JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EXTREME VALUE DISTRIBUTION AND ITS APPLICATIONS TO RAINFALL DATA

[1] M. P. Jeena
[1]Research Scholar
[1]Department of Statistics,
[1]University of Calicut, Malappuram, India.

Abstract : Extreme value theory supplies the asymptotic distributions in describing extremes, providing flexible and simple parametric models for fitting tail-related distributions. The objective of this work is to review the existing probabilistic model for extremes available in the literature in the independent and identically distributed case and their statistical applications. An attempt is made to fit two of this model to a real life data.

Keywords: Extremal types theorem, Generalized extreme value distribution, Generalized Pareto Distribution, k-th largest method, Maximum likelihood estimation, Return levels.

I. INTRODUCTION

The classical theory of extremes is concerned primarily with the asymptotic distribution and related properties of extremes from a sequence of random variables. The important roles were played by E. L. Dodd and M. Frechet in providing the conceptual framework to develop the asymptotic theory of extremes. It was Dodd in 1923, who was the first to relate the asymptotic growth of the maximum of *n* independent and identically distributed (i.i.d) random variables to the rate at which the right tail of the underlying probability density function falls off to zero and Frechet in 1927, who introduced the key idea that if there exists a limit law for maxima (minima), then it must be stable in the sense that the distribution of the maximum (minimum) of *n* i.i.d. observations drawn from it must be of the same type, except for a linear transformation. The mathematical foundation of extreme value theory is a limit theorem, first derived by Fisher and Tippett (1928) and these results were proved in complete generality by Gnedenko (1943). After the pioneering work of Gnedenko (1943), and the later works of De Haan (1976, 1984), the limit theorem is in a neat structure. The key result of Fisher-Tippett and Gnedenko is that there are three types of limit distributions for maximum. The three types may be combined into a single class, called Generalized Extreme Value Distribution (GEVD) which was proposed by Von Mises (1936) and Jenkinson (1955).

The theory of extreme values and extreme value distributions play an important role in theoretical and applied statistics. The books by Leadbetter, Lindgren and Rootzen (1983), Galambos (1987), Resnick (1987), Embrechts, Kluppelberg and Mikosch (1997), Kotz and Nadarajah (2000) and Reiss and Thomas (2007) gives the probabilistic aspects of extreme value theory in a neat structure. Coles (2001) give the probabilistic and statistical aspects of extreme value theory. The statistical application of the extremal types theorem was initiated by Gumbel (1954). This method is known as the Gumbel method or block method. Smith (1990) gives a comprehensive account of statistical aspects, especially maximum likelihood methods on parameter estimation. Extreme value theory is applied in the study of day to day market risk, the size of freak waves, size effect on material strengths, the occurrence of floods and droughts etc. The *k*-th largest method is suggested as an alternative to the Gumbel method and is based on *k* upper order statistics for a fixed integer *k*. This method was initiated by Weissman (1978) and developed by Gomes (1978, 1981), Smith (1986) and Tawn (1988).

To model extremal events in a univariate context, usually the generalized extreme value distribution is adopted. A related approach in univariate context is Peek Over Threshold approach first given by Pickands (1975). The main aim in this approach is to fit a Generalized Pareto Distribution (GPD) to excesses of a high threshold by a random variable, under the condition that sufficient data are available above the threshold.

In this paper we review basic results of classical extreme value theory. This includes the extremal types theorem for maximum, statistical application of extremal types theorem and theory of domain of attraction. Generalized extreme value distributions and its inference procedures will discuss in section 3. Section 4 provides *k*-th largest method to overcome one of the difficulties of Gumbel method. This contains limit behaviour of *k*-th maximum, the joint asymptotic distribution of largest maxima and its modeling procedure. Section 5 deals with the peak over threshold method and the inference procedure for the generalized Pareto distribution.

The objective of section 6 is to fit the generalized extreme value distribution and generalized Pareto distribution using a daily rainfall data of Alwar station in Rajasthan state over the period 1957-2012.

II. BASIC CONCEPTS OF EXTREMES

In this section we discuss the definition of extremes and some examples that illustrate the application of extremes in different real life situations.

Let $\{\xi_n, n \geq 1\}$ be a sequence of i.i.d. r.v.s on a probability space (Ω, F, P) with distribution function (d.f) F. Then for a fixed integer n, we define the order statistics of $\xi_1, \xi_2, \ldots, \xi_n$ as follows;

$$\begin{aligned} M_{n:1} &= \max(\xi_1, \xi_2, \ldots, \xi_n) \\ M_{n:2} &= \max(\{\xi_1, \xi_2, \ldots, \xi_n\} - \{M_{n:1}\}) \\ &. \\ M_{n:n} &= \min(\xi_1, \xi_2, \ldots, \xi_n) \end{aligned}$$

In this paper, $M_{n:1}$ and $M_{n:n}$ are denoted by M_n and m_n respectively. We call the maximum M_n and the minimum m_n as extreme order statistics. Other order statistics can also be interpreted as extremes. For example, $M_{n:n-(k-1)}$ and $M_{n:k}$ are called the k-th lower extreme and k-th upper extreme respectively. The minimum m_n can be written in terms of maximum of M_n as follows.

$$\begin{aligned} m_n &= \min(\xi_1, \xi_2, \ldots, \xi_n) \\ &= -\max(-\xi_1, -\xi_2, \ldots, -\xi_n). \end{aligned}$$

Before discussing the asymptotic properties of M_n and m_n we discuss some practical situations were M_n and m_n are the only statistics needed to study.

1. If ξ_j is the water level of a river at a given location on day j (starting point is immaterial), then $M_n = \max(\xi_1, \xi_2, \ldots, \xi_n)$ is the only statistic used for the analysis of floods in that location, and $m_n = \min(\xi_1, \xi_2, \ldots, \xi_n)$ is the only statistic used for the analysis of droughts in that location.

