Parameter Estimations of the Generalized Extreme Value Distributions for Small Sample Size

RaziraAniza Roslan[*]**, Chin Su Na, Darmesah Gabda**

Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Malaysia

Abstract The standard method of the maximum likelihood has poor performance in GEV parameter estimates for small sample data. This study aims to explore the Generalized Extreme Value (GEV) parameter estimation using several methods focusing on small sample size of an extreme event. We conducted simulation study to illustrate the performance of different methods such as the Maximum Likelihood (MLE), probability weighted moment (PWM) and the penalized likelihood method (PMLE) in estimating the GEV parameters. Based on the simulation results, we then applied the superior method in modelling the annual maximum stream flow in Sabah. The result of the simulation study shows that the PMLE gives better estimate compared to MLE and PMW as it has small bias and root mean square errors, RMSE. For an application, we can then compute the estimate of return level of river flow in Sabah.

Keywords Extreme Value Theory (EVT), Generalized Extreme Value (GEV), Maximum Likelihood Estimation (MLE), Probability Weighted Moments (PWM), Penalized Maximum Likelihood (PMLE), L-Moment

1. Introduction

Extreme Value Theory (EVT) is a statistics field that concentrates on any possible event that can be led to more extreme than it is normally happening. Usually, EVT is used to measure safety during catastrophic events, sometimes if we do not pay attention to the risk of an event because it just has a low occurrence it will cause huge losses. Therefore we can use EVT in a specific location to estimate the frequency and cost of such events over a period of time. EVT has been widely used in various fields such as geophysical variable, insurance, risk management and hydrology [19]. There are two approaches used when it comes to analyzing the extreme value, which is Block maxima (BM) and peak over the threshold (POT). In BM, the period will be divided into equal section and the maximum of each will be selected. The approach is usually going to pair with generalized extreme value (GEV). While POT will select every value that exceeds a certain threshold and this approach leads to generalized Pareto distribution (GPD) [1].

GEV distribution was introduced by Jenkinson [3] and has been used in many research areas such as in civil engineering design [4], in hydrology [2], to estimate air quality [15] and also in finance [14]. The GEV distribution consists of three parameters; shape (ξ), scale (σ) and location µ. This parameter estimation of GEV distribution can be obtained using several statistical methods such as the Maximum Likelihood Estimator (MLE), Probability Weighted Moments (PWM), Penalized Maximum Likelihood Estimator (PMLE) and L-moment. The aim of this study is to model the annual maximum stream flow using the GEV distribution focusing on small sample size data. We apply several methods to estimate the GEV parameters.

Each method of parameter estimation has its advantages and disadvantages. But to get ideal parameter estimation, it can be explained in terms of unbiasedness, efficiency and consistency. It is said that the parameter estimation must be unbiased where the estimated parameter closed to the true parameter and the parameter estimation is efficient. The method with the smallest of the root mean square error (RMSE) shows an efficient estimator. Other than that, the parameter estimation must be consistent where the function of estimation is well converged [12].

MLE is the method that is mostly used to estimate the GEV parameter because MLE has good asymptotic properties such as consistency and efficiency. MLE is easy to adapt to model change [12]. Besides that, MLE can be used in a complex model such as the non-stationary model, temporal dependence and covariate effect. However, this parameter estimation can only be used in large sample data and the result will become uncertain if the data is less than

50 values (minimum) [7]. This is confirmed by Hosking et al. (1985) that MLE shows a poor performance due to the small sample size. Considering MLE cannot perform well in small sample size, Coles & Dixon [18] show an investigation about how to improve MLE by proposing an alternative method called PMLE. Their study stated that PMLE will not only maintain model flexibility and large sample optimality of MLE, also help to improve it on small properties. This may be concluded that PMLE is given an improved smoother estimation along with better accuracy thandirect estimate without penalties [20].

PWM was probably advantageous for small set data because it has smaller uncertainty than ordinary moment [19] and has lower variance than others [12]. But when the shape parameter is large, this parameter estimation performs poorly [18] and upper quantile will show PWM is biased. But PWM is still preferable than MLE for small sample size data. On the other hand, MLE is more flexible than PWM because covariate can be easily added in parameterization [11].

