Modeling the Extreme Rainfall Data of Several Sites in Sabah using Sandwich Estimator

Siow Chen Sian[1*] and Darmesah Gabda[1]

[1] *Department of Mathematics with Economics, Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah*
[*] *Corresponding author. E-mail: siowcs1994@gmail.com*

Received: Dec. 07, 2020; Accepted: Mar. 08, 2021

When the extreme data were obtained from several sites in a region, spatial extreme analysis is always been considered. In this paper, we model the annual maximum rainfall data by using generalized extreme value distribution. We fit the model independently for each site to prevent extreme value complex modeling. However, it also cause the statistical assumption of dependency between sites been violated. Therefore, we applied the sandwich estimator to correct the variance of the model. We also consider an analysis of small sample sizes of the observed data. The method of penalized maximum likelihood estimation was carried out to improve the inference of the model. In the end, the return levels of the annual maximum rainfall data were computed by using the corrected model.

Keywords: Generalized Extreme Value (GEV) distribution, Penalized Maximum Likelihood Estimation (PMLE), Sandwich Estimator, Return Level

http://dx.doi.org/10.6180/jase.202206_25(3).0007

1. Introduction

Spatial extreme analysis has been proposed in many previous studies to model spatial dependency within extreme events in continuous space using recorded observations [1–3]. When studying the extremes of two or more processes, each individual process can be modeled using univariate techniques which is generalized extreme value distribution, but there also strong arguments for studying the extreme value inter-relationship. Besides that, the dependency between variables are modeling by using multivariate extreme value distribution. However, modeling multivariate extreme lead to some issues, which creates high dimensionality difficulties for both model validation and computation. There are a few examples of multivariate modeling with two dimensional, bivariate and other analyses for environmental dependence data can be found in [4]. To avoid the model misspecification and for efficient computation, a marginal estimation is a good alternative method for modeling multivariate extremes. Therefore, the sandwich estimator needed to be the standard error modification to capture the data dependency. The properties and advantages of sandwich estimator were discussed in [5].

In this study, we model the annual maximum rainfall data independently by using generalized extreme value distribution. It was well recognized that many previous studies have been applied the Generalized Extreme Value (GEV) distribution in extreme events especially in hydrology [1, 6, 7]. The GEV distribution used to model the annual maximum series (AMS) and partial duration series (PDS) [8]. Since the statistical assumption of dependency between sites is violated if the annual rainfall data were modeling independently, therefore, sandwich estimator is applied to correct the variance of the model. The application then used to compute the return level of the rainfall data. Since there are small sample size issues in maximum likelihood estimation (MLE), therefore the Penalized maximum likelihood estimation (PMLE) which proposed by [4, 9] is applied to avoid the problem.

2. Methodology

This section discusses the model fitting by using GEV distribution to annual maximum rainfall data and its parameter estimation. Then, the sandwich estimator was applied as a statistical modification to give a more appropriate estimates of standard error. In this study, R software used for computational purpose with our own written code.

2.1. Generalized Extreme Value

The cumulative distribution function (CDF) of Generalized Extreme Value (GEV) distribution is:

$$G(z,\mu,\sigma,\xi) = exp\left\{-[1+\xi(\frac{z-\mu}{\sigma})]^{\frac{1}{\xi}}\right\} \; if \; 1+\xi(\frac{z-\mu}{\sigma}) > 0 \tag{1}$$

where μ, σ and ξ represent location parameter, scale parameter and shape parameter, respectively. The GEV consists of three families of distribution that can be determined by the shape parameter; Gumbel (ξ=0), Frechet (ξ>0) and negative Weibull distribution (ξ<0). If choosing either one from GEV distribution, uncertainty would be ignored, and it may cause a biased fit. Therefore, the GEV is the appropriate model for the extreme data since it combines the three families into a single distribution. The data themselves will determine the most appropriate distribution through inferences of shape parameters [4].

