Available online at http://www.iorajournal.org/index.php/ijgor/index

International Journal of Global Operations Research

e-ISSN: 2722-1016
p-ISSN: 2723-1739

Vol. 3, No. 2, pp. 46-51, 2022

Estimation of Generalized Pareto Distribution Parameters in Traffic Accident Loss Data Modeling

Rabiu Hamisu Kankarofi [1], Galang Hawy Alfarisi [2], Moch Panji Agung Saputra[3,*]

[1]*Department of Mathematics, Yusuf Maitama Sule University, Kano, Nigeria*
[2,3]*Mathematics Study Program, Faculty of Mathematics and Natural Science Universitas Padjadjaran, Indonesia*

**Corresponding author email: moch16006@mail.unpad.ac.id*

Abstract

The problem of traffic accidents in Indonesia has a high level of risk. In an effort to minimize losses due to traffic accidents, it is necessary to study the data and characteristics of traffic accidents and identify these events as extreme events. This study was conducted to find out how to estimate the shape and scale parameters using Maximum Likelihood Estimation (MLE), and to explore data on traffic accident losses in Indonesia. The method used to analyze the extreme value of traffic accident losses is the Extreme Value Theory. One approach to identify extreme values is Peaks Over Threshold which follows the Generalized Pareto Distribution (GPD). Traffic accident loss data is divided into three types based on the cause, namely driver negligence, vehicle quality, and other external factors in the period (2008-2017). Estimation of shape and scale parameters is obtained through MLE which is then solved by Newton Raphson, because it produces equations that are not closed form. This study resulted in an estimate of the shape and scale of the GPD distribution parameter, as well as a confidence interval (1-α) of 100% with of 5%. In addition, it is concluded that the parameters obtained from the estimation have the same characteristics for each type of risk analyzed, but have different parameter values. Based on parameter estimation, GPD distribution is obtained from each risk which is expected to be useful for related parties in analyzing the number of traffic accident losses in the next period to consider steps that can be taken to reduce losses due to traffic accidents.

Keywords: Traffic Accidents, Extreme Events, Maximum Likelihood Estimation, Extreme Value Theory, Generalized Pareto Distribution

1. Introduction

The incidence and loss of traffic accidents in Indonesia has a high level of risk with data that is quite high. Reported on the Central Bureau of Statistics (also known as BPS in Indonesian) website, it can be concluded that the data on the incidence and loss of traffic accidents in Indonesia has an increasing trend of data in the period 2008 - 2017. The incidence of traffic accidents can be categorized based on the type of risk that causes the incident, namely driver negligence, vehicle quality, and other external factors. Based on these risks, appropriate countermeasures are needed to reduce the number of incidents and losses of traffic accidents. One of the first steps in risk management efforts is the need for knowledge of the identification of the loss value based on the identification of its characteristics, namely using the estimation of distribution parameters which are analyzed based on the distribution of loss data for each risk (Hubbert, 2012).

The Extreme Value Theory (EVT) method can be used to identify extreme values by looking at changes in the distribution between time periods, as well as the distribution parameters. Therefore, the distribution parameter estimation method has an important role in this study. One of the studies that has applied this theory is a study using the Peaks Over Threshold approach by Yustika (2011) by estimating the parameters of the Generalized Pareto Distribution in the case of identifying values in rice production centers. The study shows the differences in the characteristics of the value of rice production based on the distribution of GPD obtained for each period.

Several other previous studies have also discussed the problem of analyzing a risk using the Extreme Value Theory method. Gourier et al. (2019) has conducted data modeling that has a large data tail using Extreme Value Theory and introduced Copula theory. Then show that Value-at-Risk is a measure of the risk that occurs. The results of this study indicate that the possibility of diversification is not appropriate when a mean-infinite distribution is involved. Baran and Witzany (2011) conducted a study that compared Extreme Value Theory with standard estimation methods

(variance, covariance, historical simulation) to produce Value-at-Risk. These different Value-at-Risk search methods are compared with the back testing procedure and result in the volatility of returns that vary over time.

