Univariate Extreme Value Theory

Ji Won Min

March 22, 2021

Table of Contents

Introduction

Block Maxima Approach

Peak-over-Threshold (POT) Approach

EVT in Wildfire Modeling

Introduction

- Block Maxima Method

  $M_n = max\{X_1, \ldots, X_n\}$, for $n \to \infty$

  M_n follows a Generalised Extreme Value (GEV) distribution

- Peak over Threshold Method

  … for a very large threshold u_

  follows a Generalised Pareto distribution (GPD)

Block maxima approach

Model Formulation

The model focuses on the statistical behaviour of $M_n = max\{X_1, \ldots, X_n\}$, where $X_1, \ldots, X_n$ is a sequence of i.i.d. random variables with distribution function F.

$$Pr(M_n \leq x) = Pr(X_1 \leq x) \ldots Pr(X_n \leq x) = F^n(x)$$

So we look at the behaviour of F^n as $n \to \infty$.

But for any $z < z^+$, where z^+ is the upper end-point of F, $F^n \to 0$ *as* $n \to \infty$, so the distribution of M_n degenerates to a point mass on z^+.

==> Linear normalization:

$$M_n^* = \frac{M_n - b_n}{a_n}\ , \quad \text{for sequences of constants } \{a_n > 0\} \text{ and } \{b_n\}$$

Extremal Types Theorem

Theorem : If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{(M_n - b_n)/a_n \le z\} \to G(z) \qquad \text{as } n \to \infty$$

where G is a non-degenerate distribution function, then G belongs to one of the following families:

$$\mathrm{I}: G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \qquad -\infty < z < \infty;$$

$$\mathrm{II}: G(z) = \begin{cases} 0, & z \le b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b; \end{cases}$$

$$\mathrm{III}: G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b, \\ 1, & z \ge b, \end{cases}$$

for $a > 0$ (scale parameter), b (location parameter) and, in the the case of families II and III, $\alpha > 0$ (shape parameter).

- The families labeled I, II and III are known as

  I : Gumbel

  II : Fréchet

  III : Weibull

  families, and are collectively called **extreme value distributions**.

- These three types of extreme value distributions are the only possible limits for the distributions of the M_n^* regardless of the distribution F for the population.

