Application of Peak Over Threshold Approach to Model Extreme Motor Insurance Claims: A Case Study

P.A.D.A.N. Appuhamy[1], N.K. Borelessa[1], E.M.P. Ekanayake[1]

[1]Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka

ABSTRACT

Prior to the economic recession in Sri Lanka, the motor insurance business grew significantly due to the excessive importation of vehicles. More vehicles on the road and reckless driving increase the risk of extreme claims, which creates a negative impact on the industry. In order to mitigate this issue, researchers attempted to model extreme claims and thereby to provide information for better management of business. The objective of this study is to identify the best fit model for tail of the claim distribution based on data obtained from a pioneer insurer in Sri Lanka from July to December of 2021. The Peak Over Threshold approach of the Extreme Value Theory was applied to model the extreme claims. The claims at 20 percentiles between 79% and 98% were considered as tentative thresholds and the excessive amounts over each of these thresholds were modeled separately as Generalized Pareto Distributions (GPDs) using four different parameter estimation methods. Then the Mean Squared Error (MSE) at each threshold for each parameter estimation method was examined to compare their performances. The threshold and the parameter estimation method with the minimum MSE were selected as their optimum values while identifying the GPD fitted as the best model. The Bootstrap goodness of fit measured the validity of modelling. The extent of claims varied from Rs. 2167.00 to 193,065.00 during the study period with a positive skewness of 2.45 and leptokurtic, which confirmed the existence of a heavy tailed distribution for claims. The best fitted model was the GPD with the shape and scale of 1.02 and 92.09 respectively, which was attained at the optimal threshold of 91st percentile using the Biased Probability Weighted Moment method. The information on the tail helps review existing strategies for the better management of risk due to such extreme claims in future.

KEYWORDS: *Claims, Generalized Pareto Distribution, Insurance, Percentile, Threshold.*

1 INTRODUCTION

With the rapid development of the automobile technology and the fast-paced life styles, travelling by motor vehicles has become an indispensable part of life, but travelling on the road comes with its own share of risks. Due to surge in unpredictable circumstances in day-to-day life, people have been in search of compensation and protection of their vehicles to avoid or minimize the financial risk associated with it. As a remedial measure, the motor insurance business got established under the non-life insurance category. According to the rules and regulations imposed by the Government of Sri Lanka, it is mandatory for every vehicle to possess at least a third-part policy while driving on roads without which it is considered as a serious offence. Over the past few decades, the motor insurance business has grown significantly in Sri Lanka due to the excessive importation of vehicles while becoming one of the most profitable businesses over time.

Non-insurance claims are usually known to be heavy tailed (Beirlant et al., 2004). Extreme motor insurance claims are rare in frequency but there is no assurance that it would not occur in any given instance due to uncertainties embedded in everyday life. Some unexpected events make the motor claim sizes enormous and cause negative consequences on a company's finances which could even lead to its bankruptcy. Many of such incidents are caused by the failures in risk management systems and lack of awareness in implementing amendments to existing strategies for accommodating current demands and trends in a country (Gaio et al., 2015). This urges insurance companies to look for new policies and strategies to handle financial loss due to such extreme claims. For this purpose, knowledge on the

https://doi.org/10.54389/QMCS8830

distribution of extreme claims that occur in the right tail area is required by actuaries to determine appropriate level of critical estimates such as insurance premiums, reserves, re-insurance and thereby to establish proper risk management systems. This stimulated researchers to engage in modelling the tail rather than the body of claim distributions. For many decades, the bulk and the tail data were modeled together using Gamma, Weibull, Exponential and Log-normal distributions. These models seemed to be appropriate for small and moderate claim sizes. However, it was highlighted that these distributions largely overestimate or underestimate the tail probabilities (Lee, 2012).

Accurate modeling of the tail of the claim distribution and separation of extreme claims from non-extreme data are challenging. However, it is an important area of research to assure a profitable non-life insurance business in future.

