MJFAS MALAYSIAN JOURNAL OF FUNDAMENTAL AND APPLIED SCIENCES
PRINT ISSN: 2289-5981 | ONLINE ISSN: 2289-599X

RESEARCH ARTICLE

A penalized likelihood approach to model the annual maximum flow with small sample sizes

Nur Farhanah Kahal Musakkal*, Chin Su Na, Khadizah Ghazali, Darmesah Gabda

Faculty of Science and Natural Resources , University Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia

* *Corresponding author: nurfarhanahkahalmusakkal@ymail.com*

Article history
Received 19 February 2017
Accepted 1 October 2017

Abstract

The aim of this study is to model the annual maximum flow of several sites in Sabah with small sample sizes using the generalized extreme value (GEV) distribution. Previous studies have shown that the standard method of maximum likelihood estimates would give a poor estimation of the GEV parameters and quantiles for the small data set. This study will consider the penalized likelihood estimates as an alternative method to improve the inference over the standard method and retains the modeling flexibility. As for comparisons, we will illustrate the results of both methods to model the annual maximum flow in Sabah. The results show the implementation of the penalty function had the same effect to the GEV parameter estimates as suggested by previous studies.
.

***Keywords*:** Generalized extreme value, Penalized likelihood, Extreme value theory, Small sample size

INTRODUCTION

Generalized Extreme Value (GEV) distribution have been applied for numerous applications in many areas for example in hydrology (Coles, 2001), civil engineering design (Castillo, 2005) and in finance (Embrechts *et al*., 1997). Therefore, the properties of estimation parameters need to carefully consider. There are various approaches available for GEV parameters estimation such as Bayesian, estimation based on moments method, estimation based on likelihood method and much more. All available approaches have their own pros and cons. An ideal estimation should consist the properties of unbiased (the estimator is said to be unbiased when the expected parameter equal or almost identical to the true parameter value), efficient (the estimator is said to be efficient when the estimator has the minimum mean square error) and consistent (the estimator is said to be consistent when the function is well converge) (Soukissian and Tsalis, 2015).

Generally, the Maximum Likelihood Estimation (MLE) method is one of the most popular estimation methods in extreme value theory (EVT) for having good asymptotic properties such as asymptotic consistency, asymptotic efficiency, asymptotic normality and has unique properties in their capability to adapt to model change (Smith, 1985). Smith (1985) also stated that MLE can be applied to complex modeling situation such as non-stationarity, temporal dependence, and covariate effects. There are a number of previous studies using MLE to estimates parameters in GEV distribution such as Phien and Fang (1988), Nadarajah and Shiau (2005), Nadarajah and Choi (2007), Soukissian and Tsalis (2015) and many more.

There are many advantages of MLE which have been mentioned earlier but the justification is based on large sample theory, there still little application on small samples sizes (Hosking *et al.*, 1985). The poor performance of MLE in small samples issues is still a serious discussion. The most popular study of small sample sizes using MLE was done by Hosking *et al.*, (1985) and was then extended by Coles and Dixon (1999). Hosking *et al.*, (1985), performed a simulation study concentrated on small, moderate and large sample sizes such as n=15,25,50,100 for $\xi = -0.4, -0.2, 0, 0.2, 0.4$ to study the Probability Weighted Moment (PWM) estimators of the GEV distribution and to compare with MLE and sextiles estimators. For smaller data set the PWM estimator shows lower variance than others. They also compared quantile (the inverse cumulative distribution function, CDF of GEV distribution) estimators. For small data sets, the upper quantiles show PWM method is biased, however PWM still better compared to ML estimators, which shows very large biases and variances. They also stated that when estimating extreme quantiles with $\xi > 0$ in small samples all the methods are very inaccurate.

Coles and Dixon (1999) conducted a study to focus on the comparison between MLE and PWM for estimating parameters of the GEV distribution with small sample datasets. They found that for small sample sizes of extreme event, the MLE shows poor performance and this is confirmed the result of Hosking *et al.*, (1985). They performed a study to investigate more detailed the behavior of MLE with small sample size and found that MLE was not performed well in that case. From the study, they also found some considerations on how to modify MLE in order to improve its performance in small sample issues, while maintaining its flexibility characteristics and properties. They proposed an alternative method to estimate GEV parameters involved small sample sizes of an extreme event by introducing a penalty function to the standard method of MLE called penalized maximum likelihood estimator (PMLE) method and proved that their method performed well compared to existing methods. The penalized maximum likelihood methods also have been proposed by

other studies for different focus; Martin and Stedinger (2000) improved the GEV parameter estimates for small sample sizes; Zheng *et al.*, (2014) used penalty function to allowed parameters varies smoothly between neighbouring sites.

