JBAS Journal of Business and Administrative Studies
Home page: JBAS (aau.edu.et)

Extreme Value Theory For Vehicle Insurance Data

Edossa Merga Terefe [1, 2]

ABSTRACT

KEY WORDS
claim size; expected shortfall; GEV; Value-at-Risk; POT.

We study the Ethiopian vehicle insurance dataset using the models block maxima and peaks-over-threshold based on extreme value theory for estimating the risk measures, Value-at-Risk and Expected Shortfall. The extreme observations are fitted to the generalized extreme value distribution and the generalized Pareto distribution using maximum likelihood estimation. When estimating the model parameters and risk measures, the difference in estimates between the models is observed.

1. Introduction

The aim of this paper is to illustrate the tail distribution estimation of a claim size for Ethiopian vehicle insurance dataset. The illustration focuses on the graphical visualization of the claim size, and providing point and confidence intervals of estimates. We considered the right tail of the claim size distribution, and it is considered as a negative returns or losses. Then we used the estimation results to quantify the risk measures. The two risk measures considered are Value at Risk (VaR) and Expected Shortfall (ES). These two measures are used to predict how much can the claim size can rise. VaR is equal to the smallest claim size such that the probability of obtaining a greater claim size, is less than or equal to some predetermined probability α. Further, ES can be summarized as the average of the claim size that are greater than VaR. Hence, when calculating VaR, a lower limit of “the worst claim size” is obtained, while when calculating ES the average of these ”worst claim sizes” is produced [see McNeil et al. (2005) and Hull (2018)].

A number of models exist for computing VaR and ES. Here, we focus on two different models based on extreme value theory. Extreme value theory is used to analyze events that happen rarely, i.e., extreme events. In our setting, rare events consist of large claim sizes in the vehicle insurance dataset. The two models based on extreme value theory are called block maxima and peaks-over- threshold (POT). Both models have the same objective; fit a distribution to the sample of extrembservations. However, the models assume that the data follow different distributions. Also, which observations from the original sample that should be considered as extreme, differs in the two models [see Coles (2001) and Dowd (2005)].

We first consider the block maxima method using GEV distribution, which allows the determi- nation of the VaR and ES. Second, we model the exceedances over a given threshold using GPD, which enables us to estimate high quantiles of the claim paid distribution and the corresponding ES. GEV and GPD are two different distributions, but they have the same purpose: model the distribution of the extreme claim size. In particular, we can note that shape parameter, denoted ξ, is contained in both distributions, and it should therefore take similar values (and same sign) in the two distributions (Coles, 2001). In this paper, a positive ξ is obtained in both models, which is the
case in Gilli and Kellezi (2006). Dowd (2005) and McNeil et al. (2005) express that the case $\xi < 0$ is often not of great interest since most of insurance data are more heavily tailed.

[1] Research Center for Statistics, University of Geneva, Switzerland
[2]Statistics Department, Hawassa University, Ethiopia
edossa.terefe@unige.ch / edossamerga@gmail.com

The remaining of this paper is summarized as follows. We start with a Section 2 of the brief review of extreme value theory and risk measures. Before the main results are given in Section 4, the data used and exploratory data analysis are presented in Section 3. We close the paper with a discussion and conclusion of the results in Section 5.

2. Background

2.1. Extreme Value Theory

Assessing the probability of extreme events in summarizing its distribution with a risk measure is an important issue mainly in managing a risk of financial portfolios, since the viability of the insurance industry depends on probabilistic calculations of risk. Extreme Value Theory (EVT) has recently become one of the main theories in developing statistical models for extreme insurance losses and can be useful in defining supplementary risk measures, because it provides more appropriate distributions to fit extreme events. The heavy-tailed nature of insurance claims requires that special attention be put into the analysis of the tail of a loss distribution. Since a few large claims can significantly impact an insurance portfolio, statistical methods that deal with extreme losses have become necessary for actuaries. For example, in insurance a typical problem might be pricing or building reserves for products which offer protection against catastrophic losses, such as excess of loss reinsurance in order to price certain reinsurance treaties, which is often necessary to model losses in excess of some high threshold value, i.e., to model the largest r upper order statistics. There are two principal kinds of model for extreme values.

2.1.1. Block Maxima

These are models for the largest observations collected from large samples of identically distributed observations. Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables with distribution $\mathrm{F}(\mathrm{x})$. The inference is generally focused around the maximum

$$\mathrm{M_n} = \max(\mathrm{X_1}, \mathrm{X_2}, \ldots, \mathrm{X_n}) \qquad (2.1)$$

the sequence. The distribution of (2.1) is easily derived by applying the rules for independent and identically distributed random variables as

$$\begin{aligned} F_{M_n}(x) &= P(X_1 < x, X_2 < x, \ldots, X_n < x) \\ &= \mathrm{P}(X_1 < x)\mathrm{P}(X_2 < x)\ldots\mathrm{P}(X_n < x) \\ &= \textstyle\prod_{i=1}^{n} F(x) = [F(x)]^n \qquad (2.2) \end{aligned}$$

An asymptotic approximation to $[\mathrm{F(x)}]^n$ is based on the Fisher - Tippet theorem (1928). Given that $x < x^+$, where x^+ is the upper end-point of F (that is, the smallest value of x such that F (x) = 1), [F (x)]n → 0 as n → ∞. The asymptotic approximation is based on the introduction of sequences of normalizing constants $\mathrm{a_n}$ and $\mathrm{b_n}$ and adjusting the distribution in (2.1) such that

$$\mathrm{P}\left(\frac{\mathrm{M_n} - \mathrm{b_n}}{\mathrm{a_n}} \leq x\right) = [\mathrm{F}(\mathrm{a_n}x + b_n)]^{\mathrm{n}} \rightarrow \mathrm{G}(x) \qquad (2.3)$$

The Fisher and Tippet (1928) theorem states that if G(x) converges to some non-degenerate distribution function, then G(x) is said to be belong to a three-parameter family, Generalized Extreme Value Distribution (GEV) of the form

$$\mathrm{G}(x/\mu, \beta, \xi) = \begin{cases} \exp\left(-\left[1 + \xi\dfrac{x-\mu}{\beta}\right]_+^{-\frac{1}{\xi}}\right), & \text{if } \xi \neq 0 \\ \exp\left(-\exp\left(-\xi\dfrac{x-\mu}{\beta}\right)\right), & \text{if } \xi = 0, \end{cases} \qquad (2.4)$$

where $\mathrm{x}_+ = \max(\mathrm{x}, 0)$.

The GEV, $\mathrm{G}(\mathrm{x}/\mu, \beta, \xi)$ distribution with $\beta > 0$ scale parameter, $\mu \in \mathbb{R}$ location parameter and $\xi \in \mathbb{R}$ shape parameter is defined on $\{\mathrm{x} : 1 + \xi(\mathrm{x} - \mu)/\beta > 0\}$.