Having identified the importance of modeling the extreme data, Hogg & Klugman (1984) attempted to fit the tail of loss distribution as a truncated Pareto distribution using two parameter estimation methods viz. Maximum Likelihood Method and the Method of Moment, which is well-recognized and a pioneering research work carried out under the area of modeling the tail. However, according to a subsequent study by Boyd (1988), this distribution substantially underestimated the tail region of the loss distribution.

Several studies have shown that Extreme Value Theory (EVT) is a better approach to deal with the challenges in modeling the tail segment separately from the bulk of the loss severity distribution, as it provides a firm theoretical foundation in modelling extreme events (McNeil, 1997 ; Resnick, 1997). There are two types of approaches under EVT namely, Block Maxima (BM) and Peak-Over-Threshold (POT) method. In the first approach (BM), data are divided into blocks and the block maxima is modelled by fitting it into Generalized Extreme Value (GEV) distribution (Fisher & Tippett, 1928). In the other approach (POT), extreme data that exceed a sufficiently high threshold are considered for modelling. According to Pickands (1975), the cumulative distribution of the exceedances over sufficiently high threshold approaches the Generalized Pareto distribution (GPD). It was observed that the BM method discarded some extreme values that carried certain vital information as it considered only the maxima of each block for modelling, which was seen as a waste of data (Cole, 2001). On the other hand, generally, the POT approach is preferred and more efficient as it considered more extreme data for modelling than the BM, which is especially useful in case of dearth of information in the tail area (Reiss & Thomas, 2002). This approach is widely used in the non-life insurance sector to derive critical estimates (Wang et al., 2020).

In the POT approach, the selection of sufficiently high threshold is critical and challenging. Threshold must be sufficiently high to ensure the reliability of the GPD approximation. However, the high threshold reduces the sample size for modelling and increases the variance of the parameter estimation. This led to conclude that the choice of threshold should strike a balance between bias and variance (Scarrott & MacDonald, 2012). Over the past decades, several approaches have been introduced to select optimal threshold for POT. Scarrott & MacDonald (2012) categorized various techniques used for this purpose into graphical method, rule of thumb, probabilistic approach, computational approach, and mixture models. Dupuis (1999) introduced a threshold selection method based on robustness consideration. Moreover, Bayesian approach was used by Tancredi et al., (2006) who discussed the ways of incorporating threshold uncertainty in the inferences. Thompson et al., (2009) showed that these methods are computationally demanding and complicated to implement in practice. Zakaria et al., (2017) applied the POT method with two approaches to select threshold viz. the rule of thumb and graphical method. Further, Gharib et al., (2017) used the Mean Residual Life plot to select tentative set of thresholds from which the optimal was selected using the Square Error method. However, the use of graphical methods is subjective as it requires substantial expertise to interpret the plots like Mean Residual Life plot (Solari et al., 2017; Thompson, 2009). On the other hand, various methods have been introduced to automate the optimal threshold selection, for instance, Thompson et al., (2009) and years later Solari et al., (2017) introduced a technique for automatic threshold selection based on the Anderson-Darling EDF-statistic and goodness of fit test, which estimated the uncertainty associated with threshold estimation lacking in graphical method. Though there are numerous techniques in literature to choose the optimal threshold for POT approach, Davison & Huser (2015) mentioned that threshold selection is a long standing issue that still remains unresolved.

In the model building process parameter estimation is an essential part. The choice of the parameter estimation methods depends mostly on the sample size available. Incorrect choice of the

