*Mausam*, (1990), **41**, 4, 627-630

551.577.37

# Finite mixture of extreme value distribution for rainfall frequency analysis

D. S. UPADHYAY and SURINDER KAUR

*Meteorological Office, New Delhi*

(*Received 24 July 1987*)

**सार** — 3-4 बहिर्शासी (आऊटलायर्स) से युक्त भुज स्टेशन (गुजरात) में (70 वर्षों) के 24 घंटेवार अधिकतम वर्षा अनुक्रमों के आवृत्ति विश्लेषण के लिए ई वी$_1$ और ई वी$_2$ बंटनों के मिश्रण का प्रयास किया गया । दूसरे मामले के अध्ययन में ई वी$_1$ और ई वी$_2$ के मिश्रम का प्रयोग, दिन हाटे (पश्चिम बंगाल) की (58 वर्षों) चरम वर्षा की व्याख्या करने के लिए किया गया । ई वी बंटनों के परिमित मिश्रण उन सब मामलों में सही बैठते है जहां पर ई वी$_1$ अकेला प्रत्यागमन काल मानों का अधिक या कम आकलन करता है ।

ABSTRACT. A mixture of EV$_1$ and EV$_2$ distributions has been tried for the frequency analysis of 24-hourly maximum rainfall series (70 years) at Bhuj containing 3-4 outliers. In a second case study, a mixture of EV$_1$ and EV$_2$ is used to describe extreme rainfall (58 years) of Dinhata. A finite mixture of EV distributions porvides a better fit in all those cases where EV$_1$ alone over or under estimates return period values.

## 1. Introduction

The main distributions used for analysis of annual maximum precipitations series are log normal, extreme value types 1,2 & 3 (EV$_1$ , EV$_2$ & EV$_3$), Pearson and log Pearson type III. Though the observed series may satisfy the criteria for accuracy, consistency, homogeneity, stationarity, randomness and sufficiency of record length, a problem which affects the applicability of these distributions is the presence of outliers. In the present paper, attempt has been made to combine EV$_1$, linearly with EV$_2$ to tackle the outliers and with EV$_3$ to tackle the cases exhibiting small range of extremes. The main papers consulted are by Boes (1966), Chow (1953), Gumbel (1941) and Jenkinson (1955).

## 2. Mixture of two distributions

*List of symbols used*

$E_1(x)$ —Distribution function of extreme value type 1

$E_2(x)$ —Distribution function of extreme value type 2

$E_3(x)$ —Distribution function of extreme value type 3

$E(x)$ —Distribution function of mixed distribution

$E_n(x)$ —Sample distribution function

$\theta$ —Mixing proportion

$u,\alpha,k$ —Parameters of extreme value distribution

$y$ —Reduced variate of EV distribution

$x_1$, $x_2$, $x_3$, $x_T$ —$T$-year value in respect of EV$_1$, EV$_2$, EV$_3$ and mixed distribution respectively

$v(x_1)$, $v(x_2)$, $v(x_3)$ & $v(x_T)$ — Variances of statistics $x_1$, $x_2$, $x_3$ & $x_T$

SE $(x_T)$ — Standard error of $x_T$.

2.1. If $E_1$ $(x)$ and $E_2$ $(x)$ are two distribution functions, a linear combination may be defined as :

$$E_\theta (x) = \left( \tfrac{1}{2} + \theta \right) E_1 (x) + \left(\tfrac{1}{2} - \theta\right) E_2 (x) \qquad (1)$$

where, $-\frac{1}{2} \leqslant \theta \leqslant \frac{1}{2}$

$$\therefore \theta = \frac{E_\theta (x) - \frac{1}{2} [E_1 (x) + E_2 (x)]}{E_1 (x) - E_2 (x)} \qquad (2)$$

As the nature of prior information about $E_\theta$ $(x)$ is generally not known, it may be estimated by empirical distribution function $E_n(x)$ of the observed sample $(x_1, x_2, \ldots\ldots, x_n)$. An estimator for $\theta$ is :

$$\overset{\wedge}{\theta} = \frac{E_n (x) - \frac{1}{2} [E_1 (x) + E_2 (x)]}{E_1 (x) - E_2 (x)} \qquad (3)$$

$\overset{\wedge}{\theta}$ provides unbiased and consistent estimate of $\theta$ with variance of order $(1/n)$. However, when $\overset{\wedge}{\theta}$ lies outside the interval $(-1/2,\ 1/2)$, we may truncate the values going beyond $-1/2$ or $1/2$, although by doing so unbiasedness of the estimator is lost.

