Modelling extremes with R

Ilaria Prosdocimi
Ca' Foscari University of Venice, Italy

www.unive.it

@ilapros

Using R In Hydrology

vEGU 2021

Modelling extremes with R

1. Modelling extremes
2. with R

R is the language in which the main developments in Statistical modelling of extremes are implemented

Today only a brief introduction: comprehensive list of R package is available at the Extremes taskview

Today: only univariate models

Statistical models for extremes

Statistics mostly focuses on the *typical* behaviour of variables (i.e. *the mean*)

Characterisation of extremes is important for risk quantification and infrastructure design

Extremes behave somewhat differently from the rest of the distribution

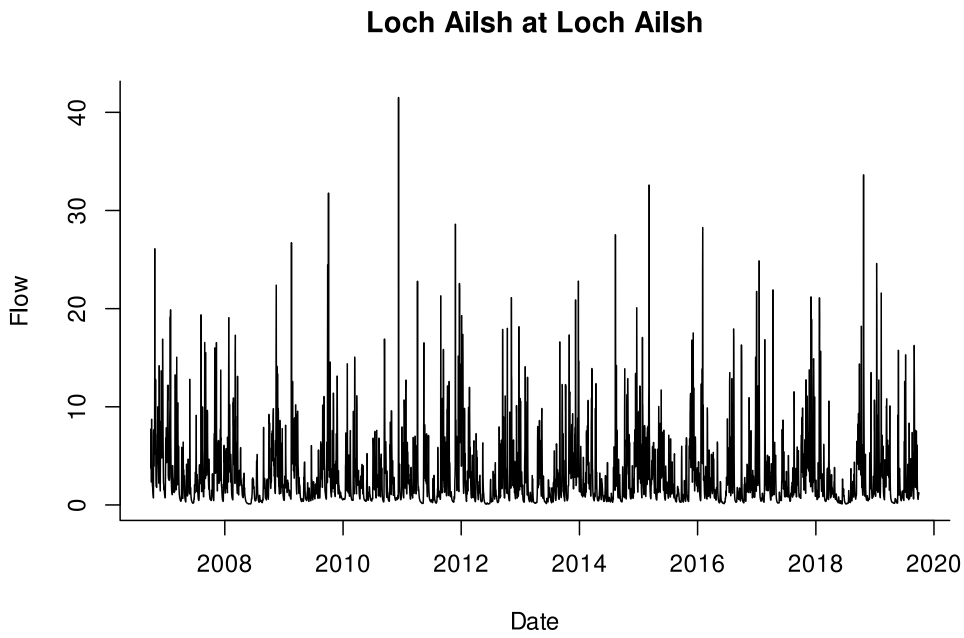
$\rightarrow$ extreme value theory gives theoretical results for extremes

Statistical literature and engineering/earth sciences literature and practice are not always aligned

First question: what is an extreme?

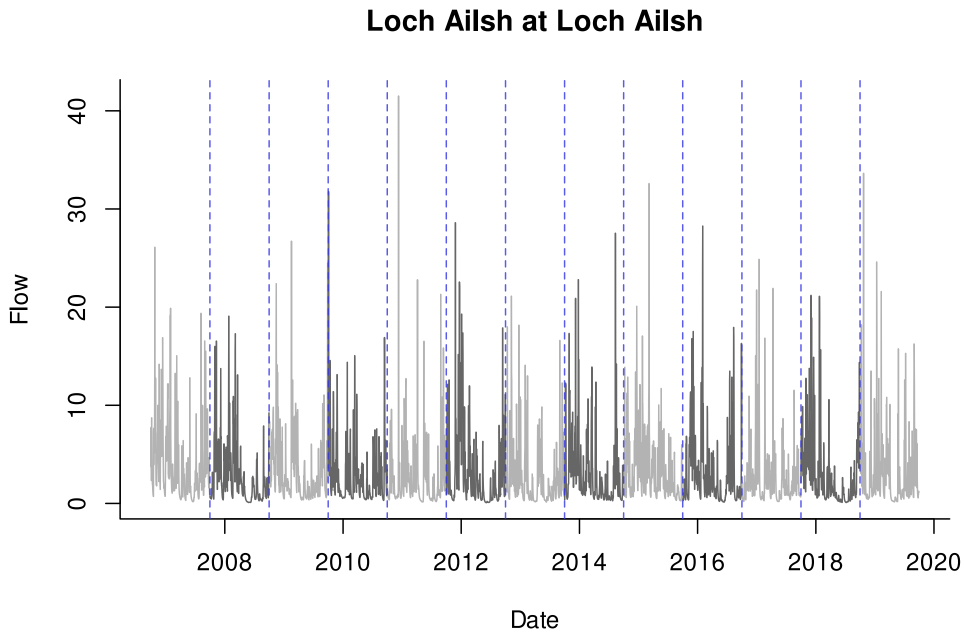
River flow dataset

```
datRiv <- rnrfa::gdf(3006)
```