2. Consider a chain made up of n links. The chain breaks when any one of its link breaks. The first link to break is the weakest link or the one that has the smallest strength. Assume that the strength of the i-th link, say ξ_i, $i = 1, 2, \ldots, n$ is a r.v with a common distribution function. Since the chain breaks when its weakest link breaks, the strength of the chain is described by the r.v., $m_n = \min(\xi_1, \xi_2, \ldots, \xi_n)$.

3. Consider an engineering or a biological system that is made up of n identical components, all of which may function simultaneously. For example, a large air plane may contain four identical engines which would be functioning simultaneously or the human respiratory system, which consists of two identical lungs. The system functions as long as at least one of the n components is functioning. Suppose the time to failure of the i-th component say, ξ_i, $i = 1, 2, \ldots, n$ is a r.v with a common distribution function. Since the system fails at the time of failure of the last component, the life length of the system is described by the order statistic $M_n = \max(\xi_1, \xi_2, \ldots, \xi_n)$.

2.1 Asymptotic Properties of M_n and m_n

In this section we discuss the asymptotic distributions of M_n and m_n. Before discussing asymptotic distribution, let us discuss the exact distribution M_n and m_n.

$$\begin{aligned} P(M_n \leq x) &= P(\xi_1 \leq x, \xi_2 \leq x, \ldots, \xi_n \leq x) = F^n(x), x \in \mathbb{R}, \\ P(m_n \leq x) &= 1 - P(m_n > x) = 1 - P(\xi_1 > x, \xi_2 > x, \ldots, \xi_n > x) = 1 - (1 - F(x))^n. \end{aligned}$$

Consider exponential distribution with mean 1 given by

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - \exp(-x), & x \geq 0. \end{cases}$$

Then the distribution of M_n becomes

$$P(M_n \leq x) = F^n(x) = \begin{cases} 0, & x < 0, \\ (1 - \exp(-x))^n, & x \geq 0, \end{cases}$$

and the distribution of m_n becomes

$$P(m_n \le x) = 1 - (1 - F(x))^n = \begin{cases} 0, & x < 0, \\ 1 - \exp(-nx), & x \ge 0. \end{cases}$$

Consider uniform distribution over the interval (a,b) given by

$$F(x) = \begin{cases} 0, & x \le a, \\ \frac{x-a}{b-a}, & a < x < b, \\ 1, & x \ge b. \end{cases}$$

Then the distribution of M_n becomes

$$P(M_n \le x) = F^n(x) = \begin{cases} 0, & x \le a, \\ \{\frac{x-a}{b-a}\}^n, & a < x < b, \\ 1, & x \ge b \end{cases}$$

and the distribution of m_n becomes

$$P(m_n \le x) = 1 - (1 - F(x))^n = \begin{cases} 0, & x \le a, \\ 1 - \{\frac{b-x}{b-a}\}^n, & a < x < b, \\ 1, & x \ge b. \end{cases}$$

One could notice that Mn is increasing and mn is decreasing in n. The point at which these Mn and m_n converging could be identified as the right end and left end point of the d.f of ξi. Motivated by these we introduce right and left end point of a d.f.. Let F be a d.f, then the right end point of F denoted by x_F is,

$$x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$$

and the left end point of F denoted by x_l is,

$$x_l = \inf\{x \in \mathbb{R} : F(x) > 0\}.$$

Consider uniform distribution over the interval (a,b) given by

$$F(x) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a < x < b, \\ 1, & x \geq b. \end{cases}$$

Then the right and left end points of F are $x_F = b$ and $x_l = a$.

Consider exponential distribution with mean 1 given by

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - \exp(-x), & x \geq 0. \end{cases}$$

Then the right and left end points of F are $x_F = \infty$ and $x_l = 0$.

Notice that properties of the d.f $F^n(x)$ of M_n for large n are governed by behaviour of $F(x)$ for large x. If F is any d.f, we can see that for all $x < x_F$, $0 < F(x) < 1$,

$$P(M_n \leq x) = F^n(x) \longrightarrow 0, \;\; as \;\; n \longrightarrow \infty$$

and we have for $x \geq x_F$, $F(x) = 1$,

$$P(M_n \leq x) = F^n(x) = 1.$$

Also for every $\epsilon > 0$,

$$\begin{aligned} P[|M_n - x_F| > \epsilon] &= P[M_n - x_F < -\epsilon \;\; or \;\; M_n - x_F > \epsilon] \\ &= P[M_n < x_F - \epsilon] + P[M_n > x_F + \epsilon] \\ &= P[M_n < x_F - \epsilon] + 1 - P[M_n < x_F + \epsilon]. \end{aligned}$$

From the above discussion, we get

$$P[M_n < x_F - \epsilon] = F^n(x_F - \epsilon) \longrightarrow 0 \;\; as \;\; n \longrightarrow \infty$$

and

$$P[M_n < x_F + \epsilon] = F^n(x_F + \epsilon) \longrightarrow 1 \;\; as \;\; n \longrightarrow \infty.$$

Therefore,

$$P[|M_n - x_F| > \epsilon] \longrightarrow 0 \;\; as \;\; n \longrightarrow \infty.$$

That is, $M_n \longrightarrow^P x_F$ as $n \longrightarrow \infty$. Also M_n is non-decreasing. Hence, $M_n \longrightarrow^{a.s} x_F$ as $n \longrightarrow \infty$. Using a similar argument one can verify that $m_n \longrightarrow^{a.s} x_l$.

Consider exponential distribution with mean 1 given by

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - \exp(-x), & x \geq 0. \end{cases}$$

Then $x_F = \infty$ and $x_l = 0$. Hence $M_n \longrightarrow^{a.s} \infty$ and $m_n \longrightarrow^{a.s} 0$.

In the examples discussed above, one can obtain the exact distribution of extremes Mn and mn in explicit form. But in many other cases the exact distribution of Mn and mn cannot be evaluated in explicit form, for example; the normal distribution. In Such situations, we go for the asymptotic distribution of extreme. In classical theory, one obtains an asymptotic normal distribution for the sum of many i.i.d. r.v.s whatever their common original d.f. A similar situation holds in extreme value theory, and in fact, a non degenerate asymptotic normalized distribution of Mn must belong to one of three possible families, regardless of the original d.f F. This result is known as extremal types theorem.