PWM is equivalent to L-moment and it performs better than MLE in terms of bias and RMSE [17,10].L moment is the summary statistic, where it provides a measure of location, kurtosis, skewness or any aspects of shape that explain about probability distribution and data sample. Although L-moment produces bias, but it is still preferable due to having a smaller variance than MLE [13] as MLE produced a very large variance and error for estimation [9]. However, L-moment parameter estimation can only be used to estimate the stationary process [8]. Therefore L-Moment and MLE can be "mixed' to produce a better result for GEV parameter estimator. The outcome of this combination helps to reduce variance and bias [13]. In this study, we will illustrate the GEV parameter estimations using simulation study. The superior method then will be applied to model the annual maximum stream flow in Sabah.

2. Methodology

The previous study has shown that PMLE is more superior to other methods. In this study, we will illustrate the GEV parameter estimation using 3 parameter estimates such as MLE, PWM, and PMLE. We conducted simulation study for methods comparison using R software with our own written code. From the result, then we will apply this method to model the annual maximum river flow in Sabah.

2.1. GEV Distribution

The GEV distribution is a family distribution consisting of three distributions called as Gumbel, Fréchet, and Weibull. These distributions can fit the extreme data set with high accuracy. Choosing only one of family GEV distribution may cause bias in data and the term of distribution, uncertainty will be ignored [12].

The GEV distribution having the non-degenerate distribution function fulfills where a_n and b_n are constant with $a_n > 0$

$$\text{P}_r \ (\text{M}_n \leq x) \approx G\left(\frac{x - b_n}{a_n}\right) = G^*(x) \tag{1}$$

The cumulative distribution function (CDF) of GEV distribution is denoted as follows [13]:

$$G(x) = \begin{cases} \exp\left[-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{1}{\xi}}\right], \xi \neq 0 \\ exp\left\{-exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, \xi = 0 \end{cases} \tag{2}$$

Where x: $1 + \left(\frac{x-\mu}{\sigma}\right) > 0$, $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$, in this model ξ, σ and μ are the parameters for shape, scale, and location. By equation (2) GEV distribution for Fréchet $\xi > 0$ and Weibull $\xi < 0$, while for Gumbel distribution $\xi = 0$ taken as $\xi \to 0$.

2.2. Maximum Likelihood Estimation (MLE)

Generally, MLE is the most popular estimation method in EVT because MLE is having good asymptotic properties such as consistency, efficiency, and normality. MLE can be applied to complex modeling situations such as temporal dependence, non-stationary and covariate effects [12]. The likelihood function can be written as

$$L(\theta/x) = \prod_{i=1}^{n} g(x) \tag{3}$$

where g is probability density function of GEV

$$L(\theta/x) =$$
$$\begin{cases} \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\left(\frac{1}{\xi}\right)-1}, \xi \neq 0 \\ \prod_{i=1}^{n} \frac{1}{\sigma} exp\left(-\frac{x-\mu}{\sigma}\right)\exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, \xi = 0 \end{cases}$$
$$\tag{4}$$

When the sample size rises to infinity, it is said that the MLE shows consistent estimator and the variance will go to zero. The asymptotic theory allows MLE to be normally distributed as the sample size rises. MLE was chosen due to the stable performance in a large sample size (n>*50*) [5]. The parameter estimation for GEV can be obtained by maximizing log likelihood function with respect to parameters.

2.3. Penalized Maximum Likelihood Estimator (PMLE)

The following Penalization Maximum Likelihood was introduced byColes & Dixon [18]. With the penalized likelihood function can be written as equation 5 [18]:

$$L_{pen}(\mu, \sigma, \xi) = L(\mu, \sigma, \xi) \times P(\xi) \tag{5}$$

Where $L(\mu, \sigma, \xi)$ is the standard likelihood function of

MLE from equation (4) and the penalty function p (ξ) is shown in equation (6):

$$p(\xi) = \begin{cases} 1 & \text{if } \xi \leq 0 \\ \exp\left(-\lambda\left(\frac{\xi}{\xi-1}-1\right)^{\alpha}\right) & \text{if } 0 < \xi < 1 \\ 0 & \text{if } \xi \geq 1 \end{cases} \quad (6)$$

Where the appropriate value for λ and α is non-negative. The PMLE will help to overcome the poor result of MLE due to the small sample size. This is supported byColes & Dixon[18], where they have conducted a study to explain the behavior of penalized likelihood. The PMLE was almost identical to the MLE for the case of **ξ** that is a negative value. But if **ξ** shows positive value PMLE will be almost the same with PWM, hence the characteristics of PMLE will inherit smaller variance at expense of negative bias[12]. Overall, PMLE has properties that will match in all sample sizes and helps to improve MLE and PWM.

