North Carolina Journal of
Mathematics and Statistics
Volume 2, Pages 45–60 (Accepted August 4, 2016, published August 19, 2016)
ISSN 2380-7539

Understanding extreme stock trading volume by generalized Pareto distribution

Shaymal C. Halder and Kumer Das

ABSTRACT. The extreme value theory (EVT) is used to assess the risk caused by extreme natural and man made events. These events exhibit clusters of outlying observations that cannot be modeled by a Gaussian distribution. The generalized Pareto distribution (GPD) have proved useful in modeling such events, in particular, it is widely used in modeling the distribution exceeding a high threshold. The GPD has uniform, triangular, exponential, and Pareto distribution as special cases. Estimating parameters of the GPD has become an important task in EVT. There are several methods for estimating parameters of the GPD such as method of moments, method of maximum likelihood, probability weighted moments, maximum Penalized Likelihood, etc. and all estimation techniques have some limitations. Even though EVT is a well-established discipline, no attempt has been made to compare all estimation techniques together. In particular, studying an appropriate method for modeling GPD in the light of stock trading volume data has not been seen. The aim of the study is manifold: first, to discuss and compare several estimation methods and their limitations; second, to investigate whether GPD can be used to model stock trading volume data; third, to compare the volatility of two stock market indexes in the light of EVT; fourth, to test the efficiency of several estimation methods for different threshold values; and finally, to obtain a required design value with a given return period of exceedance and probability of occurring extreme events. Simulated data and real financial data are considered for our study.

1. Introduction

The field of extreme values (maximums or minimums of random variables) has attracted the attention of statisticians, engineers, and economists for many years. There are two widely used approaches available to analyze extreme data, namely, the block-maxima approach and the peaks-over-threshold (POT) approach. POT plays an important role in risk management, finance, insurance, reinsurance, economics, hydrology, material sciences, telecommunications, and other industries where risky extreme events occur with very small probability. For example, POT can be used in modeling the impact of crashes or situations of extreme stress on investor portfolios. McNeil (1998) provides an interesting discussion for the 1987 crash for S&P equity data. Embrechts and Samorodnitsky (1999) review some of the basic tools from POT relevant for risk management. POT method has popularly been used to estimate return levels of significant wave height (Sterl and Caires., 2005), hurricane damage (Daspit and Das, 2012; Dey and Das, 2014, 2016b), annual maximum flood of the River Nidd at Hunsingore, England (Hosking and Willis, 1987), earthquake severity (Edwards and Das, 2016), and aviation accidents (Dey and Das, 2016a). In the POT method, a distribution is fitted to the exceedances of a variable above a high threshold. It has been shown that the GPD arises as the limiting distribution of peaks (or excesses), $X - u$, of a

Received by the editors May 20, 2016.
2010 *Mathematics Subject Classification.* 11A11; 00B12.
Key words and phrases. Extreme events, generalized Pareto distribution, peaks over threshold, estimation techniques, return level, Dow Jones Industrial Average, Dhaka Stock Exchange.

random variable X over a threshold u. Since the accuracy of extreme quantile estimation is sensitive to modeling of the tail distribution, it is important to have an efficient method of estimation of the GPD parameters. Several methods have been proposed for estimating the GPD parameters (Hosking and Willis, 1987; Davison, 1984). However, there is no universally accepted method of identifying the most appropriate methods of estimating GPD parameters. In particular, to the best of our knowledge, no attempt has been carried out in analyzing trading volume data. The central objective of this study is to evaluate the effectiveness of few of the estimation methods suggested in the literature as the threshold values vary. The problem of estimating the probability of extreme high volume in two stock markets has also been addressed.

The rest of the paper is organized as follows: Section 2 describes the basic definitions and properties of the GPD, Section 3 describes threshold selection techniques, Section 4 describes different estimation methods of GPD, Section 5 describes the simulation study using different estimation techniques, Sections 6 and 7 describe the application of the GPD on real data set, Section 8 discusses the return level and return period, and Section 9 provides a conclusion.

2. Generalized Pareto Distribution(GPD)

Let $X_1, X_2, \cdots, X_n$ be a sequence of i.i.d. random variables with marginal distribution function F and $M_n = \max\{X_1, \cdots, X_n\}$. We consider the extreme events of those of X_i that exceed some high threshold, usually denoted by u. So the stochastic behavior of extreme events is depicted by the conditional probability

$$P\left(X > x + u \mid \ X > u\right) \ = \ \frac{1 - F(x+u)}{1 - F(u)}, \text{where } x > 0.$$

If the distribution F is solved, the distribution of threshold exceedances is also be solved. In practice high values of the threshold are preferred.

If any arbitrary term in the X_i is X with distribution function F, so that for large sample size of n,

$$P(M_n \leq x) \approx G(x)^n \tag{2.1}$$

where

$$G(x) = \exp\left(- (1 + k\frac{x-\mu}{\sigma})^{\frac{-1}{k}} \right) \tag{2.2}$$

for some $\mu, \sigma > 0$ and k, then, for large threshold value u, the distribution function of $(X - u)$, conditional on $X > u$, is described by the generalized Pareto family. The model has three parameters: a location parameter, μ; a scale parameter, σ; and a shape parameter, k. The standardized cumulative distribution function (cdf) of the GPD of a random variable X can be written as :

$$G(x) = \begin{cases} 1 - \left(1 + \frac{k(x-\mu)}{\sigma}\right)^{\frac{-1}{k}}, k \neq 0, \sigma > 0 \\ 1 - \exp\left(- \frac{x-\mu}{\sigma} \right), k = 0, \sigma > 0 \end{cases} \tag{2.3}$$

The GPD defined in Equation (2.1) reduces to a 2 parameter GPD for $\mu = 0$ and for most of the practical purposes a 2-parameter GPD seems more appropriate than a 3-parameter GPD. In our study, a 2-parameter GPD is considered.

For different values of the shape parameter the GPD provides few interesting distributions. For $k < 0$ the distribution has a heavy Pareto-type upper tail. For $k = 0$, the GPD provides the exponential distribution with mean σ. And, for $k = 0.5$ and $k = 1$ the distribution is triangular and uniform respectively. When $k \leq \frac{-1}{2}$, Var$(X) = \infty$, and the rth central moment exists if and only if $k > \frac{-1}{r}$. In other words, if the random variable X has a generalized Pareto with GPD(k, σ), then the conditional distribution of $(X - u)$ subject to $X \geq u$ is also generalized Pareto with GPD$(k, \sigma + ku)$ and so the new GPD retains the same shape parameter value of k and this property is known as the threshold stability property.