The Generalised Extreme Value (GEV) Distribution

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

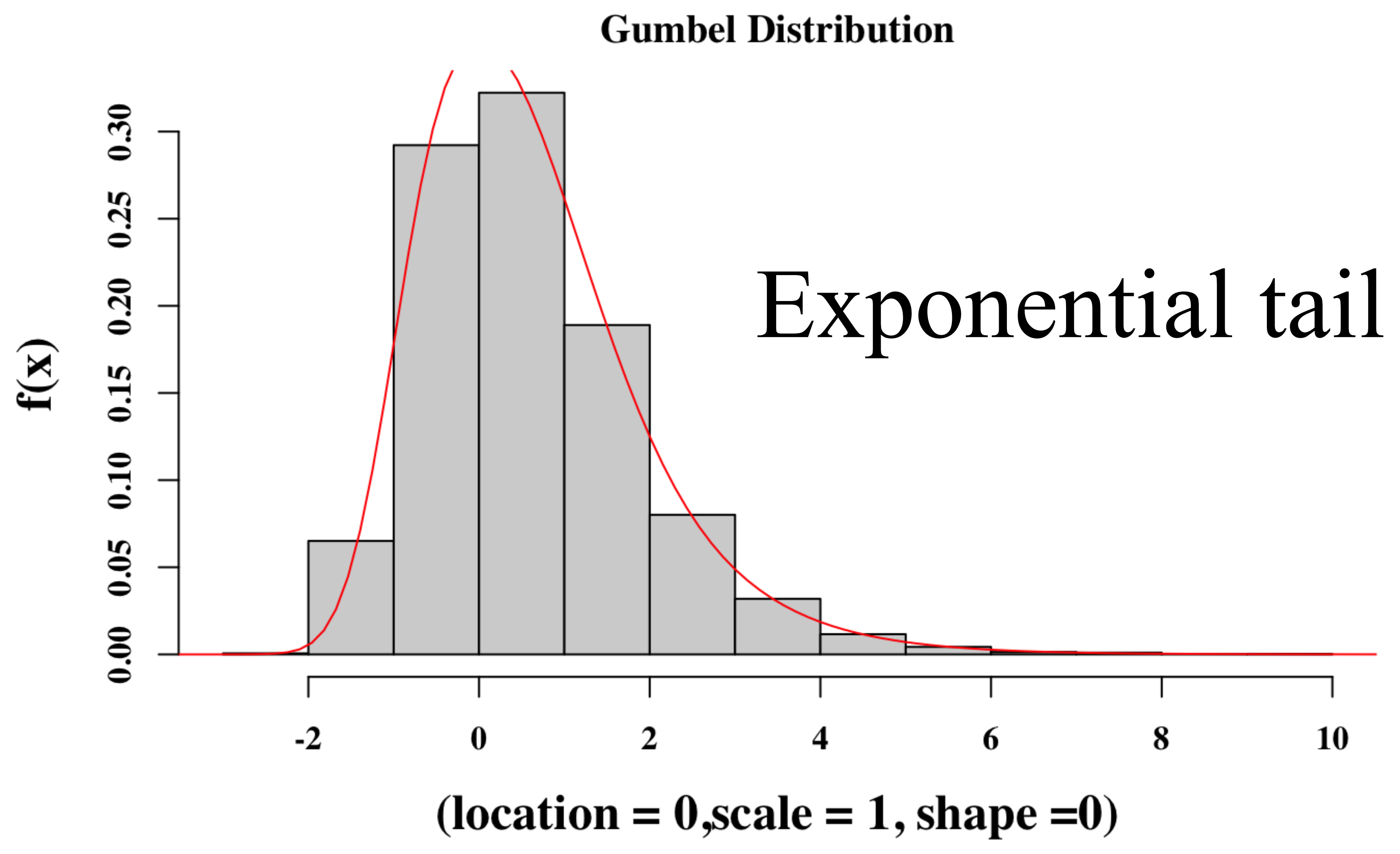
defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, where $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$.

