a purely Paretian Data, but that is generally not the case in reality. In most researches and data analysis, only a part of the data accounts for Paretian behaviour. Normally a data containing Parentian behaviour starts off as a curve and begins to straighten up when the

Paretian part kicks in. This is observed later in chapter 4 when we obtain the Zipf plots for our data sets.

3.3.2 The Mean Excess Plots

The Mean excess plot is widely used in insurance risks and extremities, one is for the validation of the generalized pareto model for the excess distribution Ghosh and Resnick (2010). The plot was introduced by Davison (1990). It uses the mean of the GPD excesses, $E(X - u \mid X > u) = \sigma_u/(1 - \epsilon)$, as a diagnostic, defined for < 1 to ensure the mean exists. This is so that, for any higher $v > u$, the expectation becomes

$$E(X - v) \mid X > v = \frac{\sigma_u + \epsilon v}{(1 - \epsilon)}$$

which is supposed to be linear in v with gradient $\epsilon/(1-\epsilon)$ otherwise the data does not follow the desired distribution, the GPD. Examples of this behaviour of the plots' function for various distributions are given by Beirlant (2004). Coles (2001) acknowledged that the interpretation of such plots can be quite challenging.

3.4 The Choice Of a Thresshold

The choice of a good or sufficient threshold can be quite difficult and compromising. t often involves balancing between bias and variance. We already mentioned how seldom rare events occur hence results in very few data points, following that fact, we will need to choose a threshold high enough so to retain the asymptotic properties of our theorem. We also need to choose a sufficiently low threshold so as to retain enough data points to obtain sufficient and efficient parameters, Coles (2001). So basically, threshold selection considers the limiting model being a sufficiently good approximation against the variance of the parameter estimates.

If the GPD is a valid model for excesses over some threshold u_0, then it is valid for the excesses over any threshold $u > u_0$, this is know as the *Threshold Stability Property* of the GPD. The shape parameter is not changed however, the scale parameter shifts to $\beta^* = \beta_u + \epsilon(u_0 - u)$. ϵ remains unchanged, so that the modified scale parameter becomes $\beta^* = \beta_u - \epsilon u$ is constant above u_0. A threshold is normally chosen and fixed before fitting is done and this choosing of a fixed threshold is commonly known as the *fixed threshold approach*.The downside of the *fixed threshold method* is that, once it is fixed the subjectivity and uncertainty associated is ignored in subsequent inferences and conclusions It is often seen when applying the threshold that there could be more than just one suitable threshold that may lead to different inferences and conclusions of tail behaviour, which is also ignored when the threshold is fixed. There are different various ways of choosing a good threshold, most of which employs diagnostic plots. Coles (2001) stated and employed three main plots.

1. Mean Residual Life Plot
2. Probability Stability Plot
3. Distribution Fit Diagnostic Plots

3.4.1 The Mean Residual Life Plot

This will be the method we will use in this thesis. this method is favoured because idea behind the theory is easy to understand and quite simple to apply. It was Introduced by Davison and Smith (1990), it uses the mean of the GPD excesses such that, $E(X - u \mid X > u) = \beta_u/(1 - \epsilon)$. so for any $v > u$, the mean becomes:

$$E(X - v \mid X > v) = \frac{\beta_u + \epsilon u}{(1 - \epsilon)} \tag{3.6}$$

Assuming the asymptotic assumptions hold, this gives us a linear relationship in *v* with gradient$\epsilon/(1-\epsilon)$ and intercept$\beta_u/(1-\epsilon)$. Beirlant (2004) shows examples of the Mean Residual Life function behaviour for various distributions. Basically, the sample mean empirical estimates of the excesses above each threshold in the range of thresholds above *u* are plotted against the range of thresholds above *u*. In other words, for a range of thresholds, we identify the corresponding mean threshold excess, then plot this mean threshold excess against the threshold.

The Mean Residual Life plot is given by the locus of points:

$$((u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x_i - u)); u < x_{max}) \tag{3.7}$$

When a suitable threshold is being selected, the lowest point above which a linear pattern or a straight line is being observed is mostly suitable, that is considered the lowest point where there is consistency with all the higher based sample excesses with a straight line, once the uncertainty of the sample is accounted for, Scarrot C. J. (2013). Coles (2001) also indicated, the interpretations of these plots can be quite challenging.

$$((u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x_i - u)); u < x_{max}) \tag{3.8}$$

3.4.2 The Probability Stability Plot

This plot puts out an idea where the exceedances of a high threshold *u* follows the GPD with parameters and β_u implies for any threshold $v > u$ the exceedances also follow the GPD with shape parameter $_v$ and scale parameter $\beta_v = \beta_u +$ $\epsilon(v-u)$. If we let

$$\beta^* = \beta_v - \epsilon_v u$$

We see the new re-parametrisation is not dependent on *v* any longer, given *u* is a reasonably high threshold. we get our resulting plot using the locus of points defined by:

$$\{(u, \beta^*); u < x_{max}\} and \{(u, \epsilon_v); u < x_{max}\}$$

x_{max} is the maximum observation. This implies, if *u* is a suitable threshold for the asymptotic assumption, then for any *v* > *u*, the parameters β^* and $_v$. the suitable threshold is then chosen point where the shape and scale parameter remains constant.

3.4.3 Distribution Fit Diagnostic Plots

This includes statistical plots such as the probability plots, quantile plots, return level plots and empirical versus fitted density comparison plots. These plots are normally used to check model fit and can also be used for choosing suitable thresholds. These plots actually provide alternative way of assessing the model performance.

There are other various ways of choosing suitable thresholds that include *The Rule Of Thumb*, where Leadbetter (1983) showed that, the threshold sequence for a population in the domain of attraction of a GPD is a function of the properties of that distribution. Ribatet (2006) Ribatet and Ouarda (2007) also proposed another method known as *The Dispersion Index Plot*, which is particularly useful when dealing with time series. It relies on the fact that the data is generated by a Poisson process.

3.5 Tail Estimations of Distribution

Setting *x* = *u* + *y* and joining equations (3.3) and (3.5), this leads us to ;

$$F(x) = (1 - F(u))G_{\epsilon,\beta}(x - u) + F(u) \tag{3.9}$$

where we had already defined u to be a suitable high threshold retaining the asymptotic properties of the GPD, then for $x > u$, our new equation (3.9) may move easily to an interpretation of the model in terms of the tail of the underlying distribution $F(x)$ for $x > u$. The goal is to use (3.9) to define and construct a *tail estimator*. we will have to find a way to estimate $F(u)$ first before we can proceed to estimate the tail. it is easier and convenient to just use obvious empirical estimator $(n-N_u)/n$. It is assumed that, there will be enough data points above u to estimate $F(u)$ but we can not use it to estimate $F(x)$ because *Empirical Estimation* is a poor method of estimation of tails of distributions where data become sparse.

Now, substituting the estimated $F(u)$ and the maximum likelihood estimates of the GPD parameters, we eventually arrive at;

$$\hat{F}(x) = 1 - \frac{N_u}{n}(1 + \hat{\epsilon}\frac{x-u}{\beta})^{-1/\hat{\epsilon}} = P(X < x \mid X > u) \qquad (3.10)$$

We should take note that this estimator is only valid for $x > u$. It can be seen as a kind of empirical estimate augmented by EVT and can be constructed whenever we believe data came from a common distribution.

So that, $P(X > x \mid X > u) = 1 - \hat{F}(x)$

From that equation, the resulting probability density function is given by

$$\hat{f}(x) = \frac{N_u}{n}\frac{1}{\hat{\beta}}(1 + \hat{\epsilon}\frac{x-u}{\hat{\beta}})^{-1-\frac{1}{\hat{\epsilon}}} = P(X = x \mid X > u) \qquad (3.11)$$

3.5.1 Maximum Likelihood Estimation

Let $x_1, x_2, \ldots, x_n$ be a sequence of independent and identically distributed random variables which exceed the threshold u. We derive the log-likelihood function by

taking logarithms of the joint density function. If N_u is the number of observations above the threshold u, then our likelihood function is given by;

$$logL(\beta_u, \epsilon \mid X) = -N_u log\beta_u - (1 + 1/\epsilon) \sum_{t=1}^{N_u} log[1 + \epsilon(\frac{x_i - u}{\beta_u})] \quad (3.12)$$

where $[1 + \epsilon(\frac{x_i - u}{\beta_u})] > 0$ and $\epsilon \neq 0.$

where = 0, the likelihood function is derived in the same way to obtain;

$$logL(\beta_u \mid X) = -N_u log\beta_u - \sum_{i=1}^{N_u} (\frac{x_i - u}{\beta_u})$$

Note: This was not derived from equation 3.12 by simply putting = 0, setting = 0 reduces the GPD to the exponential distribution, hence the Likelihood function derived.