The aim of this study is to model the annual maximum flow of several sites in Sabah with small sample sizes using the GEV distribution. In the next section, we define the GEV model follows by the discussion of the GEV parameter estimation using the method of MLE and PML method. We also show simulation studies to understand both methods before we apply the method to our data. Then, we show a comparison of MLE and PML estimators in modeling annual maximum stream flow of several sites in Sabah. Finally, we used the best method to compute the prediction of extreme river flow in Sabah.

METHODOLOGY

This section discusses the model fitting of generalized extreme value distribution to the annual maximum flow and its parameter estimation. In this study, we used R software for computational purpose with our own written code.

GEV Distribution

Modeling the extreme event is usually based on asymptotic based theory where sample of extreme event are renormalized with sequence of normalization constant. As the number of sample extreme would approach infinity, the distribution of the renormalized sample extreme converges to the GEV distribution. Recently, the GEV distribution is commonly used in extreme events for modeling and characterizing. The GEV distribution has Cumulative Distribution Function (CDF) as follows (Coles, 2001):

$$F(x;u,\sigma,\xi)=\begin{cases} exp\left\{-\left(1+\xi\left[\frac{x-u}{\sigma}\right]\right)^{-1/\xi}\right\}, & \xi\neq 0 \\ exp\left[-exp\left\{-\frac{x-u}{\sigma}\right\}\right], & \xi=0 \end{cases} \quad (1)$$

The GEV distribution consists of three parameters where $\mu \epsilon \Re$ is the location parameter, $\sigma>0$ is the scale parameter and $\xi \epsilon \Re$ is the shape parameter. The GEV (μ,σ,ξ) distribution has support on the set $\{x: 1+\xi(\mathrm{x}-\mu)/\sigma>0\}$. It will be a Gumbel distribution for $\xi=0$ (taken as $\xi\to 0$) and whereas for , $\xi<0$ and $\xi>0$ corresponds to the Negative Weibull distribution and the Frechet distribution respectively. Choosing directly one family of GEV distributions may lead to a biased fit for a given data set and ignores uncertainty in the form of the distribution. Therefore, the generalized extreme value distribution is an appropriate model for the extremes which allows for uncertainty in the selection of the three different types.

Maximum Likelihood estimation

The idea of maximum likelihood method is based on maximizing the likelihood of the observed sample (independent random variable) with respect to all the parameters which can be expressed as in equation (2).

$$L(\theta|x)=\prod_{i=1}^{n} f(x_i) \quad (2)$$

where f is the probability density function associated with distribution function in equation (1) which can be derived as $f=dF(x)/d(x)$. Therefore, the derivation of Maximum Likelihood function for GEV distribution can be obtained as in equation (3).

$$L(\theta|x)=\begin{cases} \prod_{i=1}^{n}\frac{1}{\sigma}\left(1+\xi\frac{x-u}{\sigma}\right)^{-1-\frac{1}{\xi}} e\left(-\left(1+\xi\frac{x-u}{\sigma}\right)^{-\frac{1}{\xi}}\right), \xi\neq 0 \\ \prod_{i=1}^{n}\frac{1}{\sigma}\left(exp\left(-\frac{x-u}{\sigma}\right)\right) exp\left(-exp\left(-\frac{x-u}{\sigma}\right)\right), \xi=0 \end{cases} \quad (3)$$

Penalized Maximum Likelihood Estimator

Modified MLE or Penalized Maximum Likelihood (PML) estimator is a standard application to non-parametric smoothing, in which a function that penalizes roughness is going to balance the appropriate likelihood. Penalized likelihood function is defined as $L_{Pen}=L(\mu,\sigma,\xi)\times \mathrm{P}(\xi)$. In this study we used a penalty function as follows (Coles and Dixon, 1999):

$$P(\xi)=\begin{cases} 1, & \xi\leq 0 \\ exp\left(-\lambda\left(\frac{1}{1-\xi}-1\right)^{\alpha}\right), & 0<\xi<1 \\ 0, & \xi\geq 1 \end{cases} \quad (4)$$

According to Coles and Dixon (1999), the suitable value of α and λ is equal to 1 and $L(\mu,\sigma,\xi)$ is obtained from MLE of the GEV distribution. Therefore, small sample cases of MLE can be overcome by applying methods of PML estimator. The MLE of GEV distribution yield the standard asymptotic result which all applicable to the PML estimator. Coles and Dixon (1999) conducted simulations to study the comparison of small sample behavior (n=25) of these PML estimators with the classical ML and PWM estimators. They found that when ξ is negative the behavior of PML estimator is almost identical to MLE estimator. However, when ξ is positive, the behavior of PML estimator is almost identical PWM) estimator, therefore the PML estimator will have the characteristics of smaller variance at the expense of negative bias. Thus, in the context of bias and variance the PML estimator seems to be good as the PWM estimator.