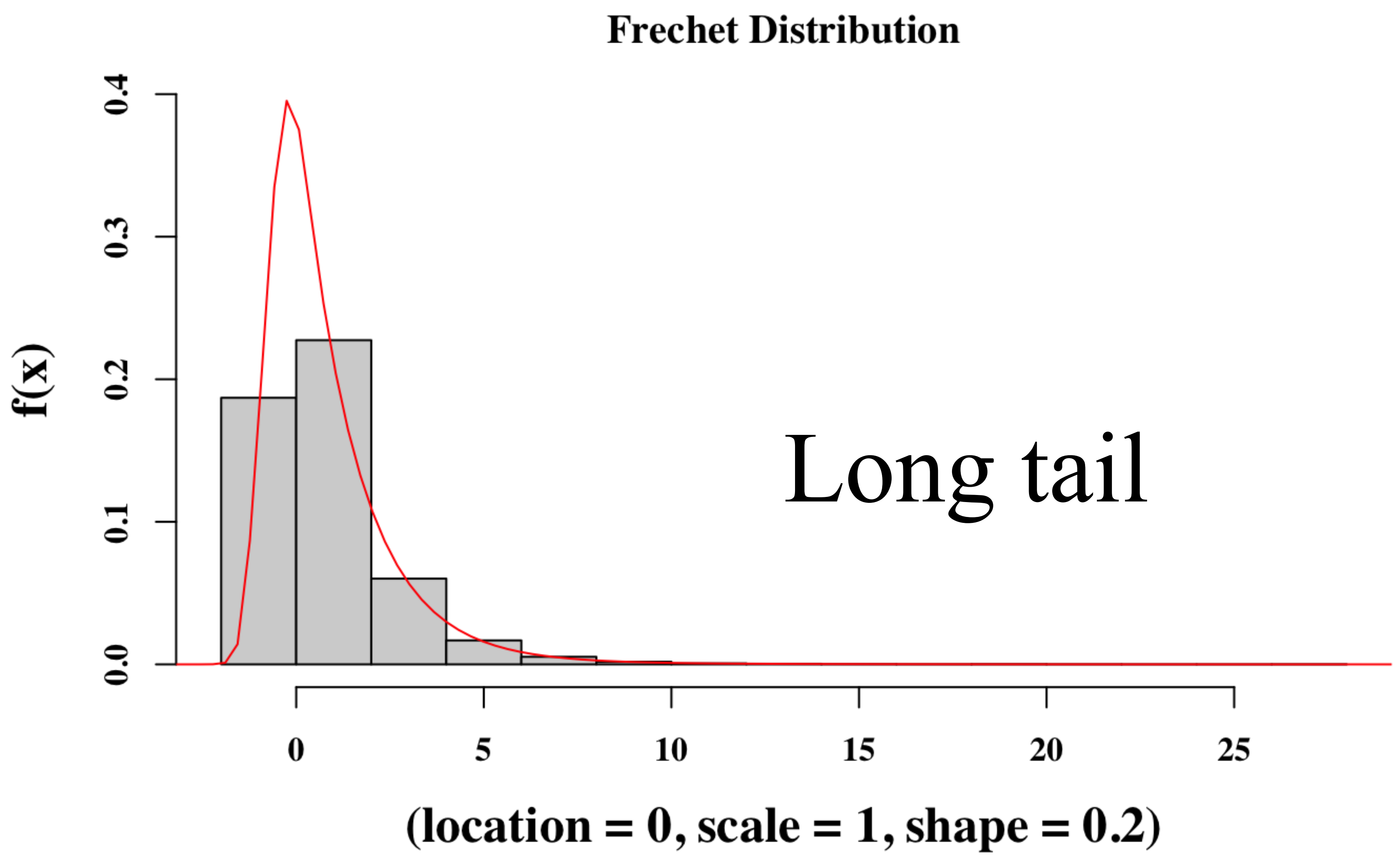
I (Gumbel) : $\xi = 0$ (interpreted as $\xi \to 0$)