parameter estimation method would seriously affect the inferences drawn. This makes researchers more cautious on parameter estimation methods under the POT framework because in many practical situations only a few observations are available for modelling the tail. The accurate estimation of the shape (ξ) and scale (σ) parameters of the GPD is as important as the optimal threshold selection. In literature, Maximum Likelihood Estimation method (MLE), the Probability Weighted Method (PWM), and the Method of Moments (MOM) have commonly been used for the estimation purpose (De Zea Bermudez & Kotz, 2010). MLE performs well when the sample size for estimation is large (Deidda & Puliga, 2009; Kang & Song, 2017). Hosking & Wallis (1987) pointed out that Method of Moment performs poorly when the shape parameter exceeds 1. The PWM method performs well for the shape parameter between $0 \leq \xi \leq 1$ and shows excellent performance when $\xi \leq 0.5$ (De Zea Bermudez & Kotz, 2010). Moreover, according to Castillo & Hadi (1997), PWM performs well when the sample size for modelling is small. Rydman (2018) applied the Unbiased Probability Weighted Moment (UPWM) method for parameter estimation of GPD and found that UPWM method is more efficient when the number of exceedances over threshold is small. Zhao et al., (2019) compared the performances of Maximum Likelihood Estimation (MLE), L-Moment, Weighted Nonlinear Least Square Likelihood Moment (WNLLSM), and Weighted Nonlinear Least Square Moment (WNLSM) in estimating the parameters of GPD under the POT method. Kang & Song (2017) used six different parameter estimation methods in fitting GPD and revealed that non-linear least square based method outperforms others. More often, performance of the GPD parameter estimators depends on both the sample size and the value of the GPD shape parameter (Gharib et al., 2017; Kang & Song, 2017). This leads to assume that the methods available for parameter estimations perform well in some situations but otherwise in some circumstances.

Moreover, it is evident from literature that, though the application of POT method is common in the field of non-life insurance, its application is very limited in a country like Sri Lanka especially for motor insurance claims. Therefore, this study focused on the application of the POT method of extreme value theory together with the best parameter estimation method to model tail and estimate the parameters of GPD to get better understanding on the behavior of tail in motor claims distributions. More specifically, this study identify the optimal threshold out of several sufficiently high tentative thresholds required for POT approach based on the mean squared errors of percentiles estimated through fitted GPDs for extreme claims with four different parameter estimation methods and thereby identify the best fit distribution for tail of the claim distribution.

2 METHODOLOGY

This study aimed at comparing the performances of four different parameter estimation methods at sufficiently high tentative thresholds to identify the best fit model for extreme motor claims under the POT framework. The motor claims received by a pioneering insurance company in Sri Lanka from July to December of 2021 were used for this study. At first, descriptive statistics were examined to get an idea on how the claims were distributed over the period of study. Then, the POT approach of the EVT was applied for model fitting as it incorporates more extreme data than the BM method. Under the POT framework, excess claims over a sufficiently high threshold can be approximated by the Generalized Pareto distribution. Therefore, the next step was to find the optimal threshold which separated extreme claims from the bulk. For this purpose, the study initially considered claims at 20 different percentiles ranging from 79% to 98% as tentative thresholds, thus covering the range of percentiles proposed for threshold in literature under the rule of thumb method (Scarrott & MacDonald, 2012). Since a sufficiently high threshold ensures a better fit of GPD, the process was initiated by taking the 98^{th} percentile as the first tentative threshold. Then the excess values above the 98^{th} percentile were modelled as the GPD and the parameters were estimated using the four different parameter estimation methods viz. Maximum Likelihood Estimation (MLE), Method of Moment (MOM), Unbiased Probability Weighted Moment (UPWM), and the Biased Probability Weighted Moment (BPWM). Subsequently, the accuracies of the fitted GPD models under the four different parameter estimation methods were evaluated separately by calculating the mean squared error statistics for which simulated data obtained from the fitted GPDs and the actual data at 10 different percentiles were used. Next, the same procedure was carried out by taking the 97^{th} percentile as the second tentative threshold, a percentile less than the previous threshold. Similarly, the process was continued until the minimum mean squared error for a particular threshold was

attained which balances the bias and variance involved in model fitting. Further, in order to ensure that there is no any other threshold which yield the minimum mean squared error, the process was executed up to the 79th percentile. The threshold and the parameter estimation method, which yielded the minimum mean squared error in fitting the GPD, were selected as the optimal threshold and the best parameter estimation method while the model fitted under these conditions was selected as the best fit for extreme motor insurance claims. The optimal threshold selection approach proposed in this study was simple and efficient compared to the one proposed by Thompson et al., (2009) in which 100 tentative thresholds between median and the 98th percentile of data were considered which required considerable computational time. Finally, the Bootstrap Goodness of fit test was conducted to validate that excess values over each threshold follow a GPD. The null hypothesis tested under this test is as follows.