The Maximum Likelihood Estimators have certain regularity conditions that must be met to exhibit the usual asymptotic properties. It is normal and important to note that, ML estimates are not always valid and the normal regularity conditions are not always existing. The dependence of the upper endpoint of the GEV on the parameters gives rise to these invalidities in regularity conditions, where the endpoints are parameter value of $u - \beta_u/\epsilon$ where

$\epsilon < 0.$

Smith (1985) outlined results showing that maximum likelihood estimators generally exist and have usual asymptotic properties when $\epsilon > -0.5$, they generally exists but do not have usual asymptotic properties when $-1 < \epsilon < -0.5$ and they generally do not exist when $\epsilon < -1.$

3.5.2 Estimation Of Return Levels

The chances and likelihood of rare events are normally carried along with the concept of return levels and return periods. Having concluded the GPD with parameters and β is a good model with suitable threshold u with the respective exceedances, then the chances of the exceedances of the random variable X over a suitable high threshold u is given by,

$$P\{X > x \mid X > u\} = [1 + \epsilon(\frac{x-u}{\beta})]^{-1/\epsilon}$$

Now, for $x > u$ and $\epsilon \neq 0$, let $\zeta_u = P\{X > u\}$, ζ_u is the probability of an observation exceeding the threshold u. Substituting ζ_u and applying various rules of conditional probabilities,

$$P\{X > x\} = \zeta_u[1 + \epsilon(\frac{x-u}{\beta})]^{-1/\epsilon}$$

The subscript u on ζ_u tells us how dependent the value is on the chosen threshold u.

The level x_m exceeded once every m-observations is then given by;

$$\zeta_u[1 + \epsilon(\frac{x-u}{\beta})]^{-1/\epsilon} = \frac{1}{m}$$

Rearranging and making x the subject, we have;

$$x_m = u + \frac{\beta}{\epsilon}[(m\zeta_u)^{\epsilon} - 1] \qquad (3.13)$$

This is known as the m-observations return level x_m. However, we are often interested in duration indexed return levels. Let n_y be the number of observations per year, then our N-year return level is given by, $m = N \times n_y$ and

$$x_N = u + \frac{\beta}{\epsilon}[(Nn_y\zeta_u)^{\epsilon} - 1] if \epsilon \neq 0 \tag{3.14}$$

$$x_N = u + \beta(Nn_u\zeta_u) if \epsilon = 0$$

We see from the equation that, we need $\beta,\ \epsilon$ and ζ_u to calculate our N-year return levels. We expect ζ_u to follow the Poison distribution given that our exceedances are rare events. We also expect the count of these exceedances to follow the Binomial distribution(n,ζ_u) as suggested by Coles (2001). The Poisson Distribution is described by the rate parameter λ, which is basically the mean of exceedances above the high threshold u per unit time.ζ_u can be estimated as

$$\hat{\zeta_u} = \frac{\hat{\lambda}}{n_y}$$

and unbiased estimate of λ is given by

$$\hat{\lambda} = \frac{N_u}{M}$$

where our N_u is the exceedances over our threshold u and M is the number of years of records. Rewriting equation (3.13) in terms of λ, we have

$$x_N = u + \frac{\beta}{\epsilon}[(N\lambda)^{\epsilon} - 1] if \epsilon \neq 0 \tag{3.15}$$

$$x_N = u + \beta log(N\lambda) if \epsilon = 0$$

3.5.3 Estimating Value At Risk(VaR)

In mathematics, measures of extreme risks are defined in terms of the loss distribution F. Say $1 > q > 0.95$, the Value-at-Risk is relatively a high quantile of

the loss distribution F and in this case the qth quantile given by the inverse of the loss distribution as;

$$VaR_q = F^{-1}(q)$$

Given a probability $q > F(u)$ the VaR estimate is calculated by inverting the tail estimation equation (3.10), we have:

$$\hat{VaR}_q = u + \frac{\hat{\beta}}{\hat{\epsilon}}((\frac{n}{N_u}(1-q))^{-\hat{\epsilon}} - 1) \quad (3.16)$$

In standard statistics, this is known as a quantile estimate, where required unknown parameter is also of an unknown underlying distribution. $\hat{VaR}_q$ can possibly be given a confidence interval using the profile likelihood approach yielding an interval which is asymptotic and also in which there is confidence that VaR lies. The asymmetric interval reflects a fundamental asymmetry in the problem of estimating a high quantile for heavy-tailed observations. Giving bounds to the interval below than to bound it above is the easier approach.

3.5.4 Estimating Tail Value at Risk Or Expected Shortfall(ES)

When losses could possibly exceed the VaR, then the need to measure these losses becomes crucial and these possibly expected losses is what is known as the Expected Shortfall i.e what is expected in the worse possible case in $(1 - p)\%$ cases, where p is the confidence level. Said differently, it gives the expected value of an investment in the worst $q\%$ of the cases.

$$ES_q = E[X \mid X \geq VaR_q]$$

applying some basic rules of conditional probability, we have

$$ES_q = VaR_q + E[X - VaR_q \mid X \geq VaR_q] \tag{3.17}$$

The second term in the above equationg is simply the mean of the excess distribution of the excesses derived earlier but this time over the threshold VaR_q. The resulting model for the distribution of excesses above the threshold u has a good stability property. If for instance we take any higher threshold, such as VaR_q for $q > F(u)$, then the excess distribution is also a GPD with the same shape parameter, but a different scaling. It is easily shown that a consequence of the model equation (3.5) is

$$F_{VaR_q}(y) = G_{\epsilon,\beta+(VaR_q-u)}(y) \tag{3.18}$$

This equation is unique in the sense that it gives us a simple explicit model of the losses over the threshold VaR. This enables us to calculate several values of the losses above or beyond VaR. Foe < 1, we note that the mean of the distribution in (3.17) is $(\beta + \epsilon(VaR_q - u))/(1 - \epsilon)$ so we can calculate our expected shortfall as

$$\frac{ES_q}{VaR_q} = \frac{1}{1-\epsilon} + \frac{\beta - \epsilon u}{(1-\epsilon)VaR_q} \tag{3.19}$$

It is worth taking note of this ratio keenly in the case where the end point of the underlying distribution is infinite. It is observed in this case that the ratio is largely determined by the factor $1/(1-\epsilon)$ and the second term on the right hand side becomes negligibly small as q approaches 1. This asymptotic observation which is asymptotic points out the relevance of the shape parameter in the tail distribution. This determines how our two risk measures, VaR_q and ES_q vary in the extreme regions of the loss distribution.

Rearranging Equation (3.18) and substituting the data-based estimates, we can calculate our Expected Shortfall by

$$\hat{ES}_q = \frac{\hat{VaR}_q}{1-\hat{\epsilon}} + \frac{\hat{\beta} - \hat{\epsilon}u}{1-\hat{\epsilon}} \tag{3.20}$$

Chapter 4

Data Analysis

4.1 Introduction

Given the claim data, the objective is to provide an estimate for a size threshold we can set below which, say, 95% to 99% of the observations. We would also like to estimate the expected loss for claims above such a threshold (expected shortfall). Answers to these questions are very important in insurance claim modelling to help inform decisions around how to allocate capital and comply with regulatory capital requirements. There are also applications of this approach in engineering, environmental modelling, and risk modelling in equity portfolios.

Finding answers to extremes using conventional statistical methods based on a representative sample is challenging, as tail events (claims) occur so rarely. We

therefore conduct an analysis of the tail distributions, using univariate extreme value theory. In particular, we adopt the peak-over-threshold method by using the GPD (Generalized Pareto Distribution) approach for exceedances (tails), rather than the block maxima approach in theory is less appropriate in modelling insurance claim data. Two data sets are used for this particular thesis. One was simulated using some assumptions. All claims are assumed to follow the Lognormal distribution. Enough information were obtained from the website *www.fema.gov* to estimate the average claim amounts and the variance, needed for our simulation. The second data set consists of costs of flood damages which were obtained from the National Disaster Management Organisation. In all, there were 207 observations(flood cases) across the districts within the Ashanti region of Ghana

As already stated in the methodology, the POT method demands choosing a threshold. The idea is that, a suitable threshold is considered an extreme as well as all other observations larger or bigger than the threshold. The Mean Residual Life Plot (MRL) was adopted to aid in threshold selection. After the thresholds were successfully chosen, the excesses of the observations exceeding the threshold relative to the threshold was fitted to the generalized pareto distribution. Using the maximum likelihood method of parameter estimation, estimates of the scale and the shape parameters were obtained. An estimate of the rate of the extremes is also obtained by dividing the number of observations above the threshold by the total number of observations. The tail distribution is then obtained with all the information obtained. Using the estimated tail distribution, the return levels and the risk measures were also estimated. The statistical software ”R” was the main tool used in simulating and data analysis.

