Using Statistical Approaches to Model Natural Disasters

Audrene S. Edwards, Kumer Pial Das, Ph.D.

Department of Mathematics, Lamar University, Beaumont, TX

*Students: *audreneedwards@yahoo.com*
Mentor: kumer.das@lamar.edu

ABSTRACT
The study of extremes has attracted the attention of scientists, engineers, actuaries, policy makers, and statisticians for many years. Extreme value theory (EVT) deals with the extreme deviations from the median of probability distributions and is used to study rare but extreme events. EVT's main results characterize the distribution of the sample maximum or the distribution of values above a given threshold. In this study, EVT has been used to construct a model on the extreme and rare earthquakes that have happened in the United States from 1700 to 2011.The primary goal of fitting such a model is to estimate the amount of losses due to those extreme events and the probabilities of such events. Several diagnostic methods (for example, QQ plot and Mean Excess Plot) have been used to justify that the data set follows generalized Pareto distribution (GPD). Three estimation techniques have been employed to estimate parameters.The consistency and reliability of estimated parameters have been observed for different threshold values.The purpose of this study is manifold: first, we investigate whether the data set follows GPD, by using graphical interpretation and hypothesis testing. Second, we estimate GPD parameters using three different estimation techniques.Third, we compare consistency and reliability of estimated parameters for different threshold values.Last, we investigate the bias of estimated parameters using a simulation study. The result is particularly useful because it can be used in many applications (for example, disaster management, engineering design, insurance industry, hydrology, ocean engineering, and traffic management) with a minimal set of assumptions about the true underlying distribution of a data set.

KEYWORDS
Extreme Value Theory; QQ Plot; Mean Excess Plot; Mean Residual Plot; Peak Over Threshold; Generalized Pareto Distribution; Maximum Likelihood Method; Method of Moments; Probability-Weighted Moments; Shapiro-Wilk test; Anderson-Darling Test

1. INTRODUCTION

Extreme Value Theory (EVT) can be used to analyze data that deviates from the median of probability distributions, and can be used as a tool for analyzing the risk for events that happen seldom, for example events such as the California earthquake of 1906, and Hurricane Katrina from 2005. Usually the earthquakes that occur in America are non catastrophic, resulting in minor damages, and even minor expenditures when it comes to restoring the city, or town that the earthquake happened in. But then there are earthquakes with higher than normal magnitudes that are catastrophic, resulting in major damage of the infrastructure within a city or town, as well as high expenditures to reconstruct its infrastructure. Catastrophic earthquakes rarely happen in comparison to non catastrophic earthquakes. Since earthquakes with high magnitudes happen seldom, analyzing or modeling these rare events presents a unique and important challenge. These events can be analyzed and modeled by EVT which provides tools to focus on extreme and rare events that occur. For this study, EVT will be used to focus on the extreme and rare earthquakes that have happened in the United States from 1700 to 2011.

In this paper, starting in Section 2, we will discuss two methods, namely, block maxima and peak over threshold (POT), which can be used to analyze a data set containing extreme values. Proceeding to Section 3, we will discuss definition, formulation and properties of three estimation techniques namely, maximum likelihood estimation (MLE) method, method of moments (MOM), and method of probability weighted moments (PWM). In Section 4, we describe the data set and use graphical procedure in modeling the right-hand tail. In Section 5, we compare GPD estimators obtained by three estimation techniques: MLE, MOM and PWM, as the threshold increases. In Section 6, we see if GPD fits into the data by performing the Anderson-Darling test. In Section 7, a simulation study has been conducted to examine bias and root mean square error (RMSE) for each estimator produced by the three techniques. Mis-specification bias has also been investigated in this section. And finally, conclusions appear in Section 8.

2. EXTREME VALUE THEORY

Extreme value analysis is a branch of statistics focusing attention on issues for modeling extreme values.[1] It deals with the extreme deviations from the median of probability distributions and seeks to assess, from a given ordered sample, the probability of events that are more extreme than a certain large value. EVT states that the shape of the distribution below this certain large value is not important. From an actuarial point of view, the distributions of large losses are very important since the insurer (or re-insurer) makes a payment only when the loss exceeds a certain large value. There are two common approaches to analyze an extreme data set: the block maxima modeling technique, and the peak over threshold (POT) technique. While both are used for modeling extreme events, each has more specific uses, especially when deciding which model would be preferable to work with the data set.

2.1 *Block Maxima Method*

In block maxima modeling, the data is grouped into *blocks* of equal length and fit the data to the maximums of each block, for example, annual maximum of daily rainfall amounts. The block maxima approach is closely associated with the use of the generalized extreme value distribution. The choice of block size can be critical to the entire study as blocks that are too small can lead to bias and blocks that are too large generate too few block maxima. Moreover, it is very often more useful to analyze the values of random variables that exceed or fall below a given threshold value. For example, in this study, it would be better to have earthquake data that exceed a certain magnitude than to have data that consist of only yearly maximum over a period of n years. Maxima in some years can be much below than several high-order statistics in other years. Thus, POT approach has been applied in this study.

2.2 *Peaks Over Threshold (POT) Method*

Let $Y_1, Y_2, \cdots, Y_n$ be a sequence of independent and identically distributed (iid) random variables with common distribution function $F(y)$.

To model the upper tail of $F(y)$, consider k exceedances of Y over a threshold u and let $X_1, X_2, \cdots, X_k$ denote the excesses (or peaks). POT is used when taking these peak values that occurred during any period of time, from a continuous record. POT depends on the threshold u, and is defined by X_i, where

$$X_i := Y_i - u | Y_i > 0 \qquad \textbf{Equation 1.}$$

are the exceedances over u for $i = 1, 2, \cdots, k$, which are asymptotically distributed and follow a GPD.[2]

2.3 *Generalized Pareto Distribution*

The cumulative distribution function (cdf) of GPD is defined as:

$$F(x) = \begin{cases} 1 - \left(1 - \frac{kx}{\alpha}\right)^{\frac{1}{k}} & \text{if } k \neq 0 \\ 1 - \exp\left(\frac{-x}{\alpha}\right) & \text{if } k = 0 \end{cases} \qquad \textbf{Equation 2.}$$

where, the scale parameter, $\alpha > 0$ and for the shape parameter $k \geq 0$ the support is $0 \leq x < \infty$ while for $k < 0$ the support

is $0 \leq x \leq -\alpha/k$.[3]
The probability density function (pdf) of GPD is defined as:

$$f(x) = \begin{cases} \frac{1}{\alpha}\left(1 - \frac{kx}{\alpha}\right)^{\frac{1}{k}-1} & \text{if } k \neq 0, \alpha > 0, \left(1 - \frac{kx}{\alpha}\right) \geq 0 \\ \frac{1}{\alpha}\exp(\frac{-x}{\alpha}) & \text{if } k = 0, \alpha > 0, x \geq 0. \end{cases} \qquad \textbf{Equation 3.}$$