The shape parameter, ξ of the GEV distribution defines a type of distribution, meaning a family of distributions specified up to location and scaling. The GEV subsumes three types of extreme value distributions which are known by other names according to the value of ξ: when $\xi > 0$ the distribution is a Fréchet distribution; when $\xi = 0$ it is a Gumbel distribution; when $\xi < 0$ it is a Weibull distribution.

$$\text{Weibull: } \Phi(\mathrm{x}/\alpha) = \begin{cases} \exp[-(-\mathrm{x})^{\alpha}] & \text{if } \mathrm{x} \leq 0 \\ 1 & \text{if } \mathrm{x} > 0' \end{cases} \quad \alpha > 0$$

$$\text{Gumbel: } \Lambda(\mathrm{x}/\alpha) = \exp[-\mathrm{e}^{-\mathrm{x}}], \quad \mathrm{x} \in \mathbb{R}$$

$$\text{Fre´chet: } \Psi(\mathrm{x}/\alpha) = \begin{cases} 0 & \text{if } \mathrm{x} \leq 0 \\ \exp[-\mathrm{x}^{-\alpha}] & \text{if } \mathrm{x} > 0' \end{cases} \quad \alpha > 0$$

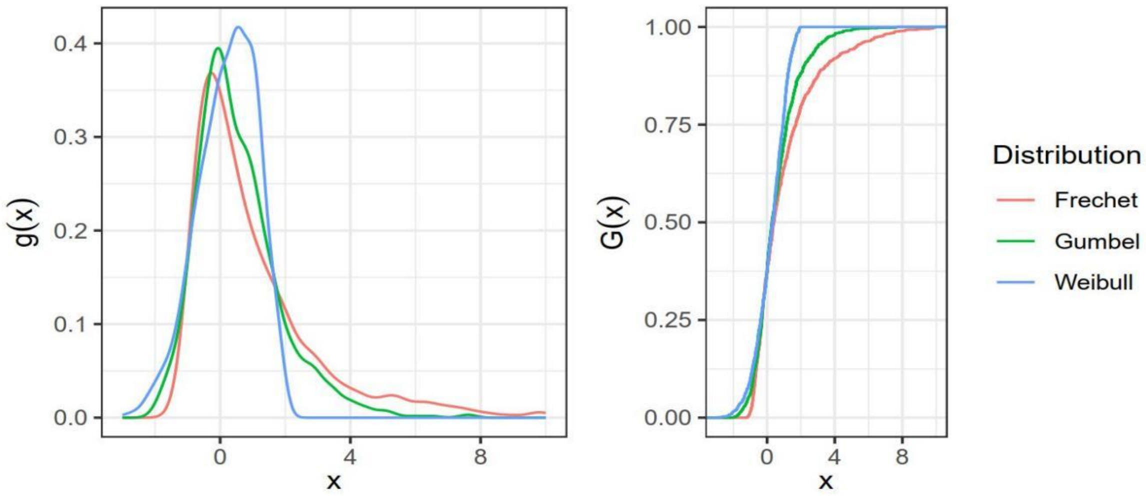

Figure 1: The density function of a standard GEV distribution in three cases: Weibull (ξ = −0.5); Gumbel (ξ = 0); and Fréchet (ξ = 0.5), and their corresponding distributions. In all cases μ = 0 and β = 1.

The density and distribution function of the GEV distribution are shown in the left and right panels of Figure 1, respectively for the three cases $\xi = -0.5, \xi = 0$ and $\xi = 0.5$, corresponding to Weibull, Gumbel and Fre´chet types, respectively. Observe that the Weibull distribution is a short-tailed distribution with a so-called finite right endpoint. The right endpoint of a distribution will be denoted by $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$. The Gumbel and Fre´chet distributions have infinite right endpoints, but the decay of the tail of the Fre´chet distribution is much slower than that of the Gumbel distribution.

The estimates of unknown parameters of GEV are obtained by minimizing the negative log likelihood with respect to parameter vectors (μ, β, ξ). The log negative likelihood of GEV can be written as

$$\ell(\mu,\beta,\xi) = n\log\beta + \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left[1+\xi\left(\frac{x_i-\mu}{\beta}\right)\right] + \sum_{i=1}^{n}\left[1+\xi\left(\frac{x_i-\mu}{\beta}\right)^{-\frac{1}{\xi}}\right],$$

provided that $\left[1+\xi\left(\frac{x_i-\mu}{\beta}\right)\right] > 0$ for $i = 1,2,\ldots,n$.

2.1.2. Peaks-Over-Threshold

The block maxima method has the major defect that it is very wasteful of data. To perform an analyses only the maximum losses in large blocks are retained. For this reason it has been largely superseded in practice by methods based on threshold exceedances, where all data that are extreme in the sense that they exceed a particular designated high level are used. Therefore, Peaks-Over- Threshold (POT) models are generally considered to be the more useful for practical applications, due to their more efficient use of the (often limited) data on extreme values in modeling the behavior of extreme values above a high threshold. An additional advantage of POT is that it provides with risk estimates that are easy to compute. Within the POT class of models one may distinguish two styles of analysis. These are the semi-parametric models built around the Hill estimator and the fully parametric models based on the generalized Pareto distribution (GPD). This paper highly concentrate more on the latter style of analysis for a reasons of relative simplicity in giving statistical estimates error using the techniques of maximum likelihood inference. The excess distribution above the threshold u can be defined as the conditional probability

$$F_u(y) = P(X-u \le y \mid X > u) = \frac{F(y+u)-F(u)}{1-F(u)}$$
$$= \frac{F(x)-F(u)}{1-F(u)}, y > 0. \qquad (2.5)$$

The methodology is based on the asymptotic approximation of $Y = X - u$ to the GPD, the rescaled excesses above a suitably high level u, should the non-degenerate limiting distribution exist Pickands (1975). For those distributions F that satisfy that the distribution in (2.3) converges to (2.4), it can be shown that for large enough u there exists a positive function σ, such that (2.5) is well approximated by the cumulative distribution function of the GPD that takes the form

$$G(y/\sigma,\xi) = \begin{cases} 1-\left(1+\frac{\xi}{\sigma}y\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1-\exp\left(-\frac{1}{\sigma}y\right) & \text{if } \xi = 0 \end{cases} \qquad (2.6)$$

where $x_+ = \max(x,0)$, defined on $\{y : y > 0\}$ and $(1 + \xi y/\sigma > 0)$ with shape parameter $\xi \in \mathbb{R}$ - known as the Extreme Value Index (EVI) and threshold dependent scale parameter σ.

The GPD is applicable to very wide classes of underlying distributions of Y according to Smith (2009). The three cases $\xi < 0, \xi = 0$ and $\xi > 0$ correspond to different types of tail behavior. The case $\xi < 0$ arises in distributions where there is finite upper bound on the claims which are possible in generally speaking and it might be thought that this case would apply in practice. It would be expected to detect such a limit only if there is a tendency for claims to cluster near the upper limit. The second case, $\xi = 0$, which can be obtained by a formal limit $\xi \to 0$ in (2.6) typically arises in cases with an exponentially decreasing tail. This arises not only when the distribution of Y is indeed exponential, but also from many other common distributions such as gamma, Weibull, normal, lognormal etc, so it might be expected to find the estimated value of ξ close to 0 in practice. However the third case, $\xi > 0$, which is usually referred as Pareto tail is of more concern because this corresponds to a genuinely long tailed distribution and most relevant for insurance risk managers.

In the use of GPD, a critical issue in practice is the selection of an appropriate threshold u, or equivalently the selection of an adequate number of upper order statistics. There is a trade off between bias and variance in threshold selection Bawa et al. (2001). If this is set too high so that the asymptotic theorem can be considered to be essentially exact, there will not be enough data over the threshold to calculate good estimates of σ and ξ. However, it is not actually wanted u to be too low since there will not be point in basing the estimates on claims which are too small to be considered large claims, and to do so could induce a bias associated with lack of fit of the GPD.

There are numerous ways of choosing the threshold as well as quantifying the uncertainty of u. A diagnostic graphical tool which has been introduced by Davison and Smith (1990) is the *sample mean excess* plot, which is also known as the *sample mean residual life (MRL)* plot is a very helpful for the selection of the threshold u through visualizing description of the GPD behavior for different values of u This is based on the fact that the mean of a GPD distributed variable Y is given by

$$E(Y) = \frac{\sigma}{1-\xi} \qquad (2.7)$$

and for the introduced an excess u

$$E(Y-u\,|Y>u) = \frac{\sigma}{1-\xi} \qquad (2.8)$$

The mean of a GPD should theoretically have linear property which means by introducing a high threshold $z > u$ should yield

$$E(Y-z\,|Y>z) = \frac{\sigma+\xi z}{1-\xi}, \quad \sigma+\xi z > 0 \qquad (2.9)$$

which gives the average of the excesses of Y over varying values of a threshold z and is a linear transformation of (2.7). Thus the excess distribution over higher thresholds remains a GPD with the same ξ parameter but a scaling that grows linearly with the threshold z. Provided that $\xi < 1$, the mean excess function is given by

$$E(z) = \frac{\sigma+\xi(z-u)}{1-\xi} = \frac{\xi z}{1-\xi} - \frac{\sigma-\xi u}{1-\xi} \qquad (2.10)$$

where $u \le z < \infty$ if $0 \le \xi < 1$ and $u \le z \le u - \sigma/\xi$ if $\xi < 0$. The linearity of the mean excess function (2.10) in z is commonly used as a diagnostic for data admitting a GPD model for the excess distribution. It forms the basis for the following simple graphical method for choosing an appropriate threshold.

