ASM Sc. J., 13, 2020
https://doi.org/10.32802/asmscj.2020.sm26(1.12)

Modelling Extreme Rainfall Using Adjusted Sandwich Estimator

Nur Farhanah Kahal Musakkal* and Darmesah Gabda

Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Jalan UMS, 88450, Kota Kinabalu, Sabah

The Generalized Extreme Value (GEV) distribution is often used to describe the frequency of occurrence of extreme rainfall. Modelling the extreme event using the independent Generalized Extreme Value to spatial data fails to account the behaviour of dependency data. However, the wrong statistical assumption by this marginal approach can be adjusted using sandwich estimator. In this paper, we used the conventional method of the marginal fitting of generalized extreme value distribution to the extreme rainfall then corrected the standard error to account for inter-site dependence. We also applied the penalized maximum likelihood to improve the generalized parameter estimations. A case study of annual maximum rainfall from several stations at western Sabah is studied, and the results suggest that the variances were found to be greater than the standard error in the marginal estimation as the inter-site dependence being considered.

Key words: Generalized Extreme Value theory, sandwich estimator, penalized maximum likelihood, annual maximum rainfall

I. INTRODUCTION

Extreme value analysis is often applied to modelling environmental spatial data. According to Cooley et al., (2012) spatial data follow the condition of multivariate extreme value distribution since they are obtained from multiple locations. Unfortunately, modelling and computation in multivariate Extreme Value Theory (EVT) often difficult to handle compared to the univariate EVT. More discussion of multivariate EVT is studied by Tawn (1990), Coles (2001), Wadsworth and Tawn (2012).

For spatial extreme, there are a few approaches that can be used to explain the behaviour of extreme frequency while capturing the dependency between sites. Cooley *et al.,* (2012) studied the methodology for spatial extremes analysis and examine max stable process model and copula approaches for modelling spatial dependence after accounting for marginal effects. It was concluded that to fit max stable process model using the data recorded at many locations remains a challenge. The same finding was drawn by Huser and Genton (2016) where the analysis was conducted to fit the temperature dataset recorded in Colorado using the combining max-stable processes and non-stationary correlation functions. They also found that the problem of building and fitting sensible non-stationary dependence models for spatial extremes is required more effort and attention. Describing the inter-dependence structure of spatial extreme phenomena is particularly challenging (Huser & Genton, 2016). The approach used in this study was based on Smith model (1990) and was also illustrated by Gabd and Tawn (2017). This method constructed based on the wrong statistical assumption and proposed for obtaining adjusted standard error which allow for the empirical spatial dependence.

Maximum Likelihood is an estimator that having good asymptotic properties that can be applied to complex modelling situation such as temporal dependence (Nadarajah & Shiau (2005); Nadarajah & Choi (2007)). However, the performance of Maximum Likelihood Estimator (MLE) require special attention when the study involved small

*Corresponding author's e-mail: farhanah.kahal@gmail.com

sample size extreme data. Detailed study about MLE on small sample size can be obtained through Hosking *et. al.,* (1985), Coles and Dixon (1999). Therefore, the objective of this study is to use the sandwich estimator to model an extreme rainfall at several sites in Sabah. Since it is involving a small sample sizes at each sites, the Penalized Maximum Likelihood (PML) was considered to estimate the generalized extreme value parameters.

II. MATERIALS AND METHODS

A multiple rainfall station in the western Sabah was selected for this study as illustrated in Table 1. Rainfall data in millimetres (mm) were obtained from the Sabah Drainage and Irrigation Department. Total number of rain gauge station in Sabah are 74 stations. The interest is on the station that close to another station and with at least 20 years of observations. As a result, only 16 rainfall stations were selected.

Table 1. List of rainfall data events used in this study

Station Name	Station No.	Latitude	Longitude	No. of observation	Years Observation	Maximum Rainfall (mm)
Bonor	4961001	4.9698	116.1768	33	1985-2017	135
Sook	5163002	5.1479	116.3012	33	1985-2017	144
Kemabong	4959001	4.9178	115.92	33	1985-2017	115
Lanas	5364003	5.3349	116.4972	27	1991-2017	113.8
Keningau	5361002	5.3455	116.1595	33	1985-2017	165.5
Tulid	5364002	5.3224	116.4211	33	1985-2017	265.8
Beaufort	5357003	5.3539	115.7242	33	1985-2017	208
Kalampun	5060001	5.0703	116.1231	33	1985-2017	157.5
Bongawan	5558001	5.5187	115.8733	33	1985-2017	216.5
Tongod	5269001	5.2712	116.9709	33	1985-2017	189
ApinApin	5462001	5.4789	116.2658	33	1985-2017	141
Ulu Moyog	5862002	5.8714	116.2504	33	1985-2017	248.2
Tambunan	5663001	5.6298	116.3246	33	1985-2017	93
Sinua	5465001	5.4826	116.5775	33	1985-2017	173.5
Pangi	5158001	5.1329	115.8741	25	1993-2017	156
Bukit Mondou	6172001	6.1962	117.2375	27	1991-2017	287.5