There are three related distributions in the family of GPD. The distributions are: exponential, Pareto, and beta. GPD has popularly been used in analyzing extreme natural and man made events.[4-6]

3. ESTIMATION TECHNIQUES TO FIND PARAMETERS

3.1 *The Maximum Likelihood Method*

Let $X_1, X_2, ..., X_n$, be a random sample, that depends on one or more unknown parameters say, $(\theta_1, \theta_2, ...\theta_m)$. Assume that those unknown parameters $(\theta_1, \theta_2, ...\theta_m)$, are restricted to a given parameter space, Ω. Then the joint probability mass function (p.m.f) of the random sample size $X_1, X_2, ..., X_n$, is called the likelihood function, such that

$$L(\theta_1, \theta_2, ...\theta_m) = f(x_1; \theta_1, \theta_2, ...\theta_m)f(x_2; \theta_1, \theta_2, ...\theta_m)......f(x_n; \theta_1, \theta_2, ...\theta_m), \qquad \textbf{Equation 4.}$$

where $\theta_1, \theta_2, ..., \theta_m \in \Omega$, and is represented as

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta_n), \qquad \textbf{Equation 5.}$$

when regarded as a function of $\theta_1, \theta_2, ..., \theta_m$.

Consider

$$[u_1(X_1, X_2, ...X_N), u_2(X_1, X_2, ...X_N), ..., u_m(X_1, X_2, ...X_N)]. \qquad \textbf{Equation 6.}$$

We refer to $[u_1(X_1, X_2, ...X_N), u_2(X_1, X_2, ...X_N), ..., u_m(X_1, X_2, ...X_N)]$, as the m-tuple in Ω that maximizes $L(\theta_1, \theta_2, ...\theta_m)$. Then, $\hat{\theta_1} = u_1(X_1, X_2, ...X_N)$, $\hat{\theta_2} = u_2(X_1, X_2, ...X_N), ..., \hat{\theta_m} = u_m(X_1, X_2, ...X_N)$, are the unique maximum likelihood estimators of

$$[u_1(X_1, X_2, ...X_N), u_2(X_1, X_2, ...X_N), ..., u_m(X_1, X_2, ...X_N)]. \qquad \textbf{Equation 7.}$$

Let $x = \{x_1, x_2, ..., x_n\}$ be a sample from the GPD with pdf in **Equation 3**:

$$\begin{aligned} L(x_i; k, \alpha) &= -n\ln\alpha + \left(\frac{1-k}{k}\right)\sum_{i=1}^{n}\ln\left(1 - \frac{k}{\alpha}x_i\right), \text{ for } k \neq 0 \\ &= -n\ln\alpha - \frac{1}{\alpha}\sum_{i=1}^{n} x_i, \text{for } k = 0. \end{aligned} \qquad \textbf{Equation 8.}$$

The log-likelihood function for $k \neq 0$ states that the function can be made arbitrarily large by taking $k > 1$ and α/k close to the maximum order statistic $x_{n:n}$.[7] Maximum likelihood estimates of α and k can be obtained by minimizing the likelihood function above. However, in order for the MLE to perform its best for the GPD, there are certain criterion that must be present. One, the sample size n, must be large (preferably, greater than 500). Two, the values of k, the shape parameter must stay within the bounds of $\frac{-1}{2}$ and $\frac{1}{2}$. When these criterion are met, MLE would be preferred due to its effective efficiency with large samples.[4]

3.2 *The Method of Moments*

Let $X_1, X_2, ...X_n$, be a random sample size from a distribution with the following p.d.f:

$$f(x_i; \theta_1, \theta_2, ...\theta_r), \qquad \textbf{Equation 9.}$$

where the sample space, $\theta_1, \theta_2, ..., \theta_r \in \Omega$. Then $E\left(x^k\right)$, where $k = 1, 2, ...$ is the k^{th} moment of the population, and $M_k = \sum_{i=1}^{n} \frac{x_i^k}{n}$, where $k = 1, 2, ...$ is the k^{th} moment of the sample. We take $E\left(x^k\right)$, set it equal to M_k, starting with $k = 1$, and keep equating $E\left(x^k\right)$ to M_k until enough equations are provided to find the unique solutions for the parameters, $\theta_1, \theta_2, ..., \theta_r$.[7]

In this study, for GPD, MOM can be used to find the unique solutions for the parameters, k and α, by finding the first and second population moment, and equating them with the corresponding sample moments.

We will first start with finding the first and second population moments of GPD. In order to achieve the task at hand, we will use the expected value approach, instead of the moment generating function (mgf) approach, since the mgf of the GPD does not exist for all values of k. Consider $E\left(1 - \frac{kx}{\alpha}\right)^r$, where the r^{th} moment exist when $k > \frac{-1}{r}$. Assume $r = 1$ and $k < 0$, where $0 < x < \infty$. By the definition of expected value and **Equation 3**, we now have:

$$\begin{aligned} E\left(1 - \frac{kx}{\alpha}\right)^1 &= \int_0^\infty \left(1 - \frac{kx}{\alpha}\right)^1 \times \left[\frac{1}{\alpha}\left(1 - \frac{kx}{\alpha}\right)^{\frac{1}{k}-1}\right] dx \\ &= \frac{1}{\alpha}\int_0^\infty \left(1 - \frac{kx}{\alpha}\right)^{\frac{1}{k}} dx. \end{aligned} \qquad \textbf{Equation 10.}$$

using the method of substitution, we have:

$$E\left(1 - \frac{kx}{\alpha}\right) = \frac{1}{1+k}. \qquad \textbf{Equation 11.}$$

Therefore, for $r = 1$, $k < 0$, and $0 < x < \infty$, we have

$$E\left(1 - \frac{kx}{\alpha}\right)^r = \frac{1}{1+rk}, \text{ where } 1 + rk > 0. \qquad \textbf{Equation 12.}$$

For $r = 1$, the properties of expectation have been used to obtain $E(x)$ as follows:

$$\begin{aligned} E(x) &= \frac{\alpha}{1+k} \\ &= \mu. \end{aligned} \qquad \textbf{Equation 13.}$$

Similarly, letting $r = 2$ in **Equation 12**, $E\left(x^2\right)$ can be found as:

$$E\left(x^2\right) = \frac{2\alpha^2}{(1+k)(1+2k)} \qquad \textbf{Equation 14.}$$

By using $E\left(x^2\right)$ and $E(x)$,

$$\sigma^2 = \frac{\alpha^2}{(1+k)^2(1+2k)}. \qquad \textbf{Equation 15.}$$

Now that we have μ and σ^2, we can now equate the first sample moment with the first population moment, using **Equation 12**:

$$\bar{x} = \frac{\alpha}{1+k}. \qquad \textbf{Equation 16.}$$

where $\bar{x}$ is the first sample moment, and μ is the first population moment.