Based on this description, the problem raised in this study is how to estimate the parameters of the Generalized Pareto Distribution using the Extreme Value Theory method in analyzing extreme values and obtaining the distribution of GPD for each type of traffic accident risk in Indonesia. The results of estimating the parameters and distribution of GPD are expected to be useful for related parties in analyzing the number of traffic accident losses in the next period to consider steps that can be taken to reduce losses due to traffic accidents.

2. Literature Review

2.1. Traffic Accidents

The traffic accident is an incident where a motorized vehicle collides with another object and causes damage. Sometimes these accidents can result in injury or death to humans or animals. Traffic accidents are events that are difficult to predict when and where they will occur (Mohammed et al. 2019).

2.2. Traffic Accidents Risks

Risk is defined as a hazard, consequence or consequence that can occur due to ongoing processes or certain events that will occur in the future. Risk is something that is always faced by humans and its nature is very uncertain. One form of risk is operational risk. Operational risk in determining market opportunities in an insurance, namely determining the most vulnerable risks, how consumers deal with risks and the consumer's understanding of the insurance (Churchill et al. 2003).

This study focuses on operational risks from external sources, namely potential losses due to unexpected disasters. In traffic accident, the operational risk data reference is the determination of GPD data distribution model, namely the risk of traffic accidents. There are three risk factors for traffic accidents: driver negligence, vehicle quality, and other risk factors.

3. Materials and Methods

3.1. Materials

This study uses secondary data obtained from the website www.bps.go.id and downloaded in the dynamic table section with the title Number of Accidents and Accident Material Losses within a period of 10 years, namely 2008 to 2017. The data taken are data value of losses and data on accidents that occurred in Indonesia during the period mentioned.

3.2. Methods

The research method used includes the following stages and formulations.

3.2.1. Maximum Entropy Bootstrapping (MEBoot)

Bootstrap was first introduced as a data resample method by Efron in 1982. Then Vinod (2009) in the journal entitled Maximum Entropy Bootstrap for Time Series in 2009 has developed bootstrapping based on the principle of maximum entropy and is commonly referred to as (MEBoot). MEBoot is essentially a method for deriving robust estimates of standard errors and confidence intervals for estimating proportions, means, medians, odds ratios, correlation coefficients or regression coefficients.

MEBoot carried out in this research is assisted by the R software, Package Meboot to make it easier to manage the data needed. The command function used is meboot (x, reps, trim=list (trim=0.10,), reachbnd=TRUE, expand.sd=FALSE, force.clt=TRUE, scl.adjustment=TRUE).

3.2.2. Selection of Threshold Value

Thresold is the initial value at the tail of the distribution that meets the distribution of extreme values. The selection of the threshold value is basically looking for an optimal balance in order to get the model error and parameter error to a minimum. One method to determine the threshold value is the percentage method. Determination of the threshold value with the percentage method is more practical and easier to apply.

In this study, the threshold value selection method used was the percentage method, due to the practical reasons mentioned above. Based on extensive simulation studies, Chavez-Demoulin (1999) recommends selecting a threshold value such that the data above the threshold is approximately 10% of the total data.

The amount of extreme data is obtained using the equation:

Lots of extreme data is obtained used this equations:

$$m = 10\% \times n \tag{1}$$

where m is lots of extreme data and n is lots of data total observed. Then threshold value u is obtained used this equation

$$u = m + 1 \tag{2}$$

3.2.3. Identification of Extreme Values

Identification of extreme values of loss data can be done by two methods. The first is the block maxima method, which is the traditional method used to analyze seasonal data. Each block period is determined by the maximum loss. Second, the Peaks Over Threshold (POT) method uses data more efficiently by identifying extreme values that are above a maximum loss value or a certain threshold value (Hubert, 2012 and Chavez-Demoulin, et al. 2003). In this study, choose the Peaks Over Threshold (POT) method in determining extreme values.