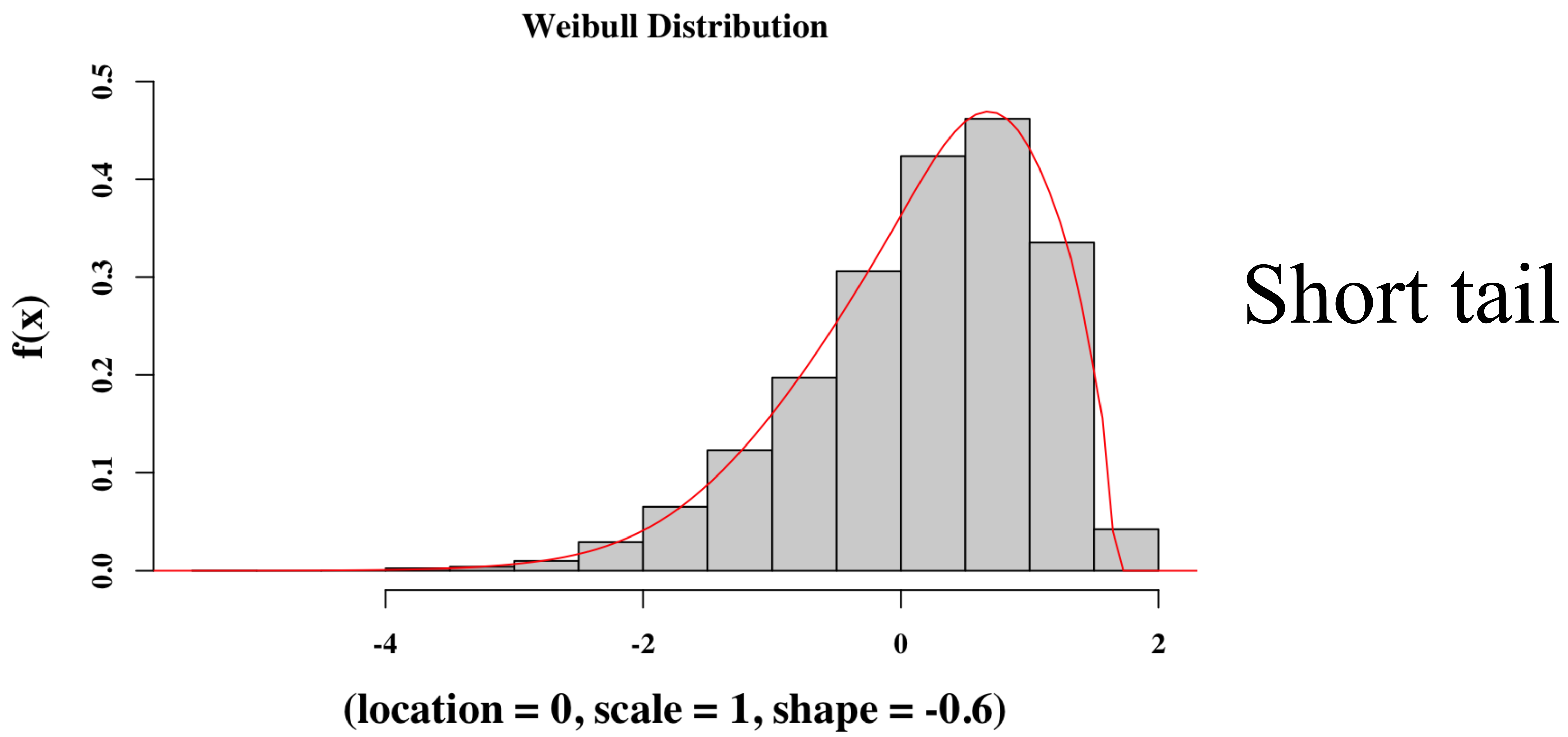
II (Fréchet) : $\xi > 0$

III (Weibull) : $\xi < 0$

ξ determines the nature of the tail distribution.


Gumbel Distribution
Exponential tail
f(x)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
-2
0
2
4
6
8
10
(location = 0,scale = 1, shape =0)


Frechet Distribution
Long tail
f(x)
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
(location = 0, scale = 1, shape = 0.2)


Weibull Distribution
Short tail
f(x)
0.0
0.1
0.2
0.3
0.4
0.5
-4
-2
0
2
(location = 0, scale = 1, shape = -0.6)

Inference for the GEV Distribution

Choice of block size:

- Small block size => approximation by the model is likely to be poor leading to bias in estimation and extrapolation

- Large block size => generates few block maxima, leading to large estimation variance

Pragmatic considerations often lead to the adoption of blocks of length one year.

Parameter estimation by maximum likelihood

Let $X_1, \ldots X_n$ be independent variables with the GEV distribution. The log-likelihood for the GEV parameters when $\xi \neq 0$ is

$$\ell(\mu, \sigma, \xi) = -m \log \sigma - (1 + 1/\xi) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right] - \sum_{i=1}^{m} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]^{-1/\xi}$$

provided that $1 + \xi(z_i - \mu)\sigma^{-1} > 0$ for $i = 1,...,m$.

The log-likelihood for the GEV parameters when $\xi = 0$ is

$$\ell(\mu, \sigma) = -m \log \sigma - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma} \right) - \sum_{i=1}^{m} \exp \left\{ - \left(\frac{z_i - \mu}{\sigma} \right) \right\}$$

Return Levels

Estimates of extreme quantiles of the maximum distribution are obtained by inverting the GEV distribution function :

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

where $G(z_p) = 1 - p$.