We can also equate our second sample moment with the second population moment as such:

$$S^2 = \frac{\alpha^2}{(1+k)^2(1+2k)}. \qquad \textbf{Equation 17.}$$

where S^2 is the second sample moment, and σ^2 is the second population moment. These two equations (**Equation 16** and **Equation 17**) will provide the unique solutions for k and α, in which we will be able to find out the MOM estimators for k and α denoted by $\hat{k}$ and $\hat{\alpha}$ respectively as follows:

$$\hat{k} = \frac{1}{2}\left(\frac{\bar{x}^2}{S^2} - 1\right) \text{ and } \hat{\alpha} = \frac{1}{2}\bar{x}\left(\frac{\bar{x}^2}{S^2} + 1\right). \qquad \textbf{Equation 18.}$$

3.3 *The Probability-Weighted Moments*

Let $x_{1:n} \leq x_{2:n} \leq \ldots\ldots.. \leq x_{i:n}$ be a random sample from GPD (k, σ), where $x_{i:n}$ is the ith order statistics of sample size n. Consider

$$\alpha_v = \frac{1}{n}\sum_{i=1}^{n}(1 - P_{i:n})^v\, x_{i:n}. \qquad \textbf{Equation 19.}$$

where $P_{i:n} = \frac{i-.35}{n}$. Then,

$$\hat{k}_{PWM} = \frac{\alpha_o}{\alpha_o - 2\alpha_1} - 2 \text{ and } \hat{\sigma}_{PWM} = \frac{2\alpha_o\alpha_1}{\alpha_o - 2\alpha_1}. \qquad \textbf{Equation 20.}$$

The following is an illustration of how the parameters $\hat{k}$ and $\hat{\sigma}$ were calculated in terms of $\bar{x}$ and t, from α_v. Note that,

$$\hat{k}_{PWM} = \frac{\bar{x}}{\alpha_o - 2\alpha_1} - 2 \qquad \textbf{Equation 21.}$$

where $\alpha_v = \frac{1}{n}\sum_{i=1}^{n}(1 - P_{i:n})^v\, x_{i:n}$.

Let $v = 0$, then

$$\begin{aligned}\alpha_0 &= \frac{1}{n}\sum_{i=1}^{n}(1 - P_{i:n})^0\, x_{i:n}\\ &= \frac{1}{n}\sum_{i=1}^{n} 1 \times x_{i:n}\\ &= \bar{x}.\end{aligned} \qquad \textbf{Equation 22.}$$

and let $v = 1$,

$$\alpha_1 = \frac{1}{n}\sum_{i=1}^{n}(1 - P_{i:n})\, x_{i:n} \qquad \textbf{Equation 23.}$$

where $\frac{1}{n}\sum_{i=1}^{n}(1 - P_{i:n})\, x_{i:n}$, will be represented by t, so $\alpha_1 = t$. From **Equation 21** we have:

$$\hat{k}_{PWM} = \frac{\bar{x}}{\bar{x} - 2t} - 2. \qquad \textbf{Equation 24.}$$

The same follows for $\hat{\sigma}_{PWM}$:

$$\hat{\sigma}_{PWM} = \frac{2\alpha_o\alpha_1}{\alpha_o - 2\alpha_1} \qquad \textbf{Equation 25.}$$

where as previously calculated $\alpha_0 = \frac{1}{n}\sum_{i=1}^{n} 1 \times x_{i:n} = \bar{x}$ and $\alpha_1 = \frac{1}{n}\sum_{i=1}^{n}(1-P_{i:n})\,x_{i:n} = t$. Then,

$$\hat{\sigma}_{PWM} = \frac{2\left(\frac{1}{n}\sum_{i=1}^{n} x_{i:n}\right)\left(\frac{1}{n}\sum_{i=1}^{n}(1-P_{i:n})\,x_{i:n}\right)}{\left(\frac{1}{n}\sum_{i=1}^{n} x_{i:n}\right) - 2\left(\frac{1}{n}\sum_{i=1}^{n}(1-P_{i:n})\,x_{i:n}\right)} = \frac{2\bar{x}t}{\bar{x}-2t}.$$

Equation 26.

So,

$$\hat{k}_{PWM} = \frac{\alpha_o}{\alpha_o - 2\alpha_1} - 2 = \frac{\bar{x}}{\bar{x}-2t} \text{ and } \hat{\sigma}_{PWM} = \frac{2\alpha_o\alpha_1}{\alpha_o - 2\alpha_1} = \frac{2\bar{x}t}{\bar{x}-2t}$$

Equation 27.

4. ANALYZING THE EARTHQUAKE DATA

To fulfill a complete understanding of EVT, this study uses a data set of earthquakes.[8] Forecasting earthquake is always a challenging problem and it has been studied by many authors.[9-14] Turning to the United States Geological Survey, the data of historic earthquakes in the U.S. and its territories were retrieved. Magnitudes of earthquake, measured in Richter scale, from 1700 to 2011 have been used in this study. Using the statistical package R the data was processed so that the following information could be found: 5 number summary, a histogram, box plot, Q-Q plot, mean residual plot and mean excess plot.