The set of all d.f.s attracted to a d.f G is said to be domain of attraction of the d.f G and is denoted by $D(G)$. That is, if $\mathbb{F}$ is the class of all d.f.s and $a_n > 0$, b_n are constants, then

$$D(G) = \{F \in \mathbb{F} : F^n(a_n x + b_n) \longrightarrow^w G(x),\ for\ some\ constants\ a_n > 0\ and\ b_n\}.$$

Next we discuss the key result in extreme value theory, the extremal types theorem for maximum.

Extremal Types Theorem for Maximum

Let $\{\xi_n, n \geq 1\}$ be a sequence of i.i.d. r.v.s on a probability space $(\Omega, \mathbb{F}, P)$ and F be the d.f of ξ_1. Let $M_n = \max(\xi_1, \xi_2, \ldots \xi_n)$. Then if for some constants $a_n > 0$, b_n, we have

$$P\{\frac{M_n - b_n}{a_n} \leq x\} \longrightarrow^w G(x) \tag{1}$$

for some non degenerate G, then G is one of the following three types.

$$Type\ I: \quad G_1(x) = \exp(-\exp(-x)), -\infty < x < \infty.$$

$$Type\ II: \quad G_2(x) = \begin{cases} 0, & x \leq 0 \\ \exp(-x^{-\alpha}), & \text{for some } \alpha > 0\,, x > 0. \end{cases}$$

$$Type\ III: \quad G_3(x) = \begin{cases} \exp(-(-x)^{\alpha}), & \text{, for some } \alpha > 0,\ x \leq 0 \\ 1, & x > 0. \end{cases}$$

Conversely, each d.f G of types discussed above may appear as a limit in

$$P\{\frac{M_n - b_n}{a_n} \leq x\} \longrightarrow^w G(x)$$

and in fact, appears when G itself is the d.f of each ξ_i.

The extremal types theorem was first proved by Gnedenko (1943). The rigorous mathematical proof is given by De Haan (1976, 1984).

The statistical application of the extremal types theorem was initiated by Gumbel (1954). We assume that for n large enough but fixed, the limit in (1) gives an approximation for G(.). To make this approximation in practice, we first divide the data into a number of blocks say on an annual basis and note down the maxi- mum M_n in each block. Then by using suitable estimation procedures, a family of extreme value distributions might be fitted to the sequence of annual maxima $\{M_n\}$. This method of fitting the extreme value distributions is known as the **Gumbel method or block method** for extremes. But there is a practical difficulty in implementing the Gumbel method described above as there are three distinct classes of limit distributions as discussed in extremal types theorem, and it is not clear which one should be fitted. Von-Mises and Jenkinson independently suggested an alternative method to overcome this difficulty. They showed that the three extreme value distributions can be unified in a three parameter family, called generalized extreme value distribution, by introducing a shape parameter. In the next section we introduce generalized extreme value distribution (GEVD).

III Generalized Extreme Value Distribution (GEVD)

A r.v ξ is said to have the generalized extreme value distribution with parameters $\sigma > 0$, μ and γ if its distribution function $G_{\mu,\sigma,\gamma}(x)$ is given by

$$G_{\mu,\sigma,\gamma}(x) = \exp\{-[1 + \gamma(\frac{x - \mu}{\sigma})]^{-1/\gamma}\}, \quad if\ \ \gamma \neq 0 \tag{2}$$

where $1+\gamma(\frac{x-\mu}{\sigma})>0$; μ and σ are the location and scale parameters and γ is the shape parameter. The generalized extreme value distribution is also known as the Von-Mises type extreme value distribution or the Von Mises- Jenkinson type distribution.

The case $\gamma=0$ is interpreted as the limit $\gamma \longrightarrow 0$, which is the Gumbel distribution with d.f

$$G(x)=\exp\{-\exp(-[(x-\mu)/\sigma]\}, \quad -\infty<x<\infty.$$

When $\gamma>0$ we have the Frechet distribution with $\alpha=1/\gamma$; when $\gamma<0$ we have the Weibull distribution with $\alpha=-1/\gamma$.

Let $\{\xi_n, n\geq 1\}$ be a sequence of i.i.d. r.v.s on a probability space $(\Omega, \mathbb{F}, P)$ with d.f F and $M_n = max(\xi_1, \xi_2, \ldots \xi_n)$. Then if for some constants $a_n>0$, b_n, we have

$$P\{\frac{M_n-b_n}{a_n}\leq x\} \longrightarrow^w G(x)$$

for some non degenerate G, then G is a member of the generalized extreme value family

$$G_{\mu,\sigma,\gamma}(x)= \exp\{-[1+\gamma(\frac{x-\mu}{\sigma})]^{-1/\gamma}\}, \quad if \;\; \gamma\neq 0$$

defined on $\{x:1+\gamma(\frac{x-\mu}{\sigma})>0\}$, where $-\infty<\mu<\infty, \sigma>0$ and $-\infty<\gamma<\infty$.

3.1 Inference Procedures for the GEVD

We first divide the data into a number of blocks say on an annual basis and note down the maximum M_n in each block. After selecting the maximum observation from each block, we estimate the parameters of generalized extreme value distribution. Many techniques have been proposed for parameter estimation such as probability weighted moment method, Bayesian method and maximum likelihood method. Commonly used method in GEVD is the maximum likelihood estimation method which was considered by Prescott and Walden (1980, 1983). This method is discussed below.

Maximum Likelihood Estimation

The maximum likelihood estimation is a method of estimating the parameters of a statistical model. This method is first introduced by R. A. Fisher in 1922. Prescott and Walden (1980, 1983) used this method to estimate the parameters of GEVD which are given below.

