Estimation of the Catastrophic Risk using Mixture Models

Zakiatul Wildani[1*]

***Abstract*— Indonesia is one of the countries in the world that is susceptible to various types of natural disasters such as earthquakes, floods, and other natural disasters. These events do not occur often, yet can result massive financial loss. This risk of loss is referred to be catastrophic risk since it impacts not only the individual but also the government while also posing a threat to insurance companies if they do not have sufficient resources to make a payment of claims. However, due to the complexity and uncertainty of natural hazards, measuring this risk is quite challenging. This study employs a method for estimating catastrophic risk in Indonesia based on the Value-at-Risk (VaR) of total loss from natural disasters. A key issue for estimating VaR is to fit an appropriate distribution. Extreme value distribution, such as Generalized Pareto Distribution (GPD) has been used to assess the tail behavior of extreme loss. However, this distribution provides no information about the central behavior that may affect the estimation of the model parameter in GPD. Therefore, this paper employs mixture models that combine the parametric form of loss distributions such as gamma, weibull, and lognormal distribution with GPD. The results reveal that VaR calculations differ significantly depending on the mixture model and confidence level used. In addition, the lognormal-GPD model is chosen as the best model that fits data best with the highest value of log-likelihood.**

***Keywords*— Catastrophic risk, Mixture models, Value-at-risk**

I. INTRODUCTION

Being located along the pacific ring of fire, Indonesia is known as one of the world's most disaster-prone countries along with the United States, China and Philippines. Landslides, earthquakes, volcanic eruptions even the potential threat of tsunami seem like an endless chain of event to hit Indonesia. Over the last few years, earthquakes and tsunamis have brought the most serious threats to people lives and property in Indonesia. The largest magnitude earthquakes in Indonesia's history around nine scales richter triggered by tsunami with waves growing as high as 30 meters occurred in Aceh, December 2004. This massive disaster caused more than 227,000 deaths in total making it the deadliest natural disaster that has occurred in the 21st century. The disaster gained international attention with offers of aid and assistance coming in from other parts of Indonesia as well as other countries. Moreover, the latest tsunami in 2018 claimed over 2,000 lives on the island of Sulawesi between September 28th and October 1st, 2018. Even worse, 5,000 people still missing long after the search for survivors was called off. In addition, a series of earthquakes shook the northern part of Lombok in the same year killed almost 300 people and injuring hundreds more. Thousands of properties were damaged and 150,000 people were left homeless. The estimated total losses are around million dollars.

Damage caused by such natural disasters is accompanied by a significant amount of financial loss. Throughout 2018, Indonesia's National Disaster Mitigation Agency (BNPB stated that the expected financial loss would be more than 2.9 billion dollars from disaster in Sulawesi and Lombok events. Risks associated with such a large-scales disaster that not only caused substantial financial loss but also cause considerable damage to the systems and infrastructures on which local communities and economies rely on are known as Catastrophic risk. This risk is crucial to be assessed as it helps to increase understanding of risk for the more extreme catastrophes. It supports the insurance industry in making decisions, which include the pricing of individual contracts, and the overall regulation of the industry [1]. However, due to the complexity and uncertainty of natural and technological hazards, assessing this risk is quite challenging.

The primary goal of this paper is to calculate the catastrophic risk based on Value-at-risk (VaR). That is, determining the maximum amount of loss with a certain confidence level. This risk measure has become a worldwide benchmark concerning risk estimation due to its simplicity [2]. Fitting a suitable distribution is a critical aspect of estimating VaR. Almost all of financial data are not normal and occasionally exhibits heavy-tailed behavior. The well-known distribution, such as normal distribution and student-t distribution failed to capture the heavy-tailed data. Then, [3] proposed the estimation of risk measures given that the distributions of losses are heavy-tailed called Extreme Value Theory (EVT).

Extreme Value Theory has been used widely in financial and insurance field (see [4] and [5]). [4] proposed a method combining GARCH model to estimate volatility and EVT for estimating the tail to estimate VaR and related risk measures of a heteroscedastic financial return series. They showed that their procedure gives better 1-day estimates of VaR than methods that ignore the heavy tails. Then, [5] presented a method to determine the type of the asymptotic distribution for the extreme changes in stock prices, foreign exchange rates and interest rates based on Generalized Pareto Distribution (GPD) and Generalized Extreme Value (GEV) distribution.