4.1 *Descriptive Statistics*

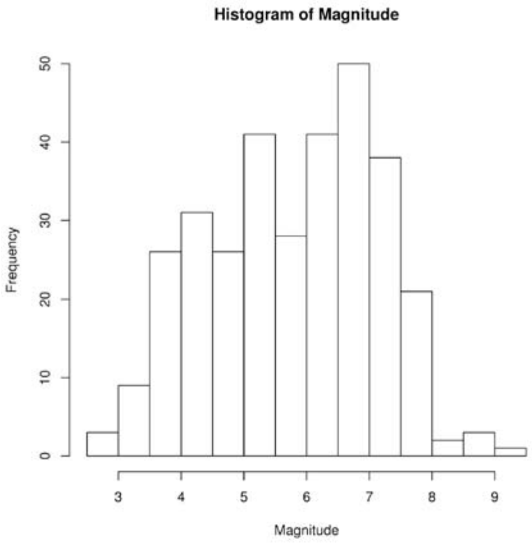
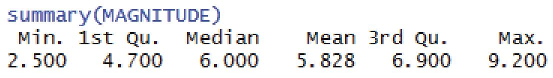


Figure 1. Histogram of Magnitude

We construct a histogram to understand the overall nature of the data. Even though the histogram does not show a very skewed pattern, it can be seen that there are some extreme values in the right tail. The box plot provides a visual for the 5 number summary plot, which includes the median of the data, the 1st and 3rd quartile, and the maximum and minimum values of the data. The 5 number summary and a box plot can be obtained by using R:



```
summary(MAGNITUDE)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.500   4.700   6.000   5.828   6.900   9.200
```

Figure 2. Five Number Summary

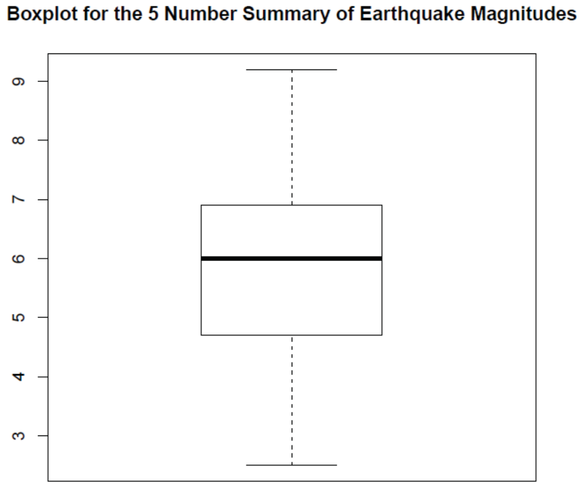


Figure 3. Box Plot for Earthquake Magnitude Data

The 5 number summary in **Figure 2** and the box plot in **Figure 3** reveals the identical information. We see that our maximum value and minimum value of the data is 9.2, and 2.5 respectively which the box plot portrays with the highest bar that is at 9.2, and the lowest bar that is at 2.5. Notice that there is a thick black line within the box. This line is the median of the data. The median of the data is 6.0, and that is where the line is placed. The space below the black line is the 1st quartile and the space above the black line is the 3rd quartile (which seems to be where most of our data lies).

4.2 *The Q-Q Plot*

The Q-Q plot is a graphical analysis of the distribution of the processed data compared to the normal distribution. When the model fits the data well, the pattern of points in Q-Q plots would exhibit a linear trend.

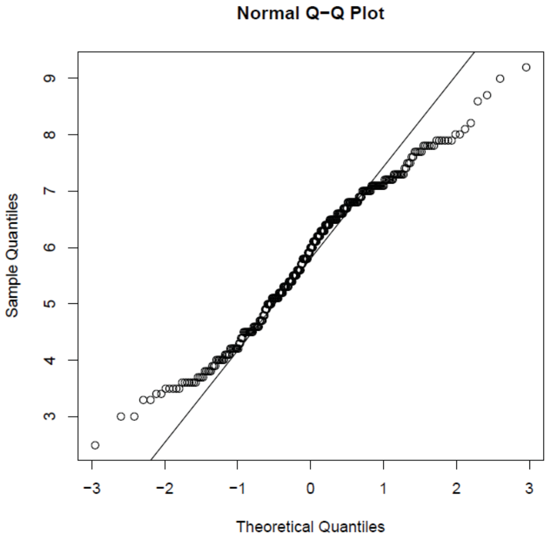


Figure 4. Normal Q-Q Plot

In **Figure 4**, the line in the plot is drawn through the first and third quartiles of the data. The circles that start to deviate from the line, are the data that do not follow a normal distribution, the data that we will need for the POT method.

4.3 *Normality Test*

The Q-Q Plot in **Figure 4** graphically shows whether the processed data, are from a normal distribution. To see if the data follows a normal distribution or does not follow a normal distribution, the Shapiro-Wilk test will be used. The Shapiro-Wilk test will be used to test the following hypotheses:

H_o: The data are normally distributed.

H_a: The data are not normally distributed.

The Shapiro-Wilk test has a test statistic W that is mathematically defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_{(i)} - \bar{x}\right)^2}. \qquad \textbf{Equation 28.}$$

where the $x_{(i)}$ represent the the order sample values, and a_i represent the constants that are generated from the mean, variance and covariance of the order statistics of a sample n, from a normal distribution.

When performing the Shapiro-Wilk test for normality in R, we get the following:

Test	Test Statistic	p-value
Shapiro-Wilk	0.97637	0.00004058

Table 1. Results of the Shapiro-Wilk Test

We use the p-value approach to draw the final conclusion of the hypothesis test. Since the p-value is less than $\alpha = 0.05$, where α is our confidence level of significance, we reject H_o in favor of H_a. The meaning of this conclusion is that there is evidence showing that the data does not follow a normal distribution.

4.4 *The Mean Excess and Mean Residual Plots*

The mean excess and mean residual plots, are ways to graphically analyze extreme values to validate a GPD model for the excess distribution, a foundation for POT modeling. Used for estimating and choosing a threshold to work with, we look for a threshold such that the plot is roughly linear. Choosing a threshold can be challenging but depending on how the threshold is chosen, the parameter estimates become sensitive to this particular threshold of choice.[15]

