The sandwich estimator approach counting for inter-site dependence of extreme river flow in Sabah

N F Kahal Musakkal and D Gabda

Department of Mathematics & Economics, Faculty of Sciences and Natural Resources, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia

E-mail: nurfarhanahkahalmusakkal@ymail.com

Abstract. Regional estimation method is often used for estimating parameters of a distribution when data are available at many sites in a region to capture inter-site dependence. In this paper, we fit generalized extreme value distribution independently to model data of river flows at each sites in Sabah to avoid extreme value complex modeling. Since our approach violated the condition of spatial analysis, we consider the adjusted standard error to correct the wrong assumption of our marginal approach. As a result, we have an appropriate corrected variance of the generalized extreme value parameters.

1. Introduction

Spatial extreme analysis is always considered when an extreme data obtained from several sites in a region [1]. Well known examples of extreme value analysis for modelling environmental dependence data can be found in [1] and [2]. In extreme value theory, the multivariate extreme value distribution is the most suitable method to analyse the extreme events in several locations because the number of location is referred as a multivariate variable as mentioned by [3] and [4] in their studies. This is because one of the properties of multivariate extreme value distribution is for modelling the dependency between variables. However, this method may lead to model complexity and computational issues.

Another approaches for spatial extreme modelling that takes the dependency into account is the joint estimation which proposed by [5]. A motivation of this study was followed the study by [6] which proposed a method that using univariate extreme model independently at each site of the spatial extreme event. The likelihood functions were applied independently at each site. This method constructed based on the wrong statistical assumption as the data dependency between sites is being ignored. An alternative method proposed by [6] was to adjust the standard error of parameters to capture the data dependency. The value of parameters estimates obtained from this method were remains unchanged. However, some modifications on the asymptotic variance which known as the sandwich estimator need to be considered. The advantages and properties of the sandwich estimator have been discussed in [7]. In this study, we will apply the sandwich estimator to model an extreme river flow at several sites in Sabah. We also consider an analysis of small sample sizes of the observed extreme at each site. This study is an extension of the study by [8].

2. Methodology
2.1. Univariate Extreme Value Distribution

The Generalized Extreme Value (GEV) distribution function is:

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

where

\( \mu = \) location parameter (\( \mu \in \mathbb{R} \))
\( \sigma = \) scale parameter (\( \sigma > 0 \)) and
\( \xi = \) shape parameter (\( \xi \in \mathbb{R} \)).

GEV distribution is always used for data that is available in block of maximum or minimum. The block is separate in equal length of time. For example, data is separated in yearly, monthly, weekly or daily. Shape parameter is the most important parameter in GEV family. GEV family consisted of three distributions as mentioned in table 1.

| Shape parameter | GEV Family         |
|-----------------|--------------------|
| \( \xi = 0 \)   | Gumbel distribution|
| \( \xi < 0 \)   | Negative Weibull distribution |
| \( \xi > 0 \)   | Frechet distribution |

According to \cite{2}, GEV distribution is highly recommended for model fitting under extremes event scenario. This is because directly pick one distribution among the GEV family may cause a biased fit.

2.2. Marginal Estimation

A marginal estimation using independent GEV distribution to each sites of an extreme river flow at Sabah is conducted for data analysis purpose. Suppose we have sites \( j = 1, \ldots, d \) consists of \( i = 1, \ldots, n \) years observation by assuming extreme values data are independent over the years. Therefore the likelihood function can be expressed as follows;

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

where \( f(x_{ij}; \theta) = \frac{dF(x)}{dx}. \) The corresponding log likelihood function is as follows;

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

In this study, penalty function \( P(\xi) \) is used to provide the likelihood with the information that the value of \( \xi \) is smaller than 1 as proposed by \cite{3} as follows:

\[ L_{pen} = \ell(\theta; x) \times P(\xi) \]  

This method is an alternative method to the standard maximum likelihood method for a case that involved small sample size of an extreme event. This method is known as penalized maximum likelihood estimation (PMLE).

