

STK4150/9150 - Environmental and Spatial Statistics

Univariate extremes

Thordis Thorarinsdottir

2024-04-18

The distribution of maxima

Assume $Y_1, \dots, Y_n$ are i.i.d. random variables with distribution function $F_Y(y)$.

The extreme value of a finite number $Y_1, \dots, Y_n$ is then $Z_n = \max\{Y_1, \dots, Y_n\}$.

The distribution of Z_n is

$$F_{Z_n}(z) = \mathbb{P}(Z_n \leq z) = \mathbb{P}(Y_1 \leq z, \dots, Y_n \leq z) = \mathbb{P}(Y_1 \leq z) \cdots \mathbb{P}(Y_n \leq z) = (F_Y(z))^n.$$

Then, we study the behaviour of $(F_Y(z))^n$ as $n \rightarrow \infty$.

Now, for any z such that $F_Y(z) < 1$, we have $(F_Y(z))^n \rightarrow 0$ as $n \rightarrow \infty$.

Thus, we need to rescale.

The distribution of maxima, part II

Let

$$Z_n^* = \frac{Z_n - b_n}{a_n}$$

for suitable sequences of constants $a_n > 0$ and b_n that are chosen to stabilize the location and scale of Z_n^* as $n \rightarrow \infty$.

Then, there are only three types of limiting distributions for the renormalized Z_n^* .

Fisher-Tippett-Gnedenko theorem (Extreme Types Theorem)

If there exist constants $a_n > 0$ and b_n such that

$$\mathbb{P}\left(\frac{Z_n - b_n}{a_n} \leq z\right) \rightarrow G(z), \quad n \rightarrow \infty,$$

where $G(z)$ is a nondegenerate distribution function, then $G(z)$ belongs to one of the following three families

$$\text{I (Gumbel)} \quad G(z) = \mathbb{1}\{-\infty < z < \infty\} \exp\left(-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right),$$

$$\text{II (Fréchet)} \quad G(z) = \mathbb{1}\{z > b\} \exp\left[-\left(\frac{z-b}{a}\right)^{-c}\right],$$

$$\text{III (Weibull)} \quad G(z) = \mathbb{1}\{z < b\} \exp\left[-\left(\frac{b-z}{a}\right)^c\right] + \mathbb{1}\{z \geq b\},$$

for parameters $a > 0$, b and $c > 0$.

Generalized extreme value (GEV) distribution

All three types of extreme value distributions can be expressed in a common form known as the generalized extreme value (GEV) distribution

$$G(z) = \exp \left(- \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right), \quad \text{for } \left\{ z : 1 + \xi \left(\frac{z - \mu}{\sigma} \right) > 0 \right\},$$

where $\mu \in \mathbb{R}$ is a *location* parameter, $\sigma > 0$ is a *scale* parameter and $\xi \in \mathbb{R}$ is a *shape* parameter.

The Gumbel case with $\xi = 0$ must be interpreted as a limiting case when $\xi \rightarrow 0$.

Generalized extreme value (GEV) distribution

All three types of extreme value distributions can be expressed in a common form known as the generalized extreme value (GEV) distribution

$$G(z) = \exp \left(- \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right), \quad \text{for } \left\{ z : 1 + \xi \left(\frac{z - \mu}{\sigma} \right) > 0 \right\},$$

where $\mu \in \mathbb{R}$ is a *location* parameter, $\sigma > 0$ is a *scale* parameter and $\xi \in \mathbb{R}$ is a *shape* parameter.

The Gumbel case with $\xi = 0$ must be interpreted as a limiting case when $\xi \rightarrow 0$.

The k th moment of G exists only if $\xi < 1/k$, and the probability of extremely large values increases as ξ increases. Typically, $\xi \in [-0.5, 0.5]$.