3. Threshold Selection and Model Validation

3.1. Mean Excess Plot

The mean excess plot (ME plot) is a tool that is used to aid the choice of a threshold. The usual practice is to adopt a threshold which is not, in one hand too small to provide a reasonable approximation to the model, or on the other hand is not too big to provide not enough data points for the model. The ME plot is also used to determine the adequacy of the GPD model of a distribution in practice. A characteristic of a fat tailed GPD with positive shape parameter is a straight line from bottom left to top right of the ME plot and a plot of a downward sloping line from top left to bottom right indicates thin tailed behavior. A straight horizontal line ME plot indicates exponential type behavior. The mean excess function of a random variable X with finite mean is defined as:

$$M(u) = E\Big(X - u \mid X > u\Big). \tag{3.1}$$

The mean excess function of the GPD is defined by Ghosh and Resnick (2010) as

$$M(u) = \frac{\sigma(u)}{1-k} = \frac{\sigma + ku}{1-k} \tag{3.2}$$

where $0 \leq u < \infty$ when $0 \leq k \leq 1$, and $0 \leq u \leq \frac{-\sigma}{k}$ when $k < 0$. If $k > 1$, the mean excess function, $M(u)$ does not exist. The mean excess function is linear function of threshold value u, that is the characterizing property of the GPD.

3.2. Q-Q Plot

Q-Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. The plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. The plot can be used to compare collections of data, or theoretical distributions. The points plotted in a Q-Q plot are always non-decreasing when viewed from left to right and if the two distributions being compared are identical, the Q-Q plot follows the 45° line $(y = x)$. If the Q-Q plot is flatter than the line $y = x$, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis. Conversely, if the Q-Q plot is steeper than the line $y = x$, the distribution plotted on the vertical axis is more dispersed than the distribution plotted on the horizontal axis. The S shaped Q-Q plot indicates that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other. How close the Q-Q plot is to this line is a measure of goodness of fit. Drift away from this line in one direction indicates that the underlying distribution may have a heavier (or lighter tail) than the fitted distribution.

4. Different Estimation Methods of GPD

4.1. Maximum-Likelihood Estimation (MLE)

The likelihood function of independent observations $X_1, X_2, \cdots, X_n$ from a 2-parameter GPD is

$$L(x_i; k, \sigma) = \prod_{i=1}^{n} f(x_i; k, \sigma), \tag{4.1}$$

where $f = dG/dx$. The MLEs are the values of k and σ, which maximize Equation (4.1). Very often, it is easier to maximize the logarithm of the likelihood function. The log-likelihood function for $k \neq 0$ states that the function can be made arbitrarily large by taking $k > 1$ and σ/k close to the maximum order statistic $x_{n:n}$. Maximum likelihood estimates of σ and k can be obtained by minimizing the likelihood function above. However, there are some samples for which no maximum likelihood solution exists. 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, then MLE would be preferred due to its effective efficiency with large samples.

4.2. Method of Moments (MOM)

MOM is a method of estimation of population parameters such as mean, variance, median, etc. (which need not to be moments), by equating sample moments with unobservable population moments and then solving those equations for the quantities to be estimated. Estimates by MOM may be used as the first approximation to the solutions of the likelihood equations, and successive improved approximations may be found by the Newton-Raphson method. Subject to the existence of the GPD moments, the mean and variance of the GPD are respectively:

$$\text{Mean} = \mu + \frac{\sigma}{1+k} \tag{4.2}$$

$$\text{Variance} = \frac{\sigma^2}{(1+k)^2(1+2k)} \tag{4.3}$$

Simplifying the above equations, the MOM estimates of the parameters are calculated as:

$$\hat{k}_{MOM} = 1/2(\frac{\bar{x}^2}{s^2} - 1) \tag{4.4}$$

$$\hat{\sigma}_{MOM} = 1/2(\bar{x})(\frac{\bar{x}^2}{s^2} + 1) \tag{4.5}$$

here $\bar{x}$ and s^2 are the sample mean and sample variance.

Castillo and Hadi (1997) and Hosking and Willis (1987) recommended MOM for $0 < k < 0.4$. Since the parameters are easy to compute, MOM estimates can also be used as the initial estimates in other estimation procedure which require numerical technique (Jockovic, 2012). When $k \leq (-1/2)$, the variance of the GPD does not exist and the rth central moment exists if and only if $k > \frac{-1}{r}$.

4.3. Probability Weighted Moments Estimators (PWMU and PWMB)

The probability-weighted moments (PWM) were introduced by Greenwood and Wallis (1979) and represent an alternative to the ordinary moments. As for the moments estimator, parameters can be expressed as a function of PWMs. The estimator is particularly advantageous for small data-sets because the probability weighted moments have a smaller uncertainty than the ordinary moments. The best performance is reached for $k \approx 0.2$ (Deidda and Puliga, 2009); for positive shape values, performances are very close to MLE ones, while for $k < 0$ (Deidda and Puliga, 2009) PWM performances become a little worse than those of MLE. Hosking and Willis (1987) used two definitions of PWM, unbiased (PWMU) and biased (PWMB), but the difference can be detected only for small samples. We only display the results involving PWMU in this paper. Hosking and Willis (1987) defined the PWM estimates of the GPD parameters as:

$$\hat{k}_{PWM} = \frac{\bar{x}}{(\bar{x} - 2t)} - 2 \tag{4.6}$$

$$\hat{\sigma}_{PWM} = \frac{2\bar{x}t}{(\bar{x} - 2t)} \tag{4.7}$$

where

$$t = \frac{1}{n}\sum_{i=1}^{n}(1 - p_{i:n})x_{i:n}, \tag{4.8}$$

with $p_{i:n} = \frac{i-0.35}{n}$ and $x_{i:n}$ is the ith order statistics of a sample size of n.

4.4. Maximum Penalized Likelihood (MPLE)

Coles and Dixon (1999) introduced a weight function for the maximum likelihood function $L(X, \theta)$ for $k > 0$ (Deidda and Puliga, 2009). MPLE corrects the tendency of MLE to diverge for small samples. They noted that the superior performance of MOM and PWM estimators to the MLEs for small sample sizes is due to the assumption of a restricted parameter space, corresponding to finite population moments (Mackay and Bahaj, 2011).

5. Simulation Study

Simulation has been performed using statistical programming R (RCoreTeam, 2013). The POT package which is an add-on package containing useful tools to perform statistical analysis for peaks over a threshold (POT) using the GPD approximation, has been used to perform the simulation in R. Simulation has been done to find out the efficiency of estimation of scale and shape parameters using different estimation methods for different sample sizes. The estimated parameters are compared by their bias and the root mean square error (RMSE). As Table 5.1 shows simulation is done for estimation of three known sets of scale and shape parameter: (1.20,0.5), (0.5,$-$0.05) and (0.5,0.2). In our simulation we have restricted the shape parameters, k in $-0.5 < k < 0.5$ because this range of values is commonly observed in practical applications (Hosking and Willis, 1987). The simulation has been repeated for $10,000$ times for the following sample sizes 30, 50, 100, and 200.