Empirically, the mean excess function is defined by the points $(u, e_n(u))$, where $e_n(u)$ is the sample mean excess function estimated as

$$e_n(u) = \frac{\sum_{i=1}^{n}(Y_i - z)1_{[Y_i>u]}}{\sum_{i=1}^{n} 1_{[Y_i>u]}} \qquad (2.11)$$

and is the sum of the excesses $(Y_1 - z),\dots,(Y_n - z)$ over the threshold z divided by the number of data points which exceeds the threshold z. The sample mean excess function describes the expected overshoot of a threshold given that exceedance occurs and is an empirical estimate of the mean excess function that is defined in (2.8). The estimated mean excess function defined in (2.9) should be linear. Whenever the points show an upward trend, it is a sign of heavy tailed behavior. Exponentially distributed data approximately would give an horizontal line and data from a short tailed distribution would show a downward trend, as noted in Corradin (2002).

Another graphical tool used to choose the threshold is the Hill graph. Let $Y_{(1)} \ge Y_{(2)} \ge \cdots \ge Y_{(n)}$ be associated descending ordered statistics of $(Y_1, Y_2, \dots, Y_n)$ which are independent and identically distributed (iid) random variables Brilhante et al. (2013). Assuming that the distribution of these random variables is heavy-tailed, the Hill estimator Hill (1975) of tail index ξ using $k + 1$ ordered statistics is defined by

$$H_{n,k} = \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{Y_{(i)}}{Y_{(k+1)}}\right). \tag{2.12}$$

Obviously$_{k,n}$, the Hill estimator is function of these extreme random variables $\{Y_{(1)}, Y_{(2)}, \dots, Y_{(k)}\}$ which depends on the chosen threshold. A Hill plot is therefore constructed by the Hill estimator of a range of k value versus the value of k or the threshold, i.e. is defined by a set of points $\{(k, H_{k,n}^{-1}), 1 \le k \le n - 1\}$. The value of Y_k above which the Hill estimator tends to be stable can be chosen as the optimal threshold u. The Hill estimator is closely related to the mean excess function. It is asymptotically equal to the reciprocal of the empirical mean excess function of $\log(Y)$ evaluated at the threshold $\log(Y^{(k+1)})$. An important feature of the Hill estimator to keep in mind is the variance-bias trade off that occurs when choosing the number of upper order statistics to use. Choosing too many of the largest order statistics can lead to a biased estimator, while too few increases the variability of the estimator.

2.2. Risk Measures

2.2.1. *Value at Risk*

Value at Risk (VaR) helps to quantify the amount of capital needed for covering loss in portfolio. It is defined as the α-th quantile of the negative returns or losses of a portfolio distribution. In other words, for some given confidence level $\alpha \in (0,1)$, the VaR of our portfolio at the confidence level α is given by the smallest number ℓ such that the probability that the loss L exceeds ℓ is no larger than $(1 - \alpha)$. Formally,

$$\begin{aligned} \text{VaR}\alpha &= \inf\{\ell \in \mathbb{R}: P(L > \ell) \le 1 - \alpha\} \\ &= \inf\{\ell \in \mathbb{R}: F_L(\ell) \ge \alpha\}, \end{aligned} \tag{2.13}$$

where F_L is defined as the distribution function. Typical values for α are α = 0.95 or α = 0.99. Note that by its definition the VaR at confidence level α does not give any information about the severity of losses which occur with a probability less than $1 - \alpha$. This is clearly a drawback of VaR as a risk measure.

2.2.2. *Expected Shortfall*

Expected Shortfall (ES) or the tail conditional expectation quantifies the average loss given that we have lost at least VaR. ES is computed by taking the average of losses that are larger than VaR. In other words, For a loss L with $E(|L|) < \infty$ and distribution function F_L the expected shortfall at confidence level $\alpha \in (0, 1)$ is defined as

$ES_\alpha = E(X / X > \text{Var}_\alpha) = \frac{1}{1-\alpha}\int_\alpha^1 q_u(F_L)\, du$, where $q_u(F_L) = F_L^{-1}(u)$ is the quantile function of F_L. Expected shortfall is thus related to VaR by

$$ES_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 \text{VaR}_u(L)\, du$$

Instead of fixing a particular confidence level α, we average VaR over all levels $u \ge \alpha$ and thus we look further into the tail of the loss distribution. Obviously ES_α depends only on the distribution of L and obviously $ES_\alpha \ge \text{VaR}_\alpha$.

3. Preliminary Data Analysis

3.1. The data

The data used for this analysis were provided from a large database of the Ethiopian Insurance Corporation, one of the biggest insurance companies in Ethiopia. It consists of policy and claim information of vehicle insurance at the individual level. The dataset originally contains n = 288, 763

unique individual contracts, represented by the observations $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ of $p = 10$ predictors of $\mathbf{X} = (X_1, \dots, X_p) \in \mathbb{R}^p$ and the response variable $Y \in \mathbb{R}$ represents claim size, from July 2011 to June 2018. The response variable is originally in Ethiopian birr and it is converted to USD at the official exchange rate at the time of data analysis. Furthermore, the analysis depends on the natural logarithmic transformation of the response variable for a better visualization.

3.2. Exploratory Plots for the Distribution of claim paid

The purpose of statistical graphics is to provide visual representations of quantitative and qualitative information. As a methodological tool, statistical graphics comprise a set of strategies and techniques that provide the researchers with important insights about the data under examination and help guide for the subsequent steps of the research process. The objectives of graphical methods are to explore and summarize the contents of large and complicated data sets, address questions about the variables in an analysis (for example, the distributional shapes, ranges, typical values and unusual observations), reveal structure and pattern in the data, check assumptions in statistical models, and facilitate greater interaction between the researcher and the data. Various graphical methods were examined to visualize data in raw and amalgamated formats. Additionally, beyond graphical exploratory data analysis, some methods to quantify fits of the data with some distributions are discussed.

Graphical visualization in this analysis starts with a distribution of large claims. Accordingly, more than 70% of the sum of all claims is created by only the 10% highest claims as shown in Figure 2.