4.2 Data Structure and Summary

For the simulated data, The annual total amount in claims for the period 1978-2016 with their corresponding number of claims was stated on the website *www.fema.gov*. With that information, the empirical estimated average amount per claim for each year was calculated by dividing the total amount for each year with their corresponding number of claims. The average of the averages was found to be $19,034.49, which is the estimate for the average claim over the time period with a corresponding variance of the averages to be $440944367.3.8507. Assuming the data follow the lognormal distribution, using the mean and variance as aids, the corresponding parameters from the normal distribution were calculated and used for the simulation. This is because, simulation requires parameters from the normal distribution and not the lognormal distribution. The corresponding normal distribution parameters estimated are 9.454 and 0.8 as the mean and standard deviation respectively.

All figures, in this chapter and the next, indicated or not for reasons of simplicity and clarity, are measured in GHS currency

4.2.1 Summary Statistics Of Data Sets

We simulate 2000 claims from a lognormal distribution using the parameters estimated.

Table 4.1: Summary Statistics od Data Sets

Statistic	Simulated	Flood Losses
Min	965.9	400.0
1st Qu	7532.0	10000.0
Median	12910.0	32000.0
Mean	18290.0	137800.0
3rd Qu	22200.0	102500.0
Max	407000.0	3298000.0

The summary shows that losses can be as low as GHC965.9 and be as high as GHC407000.0, for the claims data and GHC400.0 and GHC3298000.0 respectively for the flood losses. The difference between the quartiles(especially the 3rd quartile) and the respective maximums, shows presence of extremities in both data sets.

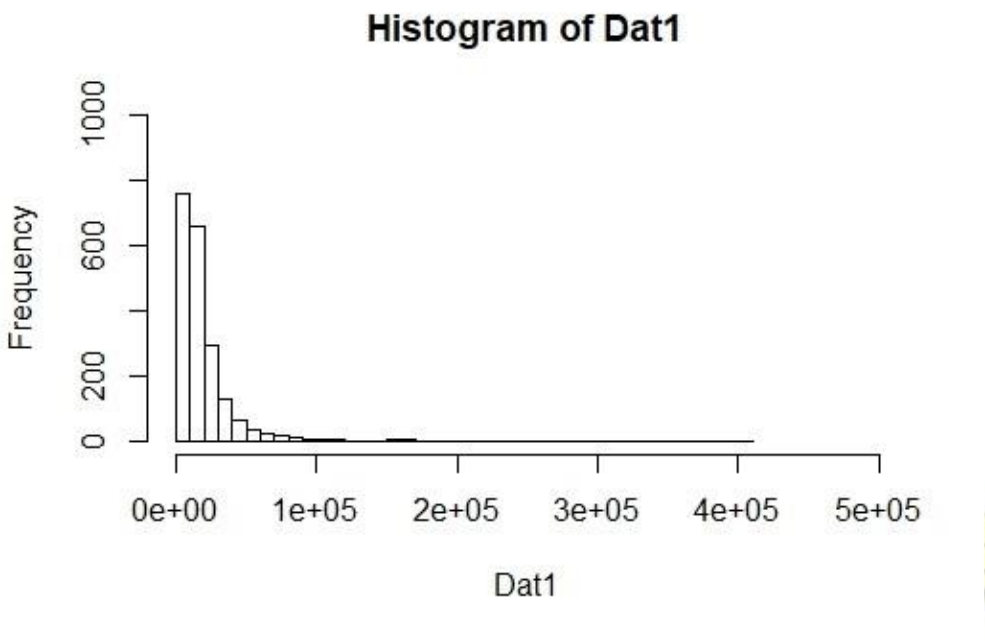


Figure 4.1: Histogram of Simulated claim Data, Dat1= Insurance Claim(GHC)

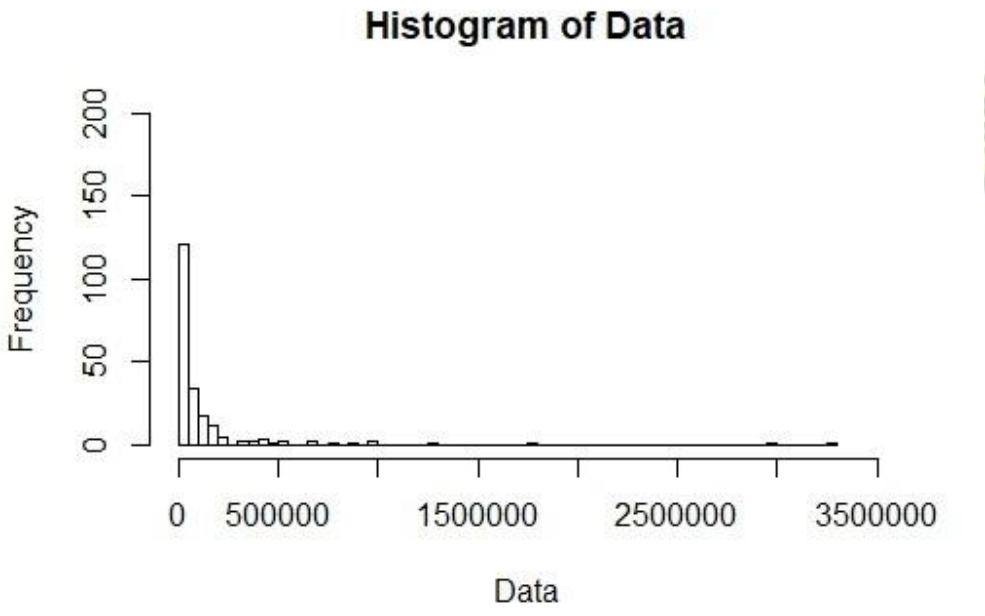
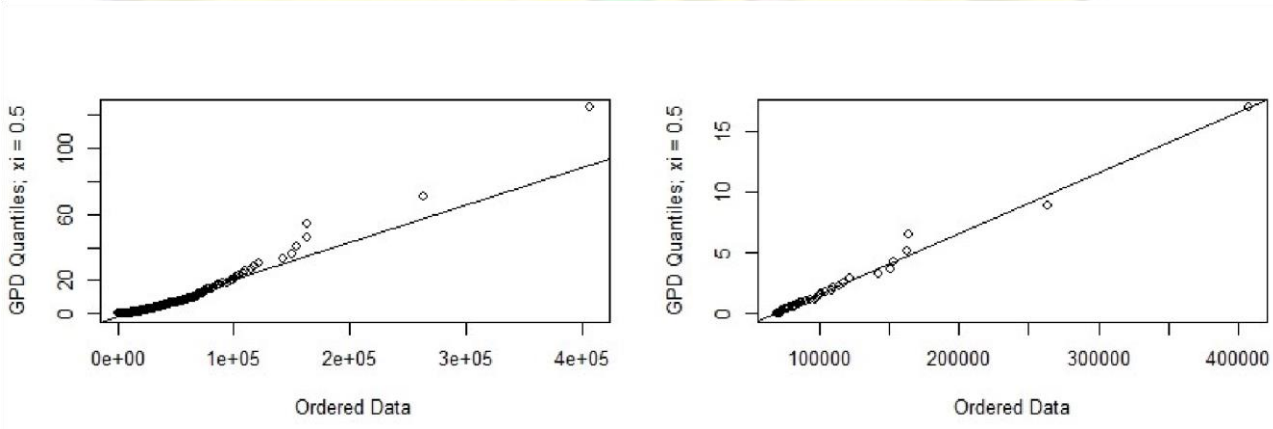


Figure 4.2: Histogram Of Real Data, Data=Cost of flood damages(GHC)

Figure (4.1) shows the histograms of the simulated Claim data, highly skewed to the right. It shows quite a significant number of our data points are below $70000, which is around the 95th quantile and very few above. The data summary and the histogram suggests a thick and pareto tail behaviour.The empirical plot also known as the *Zipfplot* was used to confirm Pareto tail behaviour in the data. Normally, a single straight line only is observed for purely Paretian data, but this is generally not the case most at times. In most empirical analysis where some Paretian behaviour is present, the Paretianity accounts for a certain amount of the data, in particular the upper tail of the distribution. The same pattern is observed in Figure (4.2), the histogram of the real data, thus the cost of flood losses.



4.2.2 The QQplots

Quantile-Quantile plots can be used to confirm the Paretian tail behaviour. For a parential data, the points plotted are expected to fall on the straight line or just around the line. this pattern is observed in both data sets above their respective thresholds as seen in Figure (4.3). and Figure (4.4).