However, since EVT only could be used to estimate the tail behavior of distributions, the central behavior are neglected despite the fact that it may alter model parameter estimation. Therefore, some of the literature used extreme value mixture models. The mixture models assume that the

[1] Departement of Business Statistics, Institut Teknologi Sepuluh Nopember, Kampus ITS Sukolilo, Surabaya, 60111, Indonesia. E-mail: zakia@its.ac.id

random variables below the threshold are drown from a loss distribution, while those above the threshold are drawn from a heavy-tailed distribution. [6] developed a mixture model that incorporates a parametric form to analyze the central behavior and GPD for the tail of the distributions. The parametric forms include Normal, Gamma, Weibull and Beta distribution. The mixture models have been used to analyze extreme data related to natural disaster as presented by [6], [7] and [8].

In mixture models, the distribution combination can take many different forms. In this study, we only consider mixture model of gamma-GPD model, weibull-GPD model and lognormal-GPD model in order to modelling the financial losses as proposed by [6] and [7]. Then, we calculate the Value-at-risk (VaR) for various confidence levels.

II. METHOD

As previously stated, the mixture model has many possible combinations. However, in this case, we only consider three combinations, gamma-GPD, weibull-GPD and lognormal-GPD. Before proceeding to the analysis, this section provides an explanation about the theoretical background of the mixture model.

2.1 Generalized Pareto Distribution

Let x denote total financial damage from natural disaster and $Y = x - u$ denote the exeedance over a certain threshold u. Therefore, [9] and [10] showed that showed that the limiting distribution of Y can be modeled by the Generalized Pareto Distribution (GPD). The cumulative probability function of GPD is given by

$$G(x|\xi,\sigma,u) = \begin{cases} 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}, if\ \xi \neq 0 \\ 1 - exp\left(-\left(\frac{x-u}{\sigma}\right)\right),\ if\ \xi = 0 \end{cases} \quad (1)$$

where ξ dan σ are the shape and scale parameters respectively. The uncertainty is involved in the choice of threshold, u. We can choose the threshold u by looking at the mean excess plot as proposed by [11]. There are some drawbacks regarding choosing the right threshold, for instance precision and bias. The GPD model only considers excesses over the threshold, but it does not provide any information below the threshold. There are many possibilities for handling both parts (below and above threshold) and for combining them. One of possibilities to handle both below and above threshold are by using a mixture model that combines a parametric form for the center or below threshold such as gamma, weibull and lognormal distribution and GPD for the tail or above threshold. By using this mixture models, inference will take into account all observations.

2.2 Mixture models

Th mixture models assume that all observation under the threshold u come from a parametric distribution denoted by $H(.|\theta_1)$ whereas those above threshold come from a heavy-tailed distribution that is GPD $G(x|\xi,\sigma,u)$. Therefore, the distribution function of F can be written as

$$F(x|\eta,\xi,\sigma,u) = \begin{cases} H(x|\eta), & if\ x < u \\ H(u|\eta) + \big(1 - H(u|\eta)\big)G(x|\xi,\sigma,u), & if\ x \geq u \end{cases} \quad (2)$$

For a sample size n, $x = (x_1, \ldots, x_n)$ and we assume the parameter vector as $\boldsymbol{\theta} = (\eta, \xi, \sigma, u)$, the likelihood function from (2) is

$$\mathrm{L}(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{x_i<u} h(x|\eta) \prod_{x_i \geq u} \big(1 - H(u|\eta)\big) \left(\frac{1}{\sigma}\left[1 + \frac{\xi(x_i-u)}{\sigma}\right]_+^{-(1+\xi)/\xi}\right) \quad (3)$$

for $\xi \neq 0$ and

$$\mathrm{L}(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{x_i<u} h(x|\eta) \prod_{x_i \geq u} \big(1 - H(u|\eta)\big) \left(\frac{1}{\sigma} exp\left[\frac{(x_i-u)}{\sigma}\right]\right) \quad (4)$$