H0: Motor insurance claims over a sufficiently high threshold follow a Generalized Pareto distribution with a positive shape parameter, ξ.

3 RESULTS AND DISCUSSION

It was revealed that the motor insurance claims were distributed with a median of Rs. 22,500.00, which means that 50% of the claims received during the study period were larger than Rs. 22,500.00. The minimum and the maximum claims received by the insurance company were Rs. 2167.00 and Rs. 193,065 respectively. Moreover, the claim distribution was leptokurtic as the kurtosis was greater than 3 and positively skewed with a skewness of 2.445. These properties indicate that the motor claim distribution is heavy tailed. The smallest (79th percentile) and the largest (98th percentile) tentative thresholds amounted to Rs.40,925.00 and Rs. 96,114.40 respectively while the number of claims above the largest and the smallest tentative thresholds were 11 and 98 respectively. That is, only 11 observations were available for modelling the tail when the threshold was set to Rs.96,114.40. The MSE of percentiles estimated through the GPD fitted to excess values over each of the 20 tentative thresholds with each parameter estimation method is presented in the following Figure 1, which enables the comparison of performances of the four parameter estimation methods with the change of sample size for tail. Moreover, Figure 1 displays the optimal threshold for POT approach which yielded the minimum MSE.

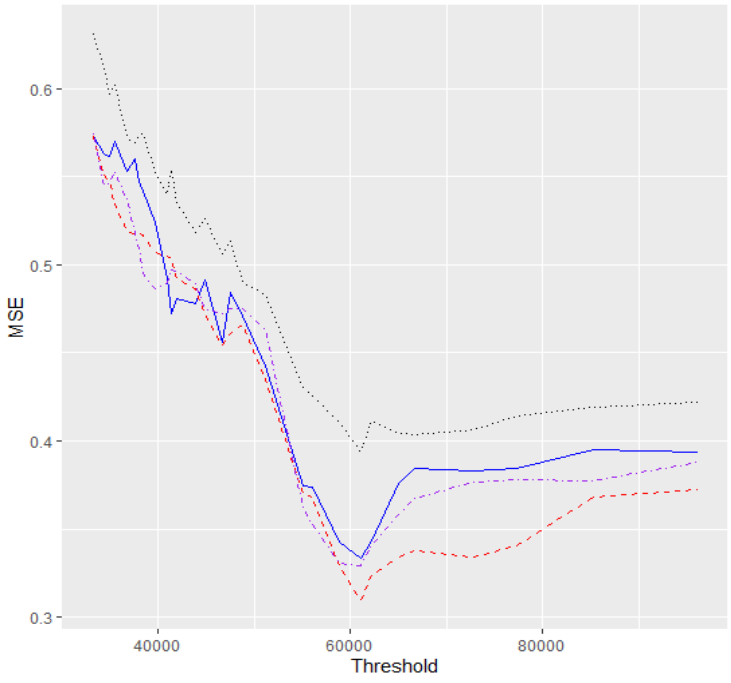

Figure 1. MSE of percentiles estimated through GPD fitted to the excess values over each of the 20 tentative thresholds with the four different parameter estimation methods.