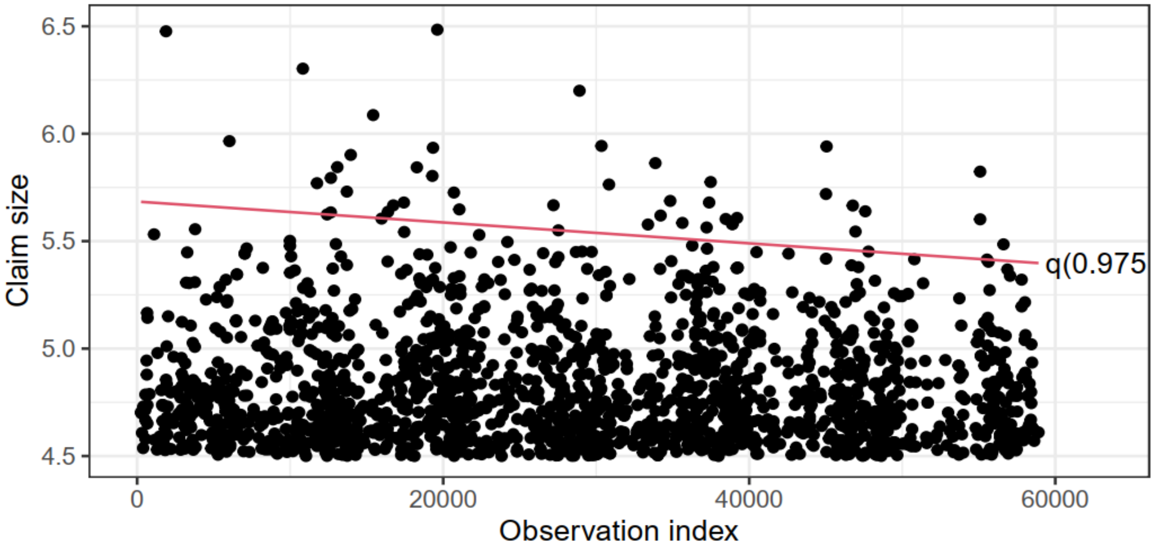

Figure 2: All claim sizes falling in an interval of $Y \in (4.5, 6.5)$ in (black points) and the 97.5% quantile (red line) of all claims.

Even for $Y \in (4.5, 6.5)$, extreme points can be seen, which indicates the important contribution of single large claims to a total risk exposure.

In Figure 2, it can be easily seen that some observations are much larger than the rest of the sample, specifically, two outliers (6.47 and 6.48) can be seen which may of course need some concern for any reason. These two largest claims could possibly happen due to a total loss, i.e. when a complete facility was lost and it may represent total losses. The second largest three claims are 6.3, 6.2 and 6.08. This shows that a particular concern should be drawn prior to the use of extreme value methods, which are applied in the rest of this section, when there is a possibility that these represent some separate process. Ideally one would like to do a separate analysis of claims resulting from total losses, but with only few such claims available, this is not practicable. The most of outliers will therefore be combined with the rest of the data for most of the analysis, but their separate origins do need to be bear in mind in interpreting the results.

The most widely recognized graphical tool to display and examine the frequency distribution and a density of a single continuous variable is the histogram. A histogram is non-parametric procedure, in a sense, constructed without assuming a statistical model and estimating its parameters from data.

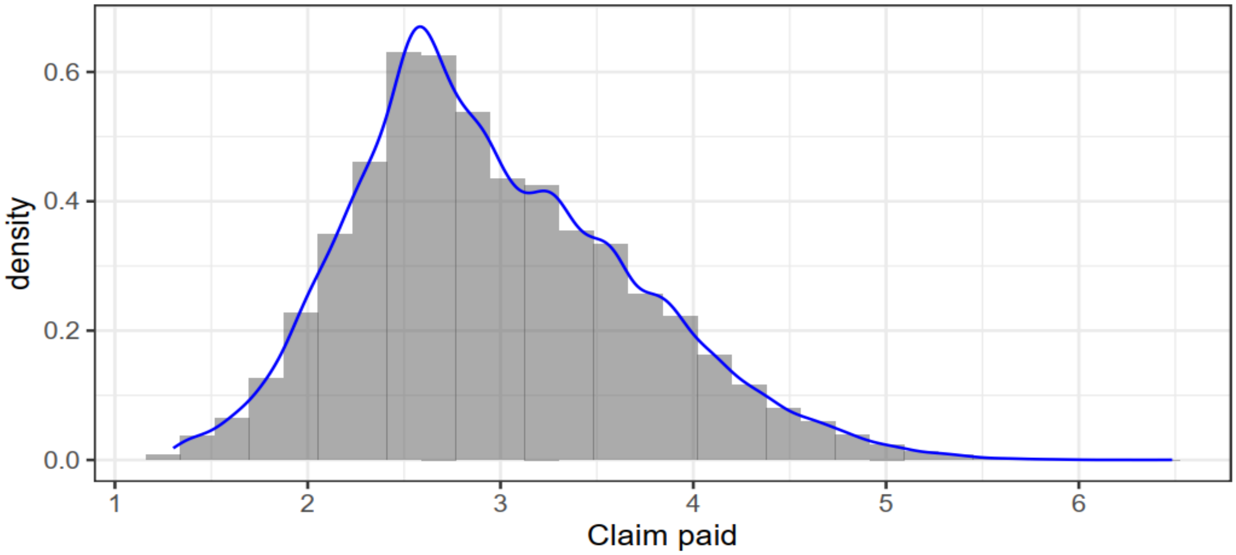

Figure 3: Frequency histogram and superimposed density plot representations of claim paid distribution.

The histogram plot in Figure 3 has a peak at the smaller values, more variability in the higher values and a bit long right tail. knowing this ahead of time is helpful to design a best fitting model later on.

Another common tool to visualize the observed distribution of data is by plotting a smoothed histogram commonly referred as empirical density, the curve superimposed on the histogram with blue line. The empirical densities overcome some of the disadvantages caused by the arbitrary discrete bins used in the basic histograms.

Although it is helpful to examine the observed data distribution, often we are examining the distribution to see whether it meets the assumption of the statistical analysis we hope to apply. As it can be seen from Figure 3, the empirical density plots shows that the claim paid variable is horizontal line (in fact some degree of smoothness is applied by default), which indicates this variable has a heavy right tail.

By proceeding claim paid examination, a quantile-quantile (Q-Q) plot can be considered as diagnostic tool to assess whether data fit or are close to a specific expected distribution. Q-Q plots can be used to judge whether observations follow a variety of distributions such as: normal, exponential and generalized Pareto distributions.

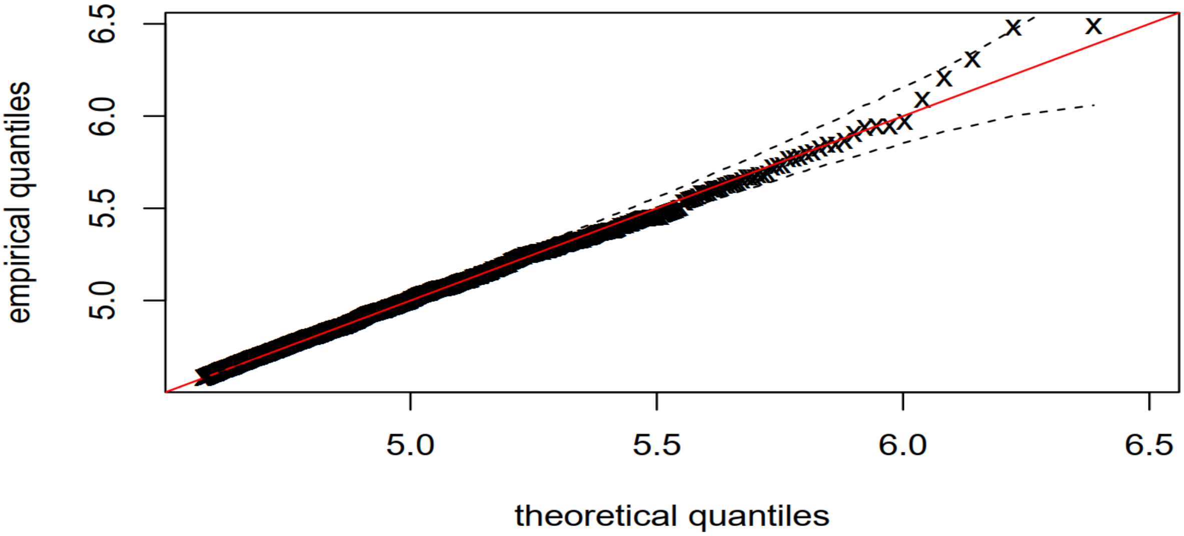

Figure 4: A Q-Q plot of claim paid empirical distribution against theoretical Generalized Pareto distribution (GPD). A threshold of $u = 4.58$ USD, above which 97.5% of claim payments fall is selected. Thus, 1,478 exceedance observations are left and the estimated parameters of GPD are: $\xi = 0.11$ and $\sigma = 0.34$.