For $\xi = 0$ where $h(x|\eta)$ is the density function of the loss distribution. Therefore, we can write the combination of the loss distribution GPD and the parametric forms, where in this case we consider gamma and weibull distribution as follows

2.2.1 Gamma distribution and GPD

The probability density function of Gamma distribution is given by

$$f_G(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp(-\beta x)\, x^{\alpha-1},\ x \geq 0 \quad (5)$$

where α is the shape parameter and β is the rate. Then, the cumulative distribution function can be written as

$$F_G(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha, \beta x),\ x \geq 0 \quad (6)$$

Based on (2) and (3) we can construct the distribution function and the likelihood mixture model as follows

$$F(x|\alpha,\beta,\xi,\sigma,u) = \begin{cases} F_G(x|\alpha,\beta), x < u \\ F_G(u|\alpha,\beta) + [1 - F_G(u|\alpha,\beta)]G(x|\xi,\sigma,u), x \geq u \end{cases}, \quad (7)$$

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \begin{cases} \prod_{x_i<u} f_G(x|\alpha,\beta) \prod_{x_i \geq u} \big(1 - F_G(x|\alpha,\beta)\big) \left(\frac{1}{\sigma}\left[1 + \frac{\xi(x_i-u)}{\sigma}\right]_+^{-(1+\xi)/\xi}\right) \\ \prod_{x_i<u} f_G(x|\alpha,\beta) \prod_{x_i \geq u} \big(1 - F_G(x|\alpha,\beta)\big) \left(\frac{1}{\sigma} exp\left[\frac{(x_i-u)}{\sigma}\right]\right), \end{cases} \quad (8)$$

where $\boldsymbol{\theta} = (\alpha, \beta, \xi, \sigma, u)$ is the parameter of the mixture model that estimated by bayesian.

2.2.2 Weibull distribution and GPD

Let λ be the scale parameter and γ be the shape parameter where $\lambda, \gamma > 0$, the density and probability function of the weibull distribution are given as follows

$$f_w(x|\lambda,\gamma) = \frac{\gamma}{\lambda}\left(\frac{x}{\lambda}\right)^{\gamma-1} \exp(-x/\lambda)^\gamma, x \geq 0 \quad (9)$$

and

$$F_w(x|\lambda,\gamma) = 1 - \exp(-x/\lambda)^\gamma, x \geq 0 \quad (10)$$

Therefore, the distribution function and the likelihood function of mixture model in (2) and (3) are rewritten as

$$F(x|\lambda,\gamma,\xi,\sigma,u) = \begin{cases} F_w(x|\lambda,\gamma), & x < u \\ F_w(u|\lambda,\gamma) + [1 - F_w(u|\lambda,\gamma)]G(x|\xi,\sigma,u), & x \geq u \end{cases}, \quad (11)$$

$$L(\boldsymbol{\theta};\boldsymbol{x}) = \begin{cases} \prod_{x_i<u} f_w(x|\lambda,\gamma) \prod_{x_i\geq u}(1-F_w(x|\lambda,\gamma)) \left(\frac{1}{\sigma}\left[1+\frac{\xi(x_i-u)}{\sigma}\right]_+^{-(1+\xi)/\xi}\right) \\ \prod_{x_i<u} f_w(x|\lambda,\gamma) \prod_{x_i\geq u}(1-F_w(x|\lambda,\gamma)) \left(\frac{1}{\sigma} exp\left[\frac{(x_i-u)}{\sigma}\right]\right), \end{cases} \quad (12)$$

Parameter $\boldsymbol{\theta} = (\lambda, \gamma, \xi, \sigma, u)$ is then estimated by bayesian approach.