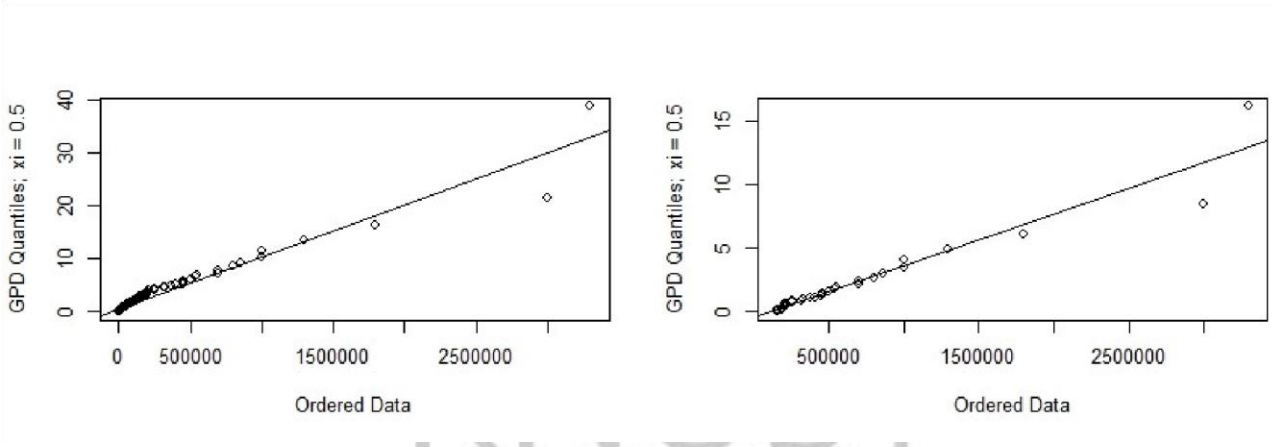


Figure 4.3: Quantile-Quantile Plots of claim data, left Plot:Data above threshold Figure 4.4: Right; Quantile-quantile plots of real data, Left Plot: Data above threshold

4.2.3 The Zipf/Empirical Plots

Figure (4.5) shows the empirical plot for the simulated insurance claim data. It is observed to be a curve until at around GHC50000 and above where we begin to observe linearity. This is not the threshold chosen, it just shows roughly around where Pareto behaviour begins to appear in the plot. Figure (4.6) shows the same pattern at around GHC200000 and above, confirming asymptotic properties of extremities.

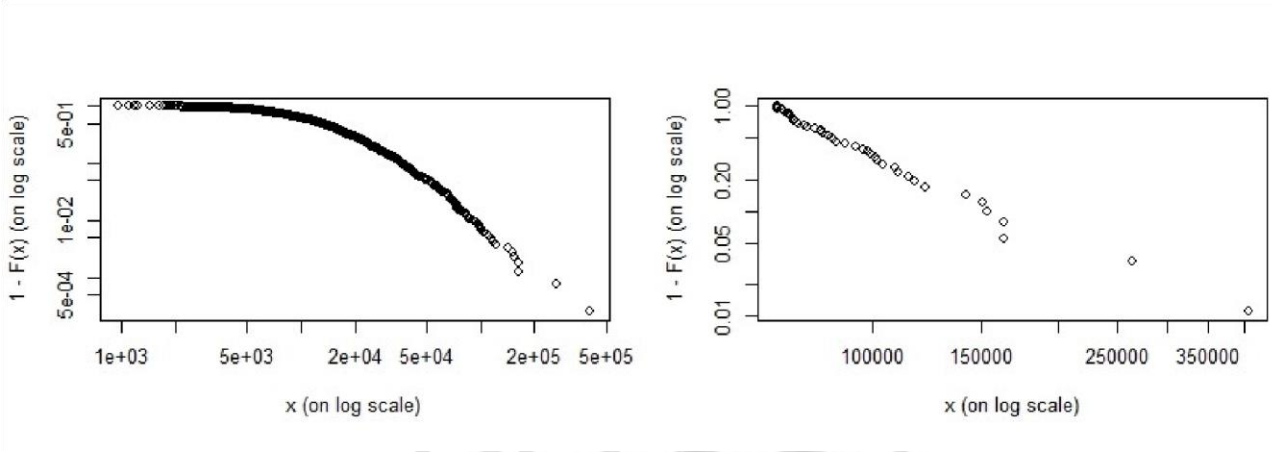


Figure 4.5: Left Plot: The Empirical Plots claim data, Right Plot: Data above Threshold

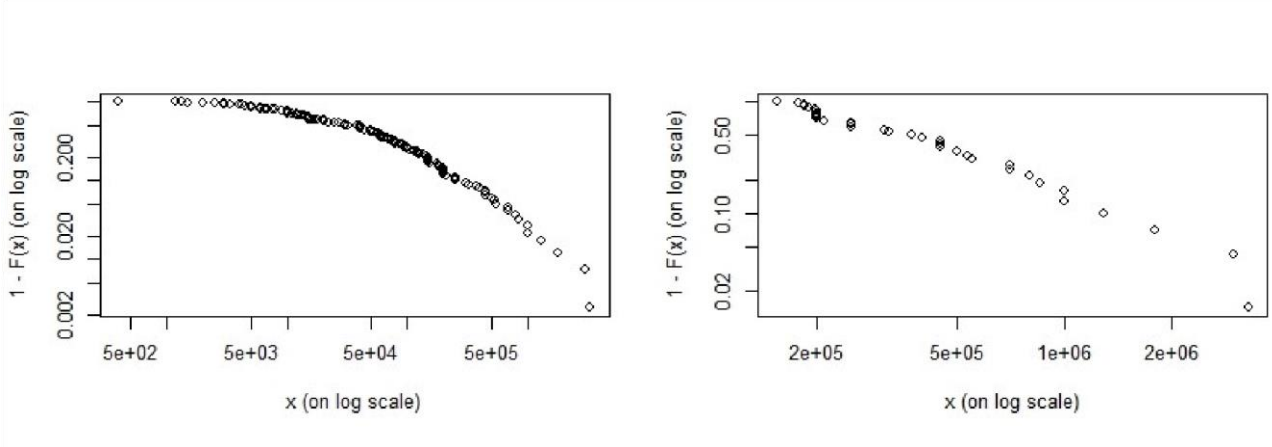


Figure 4.6: Left Plot: The Empirical Plots For real data, Right Plot: Data above Threshold

4.2.4 The Threshold And The Mean Excess Plots

The mean excess plots for the simulated insurance claim data and flood losses data are shown in Figure (4.7) and Figure (4.8) respectively. Extremely large observations are omitted in the mean excess plots to remove their distorting properties. The linearity and a positive gradient observed in the mean excess plot confirms a Pareto tail behaviour.

Figure 4.7: Left Plot: The Mean Excess Plots for claim Data, Right Plot: Data above Threshold

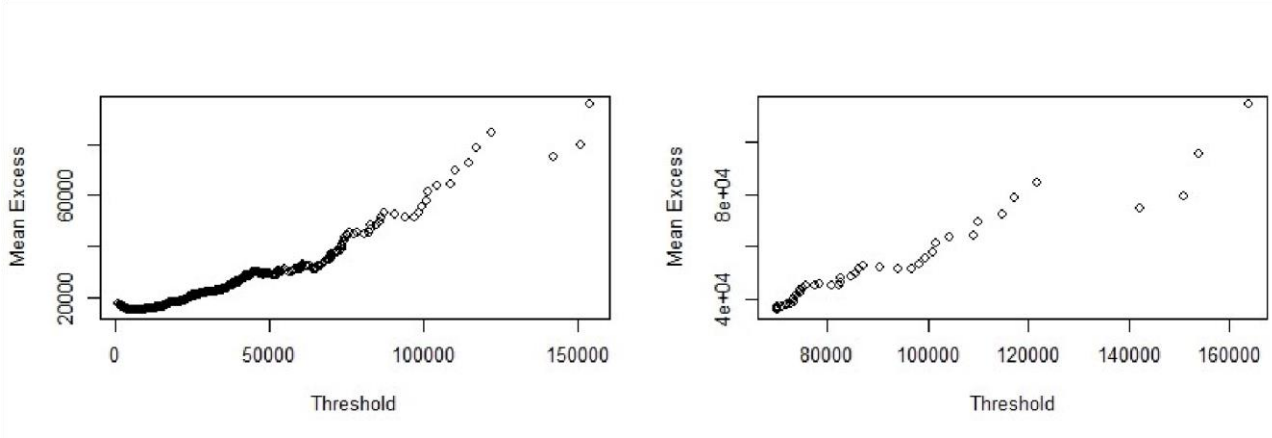
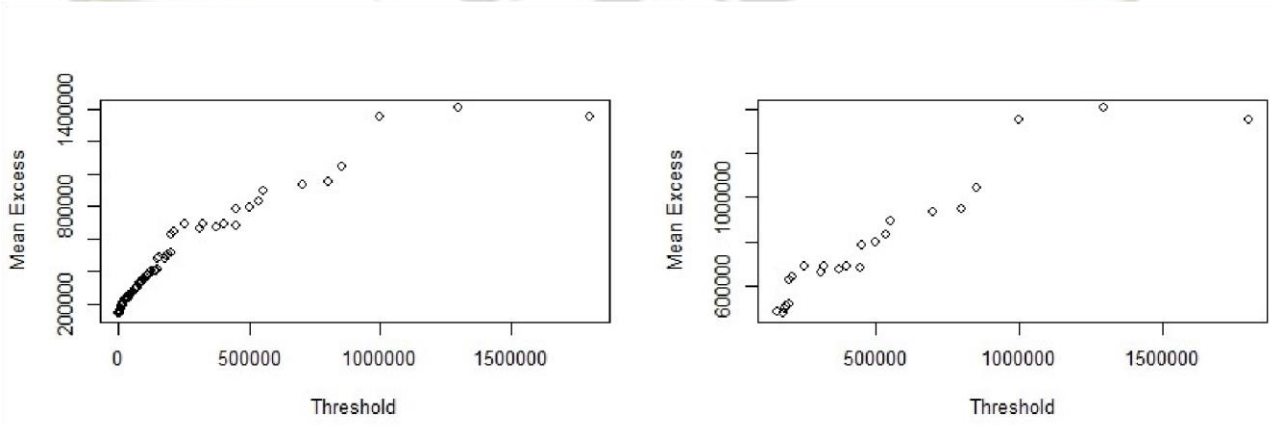