3.2.3.1. Peaks over threshold (POT)

Peaks over threshold (POT) identifies extreme values by setting a certain threshold value and ignoring the time of occurrence. Extreme values are data that are above the threshold value. Later this extreme value will be modeled for distribution. The POT method applies the Pickland Dalkema-DeHann theorem which states that the higher the threshold, the distribution for data above the threshold will follow the generalized Pareto distribution (Kang and Song, 2017 and Matthias et al. 2007). The assumption of data above the threshold that follows the GPD is obtained by looking at the tail distribution of the data away from the approximation line. The tail distribution of large data or heavy tails is seen by making a QQ-Plot on data above the threshold.

3.2.3.2. Generalized Pareto Distribution

Generalized Pareto Distribution (GPD) is defined as the distribution limit of scaled excesses above the threshold value. Suppose is a random variable of daily losses with 2 GPD parameters, the GPD distribution function of is as follows (Kang and Song, 2017 and Matthias et al. 2007).

$$g_{\xi,\beta}(x) = \begin{cases} \frac{1}{\beta}\left(1 + \frac{\xi}{\beta}x\right)^{-1-\frac{1}{\xi}}, \xi \neq 0 \\ \frac{1}{\beta}\, exp\left(-\frac{x}{\beta}\right) \quad , \quad \xi = 0 \end{cases} \tag{3}$$

Where, if $\xi > 0$ then $\beta > 0$; $x \geq 0$. If $\xi < 0$ then $0 \leq x \leq -\frac{\beta}{\xi}$.

ξ: shape parameter and β: scale parameter.

3.2.3.3. GPD Distribution Suitability Test

Distribution testing can be done using the Kolmogorov-Smirnov test. This test is done by adjusting the sample distribution function (empirical) with a certain theoretical distribution. According to Frank and Massey (1951), to get a conclusion, compare $D_{nominal}$ with $D_{1-\alpha}$ on the Kolmogorov-Smirnov table with a significance level (α). Reject H_0 if $D_{nominal} > D_{1-\alpha}$.

In this study, the process of testing the suitability of the GPD distribution for extreme data taken above the threshold value was carried out using the EasyFit software. The package or command used is Tools Goodness of Fit which takes the results of the distribution suitability test with the Kolmogorov-Smirnov test.

3.2.3.4. GPD Parameter Estimation

Davidson (1984) and Smith (1985) have discussed the Maximum Likelihood Estimation for estimating GPD parameters. The parameter estimation formula is obtained using the Maximum Likelihood Estimation (MLE) method as follows:

- Shape parameter

$$\hat{\xi} = \frac{n^2 s - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i - n\sum_{i=1}^{n} x_i} \tag{3}$$

where ξ: shape parameter, n: lots of extreme data, s: standard deviation of extreme data, and x_i: extreme data on index-i.

- Scale parameter

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

where β: scale parameter, n: lots of extreme data, and x_i: extreme data on index-i.

3.2.4. Research Steps

This research was carried out in several steps, as follows: 1) Resample data with MEBoot assisted by R software in accordance with the available packages; 2) Perform extreme data collection with equation (1); 2) Determine the threshold value with equation (2); 3) Testing extreme data using the Kolmogorov-Smirnov test on GPD assisted by

Easyfit software; 4) Calculating the estimation of GPD parameters with equation (4) for shape parameters and equation (5) for scale parameters.

4. Results and Discussion

4.1. Data Characteristics

The data used in this study is the value of traffic accident losses in Indonesia in 2008-2017 based on the risk factors; driver negligence, vehicle quality, andother risk factors. The data are briefly presented in Figure 1, Figure 2, and Figure 3.

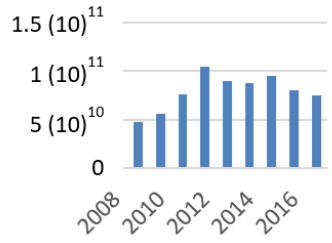

Figure 1. Traffic Accident Loss Data Risk of Driver Negligence.