=> z_p is the **return level** associated with the **return period** 1/p

Return level plot : z_p plotted against $log y_p$ where $y_p = -\,log(1-p)$.

Return level plot is convenient for model presentation and validation.

Threshold models

Model Formulation

We can also see as extreme events as those that exceed some high threshold u.

Let $X_1, \ldots X_n$ be a sequence of i.i.d. random variables with distribution function F.

A description of the stochastic behaviour of extreme events is given by the conditional probability:

$$Pr\{X > u + y \,|\, X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \quad y > 0$$

The generalized Pareto distribution

Theorem : Let $M_n = max\{X_1, ..., X_n\}$, where $X_1, ... X_n$ is a sequence of i.i.d. random variables with distribution function F. Suppose that F satisfies the extreme value theorem so that for large n,

$$Pr\{(M_n \leq z\} \approx G(z), \text{ where G is an extreme value distribution.}$$

Then, for large enough u, the distribution function of $(X - u)$ conditional on $X > u$ is approximately

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

defined on $\{y : y > 0 \; \& \; (1 + \xi y/\tilde{\sigma}) > 0\}$, where $\tilde{\sigma} = \sigma + \xi(u - \mu)$.

The family of distributions defined is called the **generalized Pareto family.**

Duality between GEV and generalised Pareto families

The shape parameter ξ is dominant in determining the behaviour of the generalised Pareto distribution:

- $\xi = 0 : H(y) = 1 - exp(-\frac{y}{\tilde{\sigma}}), y > 0$ (corresponds to exponential distribution)
- $\xi > 0$: no upper limit
- $\xi < 0$: upper bound of $u - \frac{\tilde{\sigma}}{\xi}$

Threshold Selection (Mean residual life plot)

The mean of the generalised Pareto distribution given $\xi < 1$, with threshold u_0: $E(X - u_0 | X > u_0) = \dfrac{\sigma_{u_0}}{1 - \xi}$

The generalised Pareto distribution should also valid for all thresholds $u > u_0$:

$$
\begin{aligned}
\mathrm{E}(X - u \mid X > u) &= \frac{\sigma_u}{1 - \xi} \\
&= \frac{\sigma_{u_0} + \xi u}{1 - \xi}
\end{aligned}
$$

So $E(X - u | X > U)$ is a linear function of u.

$$
\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{\max} \right\},
$$

where $x_{(1)}, \ldots, x_{(n_u)}$ are the observations that exceed the threshold u, are the points of the **mean residual life plot.**

Parameter estimation by maximum likelihood

Let $y_1, \ldots, y_k$ be k excesses of a threshold u. The log-likelihood for the generalised Pareto distribution parameters when $\xi \neq 0$ is

$$\ell(\sigma, \xi) = -k \log \sigma - (1 + 1/\xi) \sum_{i=1}^{k} \log(1 + \xi y_i/\sigma),$$

provided that $1 + \sigma^{-1}\xi y_i > 0$ for $i = 1,...,k$.

The log-likelihood for the generalised Pareto distribution parameters when $\xi = 0$ is

$$\ell(\sigma) = -k \log \sigma - \sigma^{-1} \sum_{i=1}^{k} y_i.$$

EVT in wildfire modeling

Models for fire sizes

Distribution of fire sizes tend to be heavy tailed:

- Pareto distribution

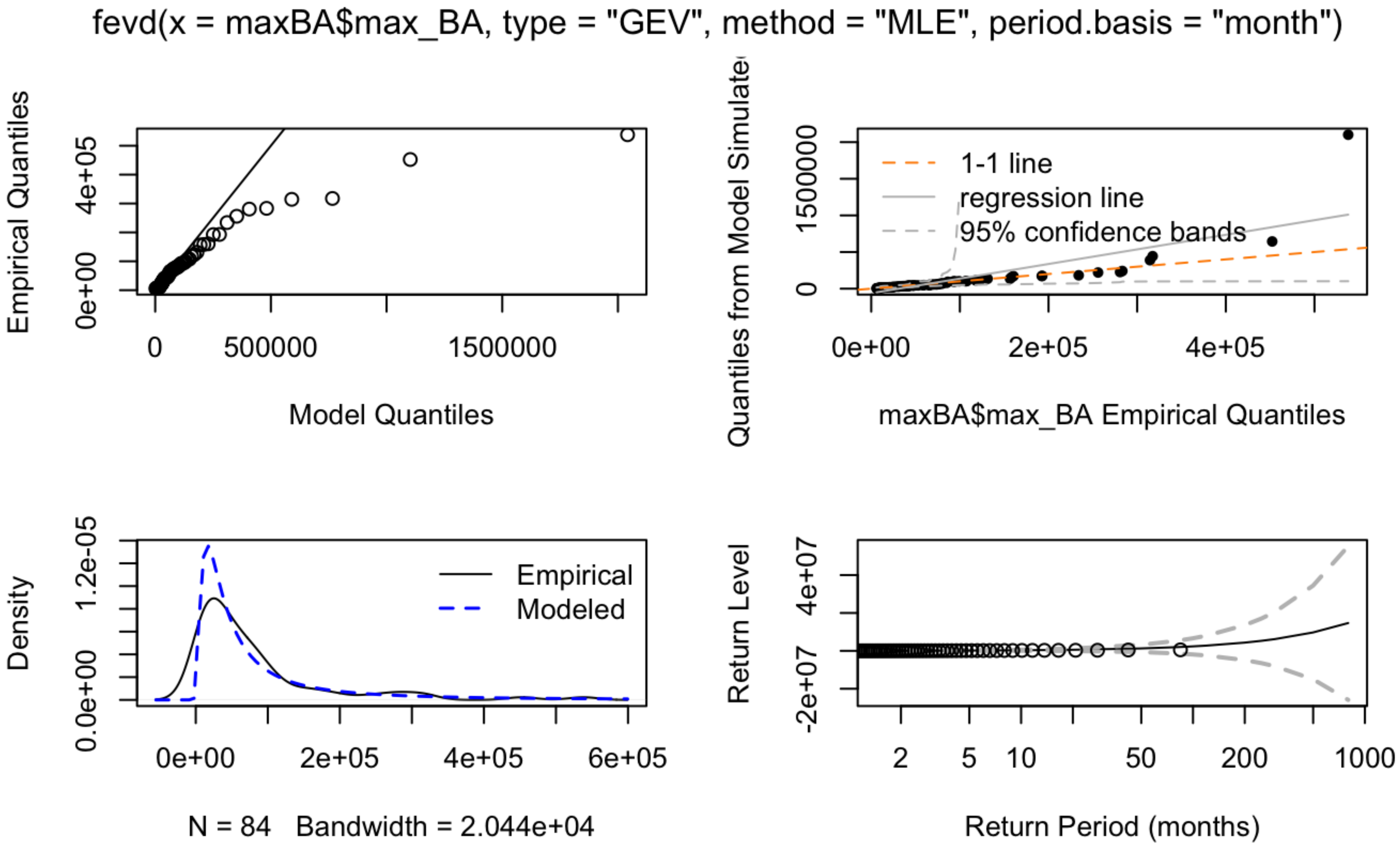
$$F(x) = 1 - (\beta/x)^{\eta}, \quad \beta \leq x < \infty$$

- Tapered Pareto distribution (smaller exponential tails; otherwise similar to the Pareto distribution)

$$F(x) = 1 - (\beta/x)^{\eta} exp(\frac{\beta - x}{\theta}), \quad \beta \leq x < \infty$$