Generalized extreme value (GEV) distribution

All three types of extreme value distributions can be expressed in a common form known as the generalized extreme value (GEV) distribution

$$G(z) = \exp \left(- \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right), \quad \text{for } \left\{ z : 1 + \xi \left(\frac{z - \mu}{\sigma} \right) > 0 \right\},$$

where $\mu \in \mathbb{R}$ is a *location* parameter, $\sigma > 0$ is a *scale* parameter and $\xi \in \mathbb{R}$ is a *shape* parameter.

The Gumbel case with $\xi = 0$ must be interpreted as a limiting case when $\xi \rightarrow 0$.

The k th moment of G exists only if $\xi < 1/k$, and the probability of extremely large values increases as ξ increases. Typically, $\xi \in [-0.5, 0.5]$.

Minima are treated by noticing that $\min\{Y_1, \dots, Y_n\} = -\max\{-Y_1, \dots, -Y_n\}$.

Block maxima

In practice, we take a long time series of observed data $y_1, y_2, \dots$, cut it into reasonably large blocks each of length n and generate a series of m *block maxima* $Z_{n,1}, \dots, Z_{n,m}$.

For example, for daily data the block size is often a year.

Here, we assume that

- the original time series is *stationary over the blocks* so that $Z_{n,1}, \dots, Z_{n,m}$ are identically distributed;
- $Z_{n,1}, \dots, Z_{n,m}$ are independent;
- n is large enough so that $\mathbb{P}((Z_n - b_n)/a_n \leq z) \approx G(z)$.

Further, we are in practice interested in $\mathbb{P}(Z_n \leq z) \approx G((z - b_n)/a_n) = G^*(z)$ which is also a GEV distribution, but with different parameters than G .

Block maxima, part II

In applications, we are often interested in the quantile z_p at probability level $p \in (0, 1)$, that is, z_p such that $G(z_p) = 1 - p$. It is given by

$$z_p = \begin{cases} \mu - (\sigma/\xi) [1 - (-\log(1 - p))^{-\xi}], & \xi \neq 0 \\ \mu - \sigma \log(-\log(1 - p)), & \xi = 0. \end{cases}$$

In environmental applications, z_p is often called the *return level* at *return period* $1/p$, meaning that z_p is expected to be exceeded, e.g., every $1/p$ year on average.

However, this is misleading in discussion of non-stationary phenomena.

Alternatively, we can say that z_p has *annual exceedance probability* $100p\%$.

Which underlying distribution F_Y yields which type of G ?

Most distributions yields a GEV distribution with $\xi \neq 0$. For example,

- the uniform distribution yields a Weibull distribution ($\xi = -1$)

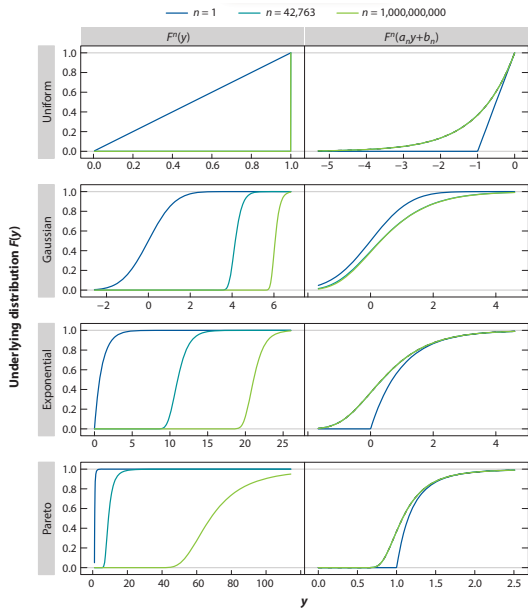
However,

- Gaussian and exponential distribution yield a Gumbel distribution ($\xi = 0$)

How fast is the convergence?

Sometimes painfully slow...

Thus, in practice, it may be convenient to use the GEV distribution rather than the individual types, as a different type may provide a better approximation of the distribution of $\max\{Y_1, \dots, Y_n\}$ for finite n .