A Q-Q plot graphs the observed data quantiles against theoretical quantiles from the expected distribution. With a Q-Q plot, if the data perfectly match the expected distribution, the points will fall on a straight line. In Figure 4, we can see that the data are reasonably Generalized Pareto distributed (GPD), as all points fall fairly closely to the straight line except for few points, which can be considered as outliers. Moreover, since the plot has a curved pattern with the slope increasing from left to right, then the data has a long right tail Keen (2010), even beyond the GPD can accommodate. Although testing whether data are consistent with a specific distribution, in this case GPD, is common, real data may be closer to many other distributions.

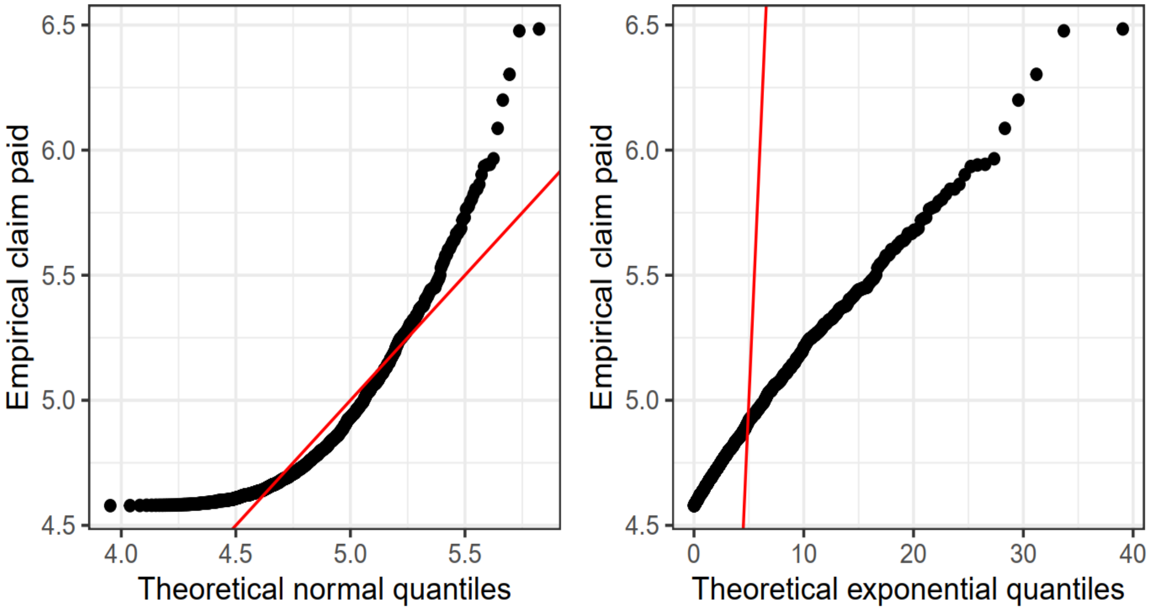

Figure 5: A normal distribution (left panel) and an exponential distribution (right panel) Q-Q plot of claim paid. The two distributions were fitted to the 1,478 exceedance observations.

The maximum likelihood estimator(s) of the parameters for normal and exponential distributions were directly computed from the empirical claim paid data. Considering a nature of claim paid variable such as its range, a Q-Q plots shown in Figure 5 are done to evaluates the fit of claim paid variable with a specified expected quantile functions from normal and exponential distributions. The plots show that the exponential theoretical distribution for claim paid is unreliable. But in the case of normal distribution, it is some how close to that of the GPD Q-Q plot as the points are seem to be symmetric with a line.

Another way to examine whether the observed distribution appears consistent with an expected distribution is to plot the empirical density against the density for the expected distribution.

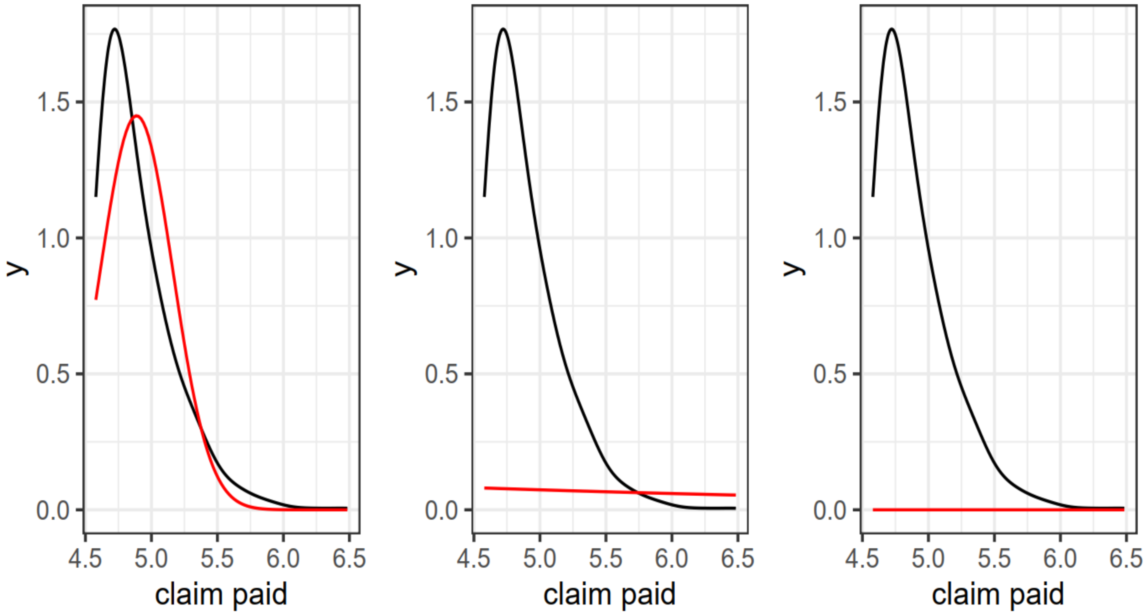

Figure 6: A theoretical curve (red line) and claim paid empirical density plot (black line) with smoothing factor of 2, from GPD (left panel), normal(middle) and exponential(right panel) distributions.

From Figure 6, it can be seen that the claim paid data appear to be close to a GPD distribution, although not perfect. For a better comparison between the normal, exponential and GPD fits, a log likelihood (LL) is employed. LL commonly used for model comparison and tells us about how likely the data are to come from that distribution with those parameters. In comparing fit of the distributions, the one that provides the higher log likelihood is a better fit for the data. Accordingly, the L L is higher for the GPD $(\mathrm{LL} = 268.39)$ than the normal distribution $(\mathrm{LL} = -190.57)$ and the exponential distribution $(\mathrm{LL} = -3823.17)$ with only one unit difference in the degrees of freedom. These results suggest that the GPD should be picked for claim paid data. More details about the GPD model and extreme value analysis are discussed in Section 2.1.

4. Results

4.1. Modeling using GEV

The maximum claim paid data across the levels of manufacturer company are shown in Figure 7. No obvious trend is observed.