Let $\{\xi_1, \xi_2, \ldots, \xi_n\}$ be sequence of i.i.d r.v.s. Then the probability density function (p.d.f) $g_{\mu,\sigma,\gamma}(x)$ for a generalized extreme value d.f G is defined in (2) with parameters μ, σ and γ is given by

$$\begin{aligned} g_{\mu,\sigma,\gamma}(x) &= \frac{dG(x)}{dx} \\ &= \frac{1}{\sigma}[1+\gamma(\frac{x-\mu}{\sigma})]^{-\frac{1}{\gamma}-1}\exp\{-[1+\gamma(\frac{x-\mu}{\sigma})]^{-\frac{1}{\gamma}}\}. \end{aligned}$$

The log-likelihood for a given data set occurring can be found by

$$\begin{aligned} l_{\mu,\sigma,\gamma}(x) &= \log\prod_{i=1}^{m} g_{\mu,\sigma,\gamma}(x) \\ &= -m\log\sigma-(1+\frac{1}{\gamma})\sum_{i=1}^{m}\log[1+\gamma(\frac{x-\mu}{\sigma})]-\sum_{i=1}^{m}[1+\gamma(\frac{x-\mu}{\sigma})]^{-\frac{1}{\gamma}} \end{aligned}$$

provided that $1 + \gamma \left(\frac{x-\mu}{\sigma}\right) > 0$ for $i = 1, \ldots, m$. Then the method of maximum likelihood works by finding the values for the parameters μ, σ and γ that maximize the value of the likelihood function. But we cannot get the maximum likelihood estimates of GEVD in concrete form. The situation for $g_{\mu,\sigma,\gamma}(x)$ when $\gamma = 0$ is even more complicated. While this represents an irregular likelihood problem, due to the dependence of the parameter space on the values of data, consistency and asymptotic efficiency of the resulting maximum likelihood estimations can be established for the case $\gamma > -1/2$. In MATLAB, the function '*gevfit*' returns maximum likelihood parameter estimates of GEVD. After estimate the parameters of GEVD, we can calculate return level for the data.

3.2 Return Levels

The theoretical return period is the inverse of the probability that the event will be exceeded in any one year (or more accurately the inverse of the expected number of occurrences in a year). Having modeled the upper tail of the distribution by fitting an asymptotically motivated model such as GEVD it remains to use such a model for inference. An event exceeding such a level is expected to occur once every N year. An N = 1/ p return level z_p is the 1 − p quantile of F such that $1 - F(z_p) = p$.

In the case where F can be modeled by a generalized extreme value distribution G then maximum likelihood estimates of the return levels z_p can be found by substituting maximum likelihood estimates of the generalized extreme value parameters into the quantile function of G. These can be calculated as follows.

$$\widehat{z_p} = \begin{cases} \widehat{\mu} - \frac{\widehat{\sigma}}{\widehat{\gamma}}\{1 - [-\log(1-p)]^{-\widehat{\gamma}}\}, & \text{for } \widehat{\gamma} \neq 0, \\ \widehat{\mu} - \widehat{\sigma}\log\{-\log(1-p)\}, & \text{for } \widehat{\gamma} = 0. \end{cases}$$

We already discussed the Gumbel method in Section 2.1. It suffers the serious drawback that it does not use the data efficiently. It causes more wastage of data. While fitting the generalized extreme value distribution, we take only the maximum value in each block and ignore all other values in that block. Further there may be observations discarded in some blocks which are greater than some other block maxima. To overcome this difficulty, two alternative models are suggested in the literature viz. the k-th largest method and the peak over threshold method. In the next section we will discuss the k-th largest method

IV k-th LARGEST METHOD

The *k*-th largest method is suggested as an alternative to the Gumbel method and is based on *k* upper order statistics for a fixed integer *k*. This method was initiated by Weissman (1978) and developed by Gomes (1978,1981), Smith (1986) and Tawn (1988). This model an extension of the annual maximum approach is to use the *k* largest observations in each fixed time period (say, one year), where $k > 1$.

Let $\{\xi_n, n \geq 1\}$ be a sequence of i.i.d. r.v.s on a probability space (Ω, F, P) with d.f F. Then for a fixed integer k, k-th maximum of $\xi_1, \ldots, \xi_n$ is defined by

$$M_{n:k} = k - th\ largest\ of\ \{\xi_1, \ldots, \xi_n\}.$$

4.1. Asymptotic Distribution of k-th Maximum

Let $\{\xi_n, n \geq 1\}$ be a sequence of i.i.d. r.v.s on a probability space $(\Omega, \mathbb{F}, P)$ with d.f F and $M_{n:k}$ be the k-th maximum, for a fixed integer k. Suppose that

$$P\{\frac{M_n - b_n}{a_n} \leq x\} \longrightarrow^{w} G(x) \; as \; n \longrightarrow \infty$$

for some non-degenerate d.f G, so that G is the generalized extreme value distribution then, for fixed k,

$$P\{\frac{M_{n:k} - b_n}{a_n} \leq x\} \longrightarrow^{w} G_k(x)$$

on $\{x : 1 + \gamma(x - \mu)/\sigma > 0\}$, where

$$G_k(x) = \exp\{-\psi(x)\} \sum_{s=0}^{k-1} \frac{(\psi(x))^s}{s!} \tag{3}$$

with

$$\psi(x) = [1 + \gamma(x - \mu)/\sigma]^{-1/\gamma}.$$

4.2. Joint Asymptotic Distribution of Largest Maxima

Considering the number of exceedances of different levels by $\xi_1, \xi_2, \ldots \xi_n$, one can obtain the *joint d.f of different extremes*. The maximum M_n has an asymptotic distribution in terms of G(.) with the same sequence of norming constants $\{a_n > 0\}$ and $\{b_n\}$. That is, when the maximum has a non-degenerate limiting d.f G for the sequence of norming constants $\{a_n > 0\}$ and $\{b_n\}$, the joint d.f of $M_{n:1}, \ldots, M_{n:r}$ exists and is in terms of $G(.)$. For $r = 2$, we state the following result. Let $\{\xi_n, n \geq 1\}$ be a sequence of iid r.vs. For given sequences of norming constants $\{a_n > 0\}$, $\{b_n\}$ and for a non-degenerate d.f G(x),

$$P\{\frac{M_{n:1} - b_n}{a_n} \leq x\} \longrightarrow^{x} G(x) \tag{4}$$

for some non-degenerate (and hence Type I,II,or III) d.f G. Then for $x_1 > x_2$,

$$P\{(M_{n:1} - b_n)/a_n \leq x_1, (M_{n:2} - b_n)/a_n \leq x_2\} \longrightarrow^{x} G(x)\{\log G(x_1) - \log G(x_2) + 1\} \tag{5}$$

when $G(x_2) > 0$ (and to zero when $G(x_2)$=0)