Parameter estimation

- Maximum likelihood (sensitive to outliers)
- Bayesian inference (the prior moderates the sensitivity of the likelihood)
- Method of moments (more robust)

The method of moments often uses *L-moments*, with the first two given by

$$\lambda_1 = \mathbb{E}[Z]$$

$$\lambda_2 = \frac{1}{2}(\mathbb{E}[Z_{2:2}] - \mathbb{E}[Z_{1:2}]) = \frac{1}{2}\mathbb{E}|Z_1 - Z_2|$$

where $Z_{k:n}$ is the k th rank among n values, or *probability weighted moments*,

$$M_{p,r,s} = \mathbb{E}[Z^p (F_Z(Z))^r (1 - F_Z(Z))^s] = \int_0^1 (F_Z^{-1}(\tau))^p \tau^r (1 - \tau)^s d\tau,$$

where p, r, s are real numbers.

Example: Annual maximum flood at Bulken (Vangsvatnet)



Example: Annual maximum flood at Bulken, part II

```
library(data.table)
## Data from sildre.nve.no/station/62.5.0
tmp <- read.csv("images_ext/BulkenData.csv",
               sep=";", skip=1)
tmp <- tmp[,1:2]
names(tmp) <- c("date", "discharge")
tmp$discharge <- as.numeric(gsub(" ", ".", tmp$discharge))

dt <- as.data.table(tmp[,1:2])
year <- as.numeric(substring(tmp$date,1,4))
month <- as.numeric(substring(tmp$date,6,7))
day <- as.numeric(substring(tmp$date,9,10))
dt <- cbind(dt, year, month, day)
n <- max(which(year==2023))
dt <- dt[1:n,]
```

```
head(dt)
```

	date	discharge	year	month	day
	<char>	<num>	<num>	<num>	<num>
1:	1892-01-01 11:00:00Z	12.10079	1892	1	1
2:	1892-01-02 11:00:00Z	11.24779	1892	1	2
3:	1892-01-03 11:00:00Z	11.24779	1892	1	3
4:	1892-01-04 11:00:00Z	10.97214	1892	1	4
5:	1892-01-05 11:00:00Z	11.24779	1892	1	5
6:	1892-01-06 11:00:00Z	11.24779	1892	1	6

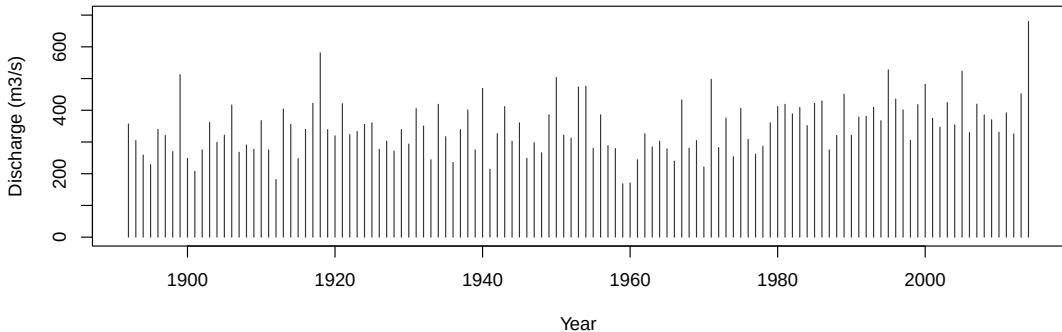
```
dt_am <- dt[,.(am=max(discharge)), by=year]
head(dt_am)
```

	year	am
	<num>	<num>
1:	1892	356.5735
2:	1893	304.5116
3:	1894	258.9497
4:	1895	228.6367
5:	1896	339.8720
6:	1897	320.5880

```
z <- dt_am$am[dt_am$year<2015]
z.year <- 1892:2014
```

Example: Annual maximum flood at Bulken, part III

```
plot(z.year, z, type="h", lty=1, ylim=c(0,700),  
     xlab="Year", ylab="Discharge (m3/s)", col="grey20")
```



Example: Annual maximum flood at Bulken, part IV

```
library(fExtremes)
library(evd)
pwm.all <- gevFit(z, type="pwm")@fit$par.ests
return.level <- function(T, xi, mu, sigma)
{
  if(xi!=0) {
    tmp <- 1 - (-log(1-1/T))^-xi
    return(mu - tmp*sigma/xi)
  }else{
    tmp <- log(-log(1-1/T))
    return(mu - tmp*sigma)
  }
}
T <- c(1:1000)
Zp.est <- return.level(T, pwm.all[1], pwm.all[2], pwm.all[3])
print(round(c(max(Zp.est), max(z))))
```

[1] 669 680

Example: Annual maximum flood at Bulken, part V

The data series has 123 years.