TABLE 5.1. GPD population cases considered in the simulation

GPD Population	Shape parameter, k	Scale Parameter, σ	Skewness
Case 1	0.5	1.2	0
Case 2	-0.05	0.5	1.73
Case 3	0.2	0.5	4.64

TABLE 5.2. Bias of scale parameter (1.2) and shape parameter (0.5)

Method	$n = 30$		$n = 50$		$n = 100$		$n = 200$	
	Scale	Shape	Scale	Shape	Scale	Shape	Scale	Shape
MOM	0.4477	-0.5381	0.3980	-0.4441	0.3208	-0.3573	0.2716	-0.2954
MLE	0.1051	-0.1739	0.0684	-0.0900	0.0275	-0.0449	0.0129	-0.0225
PWMU	0.0664	-0.2083	0.0474	-0.1437	0.0309	-0.0925	0.0183	-0.0569
MPLE	0.2200	-0.4642	0.1393	-0.3122	0.0755	-0.1814	0.0397	-0.1008

TABLE 5.3. RMSE of scale parameter (1.2) and shape parameter (0.5)

Method	$n = 30$		$n = 50$		$n = 100$		$n = 200$	
	Scale	Shape	Scale	Shape	Scale	Shape	Scale	Shape
MOM	0.5373	0.2690	0.4776	0.2221	0.3849	0.1786	0.3259	0.1477
MLE	0.1261	0.0869	0.0821	0.0451	0.0331	0.0224	0.0156	0.0113
PWMU	0.0797	0.1042	0.0568	0.0719	0.0371	0.0462	0.0219	0.0285
MPLE	0.2640	0.2321	0.1672	0.1561	0.0906	0.0907	0.0476	0.0504

The performances are evaluated by the relative bias and RMSE as defined below:

$$\text{Bias} = \frac{(\theta_{est} - \theta_{true})}{\theta_{true}} \tag{5.1}$$

$$\text{RMSE} = \sqrt{(\theta_{est} - \theta_{true})^2} \tag{5.2}$$

where θ_{est}, θ_{true} are the estimated and the true values of the parameter respectively.

Tables 5.2, 8.3 and 8.5 summarize the bias of parameters estimated by the four estimation methods for different sample sizes for three sets of scale and shape parameter such as (1.20,0.5), (0.5,-0.05) and (0.5,0.2) respectively. Similarly, Tables 5.3, 8.4 and 8.6 summarize RMSE for different estimation methods against different sample sizes for three sets of parameters discussed above in Table 5.1. Our simulation study confirms that the MLE of the GPD parameters are efficient as the sample size increases (Bias and RMSE for scale parameter estimated by MLE are larger than those of PWMU only for sample sizes $(30$ and $50)$) when there is no skewness in the data set (Case 1 of Table 5.1. The simulation study also shows that the MLEs are asymptotically efficient (as the sample size tends to infinity the MLEs achieve the Cramer-Rao lower bound for the variance of an unbiased estimator). However, in case of skewed data PWMs perform better than MLEs.

6. Discussion of Data Set

Understanding trading volume is critical because extreme volume can be a predictive measure of future price changes. Though there are still questions whether stock volumes have a finite

TABLE 6.1. Five Number Summary of DJIA (Unit : One hundred millions)

Minimum	First Quartile Q_1	Median Q_2	Third Quartile Q_3	Maximum
0.0841	1.8858	2.3032	2.7939	7.3844

TABLE 6.2. Five Number Summary of DSE (Unit : One hundred millions)

Minimum	First Quartile Q_1	Median Q_2	Third Quartile Q_3	Maximum
0.0097	0.0537	0.2011	0.4217	2.4286

variance, there is little doubt that these data are not Gaussian. Large events happen at a rate incompatible with Gaussian behavior. Mulvey (2001) discussed a number of key issues involving risk management. It is widely accepted that there's a need to correctly address issues involving market crashes, understanding stock volume should be one of them. Regulators have introduced circuit breakers to curb panic-selling and unusual volume of trading. The historical stock volumes of two stock markets (one from North America and the other from Asia) have been used to complement our simulation and theoretical argument. The Dow Jones Industrial Average (DJIA) is a stock market index that shows how 30 large publicly owned companies based in the United States have traded during a standard trading session in the New York Stock Exchange and NASDAQ. DJIA volume data are collected from Yahoo Finance for the period of December 31, 2004 through December 30, 2011. In other words, the data set comprises of $1,764$ DJIA volumes at consecutive trading days. Another set of data set has been collected from the Dhaka Stock Exchange (DSE) (`http://www.dsebd.org/recent_market_information.php`) with the same number of years for the period of December 30, 2004 through December 29, 2011. The date set comprises of $1,658$ stock share volume at consecutive trading days. DSE is one of the two stock exchanges in Bangladesh. One of the purposes of this study is to investigate whether GPD can be used to model any stock volume data.

To determine if any of the data-set is extreme, we check for the presence of outliers by using the fourth spread, f_s. The fourth spread f_s is a measure of spread that is resistant to outliers (Devore, 2010). In order to determine f_s, we first sort the n observations from smallest to largest and separate the smallest half from the largest half using the median. The median of the smallest half, the 1st quartile Q_1 is the lower fourth. The median of the largest half, the 3rd quartile Q_3 is the upper fourth. The 4th spread f_s, is given by (Devore, 2010)