2.2. The Maximum Likelihood Estimation (MLE)

The parameter estimation of the GEV can be obtained by maximizing the likelihood of the observed data (independently random variable) with respect to all the parameters [6]. The corresponding likelihood function of the GEV as shown below:

$$L(\theta|x) = \prod_{i=1}^{n} G(z,\mu,\sigma,\xi) \tag{2}$$

where f is the probability density function as in equation (2.1), which can be derived as $f = dF(x)/d(x)$. Therefore, the equation of maximum likelihood function can be shown as below:

$$ln[L(\theta|x)] = \begin{cases} \frac{1}{\sigma}\prod_{i=1}^{n}[1+\xi(\frac{x_i-\mu}{\sigma})]^{-\frac{1}{\xi}-1} \times exp[-\prod_{i=1}^{n}[1+\xi(\frac{x_i-\mu}{\sigma})]^{-\frac{1}{\xi}}],\ \xi \neq 0 \\ \frac{1}{\sigma}\prod_{i=1}^{n} exp(-\frac{x_i-\mu}{\sigma}) - exp(-\frac{x_i-\mu}{\sigma}),\ \xi = 0 \end{cases} \tag{3}$$

2.3. Penalized MLE (PMLE) / Generalized MLE

The penalized maximum likelihood or called the generalized maximum likelihood is an alternative method of standard MLE to avoid poor performance in small sample sizes. In this study, we consider two approaches of the panelized maximum likelihood estimators from [4] and [9]. These two methods introduced the penalty function to the standard method of the MLE.

2.3.1. PMLE 1

From [4], the penalized likelihood function is defined as:

$$L_{PMLE1}(\mu,\sigma,\xi) = L(\mu,\sigma,\xi) \times P(\xi) \tag{4}$$

where $L(\mu,\sigma,\xi)$ is the standard likelihood function of MLE from Eq. 3 and $P(\xi)$ is the penalty function for a range of non-negative value of α and λ which shown as below:

$$(\xi) = \begin{cases} 1,\ \xi \leq 0 \\ exp(-\lambda(\frac{1}{1-\xi}-1)^{\alpha}),\ 0<\xi<1 \\ 0,\ \xi \geq 1 \end{cases} \tag{5}$$

From [4], the suitable value of α and λ is equal to 1, therefore it can overcome the problem of small sample sizes in the MLE. A simulation study which conducted in [8] with a sample size of n=25 showed that PMLE almost better in both bias and root mean square with respect to probabilty weight moment.

2.3.2. PMLE 2

From [9], the penalty function defined as:

$$\pi(x) = (0.5+\xi)^{p-1}(0.5-\xi)^{q-1}/B(p,q) \tag{6}$$

The value of ξ is in the range of [-0.5, 0.5], with p=6 and q=9 where $B(p,q) = (\Gamma(p)\Gamma(q))/\Gamma(p+q)$. It has the mean of ξ = -0.1 and the variance is 0.015. This distribution is well behaved for small sample size since it has the smallest bias and smallest RMSE compare to other method [9].

2.4. Sandwich Estimator

The aforementioned methods maximize the likelihood function independently at each site that their statistical assumption of inter-dependency between sites being violated [7]. From [5], the estimates based on independence assumption said to misspecify if ignore the dependence. By using the sandwich estimator, the parameter values obtained independently are unchanged but the assymptotic variances will be corrected by using the following function:

Table 1. Information of each site.