What are the chances that we have never seen an event with annual exceedance probability 0.1% (or a 1000-year event) in the course of these 123 years? (assuming the annual maxima are independent)

Example: Annual maximum flood at Bulken, part V

The data series has 123 years.

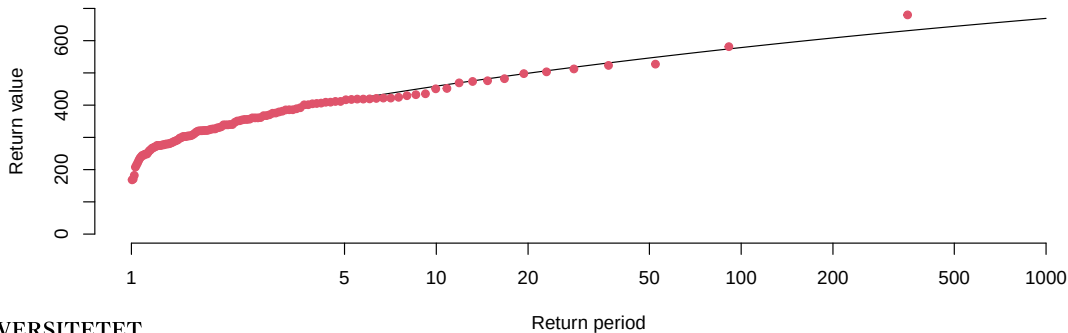
What are the chances that we have never seen an event with annual exceedance probability 0.1% (or a 1000-year event) in the course of these 123 years? (assuming the annual maxima are independent)

```
dbinom(0, 123, 0.001)
```

```
[1] 0.8842092
```

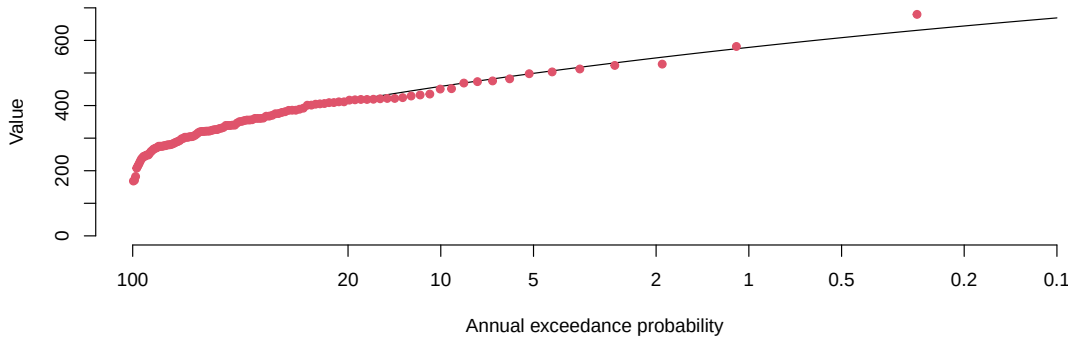
Example: Annual maximum flood at Bulken, part VI

```
plot(log(T),Zp.est,type="l",ylim=c(0,700),xlab="Return period",ylab="Return value",axes=FALSE)
axis(2); axis.vec <- c(1,5,10,20,50,100,200,500,1000); axis(1,at=log(axis.vec),label=axis.vec)
N <- length(z)
F <- ((1:N) - 0.35)/N ## use Hosking (1990) formula to plot data
T.z <- 1 / (1 - F)
points(log(T.z), sort(z), cex=0.9, col=2, pch=19)
```