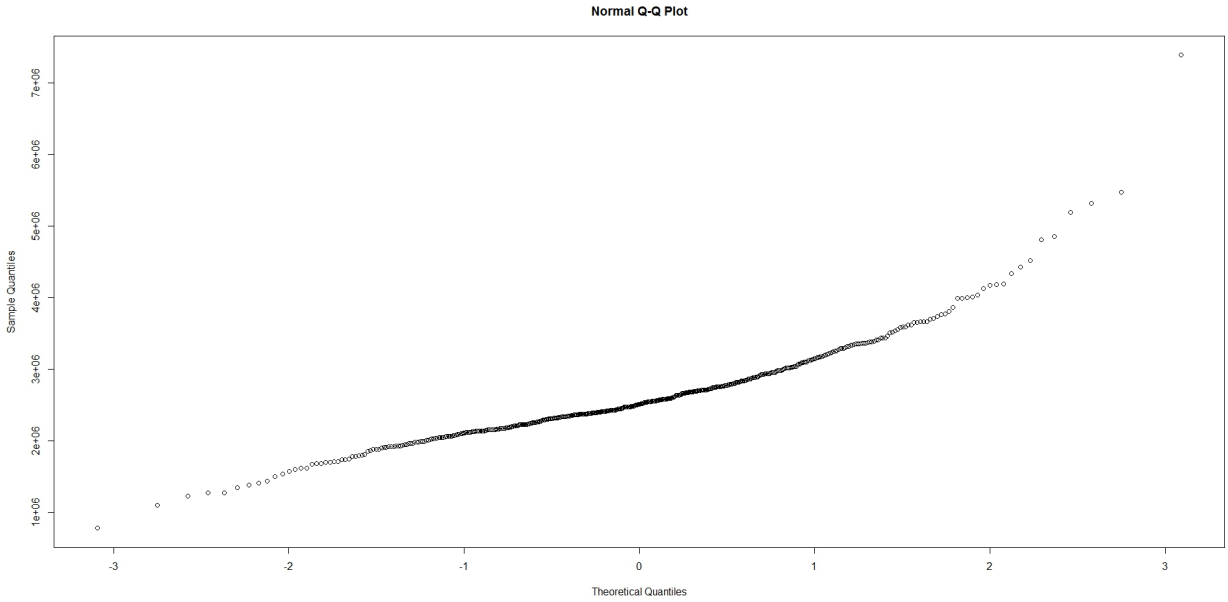
$$f_s = \text{upper fourth- lower fourth} \tag{6.1}$$

Any observation further than $1.5f_s$ from the closest fourth is an outlier. An outlier is extreme if it is more than $3f_s$ from the nearest fourth, and it is mild otherwise (Devore, 2010). For DJIA, we observe outliers in the lower end as well as in the upper end and the outliers are mild. For DSE, we do not observe outliers in the lower end, however, we observe extreme outliers in the upper end.

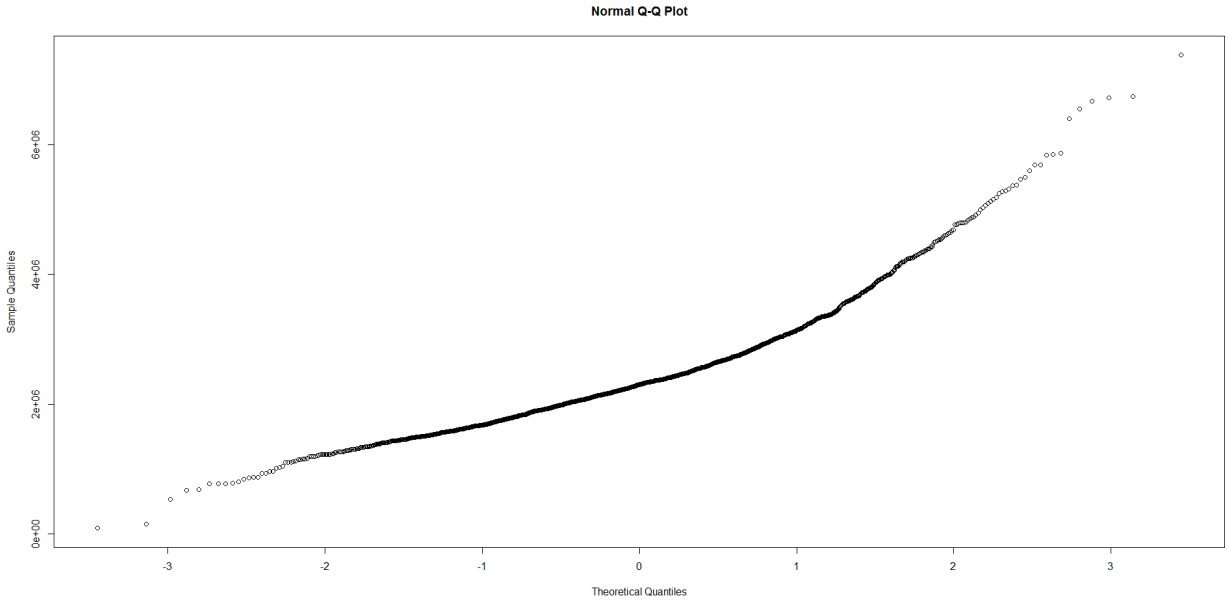
7. Fitting GPD in Data

We have used ME plot and Q-Q plot to investigate the distribution of the data sets. Figure 7.1 is drawn to test for normal Q-Q distribution and Figures 7.2, 7.3 are drawn to test for GPD distribution. All figures are drawn in the scale of one hundred millions.

Comparing figure $7.1(a)$ and $7.1(b)$ we conclude the followings about the DJIA data set:



(a) with 2 years' data set



(b) with 7 years' data set

FIGURE 7.1. Normal Q-Q plot of DJIA

1. Normal Q-Q plot for 2 years' data set replicates the normal Q-Q plot of 7 years' data set.
2. Both Q-Q plots deviate form the normal shape. Moreover, the kurtosis (which is defined as a measure of the peakedness of the probability distribution of a real-valued random variable) of the data set is 6.25. In practice, a kurtosis value greater than 5 confirms the deviation from normality of a data set.