2.4. Probability Weight Moment (PWM) & L-Moment

L-moment is a method based on a combination of PWM [7], hence PWM is equivalent to L-moment for GEV distribution [6] and this method was introduced by Hosking [10].

Random variable X for PWM and L-moment can be defined as;

$$\beta_r = M_{1,r,0} = E[X\{F(X)\}^r], r = 0,1,2,\dots \quad (7)$$

X is distribution function for F and $\hat{\beta}_r$ is the estimate of empirical distribution in;

$$\hat{\beta}_r = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}\left(\hat{F}_i\right)^r \text{Where} \hat{F}_i = \frac{(i-0.35)}{n} \quad (8)$$

Therefore parameter of GEV can be estimated using this equation;

$$\hat{\xi} = 7.8590c + 2.9554c^2 \quad (9)$$

$$\hat{\sigma} = \frac{(2b_1-b_0)\hat{\xi}}{r(1-\hat{\xi})(2^{\hat{\xi}}-1)} \quad (10)$$

$$\hat{\mu} = b_0 - \frac{\hat{\sigma}}{\hat{\xi}}\left\{\Gamma\left(1-\hat{\xi}\right)-1\right\} \quad (11)$$

$$where \; c = \frac{2b_1-b_0}{3b_1-b_0} - \frac{log2}{log3} \quad (12)$$

3. Return Level

Return level is frequently used to convey information about the likelihood of extreme events such as earthquake, flood, hurricanes, etc [19]. For the application above method, we can estimate the return level by using equation 13.

$$Z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left(\log(1-p)^{-\xi}-1\right) & \xi \neq 0 \\ \mu - \sigma log[-log(1-p)] & \xi = 0 \end{cases} \quad (13)$$

4. Simulation Study

We illustrate the comparison of GEV parameter estimations using a simulation study. For this purpose, we simulate extreme events from GEV distribution, $X\sim$GEV for (0,1,0.15) with a sample size of n=30. We repeat this simulation for 1000 times. For each case, we estimate the parameter estimation using MLE, PWM, and PML. Then we compute the bias and RMSE for method comparison.

Table 1 shows the GEV parameter estimation by MLE, PWM and PMLE. It shows that estimation is close to the actual value, $\hat{\theta} \approx \theta$.

Table 1. GEV parameter estimation of PWM, MLE and PMLE

Method	Parameter estimation ($\hat{\theta}$)		
	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
PWM	0.029251	0.959233	0.138762
MLE	0.011395	0.964506	0.145367
PMLE	0.0167963	0.971953	0.150994

Table 2 shows that the biasness is close to zero for all parameter estimation methods. As we can see from Table 3, PMLE produces smaller RMSE of ξ compared to other methods. Hence we can conclude that PMLE is superior compared to MLE and PMW as shown in a previous study Coles & Dixon [18] and Musakkal [2].

Table 2. Bias of GEV parameter estimation of PWM, MLE, and PMLE

Method	Bias		
	μ	σ	ξ
PWM	-0.000253	0.000681	-0.000123
MLE	-0.000208	0.000644	-0.000095
PMLE	-0.000242	0.000699	-0.000125

Table 3. Root mean square error (RMSE) of GEV parameter estimation of PWM, MLE, and PMLE

Method	RMSE		
	μ	σ	ξ
PWM	0.0197239	0.027801	0.017808
MLE	0.019521	0.028044	0.018097
PMLE	0.019464	0.027903	0.017759

5. Result and Discussion

5.1. Application of Data Stream Flow in Sabah

This study uses data annual maximum streamflow (m^3s^{-1}) from several stations in Sabah. Data were obtained from the Hydrology Department of Sabah. The data were collected from several stations. Table 4 shows the number of observations for each station.