Example: Annual maximum flood at Bulken, part VII

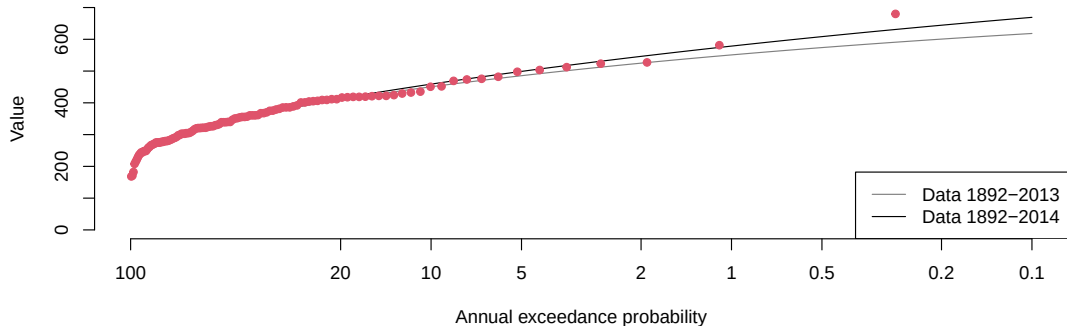
```
plot(log(T),Zp.est,type="l",ylim=c(0,700),xlab="Annual exceedance probability",
     ylab="Value",axes=FALSE)
l.new <- 100/c(1,5,10,20,50,100,200,500,1000); axis(2); axis(1,at=log(axis.vec),label=l.new)
points(log(T.z), sort(z), cex=0.9, col=2, pch=19)
```



Example: Annual maximum flood at Bulken, part VIII

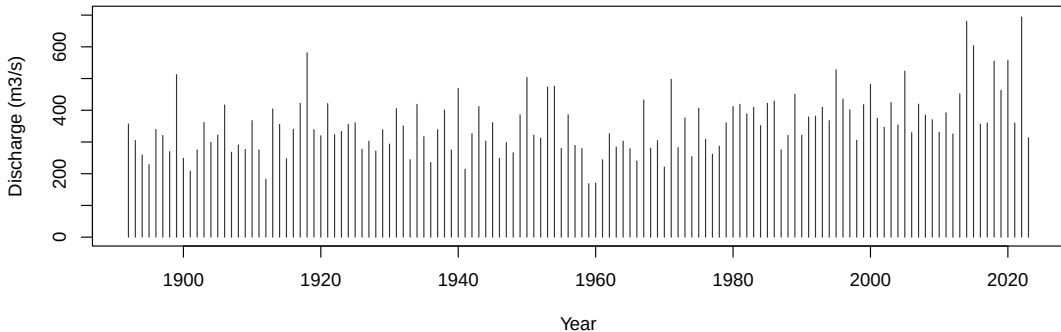
```
pwm.2013 <- gevFit(z[-123], type="pwm")@fit$par.ests
Zp.est.2013 <- return.level(T, pwm.2013[1], pwm.2013[2], pwm.2013[3])
print(round(c(max(Zp.est.2013), max(Zp.est), max(z))))
```

[1] 618 669 680

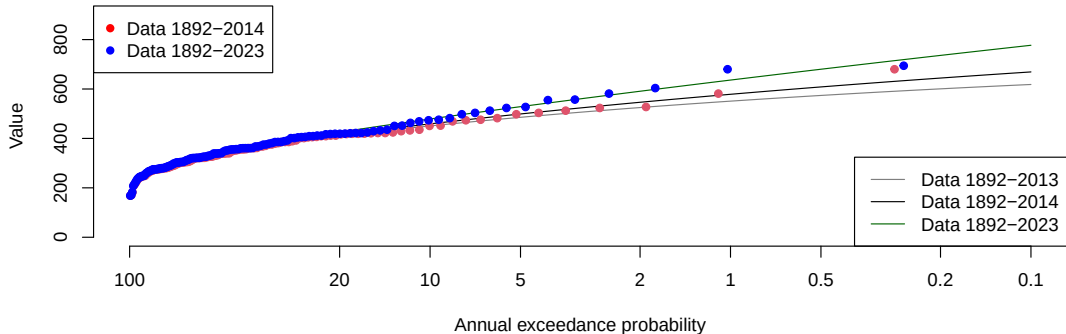


Example: Annual maximum flood at Bulken, part IX

```
plot(dt_am$year, dt_am$am, type="h", lty=1, ylim=c(0,700),  
     xlab="Year", ylab="Discharge (m3/s)", col="grey20")
```