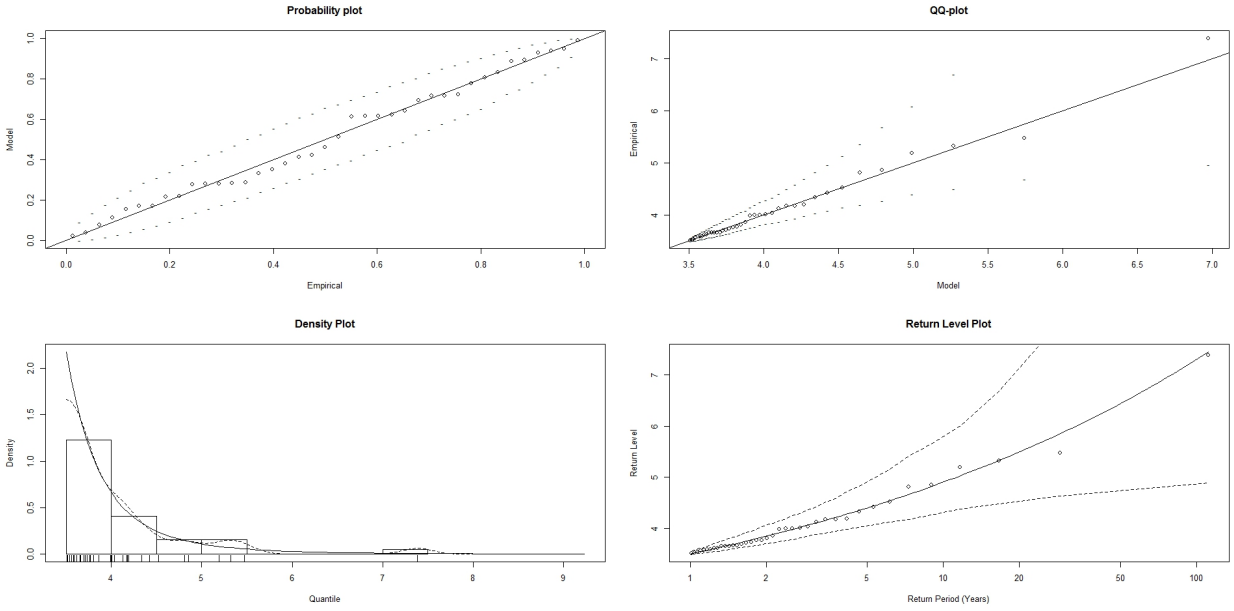


FIGURE 7.2. GPD Q-Q plot of DJIA for 2 years' data set

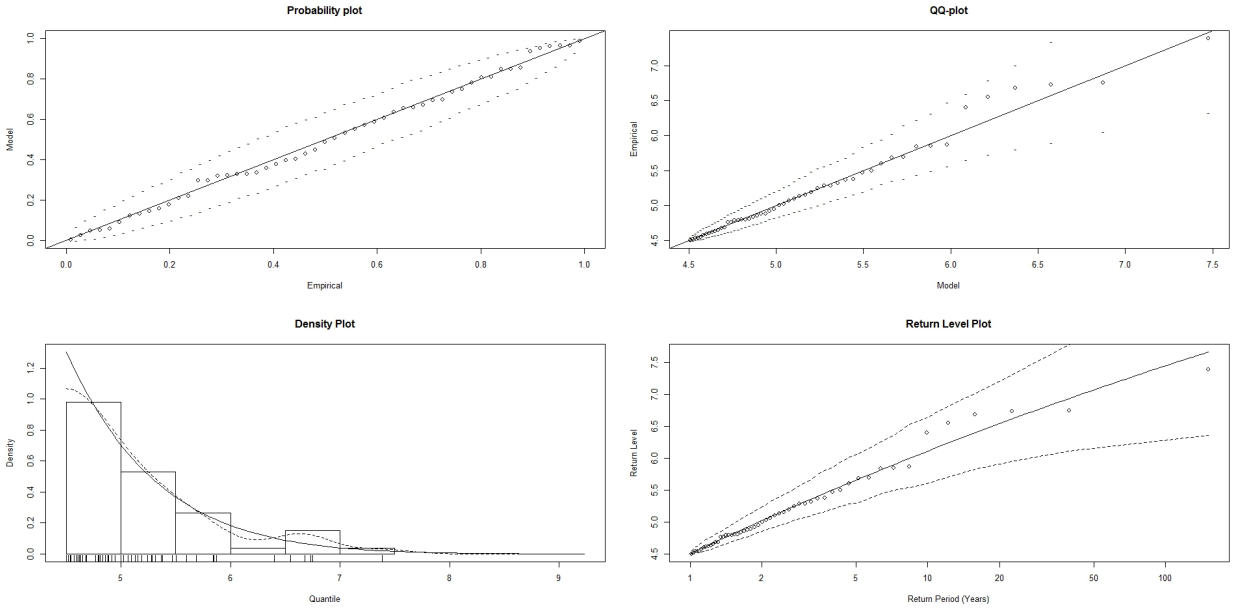


FIGURE 7.3. GPD Q-Q plot of DJIA for 7 years' data set

As seen from Figure 7.2 and 7.3, the graphs show the characteristics of the GPD as discussed in Section 3.2. The mean excess plot of the GPD is drawn in Figure $7.4(a)$ and the graph confirms the linearity as discussed in Section 3.1 from threshold value from $u = 4.0$ to $u = 6.0$. Figures $7.4(b)$, $7.4(c)$, $7.4(d)$, and $7.4(e)$ showing the mean excess plot for threshold value $u = 4.0$, $u = 4.2$, $u = 4.5$, and $u = 5.0$ respectively, show the accuracy of our assumption that GPD can be used to model trading volume data. A q-q plot of the second half of the DJIA data against the first half (Figure 7.6) shows not very significant departure from the straight line which indicates that the

TABLE 7.1. Estimated Value (EV) and Standard Error (SE) of scale parameter of DJIA with different threshold values (in hundred millions)

Method	EV,$u = 4.0$	SE	EV,$u = 4.2$	SE	EV,$u = 4.5$	SE	EV,$u = 5.0$	SE
MOM	0.842	0.117	0.710	0.111	0.743	0.142	0.731	0.195
MLE	0.850	0.120	0.720	0.123	0.768	0.161	0.801	0.252
PWMU	0.857	0.132	0.666	0.113	0.702	0.147	0.658	0.194
MPLE	0.850	0.120	0.720	0.123	0.768	0.161	0.801	0.252

TABLE 7.2. Estimated Value (EV) and Standard Error (SE) of shape parameter of DJIA with different threshold values (in hundred millions)

Method	EV,$u = 4.0$	SE	EV,$u = 4.2$	SE	EV,$u = 4.5$	SE	EV,$u = 5.0$	SE
MOM	-0.100	0.097	-0.012	0.109	-0.047	0.133	-0.065	0.185
MLE	-0.111	0.099	-0.026	0.130	-0.082	0.159	-0.161	0.251
PWMU	-0.121	0.121	0.051	0.128	0.010	0.158	0.041	0.221
MPLE	-0.111	0.099	-0.026	0.130	-0.082	0.159	-0.161	0.251

TABLE 7.3. Estimated Value (EV) and Standard Error (SE) of scale parameter of DSE data with different threshold values (in hundred millions)

Method	EV,$u = 1.0$	SE	EV,$u = 1.2$	SE	EV,$u = 1.5$	SE	EV,$u = 2.0$	SE
MOM	0.300	0.045	0.370	0.076	0.357	0.113	0.175	0.109
MLE	0.299	0.051	0.396	0.089	0.438	0.146	0.259	0.280
PWMU	0.276	0.045	0.350	0.080	0.307	0.108	0.129	0.091
MPLE	0.305	0.051	0.396	0.089	0.438	0.146	0.258	0.277

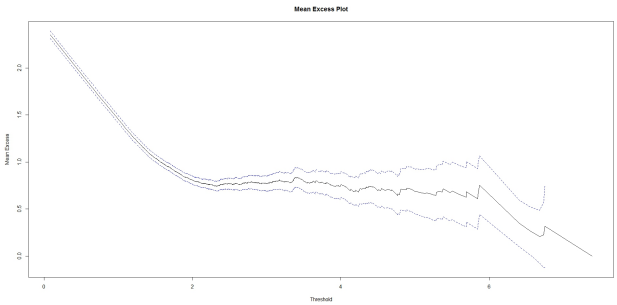
TABLE 7.4. Estimated Value (EV) and Standard Error (SE) of shape parameter of DSE data with different threshold values(in hundred millions)