4.3. Statistical Application

The k-th largest method is a method to overcome the drawback of wastage of data of the Gumbel method. Rather than fitting the asymptotic distribution of the annual maximum, it uses the k largest values from each block and fits the asymptotic distribution of the k largest order statistics to them by using suitable method like Gumbel method. After defining blocks, select k largest values from each block and then we can analyze the data set. When k = 1 the method reduces to Gumbel method.

V PEAK OVER THRESHOLD METHOD

Threshold model is based on exceedances over a high threshold and is formulated not in terms of maxima directly; but in terms of exceedances over a high threshold. This method has been developed by hydrologists for a very long and is known as the Peak Over Threshold or POT method. Theoretical aspects of this method were fully established only in 1970's. In this section, we discuss the basic results of threshold method.

Let $\{\xi_n, n \geq 1\}$ be a sequence of i.i.d. r.v.s on a probability space $(\Omega, \mathbb{F}, P)$ and F be the d.f of ξ_1. We fix some high threshold u and denote by

$$N_u = card\{i : i = 1, 2, \dots, n,\ \xi_i > u\}$$

the number of exceedances of u by $\xi_1, \xi_2, \dots, \xi_n$. We denote the corresponding excesses by $Y_1, Y_2, \dots, Y_{N_u}$, i.e., for $i = 1, 2, \dots, n$,

$$Y_i = \begin{cases} \xi_i - u, & \xi_i > u, \\ 0, & \text{otherwise.} \end{cases}$$

Then the distribution of excess values, denoted by F_u is

$$F_u(x) = P[\xi - u \leq x | \xi > u] = \frac{F(u+x) - F(u)}{1 - F(u)} \quad 0 \leq x \leq x_F - u \tag{6}$$

where $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$ is the right end point of F. In Section 3, the maximum $M_n \longrightarrow^d G(x)$, where G is a GEVD. Now assume that for some $\sigma, \gamma, 0 < \sigma < \infty, -\infty < \gamma < \infty$,

$$\lim_{u \longrightarrow x_F} \inf_{0<\sigma<\infty} \sup_{0 \leq x < \infty} |\{\frac{1 - F(u+x)}{1 - F(u+x)}\} - \exp(-\int_0^{x/\sigma} [(1+\gamma t)_+]^{-1} dt)| = 0 \tag{7}$$

where $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$ and for any y, $y_+ \equiv \max(0, y)$. For any u, x, $P[\xi - u \leq x | \xi > u]$ is the conditional probability that an observation is greater than $u + x$, given that it is greater than u. If u is large, the conditional distribution of ξ given that $\xi \geq u$ is very nearly of the form

$$1 - G(x) = \exp - \int_0^{x/\sigma} [(1+\gamma t)_+]^{-1} dt \tag{8}$$

for some $\sigma, \gamma, 0 < \sigma < \infty, -\infty < \gamma < \infty$. This can be distinguished three cases as $\gamma > 0$, $\gamma = 0$, and $\gamma < 0$.
Case 1: If $\gamma > 0$, then

$$1 - G(x) = (1 + \gamma x/\sigma)^{-1/\gamma}$$

for all x, $0 < x < \infty$. This class of distribution is known as Pareto family.
Case 2: If $\gamma = 0$, then

$$1 - G(x) = \exp(-x/\sigma)$$

for all x, $0 < x < \infty$. This is the exponential family of distribution.
Case 3: If $\gamma < 0$, then

$$1 - G(x) = (1 - |\gamma| x/\sigma)^{1/|\gamma|}$$

for all x, $0 < x \leq \sigma/|\gamma|$ and

$$1 - G(x) = 0$$

for all x, $\sigma/|\gamma| \leq x < \infty$. This is a theory about the assymptotic form of $F_u(x)$. Balkema and de Haan (1974)and Pickands(1975) modified this result is known as ***Pickands-Balkema-de Haan Theorem*** which is given below.

Let $\{\xi_n, n \geq 1\}$ be a sequence of i.i.d. r.v.s and F be the d.f of ξ_1, and let

$$M_n = \max(\xi_1, \dots, \xi_n).$$

If for some constants $a_n > 0$, b_n, so that large n,

$$P\{M_n \leq x\} \approx G(x),$$

where

$$G(x) = \exp\{-[1+\gamma(\frac{x-\mu}{\sigma})]^{-1/\gamma}\}$$

for some μ, σ and γ. Then, for large enough u, the d.f of $\xi - u$, conditioned on $\xi > u$, is approximately

$$H(y) = \ 1-(1+\gamma(\frac{y}{\overline{\sigma}}))^{-\frac{1}{\gamma}} \tag{9}$$

defined on $\{y : y > 0 and (1+\gamma(\frac{y}{\overline{\sigma}})) > 0\}$, where

$$\overline{\sigma} = \sigma + \gamma(u-\mu). \tag{10}$$

The distribution defined by equation (9) is called the Generalized Pareto Distributon (GPD). The case $\gamma = 0$ is interpreted as the limit $\gamma \to 0$, which is the exponential distribution. The above theorem implies that, if block maxima of $\{\xi_1, \xi_2, \ldots, \xi_n\}$ have approximating distribution G, then threshold excesses have a corresponding approximate distribution within the generalized Pareto family. Moreover, the parameters of the GPD of threshold excesses are uniquely determined by those of the associated GEVD of block maxima. In particular, the parameter γ in (9) is equal to that of the corresponding GEVD.