According to Figure 1, it can be seen that the MSEs of GPD fitted with the Maximum Likelihood Estimation method are somewhat higher for all the 20 tentative thresholds compared to those fitted with the other three parameter estimation methods. This is because the MLE performs well when a large sample of data is available for modelling, which was not fulfilled here. The performance of Method of Moment and Unbiased Probability Weighted Moment methods were almost similar and accurate than that of MLE. In contrast, the model parameters estimated through the Biased Probability Weighted Moment method (BPWM) showed somewhat smaller MSEs for almost all tentative thresholds than those obtained from the other three methods. The minimum mean squared errors obtained from each parameter estimation method are summarized in Table 1, which conform well with the profiles illustrated in Figure 1. Further, it can also be seen from Figure 1 that all these four minimum values were recorded at the threshold of 91st percentile equivalent to Rs. 61,056.00. Therefore, out of the 20 tentative thresholds, the 91st percentile was selected as the optimal where the claims above 91st percentile were categorized as extreme and below were non-extreme claims.

Table 1. Minimum MSE of each of the four parameter estimation method

GPD Parameters estimation method	**Minimum MSE**
MLE	0.3934
MOM	0.3290
UPWM	0.3337
BPWM	0.3091

According to Table 1, the BPWM method possessed the minimum MSE compared to other three in respect of the estimation of percentiles through the fitted GPD at the 91st percentile. Moreover, in this study, the BPWM method could be selected as the best parameter estimation method to estimate parameters of GPD as it provided more accurate and reliable estimates especially when a dearth of information prevailed in the tail area. In addition, there were 47 motor claims greater than the optimal threshold of Rs. 61,056.00 which were received during the period of study. The excess claims over the optimal threshold can be best described by the Generalized Pareto distribution with the shape and scale parameters of 1.02 and 92.09 respectively. The positive value of the shape parameter implies that the motor claims distribution is heavy tailed. Moreover, the findings confirm the existing results in the literature that BPWM method is preferred for parameter estimation of GPD when the shape parameter is positive and less than or equal to 1 and the sample size for modeling is small. The p-value ($0.997 > 0.05$) of the Bootstrap goodness of fit test confirmed that excess values over the optimal threshold follow the GPD with a positive shape parameter.

4 CONCLUSION

The demand for motor insurance business has grown significantly with growing risk associated with the increasing number of vehicles on the road and as a financial security to compensate property damage due to unforeseen accidents in future. These unexpected events can sometimes be a huge burden to insurance companies when extreme claims are received from their policyholders. A prior knowledge on the occurrence of extreme claims is really important to introduce measures for a sustainable business. It could be concluded from the study that, the motor insurance claims received during the period of study follow a heavy tailed distribution. Moreover, the extreme motor claims could be best described by Generalized Pareto Distribution with positive shape parameter. The BPWM method would be the best to estimate parameters of the GPD when there is a dearth of information in the tail area. The threshold selection method considered in this study was much simpler and easier compared to some existing approaches in literature. The distribution helps to derive some critical tail estimates like extreme quantiles, by providing useful information to review existing strategies and to introduce timely changes for better management of such risks in future.

REFERENCE

Beirlant, J., Joossens, E. & Segers,J. (2004). Generalized Pareto fit to the society of actuaries' large claims database. *North American Actuarial Journal*,*8*(2),108-111. https://doi.org/10.1080/10920277.2004.10596140

Boyd, V. (1988). Fitting the truncated Pareto distribution to loss distributions. *Journal of the Staple Inn* Actuarial Society, 31, 151–158. https://doi.org/10.1017/S2049929900010291

Castillo, E., & Hadi, A. S. (1997). Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association*, *92*(440), 1609-1620. https://doi.org/10.1080/01621459.1997.10473683

Cole, S. (2001). An Introduction to Statistical Modelling of Extreme Values, Springer Series in Statistics. https://doi.org/10.1007/978-1-4471-3675-0

Davison, A., & Huser, R. (2015). Statistics of extremes. *Annual Review of Statistics and Its Application*, 2, 203–235. https://doi.org/10.1146/annurev-statistics-010814-020133