Method	EV,$u = 1.0$	SE	EV,$u = 1.2$	SE	EV,$u = 1.5$	SE	EV,$u = 2.0$	SE
MOM	0.020	0.107	-0.106	0.144	-0.148	0.224	-0.045	0.434
MLE	0.024	0.134	-0.179	0.172	-0.375	0.257	-0.465	1.006
PWMU	0.101	0.121	-0.047	0.174	0.014	0.264	0.229	0.555
MPLE	0.006	0.127	-0.179	0.172	-0.375	0.257	-0.462	0.999

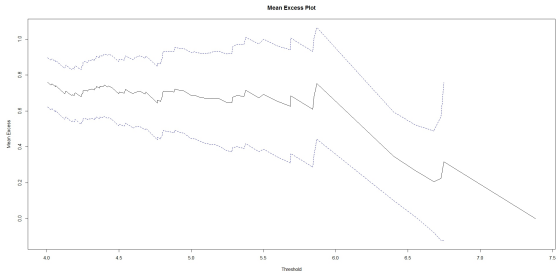
skewness remains stable throughout time. A similar trend has been observed in DSE data as well in Figure 7.7 .

Tables 7.1 and 7.2 display estimated value and standard error of scale and shape parameters of DJIA data for different estimation methods for various threshold values. It can be observed that MOM has the lowest standard error for all threshold levels with a few exception in estimating scale parameter (Table 7.3) where PWMU has the lowest standard error. This is a stark contrast to what we have observed in our simulation study where MLE has outperformed all other methods for larger sample sizes.

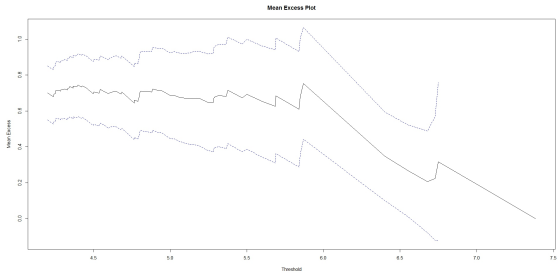
Table 7.3 and 7.4 show the estimated value and standard error of scale and shape parameters of the DSE data for different estimation methods using threshold points 1.0, 1.2, 1.5 and 2.0 respectively.Our analysis involving actual financial data shows from 7.1,7.2,7.3, and 7.4 that the standard



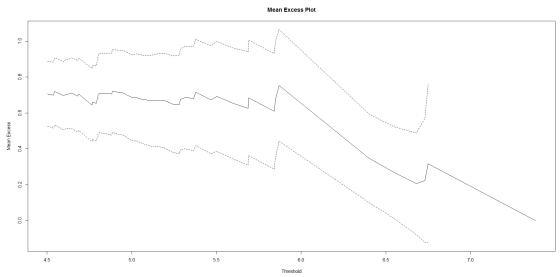
(a) ME plot with the entire data set



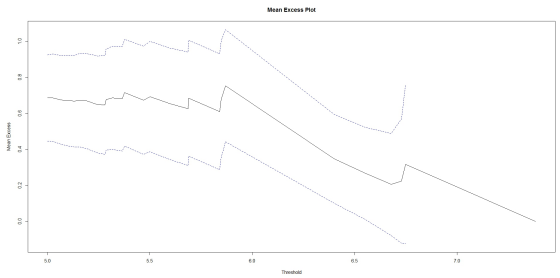
(b) ME plot with the threshold point, 4.0