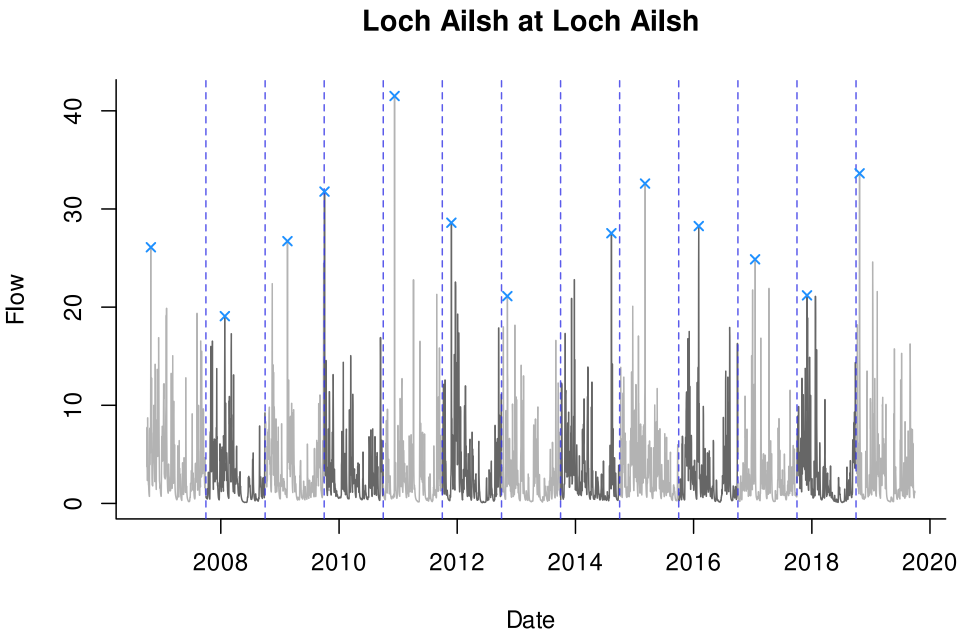
Extremes: block (annual) maxima

What is an extreme? The largest event in the water year.



Extremes: block (annual) maxima

What is an extreme? The largest event in the water year.



Extremes: theory for annual maxima

If we know that $X \sim f(x)$ - we have that
$P(max(X_1, \ldots, X_n) \leq z) = P(X_1 \leq z) \times \ldots \times P(X_n \leq z) = \{F(z)\}^n$

But $f(x)$ is unknown (and difficult to estimate).

It can be shown that $M_n = max(X_1, \ldots, X_n)$ has a limiting distribution which is the GEV:

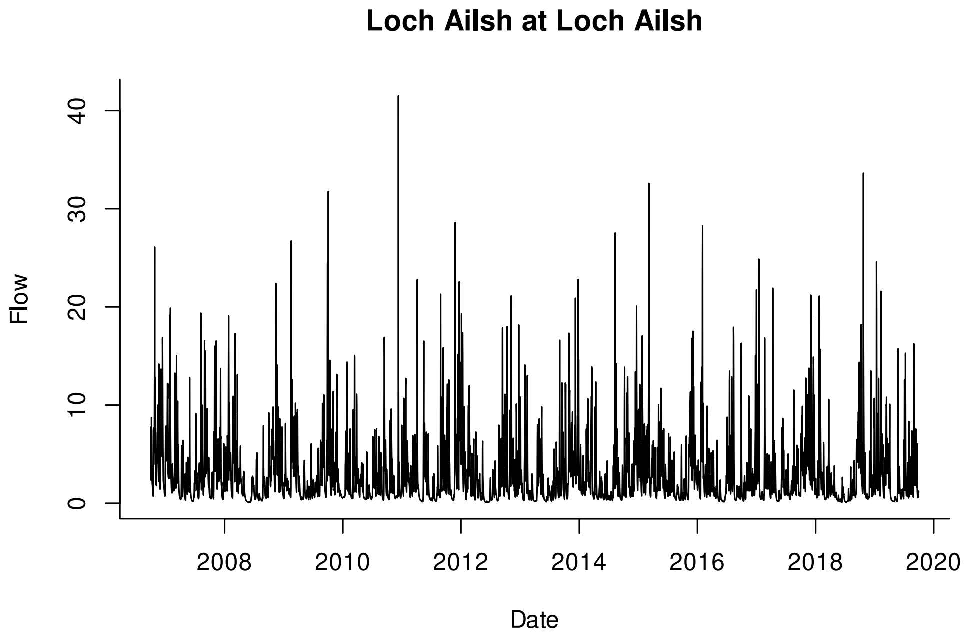
$$F(y) = \exp\left\{ -\left(1 + \xi \frac{y - \mu}{\sigma}\right)^{-1/\xi} \right\}$$

$Y \sim GEV(\mu, \sigma, \xi)$ is defined on $y : 1 + \xi(y - \mu)/\sigma > 0$ so the domain changes depending on the sign of ξ.

BUT! In engineering/hydrology $Y \sim GEV(\xi, \alpha, \kappa)$ and $\kappa = -\xi$. Software use different parametrisations - always check!

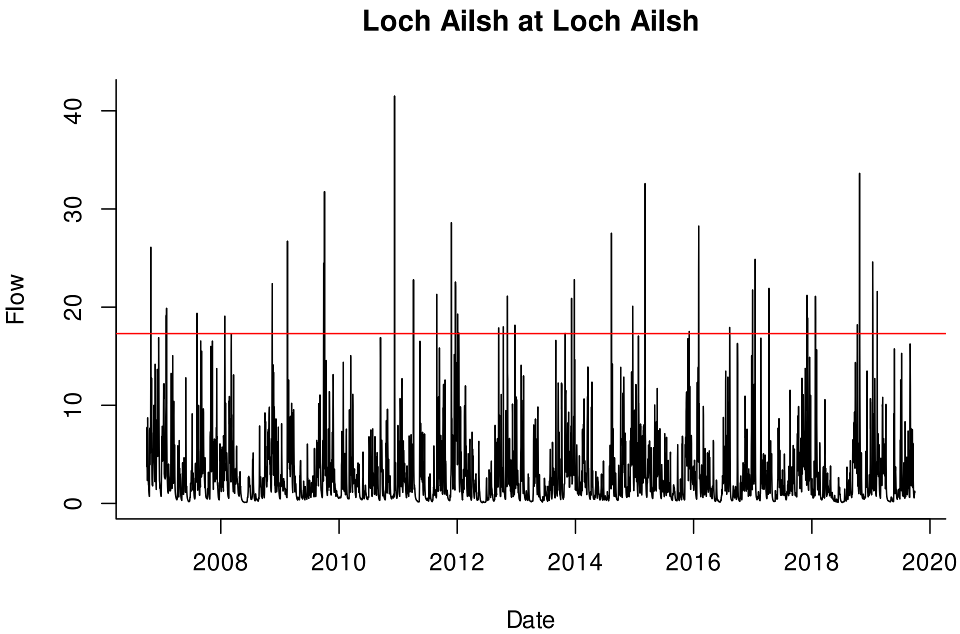
Extremes: peaks over threshold

What is an extreme? All large events in the record.