Example: Annual maximum flood at Bulken, part X



Extending the information set

An alternative to using block maxima for inference on extremes is to use threshold exceedances. There are a few reasons for this

- only looking at the block maxima may waste (precious) information as there may be more than one extreme event within a block
- we might need information which an analysis of block maxima cannot provide, e.g. the likelihood of a succession of large values (clustering of extremes)

Independent exceedances

Assume the data $Y_1, \dots, Y_n$ are iid with distribution F . Then the distribution of the rescaled maximum $Z_n^* = (\max\{Y_1, \dots, Y_n\} - b_n)/a_n$ is

$$\mathbb{P}(Z_n^* \leq z) = (F(b_n + a_n z))^n.$$

As $\log u \approx u - 1$ when $u \rightarrow 1$, this probability converges to the GEV distribution $G(z)$ as $n \rightarrow \infty$ if and only if

$$\begin{aligned} n(1 - F(b_n + a_n z)) &\approx -n \log F(b_n + a_n z) \\ &\rightarrow -\log G(z) \\ &= (1 + \xi(z - \mu)/\sigma)_+^{-1/\xi} \end{aligned}$$

Independent exceedances, part II

Further, we have

$$\mathbb{E}\left(\sum_{j=1}^n \mathbb{1}\{Y_j \geq b_n + a_n z\}\right) = n(1 - F(b_n + a_n z)).$$

It follows that $\{(Y_j - b_n)/a_n : j = 1, \dots, n\}$ converge to a Poisson process on $\mathbb{R}$ with mean measure $-\log G(z)$.

Independent exceedances, part II

Further, we have

$$\mathbb{E}\left(\sum_{j=1}^n \mathbb{1}\{Y_j \geq b_n + a_n z\}\right) = n(1 - F(b_n + a_n z)).$$

It follows that $\{(Y_j - b_n)/a_n : j = 1, \dots, n\}$ converge to a Poisson process on $\mathbb{R}$ with mean measure $-\log G(z)$.

Similarly, the two-dimensional point process

$$\{j/(n+1), (Y_j - b_n)/a_n : j = 1, \dots, n\}$$

converges to a Poisson process on $[0, 1] \times \mathbb{R}$ with measure

$$\Lambda([t_1, t_2] \times [z, \infty)) = (t_2 - t_1)(1 + \xi(z - \mu)/\sigma)_+^{-1/\xi}, \quad 0 \leq t_1 < t_2 \leq 1, z \in \mathbb{R}.$$

Independent exceedances, part III

This implies that event in the limiting process that exceed a threshold u occur according to a Poisson process of rate $\zeta_u = (1 + \xi(u - \mu)/\sigma)_+^{-1/\xi}$ and that

$$\mathbb{P}(Y > u + y | Y > u) = \frac{(1 + \xi(y + u - \mu)/\sigma)_+^{-1/\xi}}{(1 + \xi(u - \mu)/\sigma)_+^{-1/\xi}} = (1 + \xi y/\tau_u)_+^{-1/\xi},$$

where $\tau_u = \sigma + \xi(u - \mu)$ must be positive for there to be exceedances of u .

The expression on the right is the survivor function of the generalized Pareto distribution (GPD).