It may be seen that the sample observations form a discrete series. $E_n$ $(x)$ represents a step deviation function, whereas $E_1$ $(x)$ and $E_2$ $(x)$ form continuous curves. A mean curve for $E_n(x)$ should be estimated for obtaining $\theta$.

### 2.2. *Mixing of two $EV_1$ distributions with different parameters*

Situations may arise when a sample of observations is considered in two parts one drawn from population $E_1$ $(x;\ u_1,\ \alpha_1)$ and the other from $E_1\,(x;\ u_2, \alpha_2)$. The entire series may be assumed to have come from the mixed distribution:

$$E_\theta\,(x) = (\tfrac{1}{2} + \theta)\, E_1\,(x\,;\, u_1,\, \alpha_1) + (\tfrac{1}{2} - \theta)\;\; E_1\,(x\,;\, u_2,\, \alpha_2) \tag{4}$$

For large samples $E_\theta$ $(x)$ may be estimated by $E_n$ $(x)$ and $\theta$ by $\hat{\theta}$ given as :

$$\hat{\theta} = \frac{E_n\,(x) - \tfrac{1}{2}\,[E_1\,(x\,;\, u_1,\, \alpha_1) + E_1\,(x\,;\, u_2,\, \alpha_2)\,]}{E_1\,(x\,;\, u_1,\, \alpha_1) -\!\!- E_1\,(x\,;\, u_2,\, \alpha_2)} \tag{5}$$

The return period values $(x_T)$ may be obtained by:

$$x_T = (\tfrac{1}{2} + \hat{\theta})(u_1 + \alpha_1\, y_1) + (\tfrac{1}{2} - \hat{\theta})\,(u_2 + \alpha_2\, y_1) \tag{6}$$

$$y_1 = -\ln\ln\left(\frac{T}{T-1}\right),\ \text{is the reduced variate.}$$

### 2.3. *Mixing of $EV_1$ with $EV_2$ or $EV_3$*

If parameters of $\text{EV}_1$ are $(u_1,\ \alpha_1,\ y_1)$ and those of $\text{EV}_2$ are $(u_2,\ \alpha_2,\ y_2,\ k)$, then the return period values $(x_T)$ of the series are given by :

$$x_T = \left(\frac{1}{2} + \theta\right)\left(u_1 + \alpha_1\ y_1\right) + \left(\frac{1}{2} - \theta\right)$$

$$\left(u_2 + \frac{\alpha_2}{k} - \frac{\alpha_2}{k}\ y_2\right) \tag{7}$$

where, $y_2 = e^{-ky_1} \approx (1 - k y_1)$ (for small values of $k$)

$$\therefore\ x_T = (\tfrac{1}{2} + \theta)(u_1 + \alpha_1\, y_1) + (\tfrac{1}{2} - \theta)\ (\ u_2 + \alpha_2\, y_1\ )$$
$$= u^* + \alpha^*\ y_1 \tag{8}$$

where, $u^* = \bar{u} + \theta u'$,

$\alpha^* = \bar{\alpha} + \theta\alpha'$ and

$$\bar{u} = \frac{u_1 + u_2}{2}\ ,\ \bar{\alpha} = \frac{\alpha_1 + \alpha_2}{2},$$

$$u' = u_1 - u_2\ ,\ \alpha' = \alpha_1 - \alpha_2.$$

Thus, the mixture corresponds to an $\text{EV}_1$ distribution with parameters $u^*$ and $\alpha^*$.

*Variance of $x_T$*

$$v\,(x_T) = \frac{\bar{v}}{2} + 2\theta^2\,\bar{v} + \theta\, v', \tag{9}$$

where, $\bar{v} = \tfrac{1}{2}\,[v(x_1) + v(x_2)]$ & $v' = v(x_1) - v(x_2)$

and $v$ $(x_1)$ & $v$ $(x_2)$ are the variances when estimated by $\text{EV}_1$ and $\text{EV}_2$ distributions respectively.

### 3. Illustration

The mixture of two distributions is illustrated by two practical case studies in succeeding paras. These cases represent typical 24-hourly annual maximum rainfall series, where a single $\text{EV}_1$ distribution does not provide a reasonable fit to the observed sample.

#### 3.1. *Case study 1*

Daily (24-hr) annual maximum series (mm) for 70 years (1901-1970) was prepared for Bhuj (Gujarat). Maximum likelihood estimates of $\text{EV}_1$ and $\text{EV}_2$ parameters for this data set are :

$$u_1 = 60.4,\ \alpha_1 = 42.1\ \ ;$$

$$u_2 = 55.3,\ \alpha_2 = 30.6\ \ \text{and}\ \ k = -0.34.$$

#### 3.2. *Fitting of linear mixture of $EV_1$ and $EV_2$ distributions*

The distribution function of the observed sample is $E_n$ $(x) = m/n$, where $m$ is number of observations $\leqslant x$.