Probability Weighted Moments

In this study, the initial values of the parameters in both likelihood function of MLE and PML were computed using a probability weighted moments.

PWM introduced by Hosking *et al.*, (1985). For $r=0,1,2$, the plotting position estimator $b_r=\frac{1}{n}\sum_{i=1}^{n} x_i[F_i]^r$ are best evaluated at $F_i=\frac{i-0.35}{n}$. Therefore, the GEV parameters are then estimates using these equation:

$$\hat{\xi}=7.8590c+2.9554c^2, \quad (5)$$

$$\hat{\sigma}=\frac{(2b_1-b_0)\hat{\xi}}{\Gamma(1-\hat{\xi})(2^{\hat{\xi}}-1)}, \quad (6)$$

$$\hat{u}=b_0-\frac{\hat{\sigma}}{\hat{\xi}}\{\Gamma(1-\hat{\xi})-1\} \quad (7)$$

$$c=\frac{2b_1-b_0}{3b_1-b_0}-\frac{log2}{log3} \quad (8)$$

Return level

Further application of above method is for estimation of the extreme quantiles or the return level. By inverting the equation (1) the 1-p quantile of GEV distribution can be obtained by using equation (9) substituting estimates of (μ, σ, ξ) into (4) for any values of p;

$$z_p=\begin{cases} u-\frac{\sigma}{\xi}\left(1-[-\log(1-p)]^{-\xi}\right), & \xi\neq 0 \\ u-\sigma\log[-\log(1-p)], & \xi=0 \end{cases} \quad (9)$$

For example, $z_{0.01}$ corresponds to the level expected to be exceeded in every 100 years.

Simulation Study

We conducted a simulation study to explore the GEV parameter estimates method; maximum likelihood method and penalized likelihood method porposed by Coles and Dixon (1999). By using R software, we simulate extreme event of variable X for a sample size of 50 and we repeated for 5000 times. We assume that the random variable X follows a GEV ($\mu=0,\sigma=1,\xi=0.2$). We only consider $\xi>0$ as for negative value of shape parameter both methods of MLE and PML almost identical. Both methods are then compared on the bias and root mean square error of the parameter estimates and also quantiles of the GEV distribution. The results are present in Table 1

for bias, and Table 2 for root mean square error. The results show that the PML performed better than MLE in terms of bias and root mean square error, clearly seen on quantile estimates.

Table 1 The bias of GEV parameter estimates and quantile estimates of MLE and PML method

Method	μ	σ	ξ	$q_{0.01}$	$q_{0.005}$
MLE	0.01	-0.02	0.00	0.31	0.66
PML	0.01	-0.02	-0.03	-0.24	-0.23

Table 2 The root mean square error of the parameter estimates and the quantile estimates of MLE and PML

Method	μ	σ	ξ	$q_{0.01}$	$q_{0.005}$
MLE	0.16	0.13	0.13	3.04	4.74
PML	0.17	0.13	0.12	2.34	3.48

Application to stream flow data in Sabah

This study used a secondary data obtained from Hydrology Department of Sabah. We have an annual maximum streamflow (m^3s^{-1}) data from several sites in Sabah. The characteristics of the selected stations for this study are presented in the Table 3 which provide the number of the stations, the name of the river and the locations for each selected stations. The data consists of 18 sites in Sabah with a different number of observations for each site. All number of observations used in this study indicated small sample sizes (below than 50).

We fit the GEV distribution to the annual maximum flow for each site. Then we examined the goodness of fit using the Q-Q plot with 95% tolerance intervals. The results show that the GEV fit seems to be appropriate for all data. Fig. 1 shows an example of Q-Q plot with 95% tolerance interval indicates that the GEV fits the annual maximum flow data very well at Station Padas using MLE method for parameter estimations. Fig. 2 shows the same plot but by using the PML method, indicates that the GEV fits the data well.