2.2.3 Lognormal distribution and GPD

Let x be positive random variable with lognormal distribution, the probability density function can be written as follows

$$f_{LN}(x|\mu,\sigma_l) = \frac{1}{x\sigma_l\sqrt{2\pi}} exp\left(-\frac{(ln\, x-\mu)^2}{2\sigma_l^2}\right), x>0 \quad (13)$$

where μ and σ_l is the location parameter and scale parameter respectively. Then, the cumulative distribution function (CDF) can be defined as

$$F_{LN}(x|\mu,\sigma_l) = \Phi\left(\frac{(ln\, x-\mu)}{\sigma_l}\right), x>0 \quad (14)$$

where Φ is the CDF of the standard normal distribution. Based on (13) and (14) the mixture model lognormal-GPD has distribution function as follows

$$F(x|\mu,\sigma_l,\xi,\sigma,u) = \begin{cases} F_{LN}(x|\mu,\sigma_l), & x<u \\ F_{LN}(u|\mu,\sigma_l) + [1-F_{LN}(u|\mu,\sigma_l)]G(x|\xi,\sigma,u), & x\geq u \end{cases} \quad (15)$$

and the likelihood function

$$L(\boldsymbol{\theta};\boldsymbol{x}) = \begin{cases} \prod_{x_i<u} f_{LN}(x|\mu,\sigma_l) \prod_{x_i\geq u}(1-F_{LN}(x|\mu,\sigma_l)) \left(\frac{1}{\sigma}\left[1+\frac{\xi(x_i-u)}{\sigma}\right]_+^{-(1+\xi)/\xi}\right) \\ \prod_{x_i<u} f_{LN}(x|\mu,\sigma_l) \prod_{x_i\geq u}(1-F_{LN}(x|\mu,\sigma_l)) \left(\frac{1}{\sigma} exp\left[\frac{(x_i-u)}{\sigma}\right]\right), \end{cases}$$

where the parameter $\boldsymbol{\theta} = (\mu, \sigma_l, \xi, \sigma, u)$.

2.3 Value-at-Risk

Value at risk (VaR) is one of the well-known risk measures in the financial or insurance field due to its simplicity. The definition of Value-at-risk is actually a maximum of losses that will not exceed at a certain confidence level and period. This kind of risk measure is also known as quantile of the distribution function. Let x is the loss observation and α is confidence level, then we can define VaR as

$$VaR(x,\alpha) = F_x^{-1}(\alpha), \quad (16)$$

where $F_x^{-1}(\alpha)$ is the quantile of given distribution. Therefore, we can say that in the worst-case scenario, the probability that losses will exceed l is equal to (1 - α).

III. RESULTS AND DISCUSSION

3.1 Descriptive statistics and preliminary results

The dataset consists of total losses (in million US Dollar) of different types of natural disasters in Indonesia for around 54 years from 1966 to 2018. The data include total loss of natural disasters such as earthquake, volcanic activity, landslide, etc. The dataset is obtained from https://www.emdat.be/emdat db/, the International Disaster Database website based in Belgium.

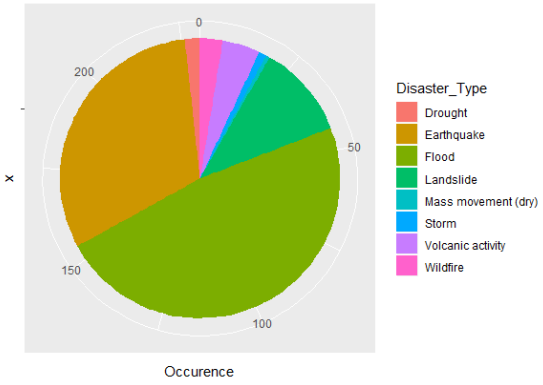

Figure 1. Natural disasters in Indonesia from period 1966-2018

Figure 1 depicts some major natural disasters that occurred in Indonesia from 1966 to 2018. Floods has the highest frequency of occurrence and then followed by earthquakes and landslides. High rainfall intensity throughout the year yields to high frequency of extreme floods [12]. For instance, extreme flood just hit East Nusa Tenggara in early April this year. Moreover, Meteorological, Climatological, And Geophysical Agency (BMKG) reported that the frequency of earthquake is increasing each year [13].

