Journal of Hydrology 530 (2015) 384–391

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Adjusting for small-sample non-normality of design event estimators under a generalized Pareto distribution

Fahim Ashkar *, Salah-Eddine El Adlouni [1]

Department of Mathematics and Statistics, Université de Moncton, Moncton, NB E1A 3E9, Canada

ARTICLE INFO

Article history:
Received 12 September 2015
Accepted 28 September 2015
Available online 9 October 2015
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Attilio Castellarin, Associate Editor

Keywords:
Generalized Pareto distribution
Design events
Confidence intervals
Maximum likelihood
Small samples

SUMMARY

The generalized Pareto distribution (GPD) is a widely used frequency model for fitting extremes in hydrology, especially to fit exceedances over a threshold in the peaks-over-threshold (POT) modeling of floods or other extreme hydrological phenomena. A key goal in fitting frequency distributions to data is to allow the estimation of distribution quantiles, which in hydrology are often used as "design events". The maximum likelihood (ML) method is a recommended method for fitting the GPD to data. To provide a measure of the statistical error involved in the estimation of design events, confidence intervals for quantiles (CIQs) have to be calculated. Hydrologists have traditionally used large-sample theory to construct such CIQs, but it is shown in the present study that this leads to inaccurate results for quantiles in the right-tail of a GPD. An improvement is therefore proposed for these classically obtained CIQs under a GPD model fitted by ML. The conventional and proposed approaches are compared through Monte Carlo (MC) simulation, and the resulting recommendations are put to use in a hydrological application.

1. Introduction

The generalized Pareto distribution (GPD), introduced by Pickands (1975), is a widely used frequency model for fitting extremes in hydrology. Among its principal uses is in the fitting of exceedances over a threshold in the peaks-over-threshold (POT) modeling of floods or other extreme hydrological phenomena (Davison, 1984; Smith, 1984; Van Montfort and Witter, 1986; Hosking and Wallis, 1987; Rosbjerg et al., 1992; Rasmussen et al., 1994; Ashkar and Ouarda, 1996, among others). More generally, the GPD is used in modeling data with a right tail and no mode in the probability density. Ashkar et al. (2004) also used the GPD to fit low stream flows below a threshold in over 100 Canadian hydrometric stations. The GPD model is also used in several other areas not related to hydrology.

The probability density function (PDF) of a GPD variable, X, is

$$f(x)=\begin{cases}\frac{1}{\beta}\left(1-\alpha\frac{x}{\beta}\right)^{1/\alpha-1} & \alpha\neq 0\\ \frac{1}{\beta}e^{-x/\beta} & \alpha=0\end{cases} \quad (1)$$

where $\beta>0$ is a scale parameter and α is a shape parameter. The range of X is $0\leqslant x<\infty$ for $\alpha<0$ and $0\leqslant x\leqslant\beta/\alpha$ for $\alpha>0$. It is noted that when $\alpha>0$, the sample space of X is a finite interval with β/α as upper bound. In this case, the GPD is short-tailed. When $\alpha<0$, the GPD has a long tail thicker than that of the exponential distribution [which corresponds to $\alpha=0$ (taken as a limit)]; the distribution in this case is sometimes simply called *Pareto* or *Pareto Type 1*.

The cumulative distribution function (CDF) corresponding to Eq. (1) is

$$F(x;\beta,\alpha)=\begin{cases}1-\left(1-\alpha\frac{x}{\beta}\right)^{1/\alpha} & \alpha\neq 0\\ 1-\exp\left(-\frac{x}{\beta}\right) & \alpha=0\end{cases} \quad (2)$$

The methods of maximum likelihood (ML), of moments (MM) and of probability weighted moments (PWM) are the principal methods proposed for fitting the GPD. When the distribution is short tailed ($\alpha>0$), it was observed by Dupuis (1996) and by Ashkar and Nwentsa Tatsambon (2007) that the MM and the PWM methods may produce estimates of the upper bound of X that are inconsistent with the observed data. The inconsistency occurs when one or more sample observations exceed the estimated upper bound $\left(\hat{\beta}/\hat{\alpha}\right)$ of X. The ML method has the advantage of not suffering from this inconsistency with the data (Ashkar and Nwentsa Tatsambon, 2007). This method is also needed in many

* Corresponding author. Tel.: +1 (506) 858 4312; fax: +1 (506) 858 4541.
E-mail addresses: ashkarf@umoncton.ca (F. Ashkar), salah-eddine.el.adlouni@umoncton.ca (S.-E. El Adlouni).
[1] Tel.: +1 (506) 858 4253; fax: +1 (506) 858 4541.

http://dx.doi.org/10.1016/j.jhydrol.2015.09.071
0022-1694/

inference procedures such as when using the Akaike (AIC) or Bayesian (BIC) information criteria, or in applying likelihood ratio tests. For these reasons, the ML fitting method will be the one to focus on in the present study.