(c) ME plot with the threshold point, 4.2



(d) ME plot with the threshold point, 4.5



(e) ME plot with the threshold point, 5.0

FIGURE 7.4. Mean excess plot of DJIA

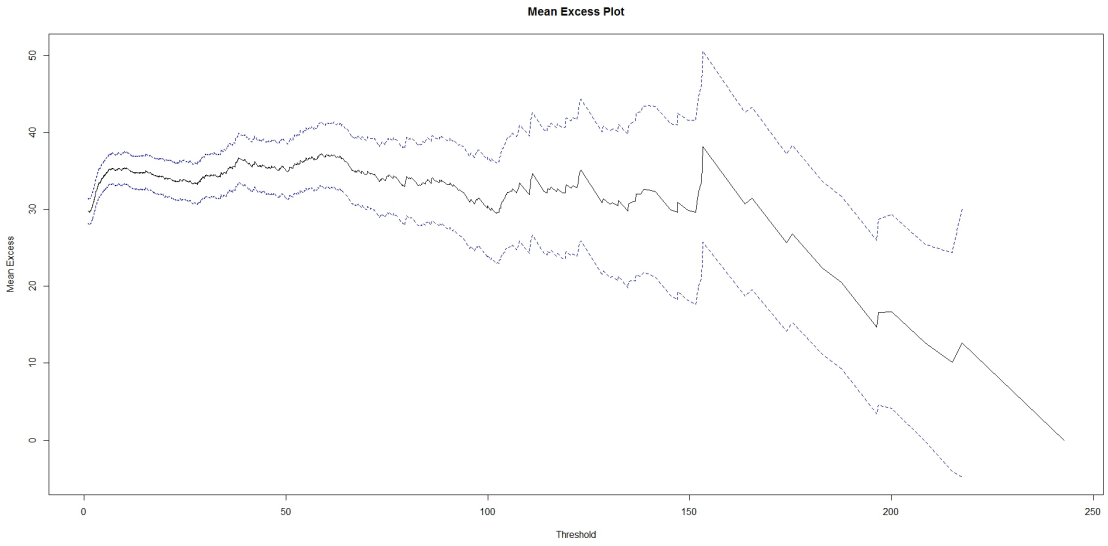


FIGURE 7.5. ME plot of 7 years' DSE data set

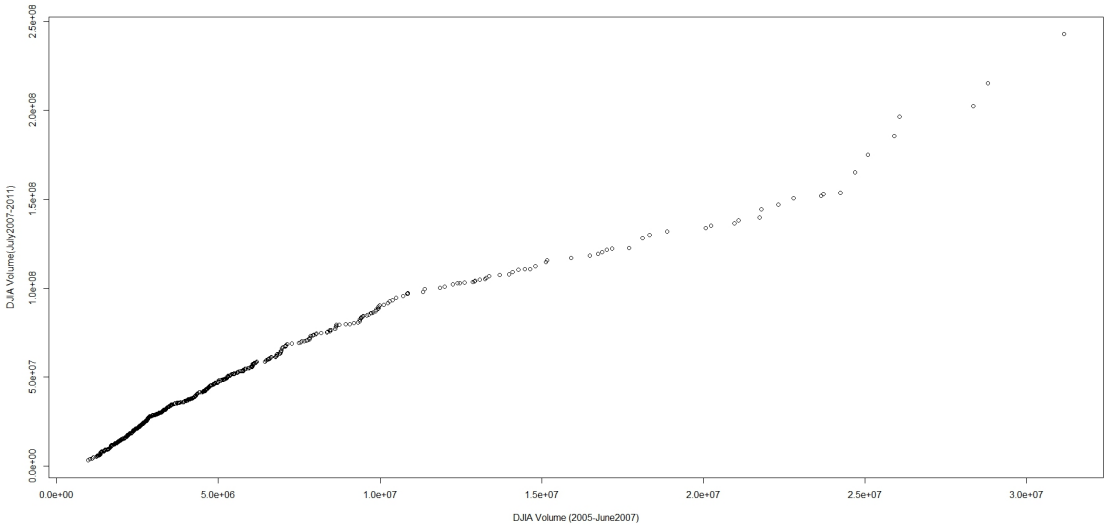


FIGURE 7.6. A Q-Q plot of the DJIA volume of 2nd half against the 1st half

errors for MOM are always smaller than MLE, in particular, for larger threshold these differences are noteworthy.

8. Estimation of Extreme Return Levels

The concepts of return period and return level are commonly used to convey information about the likelihood of extreme events such as floods, earthquakes, hurricanes etc. It is usually more convenient to interpret extreme value models in terms of return levels on an annual scale, rather than individual parameter values. The $m-$year return level is the level expected to be exceeded

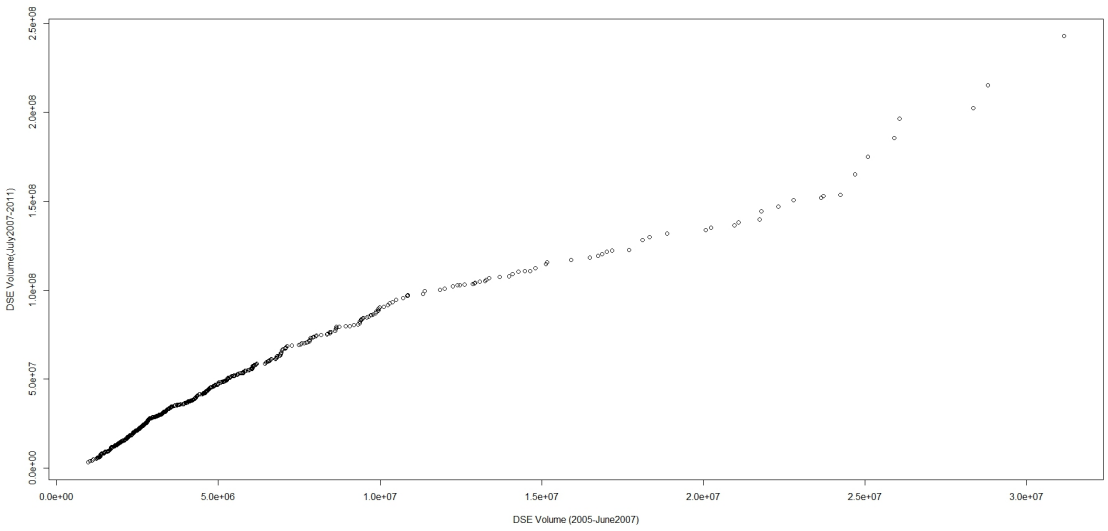


FIGURE 7.7. A Q-Q plot of the DSE volume of 2nd half against the 1st half

TABLE 8.1. Return Level and Return Period for DJIA data

Threshold (millions)	# in tail	$P(X > u)$	$\hat{\sigma}$	$\hat{k}$
400	97	0.0550	0.850	-0.111
420	80	0.0454	0.666	0.051
450	52	0.0295	0.702	0.010
500	26	0.0147	0.658	0.041

once every m years. If there are n_y observations per year, then the $m-$year return level, provided that m is sufficiently large to ensure that $x_m > u$, is defined by

$$x_m = \begin{cases} u + \frac{\sigma}{k}[(mn_y\zeta_u)^k - 1] & \text{for } k \neq 0, \\ u + \sigma \log(mn_y\zeta_u) & \text{for } k = 0, \end{cases} \tag{8.1}$$

where $\zeta_u = Pr(X > u)$, u is the threshold value, σ and k are GPD scale and shape parameter respectively. Estimation of return levels requires the substitution of parameter values by their estimates. For σ and k this corresponds to substitution by the corresponding estimates with lowest Bias and RMSE. With an exception of threshold value 1.45, MLE provides the estimates with lowest standard error and that is why we use MLE estimates in our calculations. An estimate of ζ_u, the probability of an individual observation exceeding the threshold u, is also needed. This has a natural estimator of $\hat{\zeta}_u = r/n$, the sample proportion of points exceeding u. Since the number of exceedances of u follows the binomial $Bin(n, \zeta_u)$ distribution, $\hat{\zeta}_u$ is also the maximum likelihood estimate of ζ_u (Coles and Simiu, 2003).

Tables 8.1 and 8.2 display the results involving return probability and return levels. For example, the 100 year return period states that the DJIA volume will exceed 404.2271 million in every 100 years given that the threshold value is 400 million. Table 8.1 also quantifies the probability of unusual trading volumes. For example, there is a 1.47% that the volume will ever cross 500 million.

TABLE 8.2. Return Level estimates for different Return Periods for DJIA data for threshold value of 400 millions

Return Period	Return Level (millions)
10	403.2281
15	403.4230
20	403.5561
30	403.7366
50	403.9528
100	404.2271

TABLE 8.3. Bias of scale parameter (0.5) and shape parameter (-0.05)

Method	$n = 30$		$n = 50$		$n = 100$		$n = 200$	
	Scale	Shape	Scale	Shape	Scale	Shape	Scale	Shape
MOM	0.1114	0.9109	0.0349	0.7660	0.0204	0.4054	0.0105	0.2003
MLE	0.1464	1.4131	0.0611	1.2265	0.0321	0.6275	0.0158	0.3065
PWMU	0.0619	-0.1864	0.0097	0.2322	0.0066	0.1144	0.0037	0.0578
MPLE	0.1649	1.8695	0.0671	1.3778	0.0346	0.6873	0.0168	0.3279

TABLE 8.4. RMSE of scale parameter (0.5) and shape parameter (-0.05)