5.1. Modeling Threshold Exceedances

In the classical set up the peak over threshold method could be applied as follows. First we record observations beyond a threshold u by the sequence $\{\xi_n\}$, $n \geq 1$. It is important to choose a threshold as low as possible in order to maximize the amount of observations used to make the better inference. The choice of the extreme threshold u, where the GPD model provides a suitable approximation to the excess distribution Fu, is critical in applications. Benktander and Segerdahl (1960) introduced the mean excess function as a tool popularly used for the choice of u and also to determine the adequacy of the generalized Pareto distribution model in practice. The mean excess function is defined as follows. Let ξ be a r.v with right end point x_F. The mean excess function of ξ is defined as

$$M(u) = E(\xi - u|\ \xi > u), \quad 0 < u < x_F.$$

Given an i.i.d random sample $\xi_1, \ldots, \xi_n$ from a d.f $F(x)$, a natural estimate of $M(u)$ is the empirical mean excess function $\hat{M}(u)$ defined as

$$\hat{M}(u) = \frac{\sum_{i=1}^{n}(\xi_i - u)I_{(\xi_i > u)}}{\sum_{i=1}^{n} I_{(\xi_i > u)}}, \quad u \geq 0.$$

Note: The quantity $M(u)$ is often referred to as the mean excess over threshold value u. In reliability or medical context, $M(u)$ is referred to as the mean residual life function.

The linearity of the mean excess function characterizes the generalized Pareto distribution class. Davison and Smith (1990) used this property to devise a simple graphical check that data conform to a GPD model; their method is based on the mean excess plot which is the plot of the points $\{(M_{n:k}, \hat{M}_{n:k}) : 1 < k \leq n\}$, where $M_{n:1} > \ldots > M_{n:n}$ are the order statistics of the data. If the plot is nearly linear beyond u for some u, it suggests that the generalized Pareto model is reasonably fitted to the data. Motivated by Pickand's result, the generalized Pareto distribution is fitted to the excesses using a suitable estimation procedure. Commonly we use the maximum likelihood method for parameter estimation. This concept is discussed below.

5.2. Maximum Likelihood Estimation

Let $\{\xi_n, n \geq 1\}$ be a sequence of i.i.d. r.v.s with d.f F. Having determined a threshold u by the sequence $\{\xi_n\}$, the parameters of the GPD can be estimated by maximum likelihood method. Suppose that $y_1, y_2, \ldots, y_k$ are the k excesses of a threshold u. Let $h_{\sigma,\gamma}(y)$ be the p.d.f for a generalized Pareto d.f H defined in (9) with parameters σ and γ. For $\gamma \neq 0$, the log likelihood is derived from (9) as

$$l_{\sigma,\gamma}(y) = -k\log\sigma - (1+\frac{1}{\gamma})\sum_{i=1}^{k}\log(1+\gamma\frac{y_i}{\sigma}) \tag{11}$$

provided $(1+\gamma\frac{y_i}{\sigma}) > 0$ for $i = 1, 2, \ldots, k$; otherwise, $l_{\sigma,\gamma}(y) = -\infty$.

We cannot get the parameter estimates of GPD in explicit form by analytical maximization of the log likelihood function. So numerical techniques are again required, taking care to avoid numerical instabilities when $\gamma \approx 0$ in (11), and ensuring that the algorithm does not fail due to evaluation outside of the allowable parameter space. In MATLAB, the function 'gpfit′ returns maximum likelihood parameter estimates of GPD. After estimating the parameters of GPD, we can calculate return level for the data. The return level is discussed in the next section.

5.3. Return Levels

In this section we discuss the return level of d.f F can be conditionally modeled by a GPD.

Let n_y be the number of observations in each block of data, then the N-year return level y_N is given by

$$\widehat{y_N} = \begin{cases} \frac{\widehat{\sigma_u}}{\widehat{\gamma}}[(\widehat{\zeta}_u N n_y)^{\widehat{\gamma}} - 1] + u, & \text{for } \widehat{\gamma} \neq 0, \\ \widehat{\sigma} \log(\widehat{\zeta}_u N n_y) + u, & \text{for } \widehat{\gamma} = 0. \end{cases}$$

where $\zeta_u = P(\xi > u)$.

VI. MODELING RAINFALL DATA WITH THE GEVD AND GPD

In this section we will discuss the fitting of generalized extreme value distribution and generalized pareto distribution with suitable statistical procedures by using daily rainfall data of Alwar station in Rajasthan district. For this we used a secondary data from the website waterresources.rajasthan.gov.in. The data were collected from january 1957 to December 2012. First we modeled the data with the generalized extreme value distribution.

6.1. Fitting the Generalized Extreme Value Distribution

In this section, we will fit the generalized extreme value distribution by using daily rainfall data of Alwar station over a 56 year period, i.e., from January 1957 to December 2012. The data can be shown in Figure 1. The data shows that the daily rainfall measurements of Alwar station in Rajastan district from january 1957 to December 2012. The rainfall is observed in m.m. The plot of measurements in Figure 1 suggests that the observations can be considered to be independently and identically distributed. Here we use the block maxima method to fit the GEVD parameters estimated by using maximum likelihood methods. MATLAB and Microsoft Office Excel are used to fit the GEVD. By using block maxima method, we divided the given data set into 56 year wise blocks and observed the highest values in each block. The observations are given in Table 1. The block maxima can be modeled using Generalized extreme value distribution with parameters estimated by maximum likelihood method. The parameter estimates are shown in Table 2. Table 2 shows maximum likelihood estimates of each parameter of the GEVD with associated confidence intervals and standard errors. Since the shape parameter γ is positive, the distribution of the data is of Frechet type with standard error 0.0922. The 95% confidence interval corresponding to the parameter γ is (−0.1333,0.2283). Given parameter estimates for the fitted GEVD and a good fitted model, it remains to use this model to calculate return levels for the daily rainfall data. Table 3 shows the different return levels using block maxima approach.