The study of frequency maximum rainfall is useful especially for agricultural planning. In this study, the GEV distribution is used to model the rainfall data. The cumulative distribution function of GEV is:

$$F(x) = exp\left\{-\left(1+\xi\left[\frac{x-u}{\sigma}\right]\right)^{-1/\xi}\right\} \quad (1)$$

which 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(Coles, 2001).The GEV distribution has support on the set $\{x: 1 + \xi(\mathrm{x} - \mu)/\sigma > 0\}$. According to Gumbel (1960), there are three families of extreme value distributions that can be combined in the single three-parameter family of GEV distributions. It is based on the values of the shape parameters. For $\xi = 0$, it is following Gumbel distribution (taken as $\xi \longrightarrow 0$), while for $\xi < 0$ and $\xi > 0$ it is following the Negative Weibull distribution and the Frechet distribution respectively. To avoid biased fit, it is appropriate to used GEV distribution instead of choosing directly one of the extreme value distributions. GEV distribution allows for uncertainty in the selection of the three different types. More detailed about Univariate GEV can be obtained from Coles (2001) and Haan and Ferreira (2006).

For the analysis, we conduct a marginal estimation in which we fit the GEV independently to each stations of an extreme rainfall at Sabah. Suppose each$j = 1, \dots, d$ where d is the 16 stations consist of n year observation and let the extreme values data are independent over years. Therefore, the likelihood function is:

$$L(\theta; x) = \prod_{i=1}^{n} \prod_{j=1}^{d} f(x_{ij}; \theta) \tag{2}$$

where $f(.) = dF(x)/dx$ is the density function of the GEV. The corresponding log likelihood function is as follows:

$$\ell(\theta; x) = \sum_{i=1}^{n} \sum_{j=1}^{d} log\ f(x_{ij}; \theta) \tag{3}$$

In this study penalty function $\mathrm{P}(\xi)$ (in logarithm scale) is used to provide the likelihood with the information that the value of ξ is smaller that 1 (Coles and Dixon, 1999) as follows:

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

The corresponding Penalized Maximum Likelihood (PML) estimator will be used for parameter estimation is:

$$L_{pen} = \ell(\theta; x) \times \mathrm{P}(\xi) \tag{5}$$

The PML is an alternative estimation method for a small sample sizes of an extreme event that can improve the tail behaviour. As proposed by Smith (1990), the sandwich estimator is used to modify the asymptotic variance that captured the data dependency as follows:

$$Vars(\hat{\theta}) = \left[I(\hat{\theta})\right]^{-1} J(\hat{\theta}) \left[I(\hat{\theta})\right]^{-1} \tag{6}$$

$$J(\hat{\theta}) = \sum_{i=1}^{n} \nabla\ell(\hat{\theta})_i \nabla\ell(\hat{\theta})_i^{\ T} \tag{7}$$

where $\left[I(\hat{\theta})\right] = -\mathrm{E}\nabla^2\ell(\hat{\theta})$ is the second derivative of the Penalized log likelihood, $\mathrm{E}\nabla^2$is the expected values of hessian, $\left[I(\hat{\theta})\right]^{-1}$is the inverse of hessian matrix produce covariance matrix under the independent assumption, $J(\hat{\theta})$ is the partial derivative of the log Penalized Likelihood function, $\nabla$is gradient of the log Penalized Likelihood, $\ell(\hat{\theta})$is log Penalized Likelihood.

III. RESULTS AND DISCUSSIONS

Quantile-Quantile (Q-Q) plot with 95% tolerance intervals was used to determine the suitability of the GEV model using PML method for the real rainfall data series at different location. The plot show that the GEV fit for all station. Figure 1 show well fitted of Q-Q plot with 95% tolerance intervals for four rainfall stations. Since there are 16 stations, so the stations randomly picked to illustrate the well fitted Q-Q plot.