Figure 4.8: Left Plot:The Mean Excess Plots for real data, Right: Data above Threshold

Choosing a good threshold after a Pareto tail behaviour is confirmed



becomes the next objective. The mean residual life plot is used in the study to aid in the selection of a good threshold for the simulated data. The mean residual life plot uses the same ideology as our mean excess plot. The idea in this case is to choose the value *u* above which we start observing linearity in the plot.

This becomes quite subjective but with meticulous observation skills, a suitable threshold should be chosen.

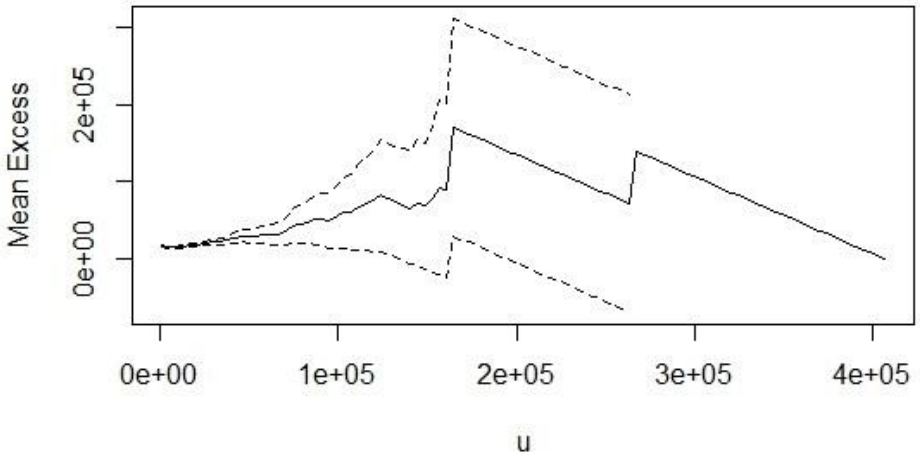


Figure 4.9: Simulated insurance claimdData Mean Residual Life Plot

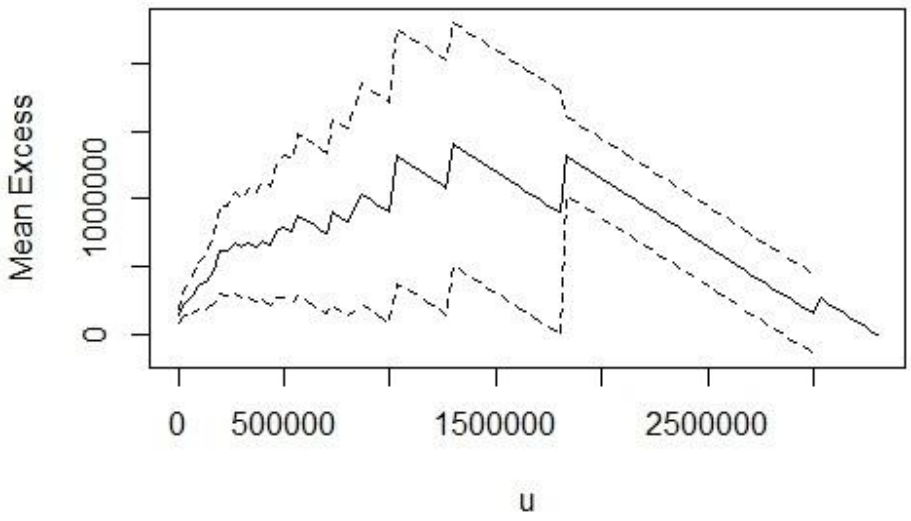


Figure 4.10: Flood losses data Mean Residual Life Plot

Figure (4.9) is the mean residual life plot(MRLplot) of the simulated data. By critical observation, we can see linearity starts occurring at around u = 70000. Choosing this as a suitable threshold, GPD was fitted to the data above the

threshold and the parameters estimated using the traditional maximum likelihood method of parameter estimation.

Table 4.2: Summary Statistics respective Data Sets above their respective thresholds

Statistic	Simulated	Flood Losses
Min	70001.4	155000.0
1st Qu	74515.0	200000.0
Median	85770.0	372500.0
Mean	105000.0	623900.0
3rd Qu	108970.0	700000.0
Max	407000.0	3298000.0

Table (4.2) above is the respective summaries of observations above their respective thresholds in each data, while Table (4.3) Below are the respective summaries of the exceedances over the respective thresholds.

Table 4.3: Summary Statistics of exceeddances above the respective thresholds of both data Sets

Statistic	Simulated	Flood Losses
Min	1.4	5000.0
1st Qu	4515.0	50000.0
Median	1577.0	222500.0
Mean	35020.0	473900.0
3rd Qu	38970.0	550000.0
Max	400000.0	3148000.0

Table 4.4: Simulated Data Parameter Estimates

	Estimates	Standard Error
Shape	0.5384743	0.2268437
Scale	17968.31	3265.7853

4.3 Estimating The Tail Distribution

The tail estimation formula was derived in Chapter 3. The excess distribution was established in equation (3.3), which was shown to be approximately same as

Table 4.5: Real Data Parameter Estimates

	Estimates	Standard Error
Shape	0.5885372	0.2291724

Scale	231839.1	5943.49

the Generalized Pareto distribution. Having fitted the data to the generalized pareto distribution, parameters obtained, and β and the estimation of $F(u)$ is substituted into equation (3.9) to obtain the Tail sstimator distribution $F\hat{}(x)$ in equation (3.10).

For the simulated claim data, there were N_u = 45 exceedances, out of n = 2000 observations. The parameter estimates are as indicated on the summary tables (4.4) and (4.5) for the simulated insrance data and flood losses data respectively. The estimator or the tail distribution function over the threshold u = 70000 for the simulated claim daa is then given by:

$$\hat{F}(x) = 1 - \frac{45}{2000}(1 + \hat{\epsilon}\frac{x - 70000}{\hat{\beta}})^{-1/\hat{\epsilon}}$$

This tail estimator is only valid for observations exceeding the threshold u, i.e $P(X < x \mid X > U)$.

Similarly for the flood losses data, N_u = 35, n = 207 and u = 150000.

The distribution function is given by

$$\hat{F}(x) = 1 - \frac{35}{207}(1 + \hat{\epsilon}\frac{x - 150000}{\hat{\beta}})^{-1/\hat{\epsilon}}$$

Using the probability density function, the results can also be obtained fo particular observations in both data sets. This equation was derived in equation 3.11,

For the simulated insurance claim Data, we have:

$$\hat{f}(x) = \frac{45}{2000}\frac{1}{\hat{\beta}}(1+\hat{\epsilon}\frac{x-70000}{\hat{\beta}})^{-1-\frac{1}{\hat{\epsilon}}} = P(X = x \mid X > u)$$

Similarly for the flood losses real data, we have :

$$\hat{f}(x) = \frac{35}{207}\frac{1}{\hat{\beta}}(1+\hat{\epsilon}\frac{x-150000}{\hat{\beta}})^{-1-\frac{1}{\hat{\epsilon}}} = P(X = x \mid X > u)$$

4.3.1 Value at Risk and Expected Shortfall

In Chapter 3, after the tail distribution was estimated, it was used to derive formulae for the value at risk and expected shortfall over a threshold, and considering all parameters estimated. Equation (3.16) is the value at risk estimator. Having obtained the necessary values, we have for q= 0.95 and q= 0.99,

Results obtained for the simulated claim data are as follows:

For q=0.95

$$\hat{VaR}_{0.95} = 70000 + \frac{\hat{\beta}}{\hat{\epsilon}}((\frac{2000}{45}(1-0.95))^{-\hat{\epsilon}} - 1) = 58338.38697$$

For q=0.99

$$\hat{VaR}_{0.99} = 70000 + \frac{\hat{\beta}}{\hat{\epsilon}}((\frac{2000}{45}(1-0.99))^{-\hat{\epsilon}} - 1) = 88270.7456$$