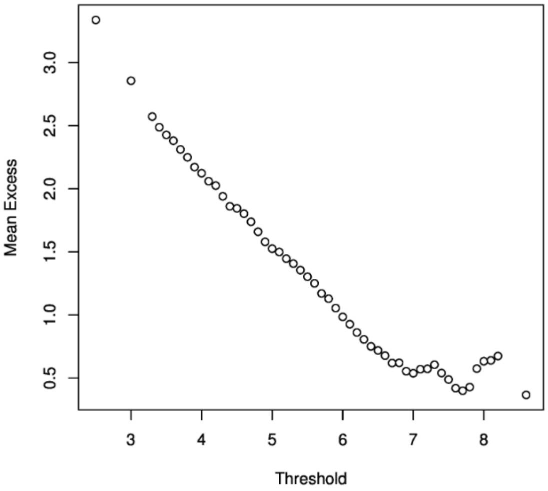


Figure 5. Mean Excess Plot

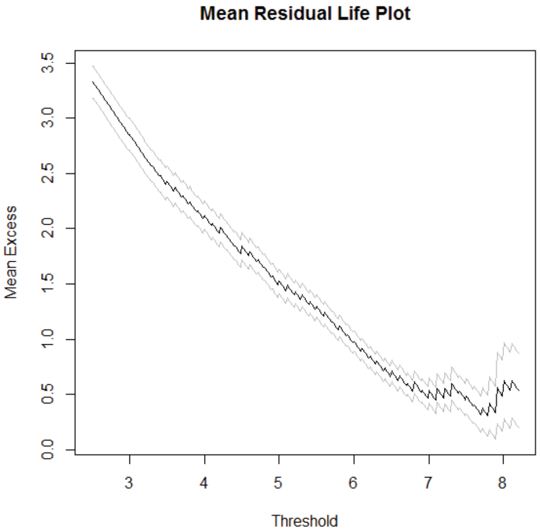


Figure 6. Mean Residual Plot

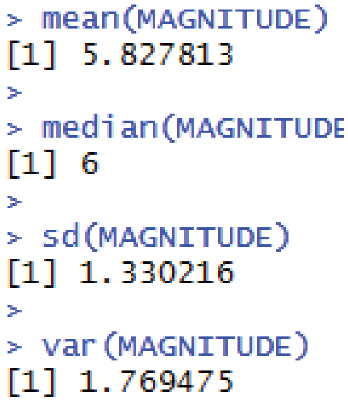
In **Figures 5** and **6**, we notice that the data for the threshold values 1-6 the data follows closely to a linear line (normal distribution) until about threshold 7 where the data starts to follow a nonlinear pattern. This is how the threshold of the data, represented by *u*, was chosen.

5. ESTIMATING GPD PARAMETERS

The MLE, MOM, and PWM estimation techniques can be used, now that a value for the threshold has been chosen. In R, with the data and threshold, the shape and scale parameters can be calculated, in order to determine which distribution within GPD would be a good fit, for the data. We need the parameters to be as unbiased as possible, so we need to see which estimation technique will be most effective when determining the scale and shape parameters. Once the parameters have been found, we will compare each technique to see which one provides the most accurate estimation for the parameters needed.

5.1 *Calculating $\hat{k}$ and $\hat{\alpha}$ from Actual Data*

The calculations needed in R, for the peak over threshold (POT) method was enabled by the POT package downloaded in R. The POT package, is a package in which tools are developed to perform most functions related to EVT, that can be used to statistically analyze POT.[16] Using the statistical package R, refer to **Figure** 7, to see the values calculated for $\bar{x}$ and S^2, where $\bar{x}$ is the mean and S^2 is the variance, were calculated:



```
> mean(MAGNITUDE)
[1] 5.827813
>
> median(MAGNITUDE
[1] 6
>
> sd(MAGNITUDE)
[1] 1.330216
>
> var(MAGNITUDE)
[1] 1.769475
```

Figure 7. Calculations of the Mean, Median, Standard Deviation, and Variance by R.

Using the equations for $\hat{k}$ and $\hat{\alpha}$, along with the following values for $\bar{x}$ and S^2 calculated in R, we find that:

$$\hat{k} = 4.924 \text{ and } \hat{\alpha} = 17.261$$ **Equation 29.**

For this study, three estimation techniques, namely, the maximum likelihood, method of moments, and probability-weighted moments, will be used to find the estimation of the shape and scale parameters of the processed data. After comparing techniques, we will be able to see which technique provides a more accurate, unbiased estimator for the shape and scale parameters needed for the GPD.

5.2 *Comparing Estimation Techniques*

Now that the threshold has been analyzed and chosen for the data set, using the package POT in R, we are able to find our shape and scale parameters for the MLE, MOM and PWM estimation techniques. We are also able to compare how accurate MLE, MOM and PWM are as the threshold increases, in **Table** 2:

		Shape (k)			Scale (α)		
u	**m**	**MLE**	**MOM**	**PWM**	**MLE**	**MOM**	**PWM**
5.5	184	-0.48 (0.04)	-1.23 (0.20)	-1.32 (0.20)	1.82 (0.14)	2.91 (0.35)	3.0 (0.36)
6.0	156	-0.39 (0.05)	-0.86 (0.15)	-0.10 (0.18)	1.32 (0.12)	1.83 (0.23)	2.0 (0.25)
6.5	115	-0.28 (0.06)	-0.46 (0.12)	-0.61 (0.15)	0.91 (0.10)	1.05 (0.14)	1.16 (0.17)
7.0	65	-0.15 (0.11)	-0.15 (0.12)	-0.18 (0.15)	0.62 (0.10)	0.62 (0.11)	0.64 (0.12)
7.5	27	-0.19 (0.18)	-0.21 (0.19)	-0.45 (0.28)	0.58 (0.15)	0.59 (0.16)	0.71 (0.21)

Table 2. Estimated Parameters

Now that m is the number of exceedances over the threshold value u; the standard errors are in parentheses. While analyzing **Table2**, we are able to recognize two findings. One, that our k stays within its restricted bounds $-\frac{1}{2} < k < \frac{1}{2}$, and two, as the value of u increases, the estimators for the scale and shape parameter, for MLE and MOM begin to take on similar values; while the values for the shape and scale parameter for PWM are sporadic when the threshold value is more than seven.

6. FITTING GPD TO THE DATA

Now that we have analyzed our data, used estimation techniques such as MLE, MOM, and PWM to estimate the scale and shape parameters, k and α, we are now able to take the data, and model it to see how accurately it follows the generalized Pareto distribution under the Peak Over Threshold (POT) method.