The distribution functions of $\text{EV}_1$ and $\text{EV}_2$ are :

$$E_1\,(x) = \exp\left[\exp\left(\frac{-x - 60.4}{42.1}\right)\right]$$

and

$$E_2\,(x) = \exp\left[-\left(1 + \frac{x - 55.3}{90}\right)\right]^{-2\cdot 94}$$

It is possible that the value of $\theta$ for some observed values of $x$ may fall outside $(-0.5,\ 0.5)$ interval. In such cases we take $\theta = 0.5$ or $-0.5$ whichever is closer to computed values. Evaluating $\theta$ from Eqn. (3) for each observed value of $x$ and taking their arithmetic mean, we get :

$$\theta = -0.1$$

The mixed distribution function, $E(x)$ is taken as :

$$E(x) = 0.4\ E_1\,(x) + 0.6\ E_2\,(x)$$

and $x_T = -8.7 + 16.8\ y_1 + 54\ e^{-0\cdot 34 y_1}$ ,

$$v(x_T) = 0.16\ v\,(x_1) + 0.36\ v\,(x_2)$$

The return period values $(x_T)$ and SE $(x_T)$ for the mixed distribution are provided in the Table 1.

All observations fall under 2 SE limits and the mixture distribution $E\,(x) = .4\ E_1\,(x) + .6\ E_2\,(x)$ fits better to the observed series than $\text{EV}_1$ alone. A comparison of $E(x)$ and $\text{EV}_1$ in respect of 3 highest recorded values is given in Table 2.

The above shows marked improvement in the recurrence interval value when we are using mixed distirbution though compared to actuals, the values are still very much higher.

TABLE 1

| Return period $T$ (yr) | EV$_1$ $x_1$ (mm) | EV$_2$ $x_2$ (mm) | Mixed distribution | | | |
|---|---|---|---|---|---|---|
| | | | $x_T$ (a) (mm) | SE($x_T$) (b) (mm) | $x_T$ ±2SF ($x_T$) (a)—2(b) (mm) | (a)+2(b) (mm) |
| 2 | 76 | 67 | 71 | 3.9 | 63 | 78 |
| 5 | 123 | 115 | 118 | 7.6 | 104 | 133 |
| 10 | 155 | 159 | 157 | 18.7 | 120 | 194 |
| 50 | 224 | 304 | 272 | 42.9 | 186 | 358 |
| 100 | 254 | 395 | 339 | 68.2 | 202 | 475 |
| 200 | 283 | 509 | 419 | 104.8 | 202 | 628 |
| 300 | 300 | 590 | 474 | 133.8 | 211 | 742 |
| 500 | 321 | 709 | 554 | 179.4 | 195 | 913 |
| 1000 | 351 | 908 | 685 | 264.3 | 157 | 1214 |
| 2000 | 380 | 1156 | 847 | 381.0 | 85 | 1509 |

TABLE 2

| $x$ (mm) | Recurrence interval (years) | | |
|---|---|---|---|
| | EV$_1$ | Mixed distri-bution | Plotting Position |
| 467 | exceeds 10,000 | 256 | 126 |
| 352 | 714 | 82 | 46 |
| 308 | 357 | 78 | 28 |

TABLE 3

| $T$ | $x_1$ (mm) | $x_3$ (mm) | $x_T$ (a) (mm) | SE($x_T$) (b) (mm) | $x_T$ ±2SE ($x_T$) (a)—2(b) (mm) | (a)+2(b) (mm) |
|---|---|---|---|---|---|---|
| 2 | 170 | 175 | 173 | 5.6 | 162 | 184 |
| 5 | 222 | 221 | 221 | 7.2 | 207 | 236 |
| 10 | 256 | 248 | 250 | 8.7 | 233 | 268 |
| 50 | 333 | 296 | 307 | 14.2 | 278 | 336 |
| 100 | 365 | 313 | 328 | 17.2 | 294 | 363 |
| 200 | 397 | 327 | 348 | 20.5 | 307 | 191 |
| 300 | 416 | 335 | 359 | 22.6 | 316 | 404 |
| 500 | 439 | 344 | 373 | 25.2 | 322 | 424 |
| 1000 | 472 | 356 | 391 | 29.0 | 333 | 441 |
| 2000 | 504 | 366 | 407 | 32.6 | 342 | 472 |
| 5000 | 546 | 378 | 428 | 37.4 | 353 | 503 |
| 10000 | 578 | 385 | 443 | 40.8 | 361 | 525 |

### 3.3. *Case study* 2

In this case the highest values exhibit very small increase in the ordered series form from 58 years' daily rainfall (mm) recorded at Dinhata; the last few values are 269, 270, 275, 278, 281 and 290 which show hardly 2 cm range.