In order to understand the distribution of the loss data, histogram and Q-Q plot are also presented as follows

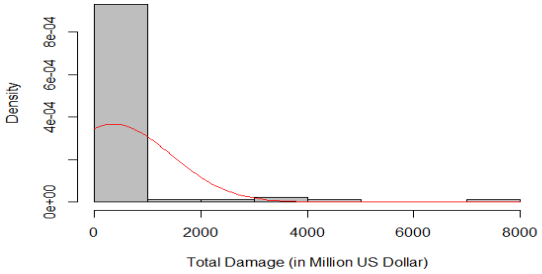

Figure 2. Histogram total loss (in million US dollars)

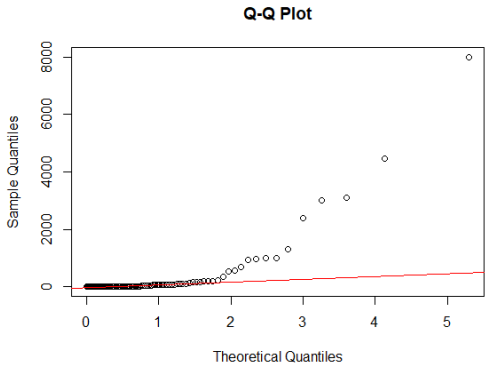

Figure 3. Q-Q Plot total loss (in million US dollars)

Figure 2 and 3 shows that the loss data may indicate skewed and heavy-tailed behavior as the histogram and QQ Plot are really far from normal. Therefore, we can use extreme value theory such as GPD to analyze the tail behavior whereas central behavior can be assessed using loss distribution such as gamma, weibull and lognormal.

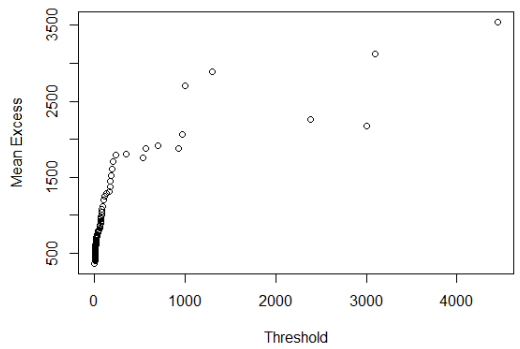

Figure 4. Mean excess plot

The mean excess plot is also given in Figure 4. This plot can be used to check whether the loss data came from a GPD model or not. The mean excess plot is roughly linear over the entire range of loss below the threshold and its upward slope, as seen in Figure 4. Then it change to gradually linear for the loss over the threshold, indicating that the data are from a GPD model with a positive shape parameter ξ [4]. As stated by [1], if the mean excess plot is close to linear for high values of the threshold, then there is no evidence against employing the GPD. Consequently, this plot also demonstrates that the loss data exhibits heavy-tailed behavior and GPD can be used to deal with this extreme behavior.

3.2 Models

The GPD model only can be used for analyzing the tail behavior of the loss data. Meanwhile, the central behavior is assessed by gamma, Weibull, and lognormal distribution (mixture models). This study fits the loss data with three mixture models, and the estimation of each model parameter is presented in Table 1.

TABLE 1.
ESTIMATION OF PARAMETERS IN THE MIXTURE MODELS

Model	Parameters				
	ξ	σ	u	α	β
Gamma-GPD	0.14	1789.14	928	0.26	1363.39
	ξ	σ	u	λ	γ
Weibull-GPD	0.15	1785.77	964.45	0.51	40.85
	ξ	σ	u	μ	σ_l
Lognormal-GPD	0.14	1789.71	936.61	2.81	2.19

Table 1 displays estimation parameters for the proposed mixture models. There are no significant differences in the estimation of GPD parameters in all models. The shape and scale parameters are aproximately around 0.15 and 1789, respectively, meanwhile the threshold is roughly 900. After model parameters have been estimated, the next step is examining and measuring the Value-at-Risk (VaR). VaR is generally defined as the maximum potential losses at given confidence levels. Table 2 summarizes the findings. These findings can aid an insurance firm or government in making risk-based decisions regarding catastrophic insurance coverage, reinsurance levels, and capital reserves.