Deidda, R. & Puliga, M. (2009). Performances of some parameter estimators of the generalized pareto distribution over rounded-off samples. *Physics and Chemistry of the Earth*, *34*(10-12), 626– 634. DOI: 10.1016/j.pce.2008.12.002

De Zea Bermudez, P. & Kotz, S. (2010). Parameter estimation of the generalized Pareto distribution – Part II. *Journal of Statistical Planning and Inference*, *140*(6), 1374-1388. https://doi.org/10.1016/j.jspi.2008.11.020

Dupuis, D.J., (1999). Exceedances over high thresholds: a guide to threshold selection. *Extremes, 1* (3), 251–261. https://doi.org/10.1023/A:1009914915709

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24, 180–290. https://doi.org/10.1017/S0305004100015681

Gaio, L.E., Junior, T.P., Lima G. F. & Bonacim C.A.G. (2015). Value-at-Risk in Times of Crisis: An Analysis of the Brazilian Market. *African Journal of Business Managemen*, *9*(5),223-232. https://doi.org/10.5897/AJBM2015.7695

Gharib, A., Davies, E.G.R., Goss, G.G., & Faramarzi, M.(2017). Assessment of the Combined Effects of Threshold Selection and Parameter Estimation of Generalized Pareto Distribution with Applications to Flood Frequency Analysis. *Water*, 9, 692. https://doi.org/10.3390/w9090692

Hogg, R., & Klugman, S. (1984). *Loss Distributions*. John Wiley & Sons. https://doi.org/10.1017/S0515036100004955

Hosking, J. R., and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3), 339–349. https://doi.org/10.2307/1269343

Kang, S. & Song, J.(2017). Parameter and quantile estimation for the Generalized Pareto distribution in peak over threshold framework. *Journal of the Korean Statistical Society*,46,487-501.

Lee,W.C.(2012).Fitting the Generalized Pareto distribution to commercial fire loss severity:evidence from Taiwan.*The Journal of Risk*,*14*(3),63-80. DOI: 10.21314/JOR.2012.244

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin*, *27*(1), 117–137.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3, 119–131. https://doi.org/10.1214/aos/1176343003

Reiss, R. D & Thomas, M. (2002). *Statistical Analysis of Extreme Values with Application to Insurance, Finance, Hydrology and other fields*, Birkhäuser,3rd Edition.

Resnick, S. I. (1997). Discussion of the Danish data on large fire insurance losses. *ASTIN Bulletin 27*(1), 139–151.

Rydman,M.(2018). Application of the Peak-Over-Threshold method on Insurance data. Project report.Uppsala University.

Scarrott, C., & MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *Revstat-Statistical Journal*,*10*(1),33-60. https://doi.org/10.57805/revstat.v10i1.110

Solari, S., Eguen, M., Polo, M. J., & Losada, M. A. (2017). Peaks Over Threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value. *Water Resources Research*, 10.1002/2016WR019426.

Tancredi, A., Anderson, C., O'Hagan, A., (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes*, 9, 86–106. DOI 10.1007/s10687-006-0009-8

Thompson,P.,Cai,Y.,Reeve,D.,& Stander, J. (2009). Automated threshold selection method for extreme wave analysis. *Coastal Engineering*, 56, 1013-1021. DOI: 10.1016/j.coastaleng.2009.06.003

Wang,Y., Haff, I. H. & Huseby, H. (2020). Modelling extreme claims via composite models and threshold selection methods. *Insurance: Mathematics and Economics*, 91, 257-268. DOI: 10.1016/j.insmatheco.2020.02.009

Zakaria, R., Radi, N.F.A. & Satari, S.Z. (2017). Extraction method of extreme rainfall data. *Journal of Physics: Conference Series*, 890. DOI: 10.1088/1742-6596/890/1/012154

Zhao, X., Zhang, Z., Cheng, W. & Zhang, P. (2019). A New Parameter Estimator for the Generalized Pareto Distribution under the Peaks over Threshold Framework. *Mathematics*,*7*(5), 406. https://doi.org/10.3390/math7050406