In hydrology, a key goal in fitting frequency distributions to data is to allow the estimation of distribution quantiles, i.e. the $100\times$ pth percentiles, by using the quantile function of X. The estimated quantiles play a key role in hydrologic design and risk analysis; hence they are commonly called "design event". Let the pth quantile of a $GPD(\alpha, \beta)$ variable be denoted by $Q(p; \alpha, \beta)$. This quantile is easily obtained by inverting the CDF of Eq (2), i.e. by substituting $p = F(x; \alpha, \beta)$, $0 \leqslant p < 1$ and calculating x as a function of p, which gives:

$$Q(p;\alpha,\beta) = \begin{cases} (\beta/\alpha)\left[1-(1-p)^{\alpha}\right] & \alpha \neq 0 \\ -\beta \ln(1-p) & \alpha = 0 \end{cases} \tag{3}$$

where $\ln(\cdot)$ denotes natural logarithm. Denoting by α_0 and β_0 the *true* parameter values of X, the pth quantile $Q(p; \alpha_0, \beta_0)$ represents the event under a GPD that is exceeded with probability $1 - p$.

In order for the quantile function $Q(p; \alpha_0, \beta_0)$ to be useful in practice, it has to be estimated from the data. In this study, the data is assumed to be in the form of a sample $\{x_i\}_{i=1}^n$ of independent and identically distributed (iid) observations. When the parameter vector (α_0, β_0) is estimated by $(\hat{\alpha}, \hat{\beta})$ and then plugged into Eq. (3), it yields the quantile (design event) estimator $Q_n(p; \hat{\alpha}, \hat{\beta})$.

Following the estimation of the design event $Q(p; \alpha_0, \beta_0)$, it is essential to provide a measure of the statistical error involved in the estimation. This is commonly done by constructing confidence intervals for the quantile (CIQs) under the chosen model and fitting method. Hydrologists have traditionally used *large-sample theory* to construct confidence intervals for design events. However, it will be shown in the present study that such CIQs are very inaccurate for quantiles in the right-tail of a GPD. The goal will therefore be to improve these classically obtained CIQs. We will begin by presenting the large-sample variance–covariance matrix of the ML parameter estimators (MLEs) $(\hat{\alpha}, \hat{\beta})$ and then review how this has traditionally been used to derive large-sample CIQs for the GPD.

The study is organized as follows. Section 2 briefly discusses GPD parameter estimation by ML and presents the basic asymptotic properties of the MLEs. Section 3 presents the approximate sampling distribution of the GPD quantile estimators. The most commonly used approach for calculating CIQs is then revisited, and an improvement to this approach is proposed. These two approaches will be referred to as the *conventional* and the *adjusted* approaches, respectively. Then, in Section 4, the conventional and adjusted approaches are compared through Monte Carlo (MC) simulation. In Section 5 the recommendations resulting from the MC simulations are put to use in a hydrological application. Finally, Section 6 presents the paper's main conclusions and presents some future research ideas.

2. Calculation of MLEs and their asymptotic properties under a GPD model

Let $\mathbf{X}_n = \{X_i\}_{i=1}^n$ denote a random sample of n iid observations from a $GPD(\alpha, \beta)$ model. A *specific observed* sample (i.e., the observed dataset) will be denoted by $\{x_i\}_{i=1}^n$. Let $x_{n:n}$ denote the largest value of the observed sample. The log-likelihood function is given by

$$\begin{aligned} l(\alpha,\beta|\mathbf{X}_n) &= \sum_{i=1}^{n} \ln\left[f(X_i;\alpha,\beta)\right] \\ &= -n\ln\beta + \frac{1-\alpha}{\alpha}\sum_{i=1}^{n}\ln\left(1-\alpha X_i/\beta\right) \end{aligned} \tag{4}$$