Results obtained for the flood losses data are as follows:

For q=0.95

$$\hat{VaR}_{0.95} = 150000 + \frac{\hat{\beta}}{\hat{\epsilon}}((\frac{207}{35}(1-0.95))^{-\hat{\epsilon}} - 1) = 562983.8452$$

For q=0.99

$$\hat{VaR}_{0.99} = 150000 + \frac{\hat{\beta}}{\hat{\epsilon}}((\frac{207}{35}(1-0.99))^{-\hat{\epsilon}} - 1) = 1836701.34$$

Following the respective values at risks, the Expected Shortfalls in each case is also computed below. Using the formula derived in equation (3.20), we have:

For the simulated data:

For $\hat{VaR}_{0.95}$ = 58338.38 and $\hat{VaR}_{0.99}$ we have,

$$\hat{ES}_{0.95} = \frac{58338.38}{1-\hat{\epsilon}} + \frac{\hat{\beta} - 70000\hat{\epsilon}}{1-\hat{\epsilon}} = 83664.87283$$

$$\hat{ES}_{0.99} = \frac{88270.7456}{1-\hat{\epsilon}} + \frac{\hat{\beta} - 70000\hat{\epsilon}}{1-\hat{\epsilon}} = 148520.1249$$

For the flood losses data:

$$\hat{ES}_{0.95} = \frac{562983.8452}{1-\hat{\epsilon}} + \frac{\hat{\beta} - 150000\hat{\epsilon}}{1-\hat{\epsilon}} = 1717147.614$$

$$\hat{ES}_{0.99} = \frac{1836701.34}{1-\hat{\epsilon}} + \frac{\hat{\beta} - 150000\hat{\epsilon}}{1-\hat{\epsilon}} = 4812731.212$$

4.3.2 Return Levels

The return level expressions were derived in Chapter 3. Return levels were noted to be levels exceeded per a stated number of observations or the level exceeded over a number of years. ζ_u was defined to be $P\{X > u\}$, the probability that an observation exceeds the threshold u. Equation (3.13) defines the level exceeded once every m-observations and denoted x_m. The N-year return level is define by setting $m = Nxn_y$, where N is the number of years considered and n_y is the number of observations per year. Implementing the poisson λ as the rate of exceedances over the high threshold, if M is the number of years of records and N_u is the exceedances, an unbiased estimate for λ is given by N_u/M. so that,

$$\hat{\zeta}_u = \frac{\hat{\lambda}}{n_y} \Rightarrow \hat{\lambda} = \hat{\zeta}_u \times n_y$$

The resulting N-year return Level is expressed in equation (3.15).

Considering the nature of the data sets, just the m-observation return limit is calculated for the simulated data since it was not simulated on yearly basis. Both the m-observations return limit and the N-year return limit is calculated for the flood losses data data.

Adopting Equation (3.13) and including all values in parameters, we estimate m-observation return levels for the next 500, 600 and 700 observations:

$$x_{500} = 70000 + \frac{\beta}{\epsilon}[(500\zeta_u)^{\epsilon} - 1] = 159477.0299$$

$$\hat{\zeta_u} = P\{X > u\} = \frac{N_u}{n} = \frac{45}{2000} = 0.0225$$

$$x_{600} = 70000 + \frac{\beta}{\epsilon}[(600\zeta_u)^{\epsilon} - 1] = 172149.3695$$

$$x_{700} = 70000 + \frac{\beta}{\epsilon}[(700\zeta_u)^{\epsilon} - 1] = 183878.2779$$

For the real data, we have:

$$x_{50} = 150000 + \frac{\beta}{\epsilon}[(50\zeta_u)^{\epsilon} - 1] = 1139331.24$$

$$\hat{\zeta_u} = P\{X > u\} = \frac{N_u}{n} = \frac{35}{207} = 0.169$$

$$x_{60} = 150000 + \frac{\beta}{\epsilon}[(60\zeta_u)^{\epsilon} - 1] = 1296014.731$$

$$x_{70} = 150000 + \frac{\beta}{\epsilon}[(70\zeta_u)^{\epsilon} - 1] = 1442256.744$$

For N-year return levels, the real dataset was recorded over the period of 2011 to 2017. Using equation (3.15), $\hat{\lambda}$ is estimatied by N_u/M where M is the number of years of records. It comes down to a 6year period data, implying $M = 6 \Rightarrow \hat{\lambda} = 35/6 = 5.89$, this is to say, on average, the threshold 150,000 is exceeded roughly about 6 times each yr..

for a 2-year and a 3-year return levels, the results obtained are as follows:

$$x_2 = 150000 + \frac{\hat{\beta}}{\hat{\epsilon}}[(2 \times 5.89)^{\hat{\epsilon}} - 1] = 1438058.752$$

$$x_3 = 150000 + \frac{\hat{\beta}}{\hat{\epsilon}}[(3 \times 5.89)^{\hat{\epsilon}} - 1] = 1891370.643$$

$$x_4 = 150000 + \frac{\hat{\beta}}{\hat{\epsilon}}[(4 \times 5.89)^{\hat{\epsilon}} - 1] = 2285309.434$$

It is observed that, the return levels increase as the number of observations increase in the m-observation return levels and also increases with increasing periods in the N-year return levels. It is also observed that they both increase at a decreasing rate but as to if it converges or not is unknown for now. But looking at it with an intuitive point of view, it should not converge as we expert worse and more catastrophic events in the future than has already occurred. When it comes to damages and loss claims pertaining to floods, the same magnitude of an event should result in more cost today than it did years back. This is because of growth in population, growth in infrastructure due to development and many more. Return levels are so therefore expected to increase without converging over increasing number of observations and years.

Chapter 5

Summary, Conclusion and Recommendations

5.1 Summary

This thesis was undertaken to explore the application of EVT with reference to flood related losses, basically, creating some kind of awareness of the intensity of floods and quantifying how much in monetary terms these losses occur. The

summary and discussions about the findings of the respective analysis performed are covered in this chapter. The conclusions on the findings and corresponding recommendations are also covered.

5.2 Discussion of Findings

The tail estimator obtained is basically the CDF ($\hat{F}(X)$) of the observations above the threshold u. The threshold was chosen by meticulously observing the Mean Excess, QQ, Zipf plots and MRL plots. Having found a range of suitable thresholds, the threshold "u" with the best possible fit and shape parameter (preferably about 0.5) is chosen. The corresponding PDFs ($\hat{f}(x)$) were also obtained by finding the first differential of $\hat{F}(X)$.

Considering the high quantiles of the tail estimator, we obtain the VaR at q=0.95 and q=0.99 for both data sets. Considering the simulated claims data, at the 95th quantile, our VaR is approximately GHC58,338.4 and at the 99th quantile, we have GHC88,270.7. Implying that, at 95% and 99% confidence respectively, we do not expect losses to exceed those limits. In other words, there is a 5% and 1% chance respectively for the losses to exceed those levels.
This is also observed for the real data where the VaR at the 95th quantile gives GHC562,983.8 and the 99th quantile gives GHC1,836,701.3.

The Expected Shortfall or the tail conditional expectation is another informative measure of risk. It estimates the potential size of the loss exceeding the VaR, in simple and more technical words, it is the expected size of the loss that exceeds the VaR. The respective Expected Shortfalls for VaRs estimated for the simulated data in chapter 4 are GHC83,664.9 and GHC148520.1 corresponding to the 95th

and 99th quantiles respectively. We also estimated GHC1717147.6 and GHC4812731.212 for the real data and also for the 95th and 99th quantiles respectively.

Return levels can be used as a measure of the maximum loss that is expected over a period of time. It is the maximum level we expect the losses to exceed over some period or over some number of occurring events. It is a rather more conservative measure as compared to the VaR. The simulated data was not defined over a period of time so the estimated return levels are based on occurring events. We calculated for the next 500, 600 and 700 observations and obtained GHC159,477.02, GHC172,149.37 and GHC183,878.28 respectively. Considering the real data, return levels for the next 50, 60 and 70 observations were calculated and GHC1,139,331.24, GHC1,296,014.73 and GHC1,442,256.74 were the respective return levels estimated. We also calculated Return levels for periods for the Real data, this is possible because the data collected over a period of 5years. For 2,3 and 4 years, the return levels estimated were, GHC1,438,058.75, GHC1,891,370.64 and GHC2,285,309.43 respectively.

It was observed from the return level plots that the return levels increase with increasing periods or number of observations but they do not seem to converge. Thinking intuitively, we actually do not expect the return levels to converge. This is because, considering loss in property damages, we expect more cost and lose in lives today than years ago assuming the same magnitude in flood hits, and this is due to growth in infrastructure and human populations over time. Basically there will be relatively larger costs in losses in the future than we experience in recent times because all settlements keep growing. Return levels are therefore not expected to converge.