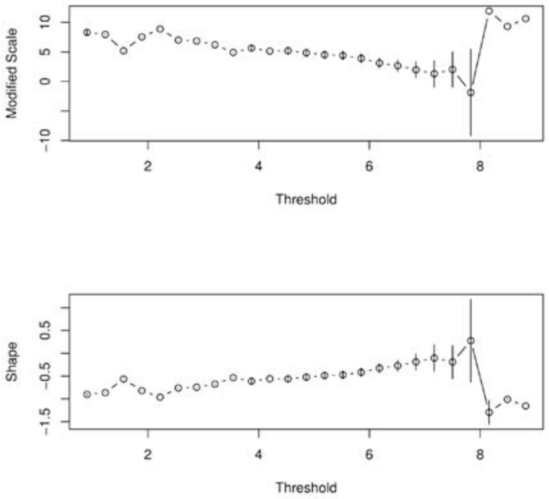
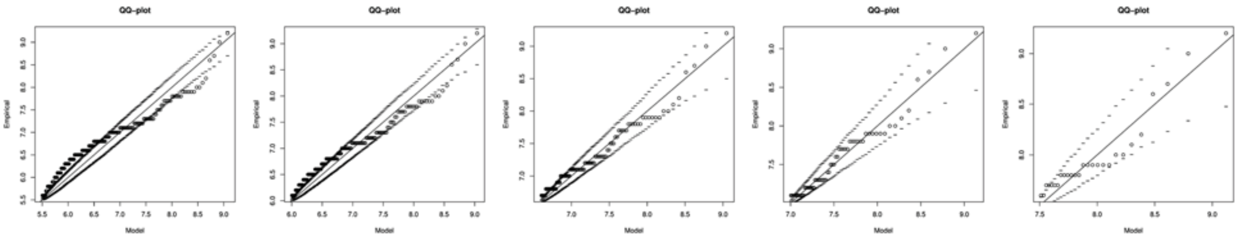


Figure 8. $E[x - u|x \geq u]$ Plot



(a) u=5.5 (b) u=6.0 (c) u=6.5 (d) u=7.0 (e) u=7.5

Figure 9. GPD qq Plots for Threshold Values 5.5-7.5

Figure 8, depicts the stability of the respective parameters, scale and shape, as the threshold value increases. Notice after the threshold value of $u = 7$, the parameter is not as stable which can be concluded by the behavior seen after the threshold value compared to the behavior before and up onto the chosen threshold value. **Figure** 9, using MLE for the shape and scale parameters, we graph the GPD qq plots for the threshold values of u from 5.5 to 7.5. Notice that as the threshold value approaches the chosen threshold value of $u = 7$ the model shows a better fit for the generalized Pareto distribution, for the POT method.

6.1 *Goodness of Fit Test*

As mentioned in the previous section, from the GPD qq plots, as the threshold value approaches the chosen threshold value of $u = 7.0$, the model shows a better fit for the data. To ensure that the data indeed follows the GPD, a goodness of fit test will be used. More specifically, the Anderson-Darling (A-D) test will be performed to test the following hypothesis:

H_o: The data follow the generalized Pareto Distribution H_a: The data does not follow the generalized Pareto Distribution

The Anderson-Darling test also has a test statistic mathematically defined as:

$$A^2 = -N - S \qquad \textbf{Equation 30.}$$

where $S = \sum_{i=1}^{N} \frac{(2i-1)}{N}[\ln F(X_i) + \ln(1 - F(X_{N+1-i}))]$ where F, is the cumulative distribution function, and X_i represents the ordered data. When performing the Anderson-Darling test for the data that exceeds the threshold, $u = 7.0$, using MLE as the chosen estimation technique in R, we get the following:

Test	**Test Statistic** (A)	p-**value**
Anderson-Darling	3.3729	0.09091

Table 3. Results of the Anderson-Darling Test

Again, we use the p-value approach to draw the final conclusion of the hypothesis test. Since the p-value is greater than $\alpha = 0.05$, where α is our confidence level of significance, we fail to reject H_o in favor of H_a. In other words, there is evidence showing that the data values exceeding the threshold $u = 7$ indeed follows the generalized Pareto distribution.

7. SIMULATION ANALYSIS

To assess the performance of the three estimation techniques a simulation study has been conducted. We have generated 10,000 GPD data points using R and we calculated estimates using all three techniques in each trial. We have performed 30 trials in this analysis.

7.1 *Bias and RMSE in Parameter Estimation*

Bias is defined as the following: a point estimator $\hat{\theta}$ is said to be an unbiased estimator if for every possible value of θ, $E\left(\hat{\theta}\right) = \theta$. If $E\left(\hat{\theta}\right) \neq \theta$, then θ is biased, and the difference $E\left(\hat{\theta}\right) - \theta$ is called the bias of $\hat{\theta}$, and is denoted as $B\left(\hat{\theta}\right)$. In practice relative bias is used to measure the efficiency of any estimation methods. Mathematically, the relative bias is defined as:

$$B\left(\hat{\theta}\right) = \frac{E\left[\left(\hat{\theta} - \theta\right)\right]}{\theta}$$ **Equation 31.**

where $\hat{\theta}$ is an estimate of θ (parameter) and $E\left(\hat{\theta}\right) = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i$

The Root Mean Square Error (RMSE) is one of the most important performance indices used in literature and is defined as:

$$RMSE(\theta) = \sqrt{E\left[\left(\hat{\theta} - \theta\right)^2\right]}$$ **Equation 32.**

The bias of each parameters estimated by the three methods is summarized in **Table** 4. Each threshold value is represented by u. In absolute terms the MOM produced the least bias of the three methods for the following threshold values: 5.5, 7.0 and 7.5. MLE provides the least bias when the threshold value is 6.0 and PWM performs the best for a threshold value of 6.5.

The RMSE of each estimation technique for each of the threshold values specified in **Table** 5. As before, each threshold value is represented by u. With an exception of the RMSE for shape parameter for threshold value 5.5, MLE performs the best among all methods for both shape and scale parameters and for all threshold values. The least RMSE values for almost all MLE estimates guarantee that MLE is the best of estimation.

u	Sample Size	Method	Shape	Scale
5.5	30	MLE	0.0069	0.0086
		MOM	-0.0003	-0.0002
		PWM	0.0003	0.0003
6.0	30	MLE	-0.0001	-0.0003
		MOM	0.0010	0.0010
		PWM	0.0005	0.0005
6.5	30	MLE	0.0007	0.0043
		MOM	-0.0021	-0.0049
		PWM	-0.0004	-0.0004
7.0	30	MLE	-0.0004	0.0032
		MOM	-0.00004	0.0018
		PWM	0.0013	0.0056
7.5	30	MLE	0.0010	0.0036
		MOM	-0.0004	-0.0006
		PWM	0.0008	0.0015