We used the maximum likelihood method and penalized maximum likelihood method for GEV parameter estimations. For comparisons, we present the results in Table 4, specifically the estimates of shape parameter estimates, $\hat{\xi}$. For a negative value of $\hat{\xi}$, the PML estimator is almost identical to ML estimator. This can be seen on sites Kinabatangan at Pagar, Labuk, Milian, Padas at Beaufort, Padas at Kemabong, Papar at Kogopon, Pegalan, Sook and Tungud as shown in the table. Our results were consistent with previous studies where the PML method shrinks the shape parameter towards zero.

For application purposes, we then computed the estimates of 100 years return value of annual maximum flow at each site. Fig. 3 shows the return value estimates of annual maximum data sites obtained by using Maximum Likelihood (MLE) and Penalized Maximum Likelihood (PML).

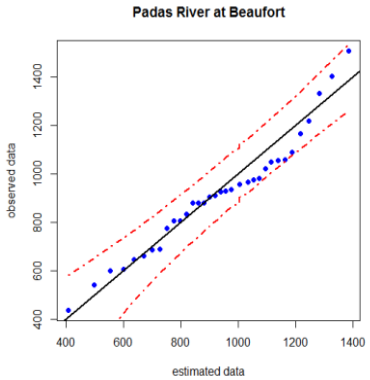

Fig. 1 Q-Q plot with 95% tolerance interval shows well fit of GEV distribution for annual maximum flow at Station Padas (MLE method)

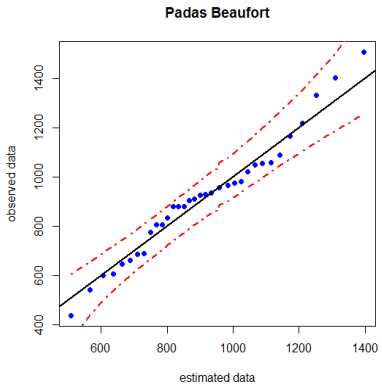

Fig. 2 Q-Q plot with 95% tolerance interval shows well fit of GEV distribution for annual maximum flow at Station Padas (PML method)

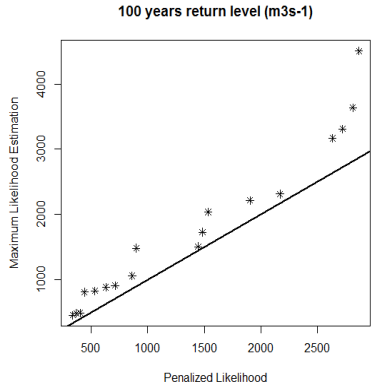

Fig. 3 Return value estimate for Sabah annual maximum data using PWM plotted against MLE.

Table 3 Stream flows data of several site in Sabah.

Bil.	No. Stn	Station Name	Lat	Long	Duration
1	6670401	Sg. Bengkoka	6.622	117.038	1972 - 2016
2	6468402	Sg. Bongan	6.443	116.813	1988 - 2016
3	5375401	Sg. Kinabatangan at Balat	5.31	117.599	1978 - 2015
4	5275401	Sg. Kinabatangan at Pagar	5.231	117.5	1986 - 2016
5	5074401	Sg. Kuamut	5.08	117.442	1969 - 2015
6	5872401	Sg. Labuk	5.855	117.352	1969 - 2015
7	4955403	Sg. Mengalong	4.992	115.577	1983 - 2015
8	5373401	Sg. Milian	5.304	117.318	1969 - 2015
9	5357403	Sg. Padas at Beaufort	5.353	115.724	1981 - 2015
10	4959401	Sg. Padas at Kemabong	4.917	115.92	1969 - 2015
11	5760402	Sg. Papar at Kogopon	5.707	116.038	1969 - 2015
12	5760401	Sg. Papar at Kaiduan	5.769	116.094	1969 - 2015
13	5261401	Sg. Pegalan	5.28	116.139	1969 - 2015
14	4764401	Sg. Sapulut	4.689	116.481	1990 - 2016
15	5261402	Sg. Sook	5.263	116.146	1969 - 2015
16	6172401	Sg. Sugut	6.196	117.237	1984 - 2015
17	6073402	Sg. Tungud	6.05	117.327	1986 - 2015
18	6364401	Sg. Wariu	6.324	116.483	1969 - 2015

Table 4 Shape parameter estimates of MLE and PML.