Table 4. Number of observations for each station

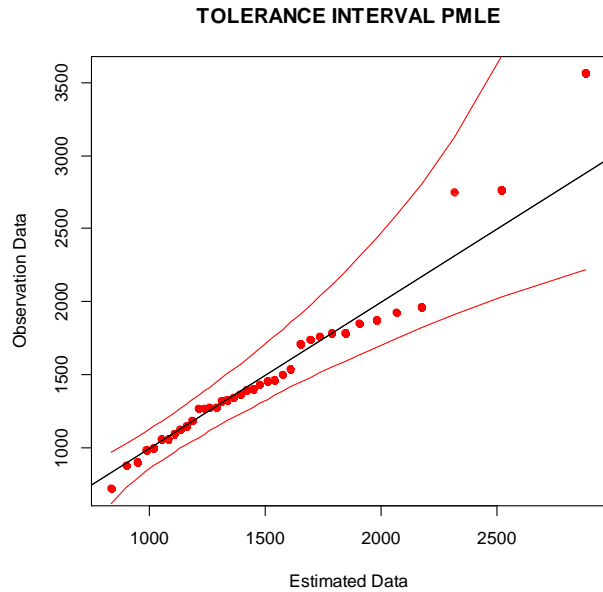| No	Station Name	Years	Duration
1	Sg Segama At Limkabong Sabah	23	1994-2019
2	Sg Balung At Balung Bridge Sabah	25	1992-2016
3	Sg Sapulut At Sapulut Sabah	29	1990-2018
4	Sg Kalabakan At Kalabakan Sabah	32	1986-2017
5	Sg Tawau At Kuhara Sabah	34	1983-2016
6	Sg Mengalong At Sindumin	37	1983-2019
7	Sg Kalumpang At Mostyn Bridge Sabah	38	1979-2016
8	Sg Lakutan At Mesapol Sabah	39	1978-2016
9	Sg Padas At Kemabong Sabah	50	1969-2018
10	Sg Kuamut At UluKuamut Sabah	50	1969-2018

We applied the result from the simulation study in modeling the annual maximum river flow in Sabah. As a result, Table 5 shows the GEV parameter estimation by using the PMLE method.

Table 5. Parameter estimation for PMLE

No	Station	PMLE		
		μ	σ	ξ
1	SgSegama At Limkabong Sabah	490.5986	365.4626	-0.0606
2	SgBalung At Balung Bridge Sabah	16.2600	10.6001	-0.0578
3	SgSapulut At Sapulut Sabah	464.0687	147.8391	-0.0144
4	SgKalabakan At Kalabakan Sabah	177.0876	196.1891	-0.0000
5	SgTawau At Kuhara Sabah	10.2473	3.6354	-0.1465
6	SgMengalong At Sindumin	205.9045	100.0220	-0.2425
7	SgKalumpang At Mostyn Bridge Sabah	173.4530	117.6743	-0.0000
8	SgLakutan At Mesapol Sabah	69.8482	35.49840	-0.0953
9	Sg Padas At Kemabong Sabah	634.7662	218.7411	-0.0003
10	SgKuamut At UluKuamut Sabah	887.36199	376.63772	0.00004

We evaluate the goodness of fit of GEV using the Q-Q plot with a 95% tolerance interval. The Q-Q plot is a useful tool to check the empirical distribution that is close or similar to the critical distribution. As a result, GEV has fitted well the annual maximum river for all stations. Figure 1 shows the example of the Q-Q plot with a 95% tolerance interval for the GEV fit of annual maximum river flow at station Segama. It can be seen that all points are scattered in a straight line with slope equal to 1 and within 95%

tolerance interval.

Figure 1. Q-Q Plot with 95% tolerance interval at station Segama (PMLE method)

We then calculated the return value of annual maximum for each site with *p*=0.01. The corresponding return value estimation for all station is shown in Table 6.