TABLE 2.
VAR AT DIFFERENT CONFIDENCE LEVELS

Model	VaR (in million US Dollar)			LL
	90%	95%	97.5%	
Gamma-GPD	1244	2577	4043	-506.88
Weibull-GPD	206	344	515	-491.03
Lognormal-GPD	276	610	1435	**-485.48**

Note: The bold values in the table refer to the highest value of LL

Table 2 shows VaR estimation under different confidence levels and mixture models. The proposed mixture models, Gamma-GPD, Weibull-GPD, and lognormal-GPD estimate VaR differently, as can be seen. Gamma-GPD model estimates VaR clearly higher than other models for each confidence level, reaching over 4000 million dollars. Furthermore, another interesting point in Table 2 is that there are no significant differences in the VaR estimation of the Weibull-GPD and lognormal-GPD model under 90% confidence level. On the other hand, the VaR estimations are significantly different for both models under 95% and 97.5% of confidence level. In addition, the riskiness depends on the choice of the confidence level. Surely, the VaR estimation at the 99% confidence level is lower than the VaR estimation at the 95% and 97.5% confidence level.

Model selection is based on the highest value of Log Likelihood (LL). The lognormal-GPD mixture model appears to be the best mixture model that best fits the loss data, as shown in Table 2. Moreover, under the selected model, we can say VaR 90% 276 million US dollars means that if an unprecedented disaster occurs in Indonesia such as earthquake, volcanic eruption, flood, etc, we are 90% confident that the loss will not exceed 276 million US dollars. In other words, the government/insurance coverage should be 276 million US dollars to cover 90% of the natural disaster's losses. This is a significant sum of money because the catastrophic risk not only affects an individual but also the entire system, infrastructure and economy as a whole. The same interpretation also applies to VaR estimation under 95% and 97.5% confidence level.

IV. CONCLUSION

Three mixture models are used to assess loss data from natural disasters in Indonesia, and the performance of these mixture models is compared using the log-likelihood value. According to the findings, the mixture model is capable of handling the center and tail behavior of loss data that is heavy-tailed. Furthermore, after receiving the estimates of the mixture models, VaR for each model is also estimated. VaR estimations vary depending on the mixture model and confidence level used. Weibull-GPD and lognormal-GPD models estimate VaR significantly lower than gamma-GPD models. For 95 percent and 97.5 percent confidence levels, VaR estimations for weibull-GPD and Lognormal-GPD reveal a substantial difference. In addition, the lognormal-GPD model is chosen as the best model that fits data best with the highest value of log-likelihood.

REFERENCES

[1] Y. Li, N. Tang, and X. Jiang, "Bayesian approaches for analyzing earthquake catastrophic risk," *Insur. Math. Econ.*, vol. 68, pp. 110–119, 2016, doi: 10.1016/j.insmatheco.2016.02.004.

[2] J. C. Hull, *Risk management e istituzioni finanziarie*. Pearson, 2008.

[3] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," in *Mathematical proceedings of the Cambridge philosophical society*, 1928, vol. 24, no. 2, pp. 180–190.

[4] A. J. McNeil and R. Frey, "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach," *J. Empir. Financ.*, vol. 7, no. 3–4, pp. 271–300, 2000.

[5] T. G. Bali, "An extreme value approach to estimating volatility and value at risk," *J. Bus.*, vol. 76, no. 1, pp. 83–108, 2003.

[6] C. N. Behrens, H. F. Lopes, and D. Gamerman, "Bayesian analysis of extreme events with threshold estimation," *Stat. Modelling*, vol. 4, no. 3, pp. 227–244, 2004.

[7] A. Frigessi, O. Haug, and H. Rue, "A dynamic mixture model for unsupervised tail estimation without threshold selection," *Extremes*, vol. 5, no. 3, pp. 219–235, 2002.

[8] S. Cabras, M. E. Castellanos, and D. Gamerman, "A default Bayesian approach for regression on extremes," *Stat. Modelling*, vol. 11, no. 6, pp. 557–580, 2011.

[9] J. Pickands III, "Statistical inference using extreme order statistics," *Ann. Stat.*, pp. 119–131, 1975.

[10] A. A. Balkema and L. De Haan, "Residual life time at great age," *Ann. Probab.*, pp. 792–804, 1974.

[11] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling extremal events: for insurance and finance*, vol. 33. Springer Science & Business Media, 2013.

[12] "Intensitas Bencana Banjir di Indonesia Selama 10 Tahun Terakhir | Databoks." .

[13] "Refleksi 2020: Lebih Dari 8.000 Gempa Terjadi di Indonesia." .