Bil	Sites	Years	$\hat{\xi}$	
			MLE	PML
1	Sg. Bengkoka	45	0.32	0.26
2	Sg. Bongan	29	0.32	0.26
3	Sg. Kinabatangan at Balat	38	0.11	0.10
4	Sg. Kinabatangan at Pagar	31	-0.27	-0.28
5	Sg. Kuamut	47	0.36	0.31
6	Sg. Labuk	47	-0.02	-0.03
7	Sg. Mengalong	32	-0.34	-0.34
8	Sg. Milian	47	-0.05	-0.05
9	Sg. Padas at Beaufort	35	-0.18	-0.19
10	Sg. Padas at Kemabong	47	-0.06	-0.07
11	Sg. Papar at Kogopon	47	0.00	0.00
12	Sg. Papar at Kaiduan	47	-0.06	-0.07
13	Sg. Pegalan	47	-0.02	-0.02
14	Sg. Sapulut	27	-0.15	-0.17
15	Sg. Sook	46	-0.19	-0.19
16	Sg. Sugut	32	0.23	0.19
17	Sg. Tungud	30	-0.56	-0.59
18	Sg. Wariu	46	0.09	0.01

Table 5 Return level estimate

Sites	$q_{0.01}$
Sg. Bengkoka	2311.80
Sg. Bongan	830.10
Sg. Kinabatangan at Balat	3309.13
Sg. Kinabatangan at Pagar	2214.74
Sg. Kuamut	4489.75
Sg. Labuk	3248.75
Sg. Megalong	456.50
Sg. Milian	2029.70
Sg. Padas at Beaufort	1496.37
Sg. Padas at Kemabong	1891.17
Sg. Papar at Kaiduan	390.35
Sg. Papar at Kogopon	1032.21
Sg. Pegalan	832.39
Sg. Sapulut	911.10
Sg. Sook	333.53
Sg. Sungud	3676.63
Sg. Tungud	910.14
Sg. Wariu	485.93

CONCLUSION

In this study, we have a data of annual maximum flow with small sample sizes at each site. The GEV distribution is an appropriate model for this extreme data. However, an appropriate method should be used to estimate the GEV parameters since its involved small sample size of an extreme event as discussed by several past studies. This study applied the penalized likelihood approach to estimate the GEV parameters since we have the small sample size of annual maximum flow at all sites. Based on application to 18 sites, it is shown that the penalized likelihood method improved the performance of the MLE. Our results show the implementation of the penalty function of the penalized likelihood method had the same effect on the GEV parameter estimates as suggested by previous studies (shrink the shape parameter estimates towards zero). This can give a realistic of the extreme quantile estimates. The penalized likelihood approach is a modification of the standard maximum likelihood method which can be easily implemented even with a complex model. For further work, this study will consider the effect of the covariate in the model.

ACKNOWLEDGEMENT

I am greatly appreciate the Hydrology Department of Sabah for providing me the data of streamflows.

REFERENCES

Castillo, E., Hadi, A. S., Balakrishnan, N., Sarabia, J. M. 2005. *Extreme Value and Related Models with Applications in Engineering and Science*. New Jersey: Wiley.

Coles, S. G., 2001. *An Introduction to Statistical Modeling of Extreme Values*. Wiley, Hoboken: Springer.

Coles, S. G., Dixon, M. J. 1999. Likelihood based inference for extreme value models. *Extremes*, 2(1), 5-23.

Embrechts, P., Klüppelberg, C., Mikosch, T. 1997. *Modelling Extremal Events for Insurance and Finance*. Berlin: Springer Verlag.

Hosking, J. R. M., Wallis, J. R., Wood, E. F. 1985. Estimation of the generalized extreme value distribution by the method of probability weighted moments. *Technometrics*, 27, 251-261.

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

Nadarajah, S., Choi, D. 2007. Maximum daily rainfall in South Korea. *Journal of Earth System Science*, 116(4), 311-320.

Nadarajah, S., Shiau, J. T. 2005. Analysis of extreme flood events for the Pachang River, Taiwan. *Water Resource Management*, 9(4), 363–374.

Phien, H. N., Fang, T. S. E. 1989. Maximum likelihood estimation of the parameters and quantiles of the general extreme value distribution from censored samples. *Journal of Hydrology*, 105, 139-155.

Smith, R. L., 1985. Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72, 67–92.

Soukissian, T., Tsalis, C. 2015. The effect of the generalized extreme value distribution parameter estimation methods in extreme wind speed prediction. *Journal of the International Society for the Prevention and Mitigation of Natural Hazards*, 78(3), 1777-1809.

Zheng, W., Zhang, J., Liu, H., Li, J. 2014. A penalized maximum likelihood approach for m-year precipitation return values estimation with lattice spatial data. *IEEE/CIC International Conference on Communications in China - Workshops (CIC/ICCC)*. 13 October 2014. Shanghai: IEEE, 16-20.