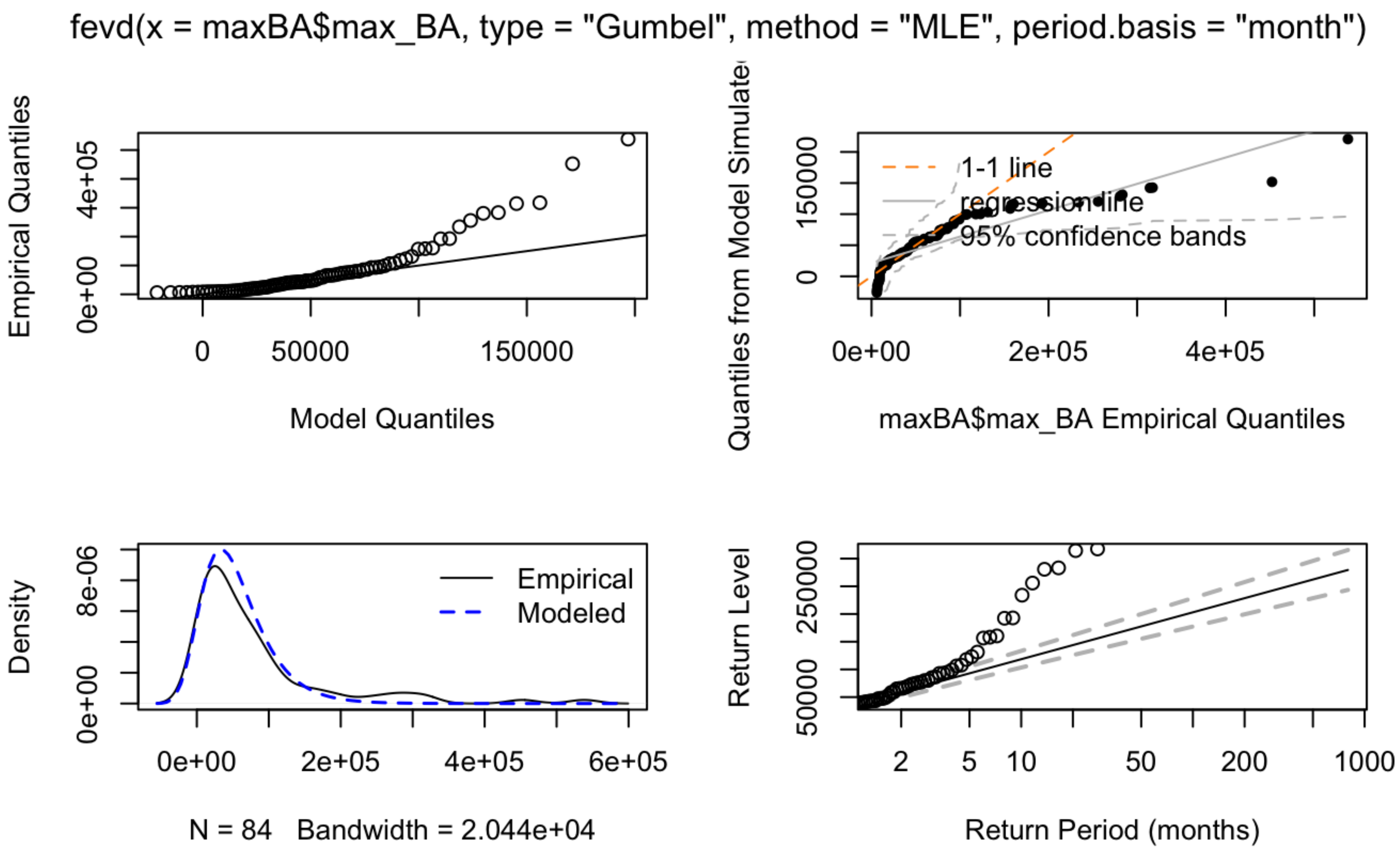
Block maxima: GEV on monthly maxima of BA (odd years)