Peak over threshold (PoT)

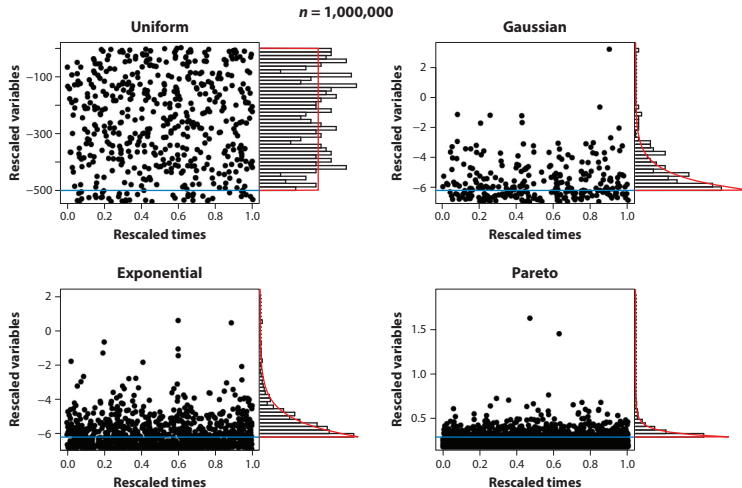
The *peak over threshold* (PoT) model for a long time series of observed data $y_1, y_2, \dots$, is that

- exceedances over a high threshold u occur as a Poisson process with *rate* parameter $\zeta_u > 0$
- sizes of excesses
 - are mutually independent
 - are independent of the Poisson process
 - follow a GPD with distribution function

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tau_u}\right)^{-1/\xi},$$

where $\tau_u > 0$ is a *scale* parameter and $\xi \in \mathbb{R}$ is a *shape* parameter.

Some examples (<https://vimeo.com/109872288>)



Connection between block maxima and PoT

For random variables $Y_1, Y_2, \dots$ representing daily values, suppose the PoT model with parameters (ζ_u, τ_u, ξ) holds for excesses over a threshold u with ζ_u as the yearly rate of the Poisson process.

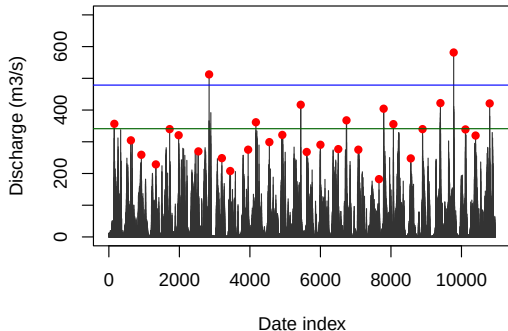
Then, the annual maxima $Z_{365,1}, Z_{365,2}, \dots$ follow a GEV distribution with parameters

$$\mu = u + \frac{(\zeta_u^\xi - 1)\tau_u}{\xi}, \quad \sigma = \tau_u \zeta_u^\xi, \quad \xi.$$

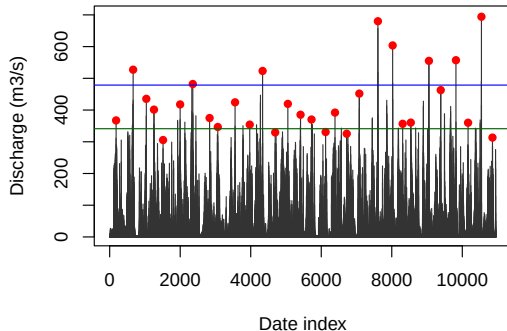
We can thus also obtain estimates of e.g. return values for certain return periods or annual exceedance probabilities from the PoT framework.

Back to Bulken

1892–1921



1994–2023



Red dots: annual maxima; green line: estimated 50% annual exceedance probability; blue line: estimated 10% annual exceedance probability

Dependent data

If the underlying data series $y_1, y_2, \dots$ is dependent, then

- under mild conditions, the limiting process for the exceedances is compound Poisson;
- clusters of exceedances of u appear according to a Poisson process of rate $\theta\zeta_u$, where $\theta \in (0, 1]$ is the *extremal index* and the mean cluster size is θ^{-1} ;
- the mean number of exceedances per unit of time does not change, but because exceedances occur in clusters, the clusters themselves necessarily tend to be further apart.