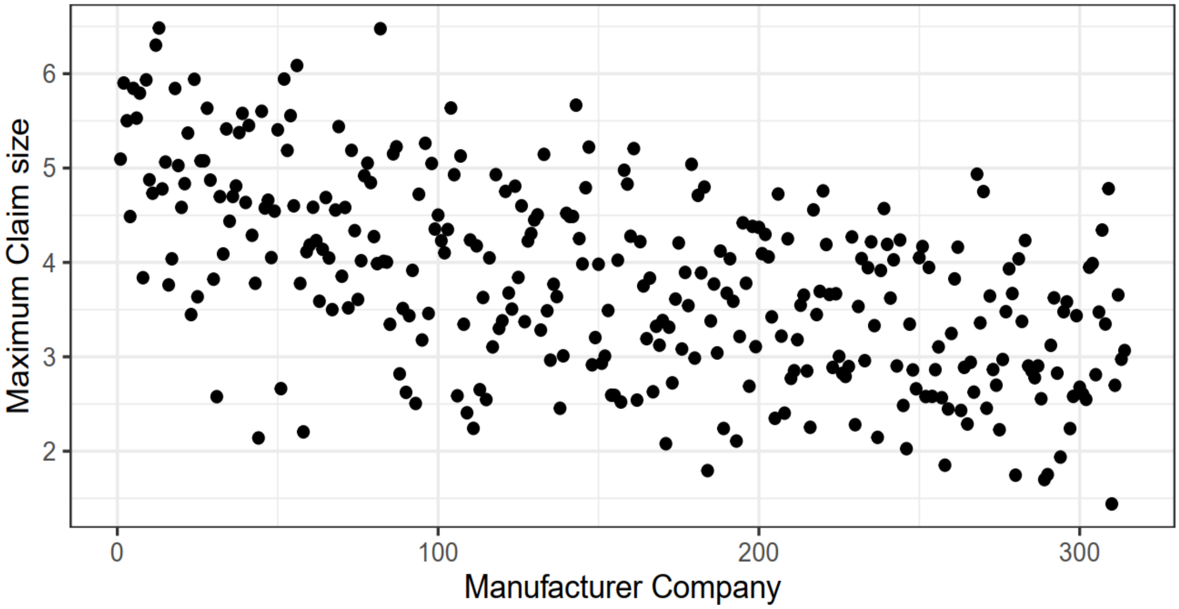

Figure 7: Scatter plot of maxima claim paid across manufacture company for Ethiopian vehicle insurance dataset.

Figure 7 represents maximum claim paid across manufacturer company, showing extreme claim paid using block maxima method. The highest four amount of claim paid, which seem to be outliers occurred in USD are 3.05m, 2.99m, 2.01m and 1.22m, that correspond to FIAT, SKY BUS, IVECO and BISHOFTU, respectively.

For GEV estimation, the Block Maxima of maximum claim paid across manufacturer company are extracted. The blocks $n = 314$ types of manufacturer company have been chosen to be reasonably large, so the GEV model is fitted to the $n = 314$ across the manufacturer's maxima using maximum likelihood estimation. The point MLE of the parameters (μ; β; ξ) for the GEV distribution and their respective 95% CI's are summarized in Table 4.1. Based on these estimates, VaR is calculated by (2.13) and VaR estimate is also presented in Table 4.1.

	Lower bound	Point estimate	Upper bound
$\hat{\mu}$	3.30	3.42	3.54
$\hat{\beta}$	0.91	0.99	1.08
$\hat{\xi}$	0.14	0.22	0.30
$\widehat{VaR}_{0.01}$	___	12.96	___
$\widehat{ES}_{0.01}$	___	14.52	___

Table 4.1: Point and 95% CI estimates of GEV parameters, and point estimates of risk measures.

The CI results shows that the confidence interval of ξ does not contains 0 and both lower and upper bounds are positive, which means the Fréchet distribution could be a more accurate model in the entire GEV family.

Since we have our estimated parameters of GEV, we can calculate the risk measures (VaR and ES), which are contained in Table 4.1. Notice that at 99% VaR and ES are 12.96 and 14.52, which indicate that, with probability 0.01, the insurance company makes claim payment of at least 12.96 and on a long position the claim payment will reach up to 14.52 on an average.

4.2. Modeling using GPD

The mean excess plot in left panel of Figure 8 is in fact fairly linear over the entire range of the claim size distribution and its upward slope leads us to expect that a GPD with positive shape parameter ξ could be fitted to the entire dataset. However, there is some evidence of a "kink" in the plot below the value 285, 000 and a straightening out of the plot above this value, so we have chosen to set our threshold at u = 285, 000 and fit a GPD to excess claim sizes of 58 observations above this threshold, in the hope of obtaining a model that is a good fit to the largest of the claim sizes. If the data really follow a GPD, then this plot should stay close to a straight line of slope $\xi/(1-\xi)$, provided $\xi < 1$ (Smith, 2009). The apparent exception to linearity of the mean excess plot is at the right-hand end of the plot, but in fact this is not such a significant matter because in this region there are very few data points- the mean excess is computed from a very small number of exceedances and hence has a lot of sampling variability. On the basis of this plot, the evidence in favour of the GPD seems good. The ML parameter estimates are $\hat{\xi} = 0.4$ and $\hat{\sigma} = 5.31$ with standard errors 0.15 and 3.62, respectively. Thus the model we have fitted is essentially a heavy-tailed, infinite-variance model. A picture of the fitted GPD model for the excess distribution $\hat{F}_u(y - u)$ is also given in right panel of Figure 8, superimposed on points plotted at empirical estimates of the excess probabilities for each claim size; note the good correspondence between the empirical estimates and the GPD curve. The Hill estimator (the reciprocal of ξ) in Figure 9 concise with the mean excess plot, indicating that the Hill estimator starts to be stable after the vertical red line, which is drawn at threshold of u = 285, 000. The estimates of tail index α obtained are between 1.25 and 2.5, suggesting ξ estimates between 0.4 and 0.8, all of which correspond to infinite-variance models for these data. The tail index α estimates based on k = 50, . . . , 120 order statistics mostly range from 1.25 to 2.5, suggesting a ξ value in the range 0.4 − 0.8, which is larger than the values estimated in with a GPD model.

In Figure 9, it can be noted that the high variability in the left region (the one determined by the largest order statistics) of the plot is not a welcome feature, since it makes difficult the proper selection of the number of upper order statistics involved in the estimation of the tail index. An important question that often arises in practice is whether one should ignore those observations, thus ignoring useful information about the behavior of the tail, or include them and get a biased estimate of α. Even though the values of α seem to be decrease as a number of exceedences increase, it can be seen that for the ideal case of setting, to a large extent the plots perform satisfactorily allowing the data analyst to identify correctly the underlying value of the tail index.

In insurance we might use the model to estimate the expected size of the insurance claim, given that it enters a given insurance layer. Thus we can estimate the expected claim size given exceedance of the threshold of USD 285, 000 or of any other higher threshold by using (2.10) with the appropriate parameter estimates.

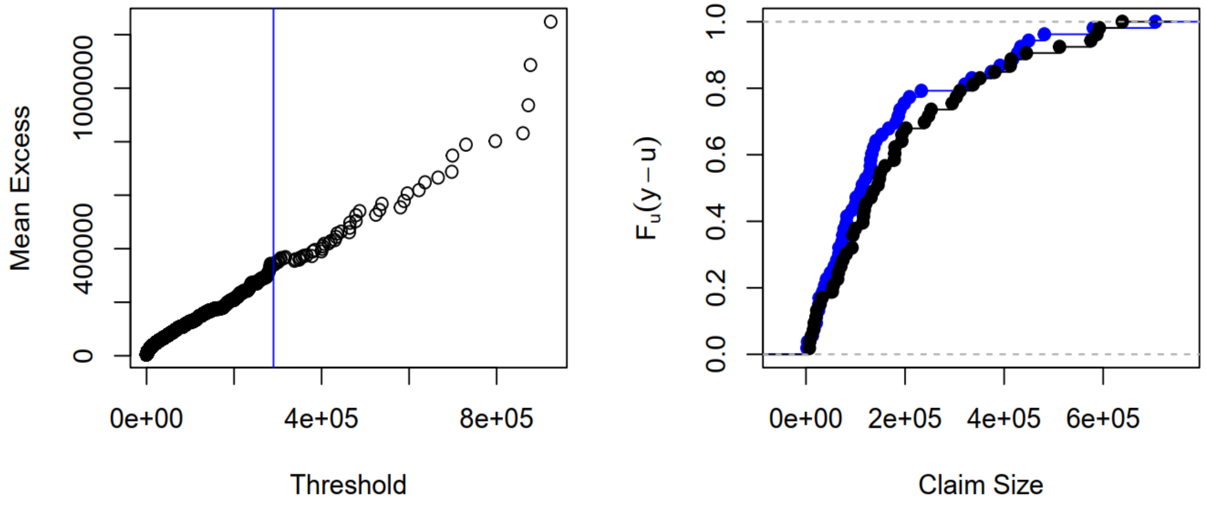

Figure 8: Sample mean excess plot (left panel) and on the right panel is Empirical distribution of excesses(black points) and fitted GPD (blue points).