2.3. Accounting for dependence
The method proposed in Section 2.2 is based on maximizing a likelihood function independently at each site which violated the statistical assumption of inter-dependency between sites. As proposed by [6], the sandwich estimator is used to modify the asymptotic variance that captured the data dependency as follows:

\[
\text{Var} s(\theta) = [I(\theta)]^{-1}J(\theta)[I(\theta)]^{-1}
\]  \hspace{1cm} (4)

where \([I(\theta)] = -\mathbb{E}[\nabla^2 \ell(\theta)]\) is defined as the second derivative of the log likelihood obtained from the equation (3) which also known as the observed Fisher Information matrix. The \(\mathbb{E}[\nabla^2 \ell(\theta)]\) is called as the expected values of hessian. The inverse of this matrix will produced covariance matrix under the independent assumption, which is equal to \([I(\theta)]^{-1}\). While for \(J(\theta)\) is the partial derivative of the log penalized likelihood function which approximating the error in likelihood estimation. The score function for \(\theta\) is the gradient, \(\nabla\) of the log penalized likelihood, \(\ell(\theta)\) with respect to \(\theta\) which can be obtained as follows;

\[
J(\theta) = \sum_{i=1}^{n} \nabla\ell(\theta)^{T} \nabla\ell(\theta)
\]  \hspace{1cm} (5)

3. Case study: An application to river flow data

This section discussed the results of modelling an annual maximum river flow at different site in Sabah. Suppose \(X_{ij}\) be an annual maximum river flow at sites \(j = 1, \ldots, d = 18\). We conduct the model fitted using Generalized Extreme Value (GEV) distribution into each data, \(X_{ij} \sim \text{GEV}(\mu_j, \sigma_j, \xi_j)\). Data recorded for all sites below 50 (< 50). Since this study consists of data recorded below than 50 years, therefore an appropriate method should be considered for parameter estimation of the GEV distribution. Therefore, the GEV parameters are estimated using the PMLE method as stated in equation 3. As a result, the estimation of positive shape parameter will approach zero which improve the tail behaviour as claimed by [8]. As for negative value of \(\xi\), the shape parameter estimation of PMLE method is close to the shape parameter estimation of maximum likelihood estimator (MLE). The results obtained in this study is in line with the previous research where the estimation of PMLE method shrinks the positive shape parameter towards zero. The illustration of shape parameter estimated obtained using MLE and PMLE can be seen in figure 1 as shown by [8].

Figure 1. Density plot of shape parameter estimated obtained using MLE and PMLE.

Since the model assumptions are violated by ignoring the dependence between sites, the sandwich estimator is then applied to correct the model. This method still produced the same values of parameter
estimates as marginal estimates, but with some modifications required on the standard error for the spatial extreme data. Table 2 below shows the results of the GEV parameter estimates and the standard error. This result is useful to predict the return value of an extreme river flow at each site in Sabah.