The equation for $x_T$ is :

$$x_1 = 152.5 + 46.2\, y_1$$

and $\text{SE}\,(x_1) = 36.7\,(1.\ 11{+}0.52 y_1{+}0.61 y_1^2)$.

For EV$_3$ ($u_3$, $\alpha_3$, $k$) distribution the maximum likelihood estimates for parameters are :

$$u_3 = 157.3,\ \alpha_3 = 48.3\ \&\ k = 0.17$$

$$\therefore\ x_3 = 448.3 - 2.91 \exp\,(-0.17\, y_1).$$

In fitting a linear mixture of EV$_1$ and EV$_3$, $\theta$ has been evaluated for each observed value of $x$ truncated to the nearest limiting value whenever it fell outside the limits (—1/2, 1/2) and then by simple arithmetic mean :

$$\theta = -0.2$$

Mixed distribution function is :

$$E(x) = 0.3\, E_1\,(x) + 0.7\, E_3\,(x)$$

$$x_T = 354.7 + 13.9\, y_1 - 198.9\, e^{-0.17\, y_1}$$

and $v(x_T) = 0.09\, v(x_1) + 0.49\, v(x_3)$.

The results are provided in Table 3.

The mixed distribution $E(x){=}0.3E_1\,(x){+}0.7E_3(x)$ provides a better fit to the observed series than $E_1\,(x)$ but in the present case, $E_3\,(x)$ appears to be equally or more efficient than EV$_1$.

## 4. Summary

In a number of cases Gumbels EV$_1$ fails to describe the observed extreme series with acceptable accuracy. Two of the types are :

(*i*) When the series has one or more outliers, whose return period as computed from EV$_1$ is several times larger than the plotting position ($T_0$) values of the outliers,

(*ii*) When the end values (arranged in ascending order) of the series show very slow increase so that their return periods are much less than $T_0$.

(2) In case (*i*) a linear mixture of EV$_1$ and EV$_2$, *i.e.*, $E\,(x) = (\tfrac{1}{2} + \theta)\ E_1\ (x) + (\tfrac{1}{2} - \theta)\ E_2\,(x)$ seems to provide a better fit to the data as illustrated by Bhuj series. This series has 3 outliers but all these have been brought below $x_T$ + 2SE ($x_T$) limit using the mixed distribution:

$$E(x) = 0.4\, E_1\,(x) + 0.6\, E_2\,(x).$$

(3) In case (*ii*) a linear combination of $EV_1$ and/or $EV_3$ is suggested. It will enhance the recu:rence interval (*T*) for the end values of the series. This is illustrated in sample study 2 using Dinhata data. The highest value of observed series (290 mm) is less $x_1$—2SE ($x_1$) when $EV_1$ alone is used for frequency analysis. But it comes under 2SE to describe the observations.

(4) From the above it may be inferred that the linear combination of $EV_1$ with $EV_2$ and/or $EV_3$ [case (*i*) and (*ii*) respectively] would improve generally the results of frequency analysis of extremes as compared those obtained by using $EV_1$ alone.

*Acknowledgements*

The authors are grateful to Dr. R. P. Sarker, the Director General of Meteorology and prof. B. L. S. Prakasa Rao, of Indian Statistical Institute for their keen interest and valuable discussions. Thanks are also due to Dr. S. N. Kathuria of I. M. D. for his help in completing the study. A special gratitude is expressed for valuable suggestions from the learned referee.

## References

Boes, D.C., 1966, 'Estimation on mixing distribution', *Ann. Math. Stat.*, **37**, pp. 177-188.

Chow, V.T., 1953, 'Freq. Analysis of Hydrological Data', Univ. Illinois, *Statistical Bulletin.*

Gumbel, E.J., 1941, *Ann. Math. Stat.*, **12**, 2, pp. 163-190.

Jenkinson, A.F., 1955, 'Freq. dist. of annual Max. values of Met. elements, *Quart. J.R. met. Soc.*, **81**, 348, pp. 158-171.

Upadhyay, D.S. and Misra, D.K., 1977, '*Prarambhic Genitiya Sankhyik*', Rajasthan Granth Academy Publication.