Table 4. Bias of Parameter Estimates

u	Sample Size	Method	Shape	Scale
5.5	30	MLE	0.0185	0.0058
		MOM	0.0153	0.0085
		PWM	0.0144	0.0080
6.0	30	MLE	0.0041	0.0016
		MOM	0.0082	0.0059
		PWM	0.0071	0.0048
6.5	30	MLE	0.0037	0.0038
		MOM	0.0059	0.0056
		PWM	0.0048	0.0046
7.0	30	MLE	0.0022	0.0025
		MOM	0.0026	0.0030
		PWM	0.0031	0.0041
7.5	30	MLE	0.0026	0.0029
		MOM	0.0024	0.0029
		PWM	0.0033	0.0049

Table 5. RMSE of Parameters Estimates

7.2 *Mis-specification Bias*

A simulation analysis has also been performed to investigate the miss-specification bias. We have miss-specified the threshold parameter to see the miss-specification bias, if any, of the scale and shape estimates. We have used estimates obtained in **Table** 2 for this miss-specification. For example, we choose MLE shape parameter (-0.48) and scale parameter (1.82) when the actual threshold is 5.5 in **Table** 4. Using these two estimated values we simulated 10,000 GPD values; however, we have chosen a different threshold value for this simulation. We continue in this process for all other threshold values. In short, we simulated value for threshold values 6, 6.5, 7 and 7.5 with estimated values obtained from threshold value 5.5. **Figures** **10** to **20** display all possible mis-specification bias in this study. It is very obvious from these graphs that MLE performs the best in all threshold values. MOM performs better than PWM in all cases.