| Parameter | Estimate | Standard Error (sandwich) | Parameter | Estimate | Standard Error (sandwich) |
|-----------|----------|---------------------------|-----------|----------|---------------------------|
| $\mu_1$   | 282.45   | 45.94                     | $\mu_{10}$ | 654.51   | 38.83                     |
| $\sigma_1$| 225.44   | 37.75                     | $\sigma_{10}$ | 203.29   | 26.01                     |
| $\xi_1$   | 0.27     | 0.88                      | $\xi_{10}$ | 0.12     | 1.54                      |
| $\mu_2$   | 114.36   | 28.28                     | $\mu_{11}$ | 122.43   | 11.42                     |
| $\sigma_2$| 80.71    | 14.96                     | $\sigma_{11}$ | 43.26    | 4.67                      |
| $\xi_2$   | 0.26     | 0.82                      | $\xi_{11}$ | 0.12     | 0.90                      |
| $\mu_3$   | 1257.71  | 84.35                     | $\mu_{12}$ | 254.77   | 22.99                     |
| $\sigma_3$| 353.20   | 39.59                     | $\sigma_{12}$ | 78.16    | 11.49                     |
| $\xi_3$   | 0.10     | 1.26                      | $\xi_{12}$ | 0.30     | 0.61                      |
| $\mu_4$   | 1265.62  | 112.98                    | $\mu_{13}$ | 221.87   | 23.36                     |
| $\sigma_4$| 369.14   | 46.40                     | $\sigma_{13}$ | 90.87    | 12.35                     |
| $\xi_4$   | -0.28    | 0.52                      | $\xi_{13}$ | 0.15     | 1.36                      |
| $\mu_5$   | 886.71   | 83.80                     | $\mu_{14}$ | 455.01   | 32.44                     |
| $\sigma_5$| 359.31   | 58.35                     | $\sigma_{14}$ | 143.08   | 17.99                     |
| $\xi_5$   | 0.30     | 0.72                      | $\xi_{14}$ | -0.17    | 0.96                      |
| $\mu_6$   | 941.14   | 105.06                    | $\mu_{15}$ | 130.64   | 11.65                     |
| $\sigma_6$| 423.70   | 46.27                     | $\sigma_{15}$ | 52.13    | 5.65                      |
| $\xi_6$   | 0.07     | 2.26                      | $\xi_{15}$ | -0.07    | 2.15                      |
| $\mu_7$   | 221.07   | 18.79                     | $\mu_{16}$ | 680.89   | 115.04                    |
| $\sigma_7$| 88.95    | 10.87                     | $\sigma_{16}$ | 403.36   | 59.71                     |
| $\xi_7$   | -0.27    | 0.55                      | $\xi_{16}$ | 0.19     | 0.85                      |
| $\mu_8$   | 860.12   | 53.74                     | $\mu_{17}$ | 477.99   | 107.94                    |
| $\sigma_8$| 287.54   | 23.89                     | $\sigma_{17}$ | 278.50   | 55.73                     |
| $\xi_8$   | -0.05    | 1.27                      | $\xi_{17}$ | -0.60    | 0.48                      |
| $\mu_9$   | 826.37   | 43.46                     | $\mu_{18}$ | 115.91   | 50.69                     |
| $\sigma_9$| 219.22   | 21.99                     | $\sigma_{18}$ | 46.58    | 5.26                      |
| $\xi_9$   | -0.19    | 0.55                      | $\xi_{18}$ | 0.22     | 0.00                      |

4. Conclusions

In this study the method which was proposed by [9] is applied in order to improve the inference of GEV parameters estimates known as the penalized maximum likelihood estimates. The result obtained is
consistent with [9] which also found the shape parameter estimates shrink towards zero. In this method, a penalty function is introduced which then added to the standard maximum likelihood method. There are another existing penalty functions (for example [10]) with different function such as a smoothing parameter [3] and smoothing function to capture the behaviour of the spatial events [11].

By using the GEV distribution, the annual maximum river flow independently at each site is modelled. As mentioned earlier this method violated the statistical assumption of dependency between sites. We applied the sandwich estimator 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 studies that used multivariate extreme value distribution for model the spatial extreme event can found in [3] and [4]. However, this method may lead to computational issues due to the high dimensionality. The size of the correction in the variance is increases as the data dependency is being considered as mentioned in study conducted by [7]. 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 river flow in Sabah.

References
[1] Buishand T A 1984 Journal of Hydrology 69 p 77-95
[2] Coles S G 2001 An introduction to statistical modeling of extreme values (Springer. Berlin Heidelberg, New York)
[3] Coles S G and Tawn J A 1991 Journal of the Royal Statistical Society 53 377-392
[4] Dixon M J and Tawn J A 1999 Applied Statistic 48 135-151.
[5] Buishand T A 1991 Journal of Hydrology 36 p 345-365
[6] Smith R L 1990 Regional Estimation from spatially dependent data (Unpublished)
[7] Gabd D and Tawn J A 2017 Inference for an extreme value model accounting for inter-site dependence AIP Conference Proceedings
[8] Musakkal K N F and Gabd D 2017 Malaysian Journal of Fundamental and Applied Sciences
[9] Coles S G and Dixon M J 1999 Extremes 2(1) 5-23
[10] Lian S and Duan Y 2016 Journal of Inequalities and Applications 1-12
[11] Cançado A L, Duarte A R, Duczmal L H, Ferreira S J, Fonseca C M and Gontijo E 2010 International Journal of Health Geographics 9 55