Sites	1	2	3	4	5
Sites	Bonor	Kalumpun	Kemabong	Pangi Dam	Sook
No. of Years	33	33	33	27	33
Maximum Observation	135.0	157.5	115.0	156.0	144.0

$$Var(\hat{\theta}) = [I(\hat{\theta})]^{-1} J(\hat{\theta}) [I(\hat{\theta})]^{-1} \quad (7)$$

where $[I(\hat{\theta})] = -E \bigtriangledown^2 l(\hat{\theta})$ is the observed Fisher Information matrix that also defined as the second derivative of the log likelihood obtained from the Eq. 4 and 6. $E\bigtriangledown^2$ is the function of the expected value of hessian. $[I(\hat{\theta})]^{-1}$, the inverse function of this matrix will produce the covariance matrix under the independent assumption [7]. While $J(\hat{\theta})$ is the partial derivative of the log penalized likelihood function which appproximating the error in likelihood estimation. The score function for θ is the gradient, $\bigtriangledown$ of the log penalized likelihood, $l(\hat{\theta})$ with respect to θ can be obtained by the following function [7]:

$$J(\hat{\theta}) = \sum_{i=1}^{n} \bigtriangledown l(\hat{\theta})_i \bigtriangledown l(\hat{\theta})_i^T \quad (8)$$

2.5. Return Level

Let p be the probability of the extreme event, estimation of extreme quantiles of the annual maximum distribution are obbtained by inverting Eq. 1:

$$Z_p = \begin{cases} \mu - \frac{\sigma}{\xi}(log(1-p)^{-\xi} - 1), \ \xi \neq 0 \\ \mu - \sigma log[-log(1-p)], \ \xi = 0 \end{cases} \quad (9)$$

Z_p is the return level with associated with the return period $1/p$, the level Z_p is expected to exceed on average once every $1/p$ years [4].

3. Case Study: An Application to Rainfall Data in Sabah

This section discusses the results of modeling an annual maximum rainfall data of 5 selected sites using the method in Section 2. The 5 selected sites are Bonor, Kalampun, Kemabong, Pangi Dam and Sook in Sabah, Malaysia. These data are obtained from Hydrology and Survey Division under Department of Irrigation and Drainage, Sabah. Table 1 shows the number of years (n) and the maximum observation for the period of time at each site.

We fitted the generalized extreme value (GEV) distribution independently to each site. The recorded observation data (number of years) for these 5 sites all below 50, which consider as a small sample sizes. Since there are small sample issues by using MLE method [10], therefore, we applied the alternative method which is the penalty function of standard MLE to estimate the GEV parameters estimation which are the penalized maximum likelihood estimation, PMLE1 and PMLE2. Both of these PMLE methods well behaved in small sample sizes compared to MLE. When the data are modeled independently and ignored the dependency between sites, the model assumptions have been violated, and it may lead to wrong conclusion [7]. The sandwich estimator is then applied to correct the variances. The method still produced the similar value of parameter estimation, but with some modifications required on the standard error for the spatial extreme data. Table 2 shows the results of GEV parameter estimates and the standard error which were modified by using sandwich estimator.

This result is useful to predict the return value of an extreme rainfall of these 5 selected sites. The return value will be calculated by using p=0.01(100-years return level estimate). Table 3 shows the corresponding return value estimation for the 5 selected sites. Upon comparing to the annual maximum data in Table 1, for both of the PMLE, site 3 and site 5 are expected to exceed the annual maximum observations on average once in every 100-year.

4. Conclusions

When modelling the spatial annual maximum rainfall data by using the generalized extreme value (GEV) distribution independently, the model assumption been violated since the dependency between sites been ignored. To solve this problem, an alernative method of multivariate extreme value distribution which is sandwich estimator approahes for model the spatial extreme event was applied. Most of the method from multivariate extreme value distributon may creates high dimensionality difficulties for both model validation and computation. Therefore, sandwich estimator can be used to refrain the high dimensional validation and computation. In conclusion, the sandwich estimator is appropriate to model the spatial extreme data to capture the dependency data, which is also an appropriate method to model the spatial extreme rainfall data in this study.

5. Acknowledgements

The authors would like to thank the Hydrology Department of Sabah for providing the data and also Universiti

Table 2. Standard Error modified using Sandwich Estimator.