5.3 Conclusion

The study was undertaken using the Peaks Over Threshold method of the extreme value theory. The aim was to study flood losses in general. Two data sets were used, one was a simulated flood insurance claim data while the other was obtained from the Natural Disaster Management Organisation. The method was applied to both data sets. The choice of the threshold influenced the maximum likelihood parameters estimated. Different techniques aided in the choice of a good threshold.

EVT is now one of the most employed techniques in risk management. It is now one of the few techniques you will find in a good risk manager's toolkit. As already mentioned previously, whenever tails of a probability distribution is considered, it is for a researcher's own good to consider applying the theoretically supported methods of EVT. The methods we already know based around the Gausian Distribution is likely to underestimate the tail risk, EVT is the most scientific approach to an inherently difficult problem, that is, predicting the size of a rare event.

This thesis covers the possibilities of losses, and creates the awareness for anyone who might find it important, so as to prepare adequately for the possibilities of extreme flood cases. Extreme losses are mostly underrated and stating the estimated figures of the losses will help prevent underrating extremes.

REFERENCES

A, T. (1992). Estimating probabilities of extreme sea-levels. *Article*.

Asumadu-Sarkodie, S., Owusu, P., and Rufangura, P. (2015). Impact analysis of flood in accra, ghana. 6:53–78.

Beirlant, J., G. Y. D. J. . T. J. (2004). Statistics of extremes: Theory and applications,. *Wiley, London*.

Beirlant, J., V. P. . T. J. L. (1996). Tail index estimation, pareto quantile plots and regression diagnostics. *J. Am. Statist. Assoc, 91, 1659-1667*.

Britannica, E. (2017). Mississippi river flood of 1927. *Encyclopaedia Britannica, Inc, https://www.britannica.com/event/Mississippi-River-flood-of-1927*.

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

Csorgo, M., P. D. . P. M. (1985). Kernel estimator of the tail index of a distriution. *Annals of Statistics 13: 1050-1077*.

Daniel Cooley, D. N. and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal*.

Danielsson, J., D. W. J. . C. G. d. V. (1996). The mothod of moments ratio estimator for the tail shape paremeter. *Com. Stat. - Th. Meth, 25(4)*.

Danielsson, J., H. P. . d. V. C. (1998). The cost of conservatism. *Risk 11(1), 101-103*.

Darehshiri, A., P. M. M. A. (2015). Identifying geochemical anomalies associated with cu mineralization in stream sediment samples in gharachaman area, northwest of iran. *Journal of Geochemical exploration. 110, 92-99*.

Davison, A. C. & Smith, R. L. (1990). Models for exceedances over high thresholds(with discussions). *J, R, Statististics Society, B, 52, 237-254*.

Dekkers, P., d. H. L. (1989). On the estimation of the extreme value index and large quantile estimation. *The Annals of statistics 17 (4), 1795-1832*.

Dodd, E. (1923). The greatesst ans least variate under general laws of error. *Transactions of the American Mathematical Society 25, 525-539.*

Drees, H. (1995). Refined pickands estimator of the extremal index. *Annals of Statistcis 23, 2059-2080*.

Embrechts, P., K. C. . M. T. (1997). Modelling extremal events for insurance and finance. *Springer*.

Embrechts, P., R. S. . S. G. (1998). Living on the edge. *Risk Magazine 11(1), 96-100*.

Fabio Rossi, P. V. (1984). Regional flood estimation methods. *Article*.

Falk, M. (1994). Efficiency of convex combinations of pickands' estimator of extreme value index. *Journal of Nonparametric Statistics 4, 133-147*.

Fisher, R. A., T. L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallesst member of a sample. *Proceedings of the Cambridge Philosophical Society 24, 180-190*.

Fraga Alves, M. (1995). Estimation of the tail parameter in the domain of attraction of an extremal distribution. *Journal of Statistical Planning and Inference*.

Ghosh, S. and Resnick, S. (2010). Discussion on mean excess plots. *Article*.

Gnedenko, B. V. (1943). *Sur la distribution limite du terme d'une serie aleatoire*. Annals of Mathematics 44.

Harter, H. L. (1978). A bibliography of extreme-value theory. *International Statistical Review / Revue Internationale de Statistique*, 46(3):279–306.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of statistics 13, 331-341*.

Hosking, J. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum(minimum) values of meteorological events. *Quarterly Journal of th Royal*.

Katz, R. W. (2010). Statistics of extremes in climate change. *article*.

Kratz, M. & Resnick, S. I. (1996). The qq estimator and heavy tails. *Communication in Statistics-Stochastic Models 12 (4), 699-724*.

Leadbetter, M. R.; Lindgren, G. . R. H. (1983). *Extremes and Related Properties of Random Sequence and Series*. Springer.

McNeil, A. (1997). Estimating the tails of loss severity distributions using evt. *ASTIN Bulletin 27, 117-137*.

McNeil, A. J. (1999). Extreme value theory for risk managers,. *British Bankers' Association, Internal Modelling and CAD II: Qualifying and Quantifyiing Risk withing a Financial Institution 93-113, RISK Books, London.*

McNiel, A. J. Frey, R. (2000). *Estimation of tail-related risk measures for heteroskedastic financial time series: an extreme value approach*. Journal of Empirical Finance 7.

Mikosch, T. (1997). Heavy-tailed modelling in insurance. *Communication in Statistcs-Stochastic Models 13(4), 799-815*.

Mises, R. V. (1936). *La distribution de la plus grande de n valeurs,*. Union Interbalcanique, 1.

Musah, A.-A. I. (2010). Application of extreme value theory for estimating daily brent crude oil prices. *Thesis*.

Pareto, V. (1897). *Cours d'Economie Politique, Vol 2*. F. Rouge, Lausanne.

Paul Embrechts, Claudia Kluppelberg, T. M. (1997). *Modelling Extremal Events*. Springer.

Pereira, T. T. (1994). Second order behaviour fo domains of attraction and the bias of generalized pickands' estimator, in "extreme value theory applications iii". *Proceedings of Gaithersburg-Conference, 1993,vol. 866, pp. 165-177.*

Pickans, J. (1975). Statistical inference using extreme order statistics. *The Annnals of Statistics 3(1), 119-131*.

Reiss, Rolf-Dieter, T. M. (2001). *Statistical Analysis of Extreme Values*. Springer.

Resnick, S. (1997). Heavy tailed modelling and teletraffic data. *The Annuals of Statistics 25 (5), 1805-1869*.

Ribatet, M., E. S. J. M. G. and Ouarda, T. B. M. J. (2007). A regional bayesian pot model for flood frequency analysis. *Stochastic Environ, Res, Risk Assess,. 21(4), 327-339*.

Ribatet, M. A. (2006). A user's guide to the pot package (version 1.4). *URL://cran.r-project.org/*.

Scarrot C. J., M. A. L. D. (2013). Boundary correction, consistency and robustness of kernel densities using extreme value theory. *Article*.

Schultze, J. & Steineback, J. (1996). On least squares estimates of an exponential tail coefficient. *Statist, Decisions 14, 353-372*.

Sheng, G. (2012). Estimation of hot and cold spells with extreme value theory. Master's thesis, Uppsala University.

Smith, L. R. (1989). Extreme value analysis of environmental time series: an application to trend detection in round-level ozone. *Statistical Science 4*.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika, 72(1): 67-90*.

Tajvidi, H. R. . N. (1997). Extreme value statistics and wind storm losses: a case study. *Scandinavian Actuarial Journal, 1, 70-94,*.

Thomas M, Lemaitre M, W. M. V. C. Y. Y. W. H. e. a. (2016). Applications of extreme value theory in public health. *PloS ONE 11(7):e0159312. https://doi.org/10.1371/journal.pone.0159312*.

Zhao, G., Wilde, S. A., Cawood, P. A., and Lu, L. (1998). Thermal evolution of archean basement rocks from the eastern part of the north china craton and its bearing on tectonic setting. *International Geology Review*, 40(8):706–721.

Zhao, X., S. C. J. R. M. . O. L. (2010). *Extreme value modelling for forcasting the market crises*. Appl. Fin. Econ, 20, 63-72.

Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. addison-wisley. *Cambridge, MA*.