Table 6. Return value estimates

No	Station	PMLE
1	SgSegama At Limkabong Sabah	1957.537
2	SgBalung At Balung Bridge Sabah	59.07734
3	SgSapulut At Sapulut Sabah	1122.095
4	SgKalabakan At Kalabakan Sabah	1079.415
5	SgTawau At Kuhara Sabah	22.4142
6	SgMengalong At Sindumin	483.1688
7	SgKalumpang At Mostyn Bridge Sabah	714.7055
8	SgLakutan At Mesapol Sabah	202.0502
9	Sg Padas At Kemabong Sabah	1640.258
10	SgKuamut At UluKuamut Sabah	2620.125

6. Conclusions

The simulation study shows that the PMLE gives a better estimate compared to MLE and PMW because it has small bias and RMSE. We then use this result for an application of modeling the annual maximum river flow in Sabah. The GEV distribution is the appropriate model for these extreme data. For the application we used 100 years return level for each of station. It shows that the theoretical distribution is similar to the empirical distribution. For

future study, we will consider the effect of the covariate in the model [14]

REFERENCES

[1] A. Fierra & L. De Haan. (2015). On the block maxima method in extreme value theory: pwm estimators. The Annals of Statistics 43, 1, 276–298.

[2] A. Z. Ismail, Z. Yusop & Z. Yusof. (2015). Comparison of flood distribution models or johor river Basin. Journal Technology 74(11): 1123-128.

[3] A.F Jenkinson. (1955). The frequency distribution of the annual maximum or minimum value of meteorological element. Quarterly Journal of the Royal Meteorological Society 87, 145-58.

[4] E. Castillo, A. S. Hadi, N. Balakrishan, & J. M. Sarabia, (2005). Extreme Value and Related Models with Applications in Engineering and Science. New Jersey: Wiley.

[5] E. Gilleland &R.W. Katz. (2006). Analyzing seasonal to interannual extreme weather and climate variability with the extreme toolkit. Research Application Laboratory, National Center for Atmospheric Research,2 (15): 1-9

[6] E.S. Martins &J. R Stedinger (2000). Generalized maximum likelihood generalized extreme value quantile estimators for hydrologic data. Water Resources Research, 36(3), 737-744.

[7] G. Lazoglou, &C. Anagnostopoulou (2017). An Overview of Statistical Methods for Studying the Extreme Rainfalls in Mediterranean. Proceedings, 1(5), 681.

[8] H B.. Hasan, N F.. Radi & S. B. Kassim.. (2012). Modeling of extreme temperature using generalized extreme value (GEV) distribution: a case study of Penang. Proceedings of the World Congress on Engineering 1. WCE 2012,

[9] J. E. Morrison, & J A. Smith (2002). Stochastic modeling of flood peaks using the generalized extreme value distribution. Water resources research. 38, 12, 1305.

[10] J.R.M. Hosking, J.R. Wallis &E. F. Wood. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moment. Technometrics 27, 3, 251-261

[11] K Engeland, H. Hidal, & A. Frigessi. (2005). Practical extreme value modelling of hydrological floods and droughts: a case study. Extremes 7, 5–30.

[12] N. F. K. Musakkal, S. N Chin., K. Ghazali. &D. Gabda (2018). A penalized likelihood approach to model the annual maximum flow with small sample sizes. Malaysian Journal of Fundamental and Applied Sciences. 13, 4, 563-566.

[13] P. Ailliot., C. Thompson, & Thomson, P (2011). Mixed methods for fitting the GEV distribution. Water Resources Research, 47(5), 1–14.

[14] P. Embrect, C. Kluppelberg & T. Mikosh. (1997). Modelling Extremal Events for Insurance and Finance. Berlin: Springer Verlag.

[15] P. Sharma, A. Chandra, S. C. Kaushik, P. Sharma &S. Jain (2012). Predicting violations of national ambient air quality standard using extreme value theory for Delhi City. Atmospheric Pollution Research, 3, 2, 170-179.

[16] R. L. Smith (1985). Maximum likelihood estimation in a class of non-regular cases. Biometrika, 72, 67–92.

[17] R. Wada, T. Waseda, & P. Jonathan. (2016). Extreme value estimation using the likelihood-weighted method. Ocean Engineering, 124(June), 241–251.

[18] S.Coles, & M. Dixon. (1999). Likelihood based inference for extreme value models. Extremes 2(1), 5-23.

[19] V. P. Singh &K. P. Das. (2016). Characterization of the tail of river flow data by generalized pareto distribution. Journal of Statistical Research. 48-50, 2, 55-70.

[20] Z. Zhou, S. Liu, H. Hua, C. S. Chen, G. Zhong, H. Lin, &C. W. Huang. (2014). Frequency Analysis for Predicting Extreme Precipitation in Changxing Station of Taihu Basin, China. Journal of Coastal Research, 68(Cm), 144–151.