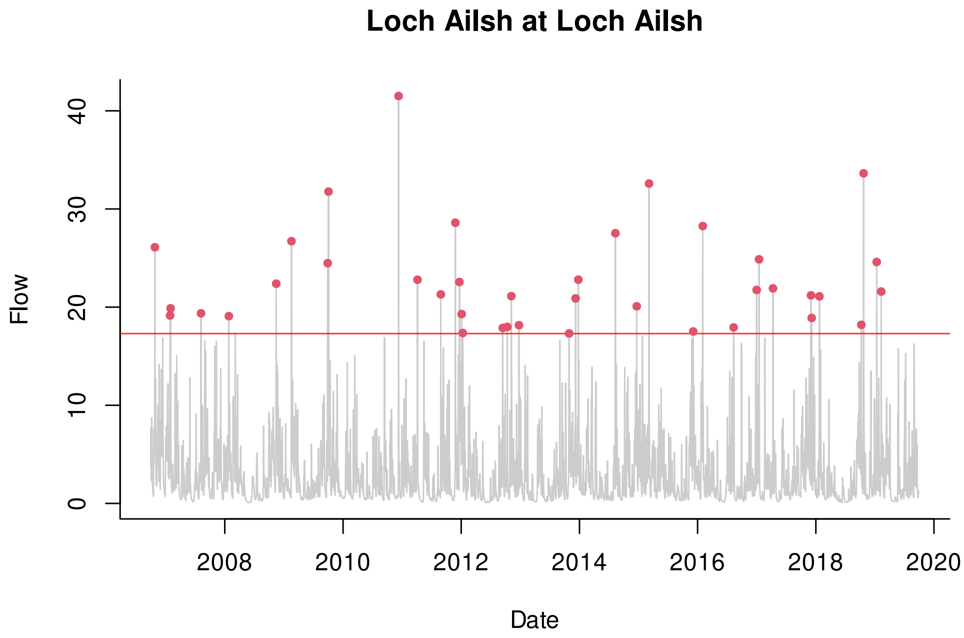
Extremes: peaks over threshold

What is an extreme? All large events in the record



Extremes: peaks over threshold

What is an extreme? All (independent) large events in the record



Extremes: theory for peaks over threshold

We define $Y = X - u$, where u is a given threshold.

The limiting distribution of threshold excesses given they are excesses $(Y|X > u)$ is a GPD:

$$F(y) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi}$$

$Y \sim GP(\sigma, \xi)$ is defined on $y : 1 + \xi(y - \mu)/\sigma > 0$ so the domain changes depending on the sign of ξ.

Number of threshold exceedances follow a Poisson distribution

Elephant in the room: choice of threshold

Extreme values - estimation

POT or AMAX: need to estimate a distribution.

Statistical estimation methods:

- L-moments
- Maximum Likelihood
- Bayesian approaches
- Generalised Maximum Likelihood - Penalised Likelihood (implicit priors)

Describing change/non-stationary models: not possible with L-moments.

Distributions which are not GEV or GPD (for example GLO or LP3) are rarely implemented.

Extreme value estimation in R

Some extreme value packages:

Package	L-moments	Bayesian	Max-Likelihood	Gen Max-Likelihood
lmom	✅	❌	❌	❌
lmomco	✅	❌	❌	❌
ismev	❌	❌	✅	❌
evd	❌	❌	✅	❌
mev	❌	❌	✅	❌
evdbayes	❌	✅	❌	❌
texmex	❌	✅	✅	❌ ✅
extRemes	✅	✅	✅	✅

(Careful when using GMLE in extRemes - the default penalty is defined on (0,0.5))

Extreme value estimation in R

Some general purpose packages:

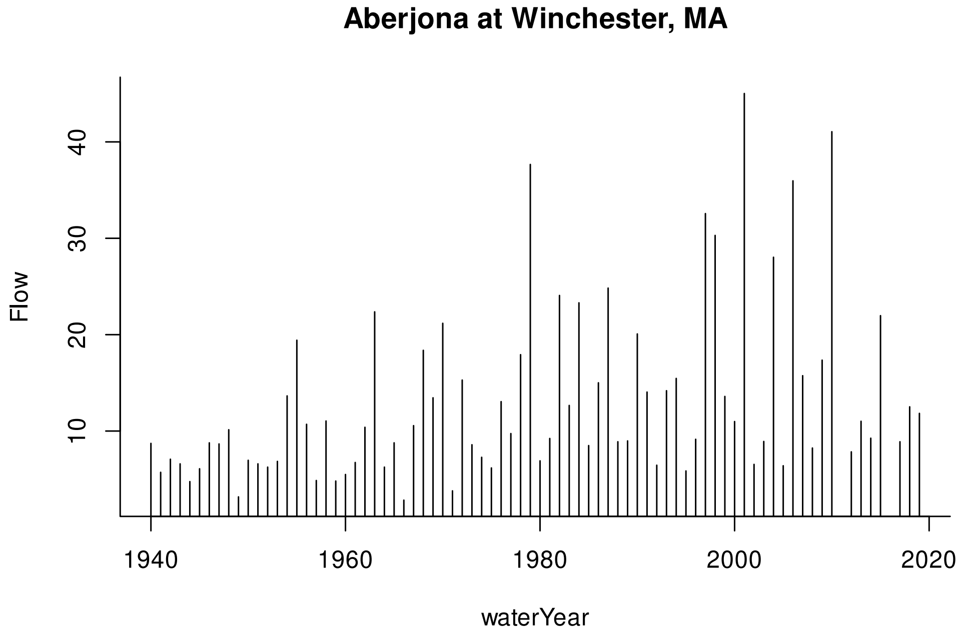
- evgam: GAM models - implements GEV, GPD (and point processes)
- gamlss: GAM models - includes several skewed distributions
- mgcv: GAM models - implements GEV distribution
- bamlss: Bayesian GAM models - includes the GEV
- rstan: general purpose Bayesian modelling - implements several distributions including the Gumbel
- brms: Bayesian regression models - includes the GEV

Statistical modelling of Extremes in R

Get some annual maxima data:

```r
library(dataRetrieval)
# The Aberjona River USGS Gage 01102500
aber <- suppressMessages
(dataRetrieval::readNWISpeak(siteNumber = "01102500"))
names(aber)[names(aber) == "peak_dt"] <- "Date"
aber <- addWaterYear(aber)
### standardise the water year
aber$swy <- scale(aber$waterYear)
## make flow into cubic meters
aber$Flow <- aber$peak_va * 0.028316846711688
with(aber, plot(waterYear, Flow, type="h", bty = "l"))
title(main = "Aberjona at Winchester, MA")
```

Statistical modelling of Extremes in R



Statistical modelling of Extremes in R

Estimates for the GEV parameters:

```
## L-moments
lmom::pelgev(lmom::samlmu(aber$Flow))
```

```
##          xi      alpha          k
##   8.5733422  4.6191646 -0.2882667
```

```
## Maximum likelihood - notice the sign of the shape parameter
ifit <- ismev::gev.fit(aber$Flow, show=FALSE)
rbind(ifit$mle, ifit$se)
```

```
##           [,1]      [,2]      [,3]
## [1,] 8.5170738 4.4605189 0.3411568
## [2,] 0.5739466 0.4980921 0.1016359
```

Statistical modelling of Extremes in R

```r
## Bayes
suppressMessages(b <- texmex::evm(Flow, data = aber,
                                  method = "simulate",
                                  family = texmex::gev,
                                  iter = 70701,burn=700,thin=7))
#  ,priorParameters = list(c(8,4,0.3),diag(c(40,3,1))))
### IMPORTANT
## check convergence of the chains!
summary(b)
```

```
## Family:       GEV
##
## Posterior summary:
##                  Posterior mean        SD
## mu: (Intercept)        8.549590 0.5918967
## phi: (Intercept)       1.525566 0.1138874
## xi: (Intercept)        0.344182 0.1038112
```

Non-stationary models

Can be implemented in `ismev`, `texmex` and `extRemes`...

```
## location parameter changes in time
library(extRemes)
exFit <- fevd(aber$Flow, location.fun = ~aber$swy , method = "MLE")
## the ci for mu1 does not include 0
ci(exFit, type = "par")
```

```
## fevd(x = aber$Flow, location.fun = ~aber$swy, method = "MLE")
##
## [1] "Normal Approx."
##
##       95% lower CI  Estimate 95% upper CI
## mu0      7.6355445 8.6740704    9.7125962
## mu1      0.6096182 1.3425945    2.0755709
## scale    3.1641869 4.0861272    5.0080675
## shape    0.1682329 0.3781068    0.5879807
```

```
## effective 30 years return level
eff30yrs <- return.level(exFit, return.period=30)
```

Non-stationary models

```
with(aber,plot(waterYear,Flow, pch =16,col = "grey40",bty = "l"))
lines(aber$waterYear, eff30yrs, col = "orange")
```