According to the number of station (16 stations) and the parameters of GEV distribution, the covariance matrix produced a 48 $\times$ 48 matrix (16$\times$3=48).

$$\begin{bmatrix} \mu_1 & \mu_1\sigma_1 & \mu_1\xi_1 & & \mu_1\mu_{16} & \mu_1\sigma_{16} & \mu_1\xi_{16} \\ \mu_1\sigma_1 & \sigma_1 & \sigma_1\xi_1 & \cdots & \mu_1\sigma_{16} & \sigma_1\sigma_{16} & \sigma_1\xi_{16} \\ \mu_1\xi_1 & \sigma_1\xi_1 & \xi_1 & & \mu_1\xi_{16} & \sigma_1\xi_{16} & \xi_1\xi_{16} \\ & \vdots & & \ddots & & \vdots & \\ \mu_1\mu_{16} & \mu_1\sigma_{16} & \mu_1\xi_{16} & & \mu_{16} & \mu_{16}\sigma_{16} & \mu_{16}\xi_{16} \\ \mu_1\sigma_{16} & \sigma_1\sigma_{16} & \sigma_1\xi_{16} & \cdots & \mu_{16}\sigma_{16} & \sigma_{16} & \sigma_{16}\xi_{16} \\ \mu_1\xi_{16} & \sigma_1\xi_{16} & \xi_1\xi_{16} & & \mu_{16}\xi_{16} & \sigma_{16}\xi_{16} & \xi_{16} \end{bmatrix}_{48 \times 48}$$

Table 2 below shows the results of the estimated standard error of GEV parameter based on the marginal approach and adjusted standard error using sandwich estimator as shown in equation 6. The marginal approach gives an underestimate of standard error. The size of the correction in the variances by sandwich estimator is increases as the data dependency increases (Gabd & Tawn, 2017). This means the inter-dependency between sites are successfully considered and the underestimate standard error is corrected. Hence the wrong choice due to the underestimate standard error for example the design rainfall of hydraulic infrastructure which may cause to infrastructure failures and other negative consequences are managed to avoid. The result obtained from Table 2 is useful to predict the accurate return level of

an extreme rainfall in western Sabah. A common of location and scale parameters of GEV can be performed for further analysis.

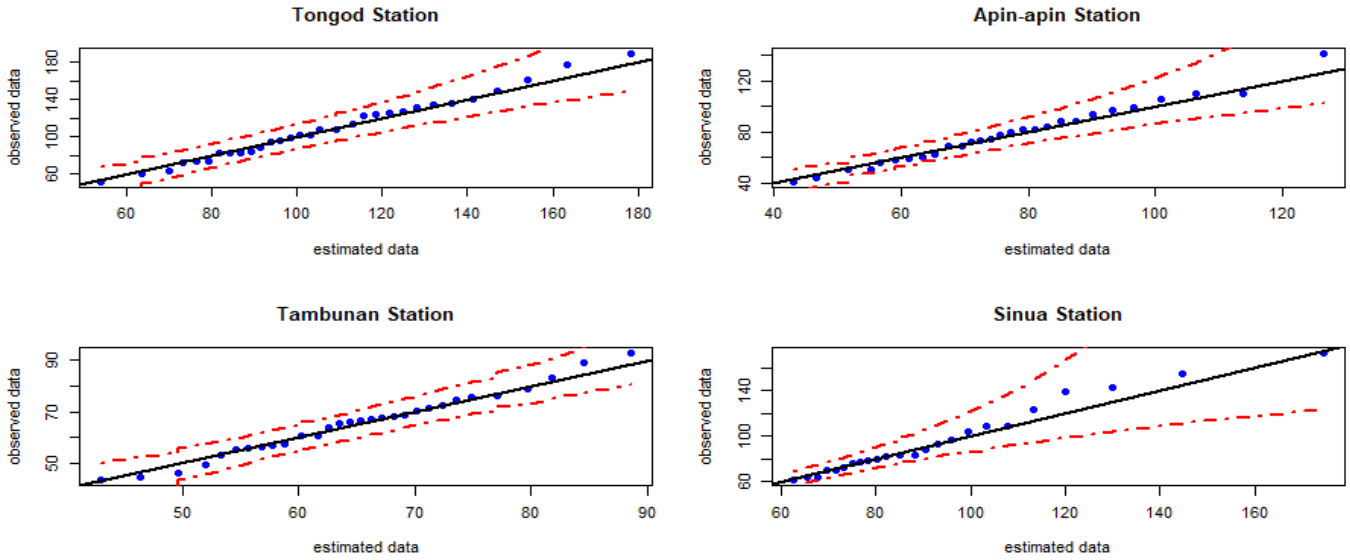

Figure 1. Q-Q plot with 95% tolerance interval shows well fit of GEV distribution for annual maximum rainfall at several stations in Sabah