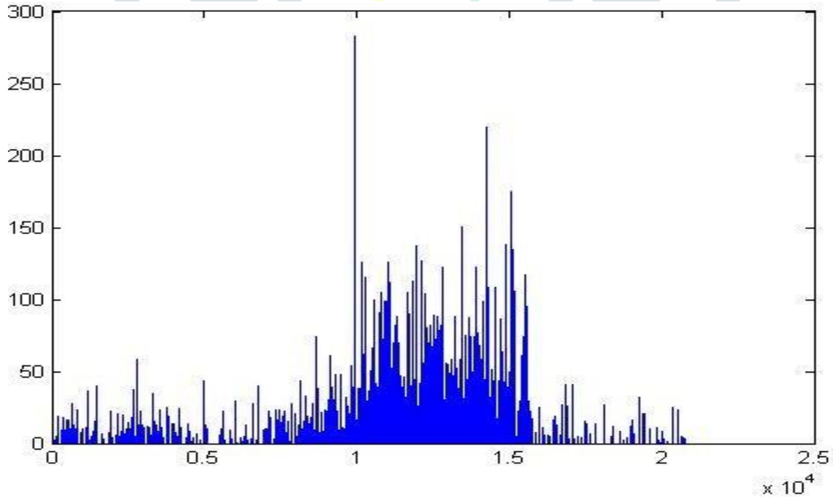

Figure 1: Daily Rainfall Data Taken From January 1957 to December 2012

Table 1: Block Maximum Rainfall Observations

Year	Rainfall (m.m)	Year	Rainfall (m.m)	Year	Rainfall (m.m)	Year	Rainfall (m.m)
1957	81.5	1971	72	1985	88.4	1999	106
1958	76.7	1972	88	1986	49	2000	46
1959	104.1	1973	99	1987	25.2	2001	38
1960	104.1	1974	99	1988	70	2002	75
1961	91.4	1975	78.8	1989	47	2003	113
1962	99.3	1976	126.1	1990	138	2004	59
1963	70	1977	112	1991	52	2005	126
1964	82	1978	108	1992	88	2006	137
1965	99.1	1979	122.6	1993	53.8	2007	81
1966	40.8	1980	52	1994	42	2008	115
1967	90.5	1981	56	1995	175	2009	47
1968	44	1982	70	1996	282.5	2010	117
1969	219.6	1983	82	1997	135	2011	95
1970	108	1984	86	1998	150	2012	127

Table 2: GEV Parameter Estimates by using Block Maxima Approach

Block maxima	μ	σ	γ
MLE	73.8933	31.7656	0.0475
Standard Error	4.7541	3.4866	0.0922
95% CI lower bound	64.5754	25.617	-0.1333
95% CI upper bound	83.2112	39.3901	0.2283

Table 3: Return Levels using Block Maxima Approach

Return levels	MLE
$\hat{z}_{1/10}$	149.3396
$\hat{z}_{1/15}$	164.4580
$\hat{z}_{1/25}$	183.6244
$\hat{z}_{1/50}$	210.0757

It gives10-year,15-year,25-yearand50-yearreturnlevelswerecalculatedusingthemaximumlikelihood estimates of the fitted model

6.2. Fitting the Generalized Pareto Distribution

In this section we fit the generalized pareto distribution by using the daily rainfall measurements of Alwar station in Rajastan district from January 1957 to December 2012 which is given in the previous section. The Generalized pareto distribution can be fitted by using the peak over threshold method. The fitting procedures are done using soft wares Microsoft Office Excel, MATLAB and SPSS.

To model the data by peak over threshold method, first we fix a threshold value u as $u = 13$ by using the mean excess plot and take all the values above the threshold u. Thus we get 857 excess values. The mean excess plot is shown in Figure 2.

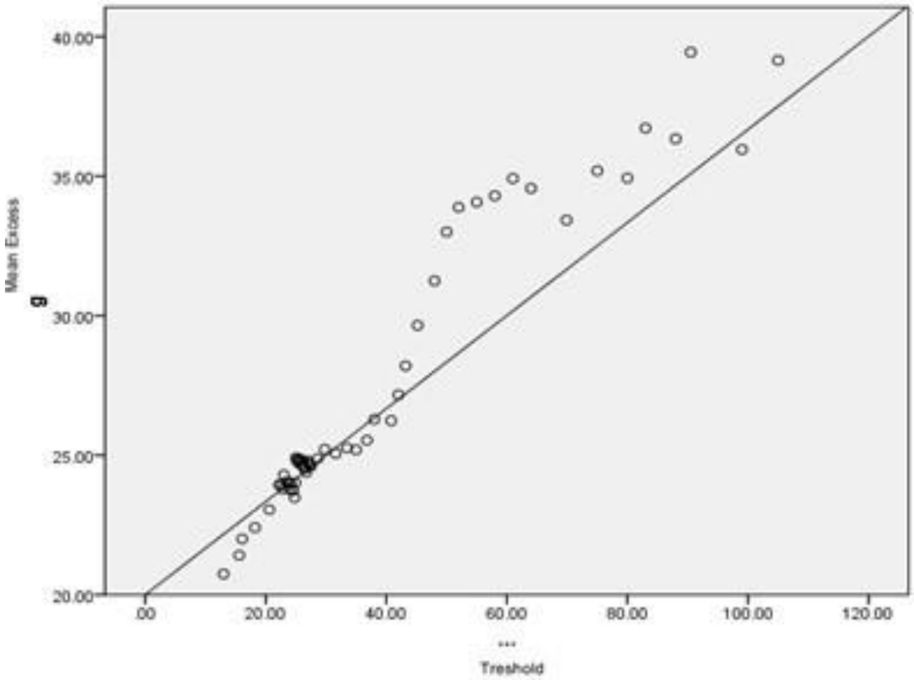

Figure 2: Mean Excess Plot

By using this observations above the threshold u, we can estimate the parameters of GPD. The parameters are estimated by using maximum likelihood estimation procedure. The parameter estimates are shown below. Table 4 shows the maximum likelihood estimates of each parameter of the GPD with associated confidence intervals and standard errors. The model fit yielded parameter estimates of $\hat{\sigma}_u = 55.6093$ and $\hat{\gamma} = -0.1653$ with standard errors 2.8672 and 0.0192 respectively. And associated confidence intervals of $\hat{\sigma}_u$ is (50.2643, 61.5226) and of $\hat{\gamma}$ is (-0.2030, -0.1277). Now we use this model to calculate return levels for the daily rainfall data. Table 5 shows the different return levels using threshold approach.