Appendix A

5.4 R codes

5.4.1 Codes for Simulated Data

```
> set.seed(200)
> Dat1 = rlnorm(2000,meanlog = 9.454,sdlog = 0.8)
> summary(Dat1)
> hist(Dat1,breaks = 50,xlim = c(0,500000),ylim = c(0,1000))
> quantile(Dat1,0.95)
> quantile(Dat1,0.99)
> cho.th = mrl.jas(Dat1)
```

```
> above.thresh = Dat1[Dat1 > 70000]
> length(above.thresh)
> summary(above.thresh)
> above.thresh
> EPi = emplot(Dat1,alog = "xy")
> EPi2 = emplot(above.thresh,alog = "xy")
> rmMEPl = meplot(Dat1,omit = 4)
> rmMePl2 = meplot(above.thresh,omit = 3)
> QQ1 = qplot(Dat1,xi = 0.5,threshold = 70000)
> QQ2 = qplot(Dat1,xi = 0.5)
> thresh.exceed = above.thresh − 70000
> summary(thresh.exceed)
> fit2 = gpd.fit(Dat1,threshold = 70000)
> diag = gpd.diag(fit2)
```

5.4.2 Real Data Codes

```
> Flood = read.csv(file.choose())
> str(Flood)
> summary(Flood)
> Data = Flood$AMOUNT
> summary(Data)
> hist(Data,breaks = 50,xlim = c(0,3500000),ylim = c(0,200))
> quantile(Data,0.90)
> cho.th = mrl.jas(Data)
> above.thresh2 = Data[Data > 150000]
> length(above.thresh2)
> summary(above.thresh2)
> epi = emplot(Data,alog = "xy")
```

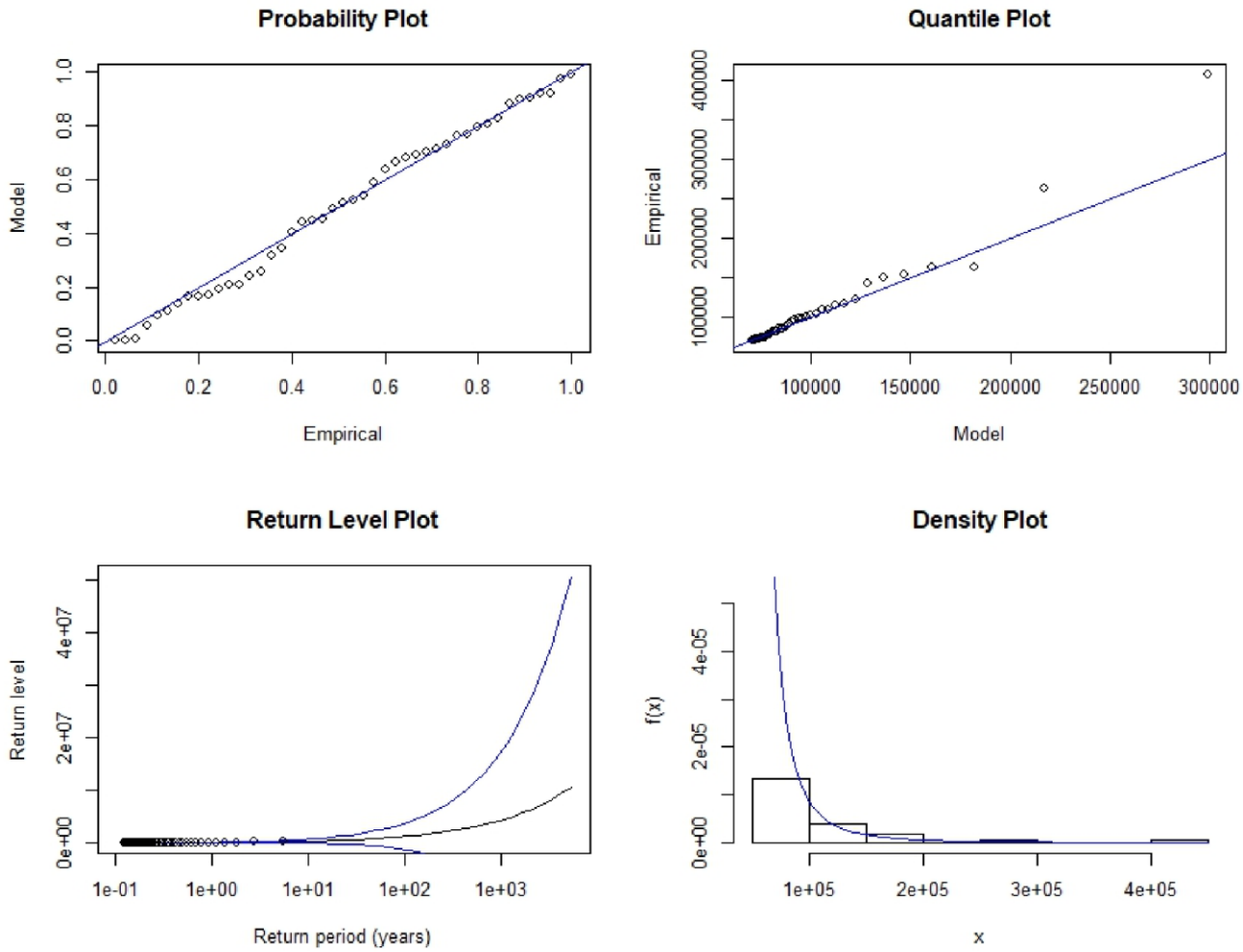
```
> epi2 = emplot(above.thresh2,alog = "xy")
> Mepl = meplot(Data,omit = 2)
> Mepl2 = meplot(above.thresh2,omit = 2)
> qq1 = qplot(Data,xi = 0.5)
> qq2 = qplot(Data,xi = 0.5,threshold = 150000)
> thresh.exceed2 = above.thresh2 - 150000
> summary(thresh.exceed2)
> FIT = gpd.fit(Data,threshold = 150000)
> plots = gpd.diag(FIT)
```

5.4.3 Mean Residual Life Plot Codes(mrl.jas)

```
> mrl.jas = function(data,umin = min(data),umax = max(data) -
0.1,conf = 0.95,
+ nint = 100)
+ {+x = xu = xl = numeric(nint)
+ u = seq(umin,umax,length = nint)
+ for(iin1 : nint){
+ cbind(x,u)
+}
>
```

Table 5.1: Structure of Flood losses

DATE	DISTRICT	AMOUNT
05/07/2010	NEW AMAKOM	GHC 6,000.00
20/08/2010	NORTH PATASE	GHC 40,000.00
22/09/2010	DICHEMSO ATEPOMYA	GHC 6,700.00
04/01/2011	ATIMATIM	GHC 64,500.00
17/07/2011	NSESE	GHC 145,000.00
17/07/2011	JUABEN ESTATE	GHC 5,000.00
18/07/2011	KONONGO-ODUMASE	GHC 150,000.00
18/07/2011	ASOKORE	GHC 60,000.00



28/07/2011	ADANSI SOUTH	GHC 60,750.00
17/07/2011	EJISU JUABEN	GHC 5,000.00
18/07/2011	ASANTE AKIM NORTH MUNI.	GHC 150,000.00
18/07/2011	SEKYERE EAST	GHC 60,000.00
28/07/2011	ADANSI SOUTH	GHC 60,750.00
17/07/2011	BOSOME FREHO	GHC 200,000.00
18/07/2011	ASANTE AKIM NORTH MUNI.	GHC 150,000.00
15/08/2011	SEKYERE AFRAM PLAINS	GHC 20,000.00

5.5 Diagnostic Plots

Figure 5.1: Diagnostic Plots of Simulated Data

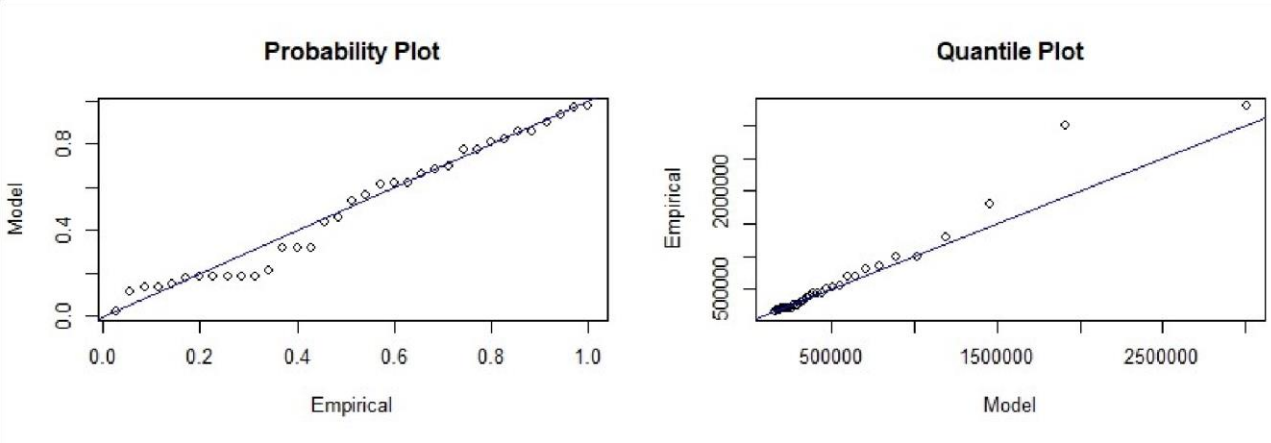


Figure 5.2: Diagnostic Plots of Real Data