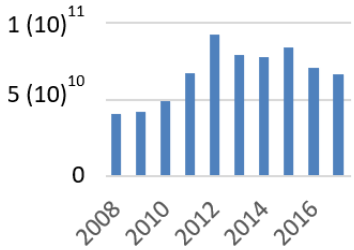

Figure 2. Traffic Accident Loss Data Vehicle Quality Risk.

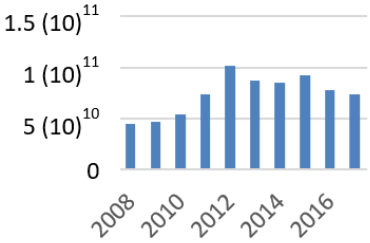

Figure 3. Traffic Accident Loss Data Other Risk Factors.

Traffic accident loss data for each risk shows characteristics that are not in accordance with the assumptions needed to identify extreme values with Peaks Over Threshold, so that in the next stage the data resample process is carried out with maximum entropy bootstrapping (MEBoot).

4.2. Maximum Entropy Bootstraping (MEBoot) Processing of Loss Data

Loss data is processed by the MEBoot process assisted by Software R. Then the threshold is taken with a percentage of 10% and a lot of extreme data above the threshold value. The summary results of the processed data are given in Table 1.

Table 1. MEBoot Data Loss Summary

Event Type	Resample	Lots of Data	Extreme Data	Threshold (IDR)
Rider Negligence	120	1200	120	IDR 100,666,312,262
Vehicle Quality	120	1200	120	IDR 88,975,241,430
Other Accident Factors	120	1200	120	IDR 7,104,331,775

The MEBoot summary results in Table 1 show the amount of traffic accident loss data to be as much as 1200 data. The MEBoot data was taken because it was in accordance with the assumption of a large data tail distribution. The tail distribution of large data is seen based on the QQ-Plot results against the MEBoot data. This QQ-Plot is used to see the suitability of the MEBoot data with the nature of the extreme data which has a large data tail as an indication of the extreme data with Generalized Pareto Distribution (GPD) distribution. QQ-Plot is done with the help of software R: QQ-Plot Package. The following are the QQ-Plot results from the MEBoot data which can be seen in Figure 4, Figure 5, and Figure 6.

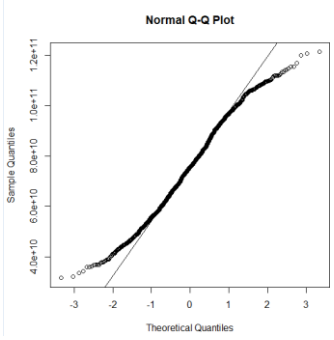

Figure 4. QQ-Plot of MEBoot Result Data Type Risk of Rider Negligence

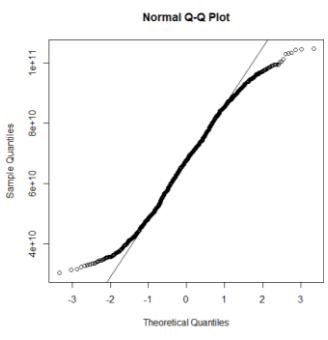

Figure 5. QQ-Plot of MEBoot Result Data Type Risk of Vehicle Quality

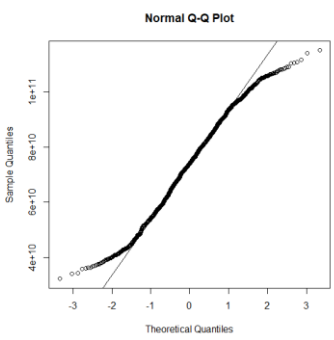

Figure 6. QQ-Plot of MEBoot Result Data Type Risk of Other Accident

Figure 4, Figure 5, and Figure 6 show results that are in accordance with the desired assumption, namely the tail of the data is large or away from the approximation of the normal line. This assumption results in an interpretation that the extreme data will fit the GPD distribution.