The table 5 gives 10-year, 15-year, 25-year and 50-year return levels were calculated using the maxi- mum likelihood estimates of the fitted model.

Table 4: GPD Parameter Estimates by using Threshold Approach

Threshold maxima	σ_u	γ
MLE	55.6093	-0.1653
Standard Error	2.8672	0.0192
95% CI lower bound	50.2643	-0.2030
95% CI upper bound	61.5226	-0.1277

Table 5: Return Levels using Threshold Approach

Return levels	MLE
$\hat{y}_{10}$	202.9353
$\hat{y}_{15}$	212.4311
$\hat{y}_{25}$	223.5230
$\hat{y}_{50}$	237.1516

VII. CONCLUSION

Extreme value theory is concerned with understanding the tails of processes which can be understood to follow a probability distribution. Standard modeling approaches are not enough to characterize the behaviour of such tails. So we suggest different modeling approaches to describe the tails. The choice of modeling approach to use is much dependent upon the nature of observations to be modeled. In the case of annual maxima of daily rainfall measurements of Alwar station in Rajasthan district, we choose the block-maxima and peak over threshold approaches to model the upper tail. When we fit the generalized extreme value distribution and generalized pareto distribution to the given data, it was found that the shape parameter of the distributions is very important in characterizing the nature of the tail. The result of fitting an appropriate model to data is to understand the extreme observation level which is expected to only be exceeded after a certain number of years. Also the return levels are a key result of extreme value theory modeling.

REFERENCES

[1] Balkema, A. A. and De Haan, L. (1974). “Residual Life Time at Great Age”, Ann. Probab. 2., 792-804.

[2] Benktander, G. and Segerdahl, C (1960). “On the Analytical Representation of Claim Distributions with Special Reference to Excess of Loss Reinsurance”, XIIth International Congress of Actuaries Brussels.

[3] Coles, S. G. (2001). “An Introduction to Statistical Modeling of Extreme Values”, Springer Series.

[4] Davison, A. C. and Smith, R. L. (1990). “Models for Exceedances over Thresholds (with Discussion)”, Journal of the Royal Statistical Society, B 52., 393-442.

[5] De Haan, L. (1974). “Weak Limits of Sample Range”, J. Appl. Probab. 11., 836-841.

[6] De Haan, L. (1976). “Sample Extremes : An Elementary Introduction”, Statist. Neerlandica 30., 161- 172.

[7] Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997). “Modeling Extremal Events for Insurance and Finance”, Springer-Verlag.

[8] Fisher, R. A. and Tippet, L. H. C. (1928). “Limiting Forms of The Frequency Distribution of the Largest or Smallest Member of a Sample”, Proc. Cambridge phil. Soc. 24., 180-190.

[9] Galambos, J. (1987). “The Asymptotic Theory of Extreme Order Statistics”, Krieger, Florida, 2nd Edition.

[10] Gnedenko, B. V. (1943). “Sur Ia Distribution Limite du Terme Maximum d’une Serie Aleatoire”, Annals of Mathematics 44, 423-453.

[11] Gomes, M. I. (1978). “Some Probabilistic and Statistical Problems in Extreme Value Theory”, Ph.D Thesis, University of Sheffield.

[12] Gumbel, E. J. (1954). “Statistical Theory of Extreme Values and Some Practical Applications”, Na- tional Bureau of Standards Applied Mathematics Series, 33.

[13] Jenkinson, A. F. (1955). “The Frequency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Events”, Quarterly Journal of the Royal Meteorological Society 81., 158-172.

[14] Kotz, S. and Nadarajah, S. (2000). “Extreme Value Distributions- Theory and Applications”, Imperial College Press.

[15] Leadbetter, M. R., Lindgren, G. and Rootzen, H. (1983). “Extremes and Related Properties of Random Sequences and Processes”, Springer-Verlag New York.

[16] Pickands, J. (1975). “Statistical Inference Using Extreme Order Statistics”, Annals of Statistics 3., 119-131.

[17] Prescott, P and Walden, A. T. (1980). “Maximum Likelihood Estimation of the Parameters of the Generalized Extreme Value Distribution”, Biometrika 67., 723-724.

[18] Prescott, P. and Walden, A. T. (1983). “Maximum Likelihood Estimation of the Parameters of the Three-Parameter Generalized Extreme Value Distribution from Censored Samples”, Journal of Statistical Computation and Simulation 16., 241-250.

[19] Reiss, R. D. and Thomas, M. (2007). “Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and other Fields”, Birkhauser Basel- Boston-Berlin.

[20] Resnick, S. I. (1987). “Extreme Values, Regular Variation, and Point Processes”, Springer Verlag, New York.

[21] Smith, R. L. (1986). “Extreme Value Theory Based on the r Largest Annual Events”, Journal of Hy- drology 86., 27-43.

[22] Smith, R. L. (1990). “Extreme Value Theory”, Handbook of Applicable Mathematics (Supplement), W. Ledermann (ed.), John wiley, Chichester.

[23] Tawn, J. A. (1988a). “Bivariate Extreme Value Theory: Models And Estimation”, Biometrika 75., 397-415.

[24] Von Mises. (1936). “La Distribution de la Plus Grande de n Valeurns”, Reprinted in Selected Papers 11, Amer. Math. Soc. Providence, R. I. (1954)., 271-294.

[25] Weissman, I. (1978). “Estimation of Parameters and Quantiles Based on the k Largest Observations”, Journal of the American Statistical Association 73., 812-815.