The ML estimation algorithm employed in the present study is one proposed by Davison (1984) and subsequently used by Choulakian and Stephens (2001). In this algorithm, a change of parameters is made: $\theta = \alpha/\beta$; $\alpha = \alpha$, so that the two-dimensional search for an ML solution $(\hat{\alpha}, \hat{\beta})$ is reduced to a one-dimensional search of the value $\hat{\theta}$ that maximizes the "profile log-likelihood function" $L^*(\theta)$ (Choulakian and Stephens, 2001), which is given by:

$$L^*(\theta) = -n - \sum_{i=1}^{n}\ln(1-\theta x_i) - n\ln\left[-(n\theta)^{-1}\sum_{i=1}^{n}\ln(1-\theta x_i)\right] \tag{5}$$

for $\theta < 1/x_{n:n}$. Suppose a local maximum $\hat{\theta}$ of (5) is found, then the ML estimates $(\hat{\alpha}, \hat{\beta})$ would be given by

$$\begin{aligned} \hat{\alpha} &= -(n^{-1})\sum_{i=1}^{n}\ln(1-\hat{\theta}x_i) \\ \hat{\beta} &= \hat{\alpha}/\hat{\theta} \end{aligned} \tag{6}$$

Using the log-likelihood function of Eq. (4), the following large-sample variance–covariance matrix for the MLEs is obtained (see, e.g., Choulakian and Stephens, 2001):

$$\Sigma = \begin{bmatrix} Var(\hat{\alpha}) & Cov(\hat{\alpha},\hat{\beta}) \\ Cov(\hat{\alpha},\hat{\beta}) & Var(\hat{\beta}) \end{bmatrix} = \begin{bmatrix} (1-\alpha)^2 & \beta(1-\alpha) \\ \beta(1-\alpha) & 2\beta^2(1-\alpha) \end{bmatrix}, \quad \alpha < 0.5 \tag{7}$$

When $\alpha < 0.5$, the MLEs have their familiar properties of consistency, asymptotic normality and asymptotic efficiency (Smith, 1984). The key asymptotic result upon which the calculation of CIQs is based, is:

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}^{n\to\infty} \sim N\left[\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \Sigma_{\substack{\alpha=\alpha_0 \\ \beta=\beta_0}}\right] \tag{8}$$

which states that asymptotically, the MLE vector $(\hat{\alpha}, \hat{\beta})$ has a (bivariate) normal distribution with mean (α_0, β_0) and variance–covariance matrix Σ (given in Eq. (7)) evaluated at $(\alpha = \alpha_0, \beta = \beta_0)$.

3. Approximate CIQs

Based on the approximate normality of the MLEs presented in Eq. (6), an approximate distribution is obtained for the pth quantile estimator $Q_n(p; \hat{\alpha}, \hat{\beta})$ and then used to construct approximate CIQs under a GPD model [i.e., confidence intervals for the *true* quantile $Q(p; \alpha_0, \beta_0)$]. The calculations are done in the Appendix; where it is shown that:

$$Q_n(p;\hat{\alpha},\hat{\beta})^{n\to\infty} \sim N\left(Q(p;\alpha_0,\beta_0),\hat{\sigma}^2\right) \tag{9}$$

with $\hat{\sigma}^2$ given in the Appendix.

3.1. Conventional CIQs

Traditionally, the asymptotic normality of $Q_n(p; \hat{\alpha}, \hat{\beta})$ [Eq. (9)] has formed the basis for constructing CIQs. In fact, from Eq. (9) the following probability statement may be written at the 95% confidence level:

$$Pr\left[-1.96 \leqslant \frac{Q_n(p;\hat{\alpha},\hat{\beta}) - Q(p;\alpha_0,\beta_0)}{\hat{\sigma}} \leqslant 1.96\right] \approx 95\% \tag{10}$$

(similar statements can be made at other confidence levels). Therefore, traditionally, the 95% CI for $Q(p; \alpha_0, \beta_0)$ has been

$$\mathrm{CIQ}_{conventional}(95\%) = \left[Q_n(p;\hat{\alpha},\hat{\beta}) \pm 1.96\hat{\sigma}\right] \tag{11}$$