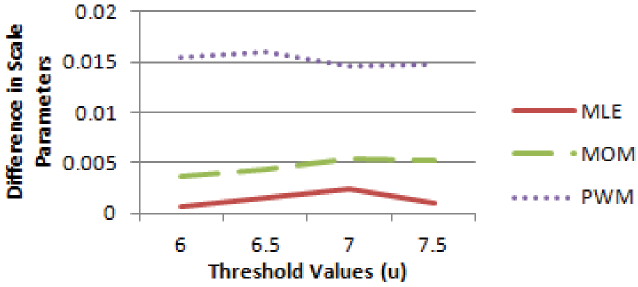


Figure 10. Actual Threshold 5.5

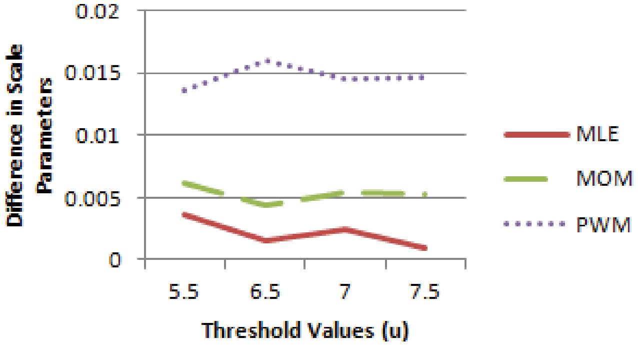


Figure 11. Actual Threshold 6.0

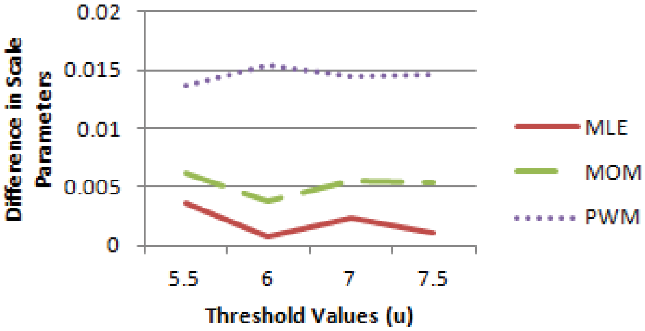


Figure 12. Actual Threshold 6.5

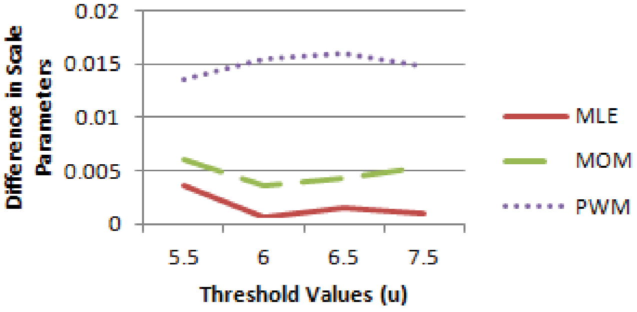


Figure 13. Actual Threshold 7.0

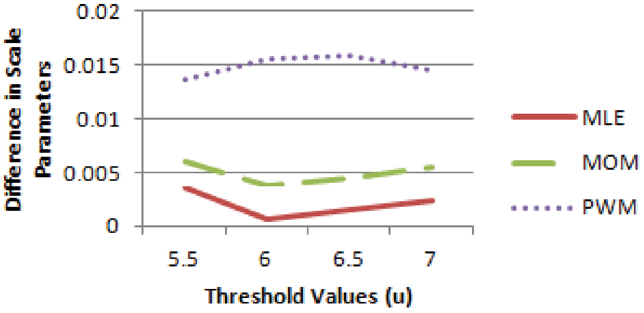


Figure 14. Actual Threshold 7.5

Figure 15. Mis-specification Bias of Scale Parameter

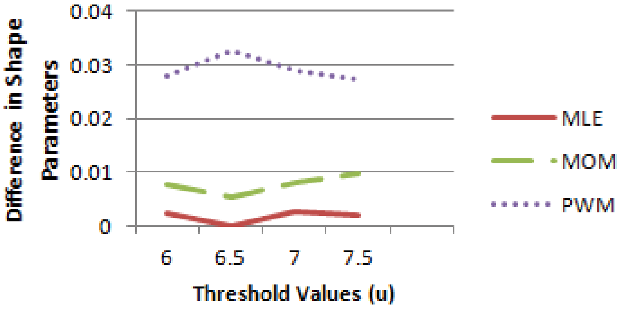


Figure 16. Actual Threshold 5.5

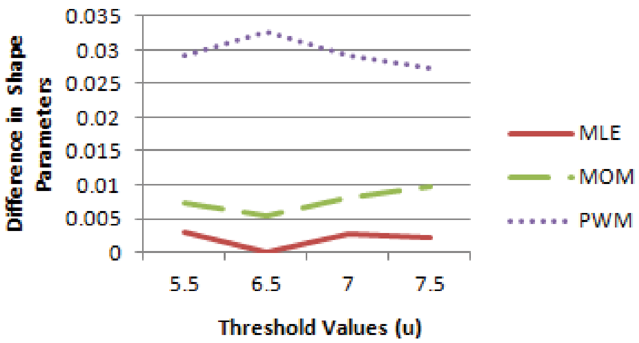


Figure 17. Actual Threshold 6.0

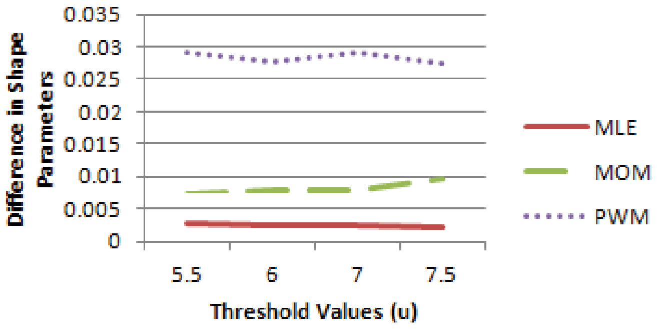


Figure 18. Actual Threshold 6.5

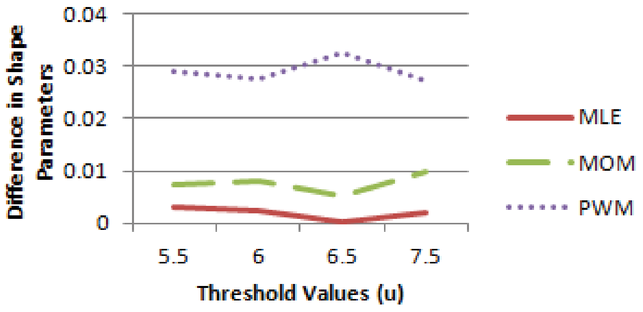


Figure 19. Actual Threshold 7.0

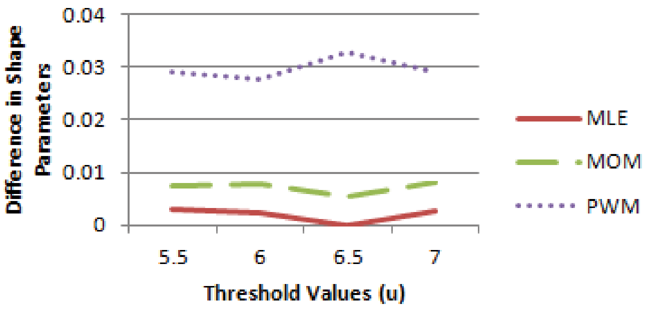


Figure 20. Actual Threshold 7.5

CONCLUSIONS

The following conclusions can be drawn from this study: (1) it is reasonable to fit GPD to earthquake data; (2) MLE, MOM and PWM offer methods for estimating the parameters of GPD; and (3) the simulation study and the mis-specification analysis show that MLE performs the best among all methods.

REFERENCES

1. Klugman, S.A., Panjer, H. H., Willmot. G.E.(2008). *Loss Models: From Data to Decisions*, Wiley.
2. Pickands, J. (1975). Statistical inference using order statistics. *Ananls of Statistics*, 3, 119-31.
3. Hosking, J.R.M, and Wallis, J. (1987). Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, 29(3), 339-349.
4. Castillo, Enrique, and Hadi, Ali,S.(1997). Fitting the Generalized Pareto Distribution to Data. *Journal of the American Statistical Association*, 92(440), 1609-1620.
5. Dey, A and Das, K., (2014), *Quantifying extreme Hurricane risk in the US Gulf Coast*, In JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association.
6. Dey, A, and Das, K., (2016), *Modeling Extreme Hurricane Damage using the Generalized Pareto Distribution*, American Journal of Mathematical and Management Sciences, 35(1), 55-66.
7. Hogg, R.V., Tanis, E.(2009). *Probability and Statistical Inference* (8th ed., pp.273-281). Upper Saddle River, NJ: Pearson.
8. Historic Earthquakes in the United States and Its Territories. *http://earthquake.usgs.gov/earthquakes/states/historical.php/* (accessed March 2014)
9. Castanos, H., and Lomnitz C.(2002). PSHA: is it science? *Engineering Geology*, 66, 315-317.
10. Cornell, C.A.(1968). Engineering seismic risk analysis. *Bulletin of the Seismological Society of America*, 58, 1583-1606.
11. Field, E. H.(2007). Overview of the working group for the development of regional earthquake likelihood models (RELM). *Seismological Research Letters*, 78(1), 7-16.
12. Jordan, T.(2006). Earthquake predictability, brick by brick. *Seismological Research Letters*, 77(1), 3-6.
13. Krintzsky, E. L.(2002). Epistemic and aleatory uncertainty: a new shtick for probabilistic seismic hazard analysis. *Engineering Geology*, 66, 157-159.
14. Pisarenko, V.F., Lyubushin, A. A., Lysenko, N.B., and Golubieva, T.V.(1996). Statistical estimation of seismic hazard parameters: maximum possible magnitude and related parameters. *Bulletin of the Seismological Society of America*, 86, 691-700.
15. Ghosh, S. and Resnick, S.(2010) A discussion on mean excess plots. *Stochastic Processes and their Applications*, 120: 1492-1517.
16. R Core Team (2006). A User's Guide to the POT Package. URL *http://cran.r-project.org/*.

ABOUT THE STUDENT AUTHOR

Audrene S. Edwards is from Groves, Texas. She graduated from Lamar University in 2014 with a B.S. in Mathematics. Audrene is presently a graduate student of Lamar University. She is currently working on receiving her Masters in Mathematics. She aspires to pursue her PhD in statistics to become a professor.

PRESS SUMMARY

The extreme value theory (EVT) is a well known discipline in statistics and in many other physical sciences and engineering disciplines. It has been developed in order to model and analyze rare but extreme events. Using data from selected earthquakes of general historic interest, EVT has been used to construct a model on the extreme and rare earthquakes that have happened in the United States from 1700 to 2011. The primary goal of fitting such a model is to estimate the amount of losses due to those extreme events and the probabilities of such events. By estimating the amount of losses and the probabilities of such events, this information can be useful in precautionary measures created for natural disasters and catastrophe bonds.