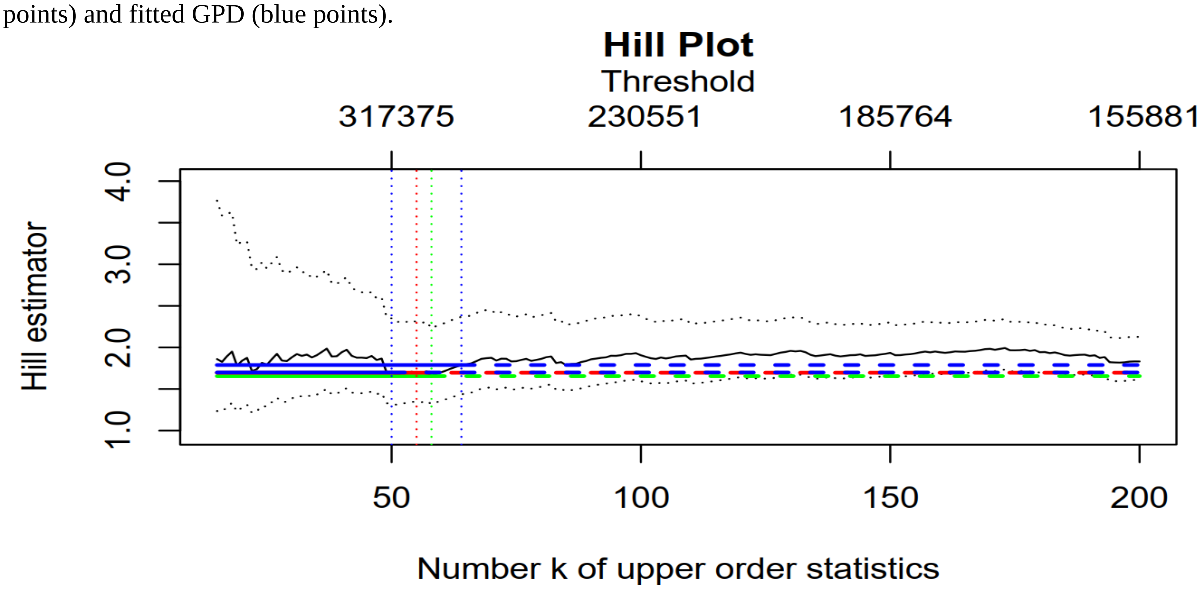

Figure 9: Hill plots of four different thresholds together with 95% confidence interval, and their respective estimated the tail indices $\alpha = 1/\xi's$. The four ξ estimators obtained are $\hat{\xi} = 0.56, 0.6, 0.59, 0.57$ by setting the corresponding thresholds $u = 280000, 285000, 300000, 315000$. The number of exceedences k is plotted on the horizontal axis while the estimation of the reciprocal of ξ is plotted on vertical axis.

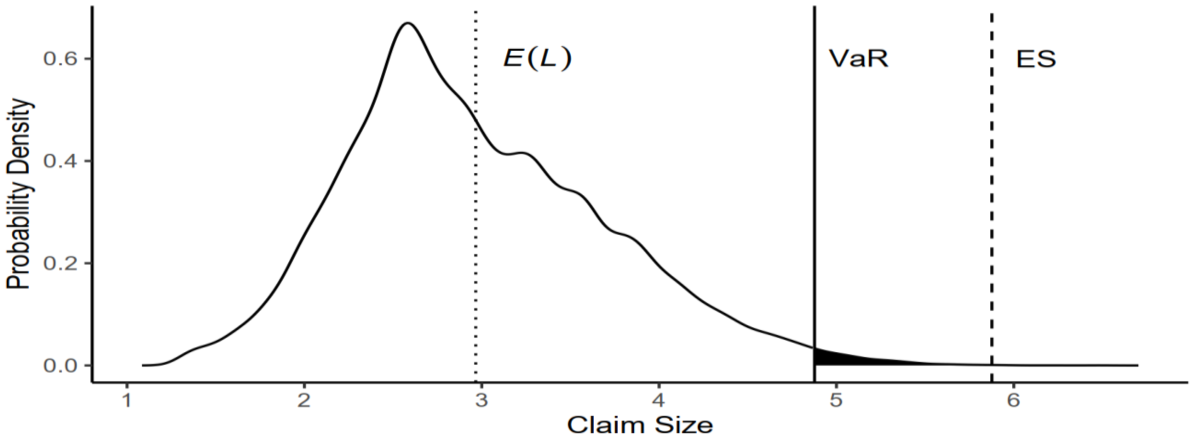

Figure 10: Claim size distribution with the 99% VaR marked as a vertical line; the mean claim is shown with a dotted line and the 99% expected shortfall is marked with a dashed line.

Figure 10 illustrates the notion of VaR. The probability density function of a claim size distribution is shown with a vertical line at the value of the 99% VaR. Note that the estimated mean claim size is $(E\ (L)\ =\ 3.7)$, while the 99% estimated VaR and ES values are approximately 5.84 and 5.87, respectively; indicating that there is a 1% chance that the insurance company makes claim payment of at least 5.84. Denoting by μ the mean of the claim size distribution, sometimes the statistic $\mathrm{VaR}^{\mathrm{mean}} = \mathrm{VaR}\alpha - \mu$ is used for capital-adequacy purposes instead of ordinary VaR. The distinction between the ordinary VaR and $\mathrm{VaR}^{\mathrm{mean}}$ is of little relevance in market risk management, where the time horizon is short and μ is close to zero. It becomes relevant in credit where the risk-management horizon is longer.

In particular, in loan pricing one uses $\mathrm{VaR}^{\mathrm{mean}}$ to determine the economic capital needed as a buffer against unexpected losses in a loan portfolio. Taking the expectation of the claim size distribution into account is also important in the growing field of asset-management risk. The 99% expected shortfall value is 5.87, which is much higher than the expected claim size value of 3.7 in this case.

5. Discussion and Conclusion

The purpose of this study was to analyze the tail distribution of claim size for Ethiopian vehicle insurance dataset using EVT. We used both GEV and GPD distributions in block maxima and peaks-over-threshold methods, respectively. GEV and GPD are two different distributions, but in this context their purpose is the same; they show the distribution of extreme claim size. Also, the shape parameter ξ is the same parameter in the two distributions. Theoretically, we should get the same estimation of ξ in the methods (Coles, 2001). A weaker hypothesis is that the sign of ξ should be the same in the models, which the case in this study. This is also the case in Gilli and Kellezi(2006). Even though we obtained $\hat{\xi} > 0$ in both models, the magnitudes are different; $\hat{\xi} = 0.22$ in GEV and $\hat{\xi} = 0.4$ in GPD.

One of the objectives of this paper is to answer the question, How much can the claim size fall beyond certain level of threshold. To answer this question, we use the two risk measures VaR and ES. However, how the risk measures should be estimated is not straightforward; there exist several methods for this purpose [see Hull (2018)]. Here, we focus on the block maxima and POT methods. Before estimating VaR and ES, the parameters of the distributions must be estimated:
$(\mu;\ \beta;\ \xi)$ in GEV and $(\sigma;\ \xi)$ in GPD. Since we obtained different estimators of ξ in GEV and GPD, we obviously get different estimators of VaR and ES. The 99% VaR and ES estimators in GEV are
12.96 and 14.52, respectively, while in the GPD, the estimators are 5.84 and 5.87, respectively. This arises an interesting question such as; which one of the the two methods that produces the more accurate estimates of VaR and ES. Performing the "Backtesting" strategy is the popular approach to evaluate the estimates of VaR and ES, which is beyond the scope of this paper and the next step research of the author.