4.3. Kolmogorov-Smirnov Test on Extreme Data against GPD

Extreme data that assumes GPD is tested for conformity with the Kolmogorov-Smirnov test assisted by Easyfit software. The results of the suitability test are shown in Figure 7, Figure 8, and Figure 9.

Goodness of Fit - Details [hide]

Gen. Pareto [#24]

Kolmogorov-Smirnov

Sample Size	120		
Statistic	0.03719		
P-Value	0.9943		
Rank	1		
α	0.2	0.1	0.05
Critical Value	0.09795	0.11164	0.12397
Reject?	No	No	No

Figure 7. Kolmogorov-Smirnov Test Results Extreme Data from the Risk of Rider Negligence with GPD

Goodness of Fit - Details [hide]

Gen. Pareto [#24]

Kolmogorov-Smirnov

Sample Size	120		
Statistic	0.05323		
P-Value	0.86769		
Rank	1		
α	0.2	0.1	0.05
Critical Value	0.09795	0.11164	0.12397
Reject?	No	No	No

Figure 8. Kolmogorov-Smirnov Test Results Extreme Data from the Risk of Vehicle Quality with GPD

Goodness of Fit - Details [hide]

Gen. Pareto [#24]

Kolmogorov-Smirnov

Sample Size	120		
Statistic	0.05132		
P-Value	0.89383		
Rank	1		
α	0.2	0.1	0.05
Critical Value	0.09795	0.11164	0.12397
Reject?	No	No	No

Figure 9. Kolmogorov-Smirnov Test Results Extreme Data from the Risk of Other Accident Factors with GPD

From the results of the Kolmogorov-Smirnov test in Figure 7, Figure 8, Figure 9 it can be concluded that the extreme data is in accordance with the GPD distribution because there is no rejection or hypothesis of the assumption of the GPD distributed data being accepted, so that it can be continued to estimate the parameters.

4.4. Estimation of GPD Parameters

The calculation of GPD parameter approximate require deviation standard (s), lots of extreme data (n), and sum of extreme data values $(\sum_{i=1}^{n} x_i)$ that taken from descriptive statistics extreme data.

Table 2. Descriptive Statistics Extreme Data

	Rider Negligence	Vehicle Quality	Other Accident Factors
Data	120	120	120
Mean	$1.06245(10)^{11}$	94582033001	$1.02985(10)^{11}$
Standard Deviation	3982109145	4060705587	4369527320
Sample Variance	$1.58572(10)^{19}$	$1.64893(10)^{19}$	$1.90928(10)^{19}$
Kurtosis	-0.680525341	0.779321773	-0.351506991
Skewness	0.542142389	0.839064183	0.686936715
Minimum	$1.00727(10)^{11}$	89125793911	97164971213
Maximum	$1.15519(10)^{11}$	$1.09502(10)^{11}$	$1.14958(10)^{11}$
Sum	$1.27495(10)^{13}$	$1.13498(10)^{13}$	$1.23582(10)^{13}$

Based on Table 2 obtained $n = 123$, $s = 54110542295.94$, and $\sum_{i=1}^{n} x_i = 14601957364231$. Then, obtained the results of shape parameter and scale parameter approximation used formula (3) and formula (4) on the following Table.

Table 3. Parameter Estimation Results for Extreme Data of Each Risk.

	Rider Negligence	Vehicle Quality	Other Accident Factors
$\hat{\xi}$	−0,029391782	−0,034890522	−0,034381579
$\hat{\beta}$	106245476096.03	94582033001.483	102985266233.95

The results of the two GPD parameters showed the distribution function with $\xi < 0$ then x that bounded of $0 \leq x \leq -\frac{\beta}{\xi}$. For example, for the risk of driver negligence, the upper limit value of x is $-\frac{\beta}{\xi} = -\frac{1946237286}{-0.113403547} = 17162049436.6654$. That upper limit value of x fit to the maximum values of extreme data in Table 3 because the

maximum values of extreme data not pass through from upper limit value. So that a Generalized Pareto Distribution can be determined that is appropriate for each risk from this traffic accident, namely:

(a) Generalized Pareto Distribution Rider Negligence:

$$g(x) = 9.4121654563\,(10^{-12})\,(1-(2.7664\,(10^{-13}))\,x)^{33.0231} \quad (6)$$

(b) Generalized Pareto Distribution Vehicle Quality Risk:

$$g(x) = 1.0572832579\,(10^{-11})\,(1-(3{,}68891(10^{-13}))\,x)^{27.6610} \quad (7)$$

(c) Generalized Pareto Distribution Other Accident Factors Risk:

$$g(x) = 9.710126861\,(10^{-12})\,(1-(3{,}3384(10^{-13}))\,x)^{28.08534} \quad (8)$$

5. Conclusion

Based on the results of data processing the value of traffic accident losses in Indonesia using the Extreme Value Theory method, it was found that the Generalized Pareto Distribution parameter has the same characteristics for each type of risk analyzed, namely the form parameter $\xi < 0$. Based on the parameter estimation, the GPD distribution is obtained in equation (6), (7), and (8) of each risk which is expected to be useful for related parties in analyzing the number of traffic accident losses in the next period to consider steps that can be taken to reduce losses due to traffic accidents.

References

Baran, J. and Witzany, J. (2011). *A Comparison of EVT and Standard VaR Estimations*. SSRN eJournals. Czech: Science Foundation grant no. 402/09/0732

Chavez-Demoulin, V. (1999). *Two problems in environmental statistics: Capture-recaptureanalysis and smooth extremal models*. Ph.D. thesis. Department of Mathematics, Swiss Federal Institute of Technology, Lausanne.

Chavez-Demoulin, V., Davison, A., and McNeil, A. (2003). A point process approach to value-at-risk estimation. *National Centre of Competence in Research Financial Valuation and Risk Management*, *134*, 1-30.

Davison, A. C. (1984). *Modelling Excesses Over High Thresholds, with an Application, in Statistical Extremes and Applications*, ed. J.Tiago de Oliveira , Dordrecht : D.Reidel. 461-482.

Efron, B. (1982). *The Jacknife, the Bootstrap and Other Resampling Plans*. Stanford, California: National Science Foundation, Division of Biostatistics.

Frank, J. and Massey, Jr. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association. Vol. 46*(253), 68-78.

Gourier, E. Abbate, D. and Farkas, W. (2009). Operational Risk Quantivication using Extreme Value Theory and Copulas: from Theory to Practice. *The Journal of Operational Risk, Vol. 4*(3).

Hubbert, S. (2012). *Essential Mathematics for Market Risk Management*. Great Britain (UK): John Wiley & Sons Ltd., Publication.

Kang, S. and Song, J. (2017). Parameter and Quantile Estimation for the Generalized Pareto Distribution In Peaks Over Threshold Framework. *Journal of the Korean Statistical Society*, *Vol. 46*(4), 487-501.

Matthias, D., Paul, E., and Lambrigger, D. D. (2007). The Quantitative Modeling of Operational Risk : Between G and H and EVT. *Astin Bulletin, Vol 37*(2), 2007, 265-291.

Mohammed, A., Ambak, K., Mosa, A., and Syamsunur, D. (2019). A Review of the Traffic Accidents and Related Practices Worldwide. *The Open Transportation Journal, 13,* 65-83.

Smith, R. (1985). Maximum Likelihood Estimation in A Class of Nonregularcases. *Journal of Biometrika,72*(1), 67–90.

Vinod, H. D. and Lacalle, L. J. (2009). Maximum Entropy Bootstrap for Time Series: The MEBoot R Package. *Journal of Statistical Software. 29*(5), 1-19.

Yustika, D. W. (2011). Estimating Total Claim Size in the Auto Insurance Industry: a Comparison between Tweedie and Zero-Adjusted Inverse Gaussian Distribution. *Journal of Brazilian Administration Review*,*8*(1), 37-47.