PMLE1			PMLE2		
Parameter	Estimations	Standard error (sandwich)	Parameter	Estimations	Standard error (sandwich)
μ_1	91.91	8.93	μ_1	91.63	8.26
σ_1	26.56	3.29	σ_1	26.20	3.03
ξ_1	-0.59	0.23	ξ_1	-0.57	0.23
μ_2	84.34	6.94	μ_2	84.10	6.79
σ_2	31.50	3.12	σ_2	31.25	2.95
ξ_2	-0.35	0.34	ξ_2	-0.34	0.35
μ_3	65.81	3.34	μ_3	65.60	3.36
σ_3	15.28	2.19	σ_3	15.12	2.09
ξ_3	-0.05	4.76	ξ_3	-0.02	10.19
μ_4	78.94	4.07	μ_4	78.82	4.11
σ_4	17.74	2.06	σ_4	17.66	2.03
ξ_4	-0.09	1.05	ξ_4	-0.08	1.28
μ_5	95.74	5.49	μ_5	95.15	5.30
σ_5	22.27	3.85	σ_5	21.66	3.43
ξ_5	-0.28	1.20	ξ_5	-0.24	1.47

Table 3. Return Value Estimation.

	Sites	Bonor	Kalumpun	Kemabong	Pangi Dam	Sook
Return	PMLE1	134.184	156.328	128.832	146.342	152.863
value	PMLE2	134.267	156.870	131.551	147.429	155.908

Malaysia Sabah for financially supporting this study under the UMSGreat Grant (GUG0209-1/2018).

References

[1] S. Yoon, B. Kumphon, and J. S. Park, (2015) *"Spatial modeling of extreme rainfall in northeast Thailand"* **Journal of Applied Statistics** ***42***(8): 1813–1828. DOI: 10.1080/02664763.2015.1010492.

[2] J. Tawn, R. Shooter, R. Towe, and R. Lamb, (2018) *"Modelling spatial extreme events with environmental applications"* **Spatial Statistics** ***28***: 39–58. DOI: 10.1016/j.spasta.2018.04.007.

[3] J. Blanchet and A. C. Davison, (2011) *"Spatial modeling of extreme snow depth"* **Annals of Applied Statistics** ***5***(3): 1699–1725. DOI: 10.1214/11-AOAS464.

[4] S. Coles and M. Dixon, (2000) *"Likelihood-Based Inference for Extreme Value Models"* **Extremes** ***2***(1): 5–23. DOI: 10.1023/A:1009905222644.

[5] D. Gabd and J. Tawn. "Inference for an extreme value model accounting for inter-site dependence". In: *AIP Conference Proceedings*. ***1830***. American Institute of Physics Inc., 2017, 70035. DOI: 10.1063/1.4980985.

[6] N. F. Musakkal and D. Gabda. "The sandwich estimator approach counting for inter-site dependence of extreme river flow in Sabah". In: *Journal of Physics: Conference Series*. ***890***. 1. Institute of Physics Publishing, 2017, 12148. DOI: 10.1088/1742-6596/890/1/012148.

[7] N. F. Kahal Musakkal, S. N. Chin, K. Ghazali, and D. Gabda, (2017) *"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. DOI: 10.11113/mjfas.v0n0.620.

[8] J. E. Morrison and J. A. Smith, (2002) *"Stochastic modeling of flood peaks using the generalized extreme value distribution"* **Water Resources Research** ***38***(12): 41–1–41–12. DOI: 10.1029/2001wr000502.

[9] E. S. Martins and J. R. Stedinger, (2000) *"Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data"* **Water Resources Research** ***36***(3): 737–744. DOI: 10.1029/1999WR900330.

[10] J. R. Hosking, J. R. Wallis, and E. F. Wood, (1985) *"Estimation of the generalized extreme-value distribution by the method of probability-weighted moments"* **Technometrics** ***27***(3): 251–261. DOI: 10.1080/00401706.1985.10488049.