It is possible that the choice of α influence the result. A small α needs to be chosen for the formulas of VaR and ES to be accurate [see Dowd (2005)]. Here, we let $\alpha = 0.01$, but an even smaller α should be even better. Since POT extract the extreme events more efficient than GEV, it is possible that POT is more sensitive to the choice of α than GEV. A solution is then to chose a smaller α . On the other hand, this would imply that fewer observations will be considered as extreme, which also can lead to poor estimates. Choosing another value of α could also bring some new light on the discussion.

References

J. Bawa, L. Trenner, S. Coles, and P. Dorazio. *An introduction to statistical modeling of extreme values*. Springer, 2001.

M. F. Brilhante, M. I. Gomes, and D. Pestana. A simple generalisation of the hill estimator. *Computational Statistics & Data Analysis*, 57(1):518–535, 2013.

S. Coles. An introduction to statistical modeling of extreme values. , 2001.

S. Corradin. Economic risk capital and reinsurance: an extreme value theory's application to fire claims of an insurance company. Sixth International Congress on Insurance: Mathematics and Economics, Lisbon, 2002, 2002.

A. C. Davison and R. L. Smith. Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society, Series B (Methodological)*, 52(3), 1990.

K. Dowd. Measuring market risk. , 2005.

R. Fisher and L. Tippet. Limiting forms of the frequency distribution of the largest or smallest member of a sample. 1928.

M. Gilli and E. Kellezi. An application of extreme value theory for measuring financial risk. *Computational Economics*, 27(207 - 228):2 – 3, 2006.

B. M. Hill. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 13, 1975.

J. C. Hull. Risk management and financial institutions. , 2018.

K. J. Keen. *Graphics for Statistics and Data Analysis with R*. Texts in Statistical Science. CRC Press, Taylor & Francis group, Chapman & Hall/CRC, NY, USA, 2010.

J. McNeil, P. Embrechts, and R. Frey. Quantitative risk management concepts, techniques and tools. *Princeton University Press Princeton and Oxford*, 2005.

J. I. Pickands. Statistical inference using extreme value order statistics. *Annals of Statistics*, 1975.

R. L. Smith. *Extreme value analysis of insurance risk*. Department of Statistics and Operations Research, University of North Carolina,Chapel Hill, NC 27599-3260, 2009.

Appendix

The following R code is used to generate all the results and figures, preferably in RStudio. motor_data.csv dataset can be provided upon request.

```
> rm(list = ls()) ; setwd(); library(fExtremes); library(ggplot2); library(tidyverse); library(cowplot); library(dplyr); set.seed(4444)

> Weibull_x <- gevSim(model = list(xi = -0.5, mu = 0, beta = 1), n = 1000)

> Gumbel_x <- gevSim(model = list(xi = 0, mu = 0, beta = 1), n = 1000)

> Frechet_x <- gevSim(model = list(xi = 0.5, mu = 0, beta = 1), n = 1000)

> dens_dat <- bind_rows(tibble(Distribution = "Weibull", x = as.numeric(Weibull_x)),

+ tibble(Distribution = "Gumbel", x = as.numeric(Gumbel_x)),

+ tibble(Distribution = "Frechet", x = as.numeric(Frechet_x)))

> dens_plt <- ggplot(data = dens_dat) + geom_density(aes(x = x, color = Distribution)) +

+ labs(x = expression(x), y = expression(g(x))) + xlim(-3,10) + theme_bw() + theme(legend.position = "none");dens_plt

> cdf_plt <- ggplot(dens_dat) + stat_ecdf(aes(x = x, color = Distribution)) + labs(x = expression(x), y = expression(G(x))) + xlim(-3,10) + theme_bw();cdf_plt
> dens_cdf_plt <- plot_grid(dens_plt,
   cdf_plt);dens_cdf_plt

> motor_claimed <- read.table("motor_data.csv", header
   = TRUE, sep = ",")

> obs.idx = 1:length(motor.claimed$CLAIM_PAID)

> high_claims_plt <- ggplot(motor.claimed,
   aes(x=obs.idx, y=CLAIM_PAID)) + geom_point() +

  xlab("Observation index") + ylab("Claim size") +
   stat_quantile(quantiles = 0.975, col=2 ) +
   ylim(c(4.5,6.5)) + annotate("text", x = 6.3e4, y = 5.4,
   label = "q(0.975)") + theme_bw();high_claims_plt

> histogram_claim_plt <- ggplot(data=motor.claimed,
   aes(x=CLAIM_PAID)) + geom_histogram(aes(y =
   ..density..),alpha = 0.5, position =
   "identity")+geom_density(

   col = "blue")+xlab("Claim
   paid")+theme_bw();histogram_claim_plt

> threshold <- quantile(motor.claimed$CLAIM_PAID,
   probs = 0.99)[[1]]

> z <- density(motor.claimed$CLAIM_PAID)

> mydf <- data.frame(x = z$x, y = z$y) %>% mutate(area
   = x >= threshold)

mycols <- list("white", "black")

> density_claim_plt <- ggplot(data=mydf, aes(x=x,
   ymin=0, ymax=y,fill=area)) +

  geom_ribbon() + geom_line(aes(y=y)) +
   theme_classic() +

  annotate("text", x =
   mean(motor.claimed$CLAIM_PAID) + 0.35, y = 0.6,
   label = "italic(E(L))==2.97", parse = TRUE) +

geom_vline(xintercept =
mean(motor.claimed$CLAIM_PAID), linetype = 3) +

  annotate("text", x = threshold + 0.25, y = 0.6, label =
   "VaR", parse = TRUE) +

  geom_vline(xintercept = threshold, linetype = 1) +
```

```
  annotate("text", x = threshold + 1.25, y = 0.6, label =
   "ES", parse = TRUE) +

  geom_vline(xintercept = threshold + 1, linetype = 2) +
   scale_fill_manual(values = mycols) +

  theme(legend.position = "none") + xlab("Claim Size") +
   ylab("Probability Density");density_claim_plt

> library(ismev); probs <- 0.975

> threshold = quantile(motor.claimed$CLAIM_PAID,
   probs = probs, na.rm = TRUE)

> gpd.ml = gpd.fit(motor.claimed$CLAIM_PAID,
   threshold=threshold)

> exc.data =
   motor.claimed$CLAIM_PAID[which(motor.claimed
   $CLAIM_PAID>threshold)]

> scale.est = gpd.ml$mle[1]; shape.est = gpd.ml$mle[2]

> library(tea)

> qq.gpd <- qqgpd(exc.data, nextremes =
   length(exc.data), scale = scale.est, shape = shape.est)

> hlplot <- hillplot(motor.claimed$CLAIM_PAID,
   orderlim = c(15,2e2), y.alpha = FALSE,try.thresh =
   c(5.47, 5.45, 5.5, 5.55),lwd = 1, main = "", ylab =
   expression("Tail Index - " ~ xi), xlab = "Number k of
   upper order statistics", legend.loc = NULL)

> max_claim <- numeric()

> for (i in 1:length(unique(motor.claimed$MAKE)))

{max_claim[i] <-
   max(motor.claimed$CLAIM_PAID[which(motor.clai
   med$MAKE==unique(motor.claimed$MAKE)[i])])}

> gev_max_claim_plt <- ggplot() +theme_bw() +
   geom_point(aes(x = seq(1,
   length(unique(motor.claimed$MAKE))), y =
   max_claim)) + labs(x = "Manufacturer Company", y
   = "Maximum Claim size")
```