Table 2. Adjusted Standard Error using Sandwich Estimator

	Std error of marginal distribution	Adjusted std error		Std error of marginal distribution	Adjusted std error
μ_1	4.9580	8.9315	μ_9	5.5401	6.0670
σ_1	3.8167	3.2838	σ_9	3.8853	3.1150
ξ_1	0.1029	0.2266	ξ_9	0.1078	0.5970
μ_2	4.6779	5.2282	μ_{10}	5.7802	6.2828
σ_2	3.7771	3.8498	σ_{10}	4.2728	3.7298
ξ_2	0.2031	1.2022	ξ_{10}	0.1414	1.5233
μ_3	3.0859	3.5815	μ_{11}	3.4824	3.9781
σ_3	2.3372	2.1929	σ_{11}	2.6317	2.4154
ξ_3	0.1575	4.7716	ξ_{11}	0.1435	14.7979
μ_4	3.1486	3.3475	μ_{12}	5.4408	6.4260
σ_4	2.3977	2.2604	σ_{12}	3.9832	3.3380
ξ_4	0.1774	0.9375	ξ_{12}	0.1099	2.5597
μ_5	2.2753	3.6554	μ_{13}	2.2949	2.4523
σ_5	1.8488	1.8308	σ_{13}	1.6491	1.4380
ξ_5	0.1234	0.7685	ξ_{13}	0.1198	0.5662
μ_6	4.1093	5.2051	μ_{14}	4.0491	5.4681
σ_6	3.0653	2.6415	σ_{14}	3.4852	3.7165
ξ_6	0.1005	1.0791	ξ_{14}	0.1814	1.0962
μ_7	3.7546	5.4463	μ_{15}	3.7319	4.1616
σ_7	3.5282	4.2831	σ_{15}	2.5781	2.0632
ξ_7	0.1848	0.8236	ξ_{15}	0.0986	1.0457

μ_8	5.9317	7.6559	μ_{16}	9.1693	12.0121
σ_8	4.0686	3.1203	σ_{16}	7.5736	7.9229
ξ_8	0.0946	0.3406	ξ_{16}	0.2049	4.4729

IV. CONCLUSIONS

This paper has demonstrated a penalty function introduced by Coles and Dixon (1999) added to the standard maximum likelihood method. By using the GEV distribution, the annual maximum rainfall is modelled independently at each site. Since this method violated the statistical assumption of spatial environmental data, the sandwich estimator applied in order to correct the variances of GEV parameters. This is an alternative method to the multivariate extreme value distribution approaches for the spatial extreme value modelling. The size of the correction in the variance is increases as the data dependency is being considered as mentioned in study conducted by Gabd and Tawn (2017). The implement of sandwich estimator in this study helps to avoid high dimensional of mathematical computation.

Therefore, it can conclude that the sandwich estimator is an appropriate method to model the spatial extreme rainfall in Sabah. A similar framework may be useful in fitting probability model for maximum rainfall in other parts of the region.

V. ACKNOWLEDGEMENTS

The authors would like to express sincere gratitude to the Sabah Drainage and Irrigation Department for providing the rainfall data. This research is funded by Research Grant SBK0263-SG-2016.

VI. REFERENCES

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

Coles, S.G. 2001, An Introduction to Statistical Modeling of Extreme Values. Springer.

Cooley, D. & Sain, S.R. 2010, Spatial hierarchical modeling of precipitation extremes from a Regional Climate Model. Journal of Agricultural, Biological, and Environmental Statistics, 15(3), 381-402.

Gabd, D. & Tawn, J.A. 2017, Inference for an extreme value model accounting for inter-site dependence. AIP Conference Proceedings.

Gabda, D., Towe, R., Wadsworth, J.L. & Tawn, J.A. 2012, Discussion of Statistical Modeling of Spatial Extremes. Statistical Science, 27(2), 189-192.

Gumbel, E.J. 1960, Multivariate extreme distributions. Bulletin of the International Statistical Institute, 39(2), 471-475.

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(3), 251-261.

Huser, R. & Genton, M.G. 2016, Nob-stationary dependence structures for spatil extremes. Journal of agricultural, biological and environmental statistics, 21(3), 470-491.

Musakkal, K. N. F. Chin, S.N., Ghazali, K. & Gabd, D. 2017, A penalized likelihood approach to model the annual maximum flow with small sample sizes. *Malaysian Journal of Fundamental and Applied Sciences*. Vol:3. 4.

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. Volume 19, Issue 4, pp 363–374.

N F Kahal Musakkal and D Gabd 2017, J. Phys.: Conf. Ser. 890.

Smith, R. (1990), Max-stable processes and spatial extremes. Unpublished manuscript.

Tawn, J.A. 1990, Modelling multivariate extreme value distributions. Biometrika, 77, 245-253.

Wadsworth, J. L., & Tawn, J. A. 2012, Dependence modelling for spatial extremes. *Biometrika*, *99*(2), 253–272.