Method	$n = 30$		$n = 50$		$n = 100$		$n = 200$	
	Scale	Shape	Scale	Shape	Scale	Shape	Scale	Shape
MOM	0.0557	0.0455	0.0175	0.0383	0.0102	0.0203	0.0052	0.0100
MLE	0.0732	0.0707	0.0305	0.0613	0.0161	0.0314	0.0079	0.0153
PWMU	0.0309	0.0093	0.0048	0.0116	0.0033	0.0057	0.0018	0.0029
MPLE	0.0825	0.0935	0.0335	0.0689	0.0173	0.0343	0.0083	0.0164

TABLE 8.5. Bias of scale parameter (0.5) and shape parameter (0.2)

Method	$n = 30$		$n = 50$		$n = 100$		$n = 200$	
	Scale	Shape	Scale	Shape	Scale	Shape	Scale	Shape
MOM	0.1498	-0.6649	0.0997	-0.4431	0.0614	-0.2662	0.0377	-0.1633
MLE	0.1195	-0.5143	0.0598	-0.2672	0.0266	-0.1182	0.0128	-0.0576
PWMU	0.0421	-0.2507	0.0178	-0.1258	0.0087	-0.0596	0.0041	-0.0294
MPLE	0.1653	-0.7897	0.0913	-0.4592	0.0444	-0.2255	0.0221	-0.1137

TABLE 8.6. RMSE of scale parameter (0.5) and shape parameter (0.2)

Method	$n = 30$		$n = 50$		$n = 100$		$n = 200$	
	Scale	Shape	Scale	Shape	Scale	Shape	Scale	Shape
MOM	0.0749	0.1329	0.0499	0.0886	0.0307	0.0532	0.0189	0.0327
MLE	0.0597	0.1029	0.0299	0.0534	0.0133	0.0236	0.0064	0.0115
PWMU	0.0210	0.0501	0.0089	0.0252	0.0044	0.0119	0.0020	0.0059
MPLE	0.0826	0.1579	0.0457	0.0918	0.0222	0.0451	0.0111	0.0227

9. Discussion and Concluding Remarks

The present paper is concerned with the tail estimation for stock volume series. Our main findings can be summarized as follows.

- Outliers are present in both data sets. Graphical interpretation confirms that stock volume data can be modeled by GPD.
- The PWM method which is a variation of the MOM, provides most efficient GPD estimates in our simulation study where the model is positively skewed. As expected MLE is preferable when the data is not skewed and when the sample size is large. Even though the MOM performs well in our stock volume data, PWMU provides better estimates on at least few occasions. It has the second least standard error in majority of the cases. Although no method is uniformly best, the simulation results and the results from the stock volume data show that estimation method based on moments performs well compared to method based on maximum likelihood. Even though the characterizations of fat-tailedness (or heavy-tailedness) are somewhat arbitrary, it is our understanding that the widely used approach based on the moments of a distribution should be helpful to understand those extreme behaviors.
- We provide an explanation and demonstration of estimating probabilities and return periods which are important to understand the occurrence of extreme stock volumes which may lead to a market crash.

References

Castillo, E. and Hadi, A. S. (1997). Fitting the generalized pareto distribution to data. *Journal of the American Statistical Association*, 92(440):1609–1620.

Coles, S. and Dixon, M. (1999). Likelihood-based inference for extreme value models. *Extremes*, 2(1):5–23.

Coles, S. and Simiu, E. (2003). Estimating uncertainty in the extreme value analysis of data generated by a hurricane simulation model. *Journal of Engineering Mechanics*, 1288:0733–9399.

Daspit, A. and Das, K. (2012). The generalized pareto distribution and threshold analysis of normalized hurricane damage in the united states gulf coast. *Joint Statistical Meetings (JSM) Proceedings, Statistical Computing Section, Alexandria, VA: American Statistical Association*.

Davison, A. C. (1984). Modeling excesses over high threshholds, with an application. *Statistical Extremes and Applications. The Netherlands: ed.J. Tiago de Oliverita, Reidel, Dordrecht*, pages 461– 482.

Deidda, R. and Puliga, M. (2009). Performances of some parameter estimators of the generalized pareto distribution over rounded-off samples. *Physics and Chemistry of the Earth*, 34:626–634.

Devore, J. L. (2010). *Probability and Statistics for Engineering and the Sciences., Eighth ed., Monterey, Calif.:Brooks/Cole Pub*, pages 40–41.

Dey, A. and Das, K. (2016a). Quantifying the risk of extreme aviation accidents. *Physica A: Statistical Mechanics and Applications*, 463:345–355.

Dey, A. K. and Das, K. (2014). Modeling extreme hurricane damage in the united states. *Joint Statistical Meetings (JSM) Proceedings, Section on Risk Analysis, Alexandria, VA: American Statistical Association*, pages 4356–4365.

Dey, A. K. and Das, K. (2016b). Modeling extreme hurricane damage using the generalized pareto distribution. *American Journal of Mathematical and Management Sciences*, 35(1):55–66.

Edwards, A. and Das, K. (2016). Using the statistical approach to model natural disasters. *American Journal of Undergraduate Research*, 13(2):87–104.

Embrechts, P., R. S. and Samorodnitsky, G. (1999). Extreme value theory as a risk management tool. *North American Actuarial Journal*, 3(2):30–41.

Ghosh, S. and Resnick, S. (2010). A discussion on mean excess plots. *Stochastic Processes and their Applications*, 120:1494.

Greenwood, J. A., L. J. M. M. N. C. and Wallis, J. R. (1979). Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water Resour. Res.*, 15(5):1049–1054.

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

Jockovic (2012). Quantile estimatation for the generalized pareto distribution with application to finance. *Yugoslav Journal of Operations Research*, 22(2):297–311.

Mackay, E. B.L., C. P. G. and Bahaj, A. S. (2011). A comparison of estimators for the generalised pareto distribution. *Ocean Engineering*, 38:1338–1346.

McNeil (1998). On extremes and crashes. *RISK*, 11:99.

Mulvey, John, M. (2001). Risk management systems for long-term investors: Addressing/managing extreme events. *Discussion paper, Bendheim Center for Finance, Princeton University*.

RCoreTeam (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.*

Sterl, A. and Caires., S. (2005). Climatology, variability and extrema of ocean waves: the web-based knmi/era-40 wave atlas. *International Journal of Climatology*, 25(7):963–977,doi:10.1002/joc.1175.

(S. C. Halder) DEPARTMENT OF MATHEMATICS AND STATISTICS, AUBURN UNIVERSITY, AUBURN, AL 36849, USA

E-mail address: sch0038@auburn.edu

(K. Das) DEPARTMENT OF MATHEMATICS, LAMAR UNIVERSITY, BEAUMONT, TX 77710, USA

E-mail address, Corresponding author: kumer.das@lamar.edu

URL: http://artssciences.lamar.edu/mathematics/faculty/kumer-das.html