Extremal index

The *extremal index* of a stationary process summarizes the degree of clustering of its extremes.

It may be defined in different ways that are equivalent under mild conditions.

Extremal index: reciprocal of the mean cluster size

We call $\{u_n\}$ a threshold sequence if $n(1 - F(u_n)) \rightarrow \zeta > 0$ as $n \rightarrow \infty$.

The blocks approach defines the extremal index as the reciprocal of the limiting expected number of exceedances of a threshold sequence $\{u_n\}$ in a block of length $r_n = o(n)$.

That is, it is defined as the reciprocal of

$$\lim_{n \rightarrow \infty} \sum_{j=1}^{r_n} \mathbb{E}(\mathbb{1}\{Y_j > u_n\} | Z_{r_n} > u_n) = \lim_{n \rightarrow \infty} \frac{r_n(1 - F(u_n))}{\mathbb{P}(Z_{r_n} > u_n)}.$$

Extremal index: based on the subasymptotic extremal index

The probability that an arbitrary exceedance of u is followed by a run of $r - 1$ values below u is

$$\theta(u, r) = \mathbb{P}(Y_2 \leq u, \dots, Y_r \leq u | Y_1 > u).$$

Then, the extremal index may be defined as

$$\theta = \lim_{n \rightarrow \infty} \theta(u_n, r_n),$$

with u_n and r_n as before.

Cluster sizes

The limiting cluster size distribution can then be defined as

$$\pi(i) = \lim_{n \rightarrow \infty} \mathbb{P}\left(\sum_{j=1}^{r_n} \mathbb{1}\{Y_j > u_n\} = i \mid Y_1 > u_n\right), \quad i = 1, 2, \dots$$

Cluster maxima vs. arbitrary exceedances

A consequence here is that cluster maxima and arbitrary exceedances have the same GPD. To see this, let

$$\theta(u_n) = \frac{\mathbb{P}(Z_{r_n} > u_n)}{r_n(1 - F(u_n))} = \frac{\mathbb{P}(Z_{r_n} > u_n)}{r_n \mathbb{P}(Y_1 > u_n)},$$

This sequence tends towards θ for any threshold sequence $\{u_n\}$.

Cluster maxima vs. arbitrary exceedances, part II

Let $y > 0$ and define another threshold sequence $u'_n = u_n + ya_n$ where $\{a_n\}$ is the renormalizing sequence for convergence of maxima.

Then,

$$\mathbb{P}\left(\frac{Z_{r_n} - u_n}{a_n} > y \mid Z_{r_n} > u_n\right) = \frac{\mathbb{P}(Z_{r_n} > u'_n)}{\mathbb{P}(Z_{r_n} > u_n)} = \frac{\theta(u'_n)}{\theta(u_n)} \frac{\mathbb{P}(Y_1 > u'_n)}{\mathbb{P}(Y_1 > u_n)} = \frac{\theta(u'_n)}{\theta(u_n)} \mathbb{P}\left(\frac{Y_1 - u_n}{a_n} > y \mid Y_1 > u_n\right)$$

and since $\theta(u'_n)/\theta(u_n) \rightarrow 1$ as $n \rightarrow \infty$, it follows from the stationarity of the process that the limiting distribution of cluster maximum that exceeds the threshold is the same as that of an arbitrary observation that exceeds the threshold.

Cluster maxima vs. arbitrary exceedances, part III

It follows that we can estimate the GPD parameters using either all of the exceedances or only the cluster maxima.

In the former, we must account for potential dependence when we assess uncertainty in the estimation.

In the latter, we must account for potential biases in the identification of cluster maxima.

References

Naess, A. (2023) *Applied extreme value statistics*. Available at https://folk.ntnu.no/arvidn/AN/BOOK_160223.pdf.

Olafsdottir, H. K., Rootzén, H., & Bolin, D. (2021). Extreme rainfall events in the Northeastern United States become more frequent with rising temperatures, but their intensity distribution remains stable. *Journal of Climate*, 34(22), 8863-8877.