Library “extRemes”
“fevd” function



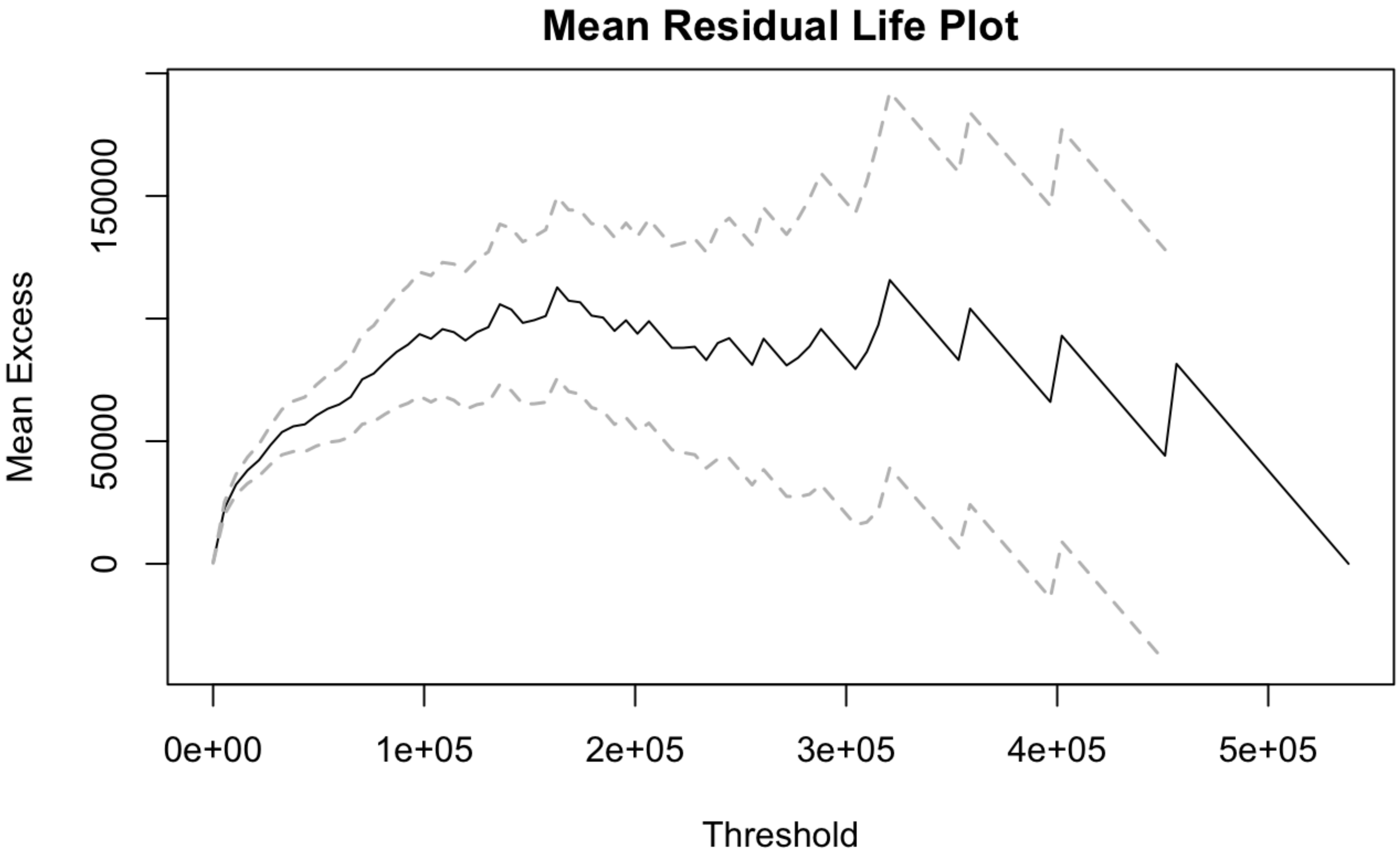
Block maxima: Gumbel on monthly maxima of BA (odd years)

Library “extRemes”
“fevd” function

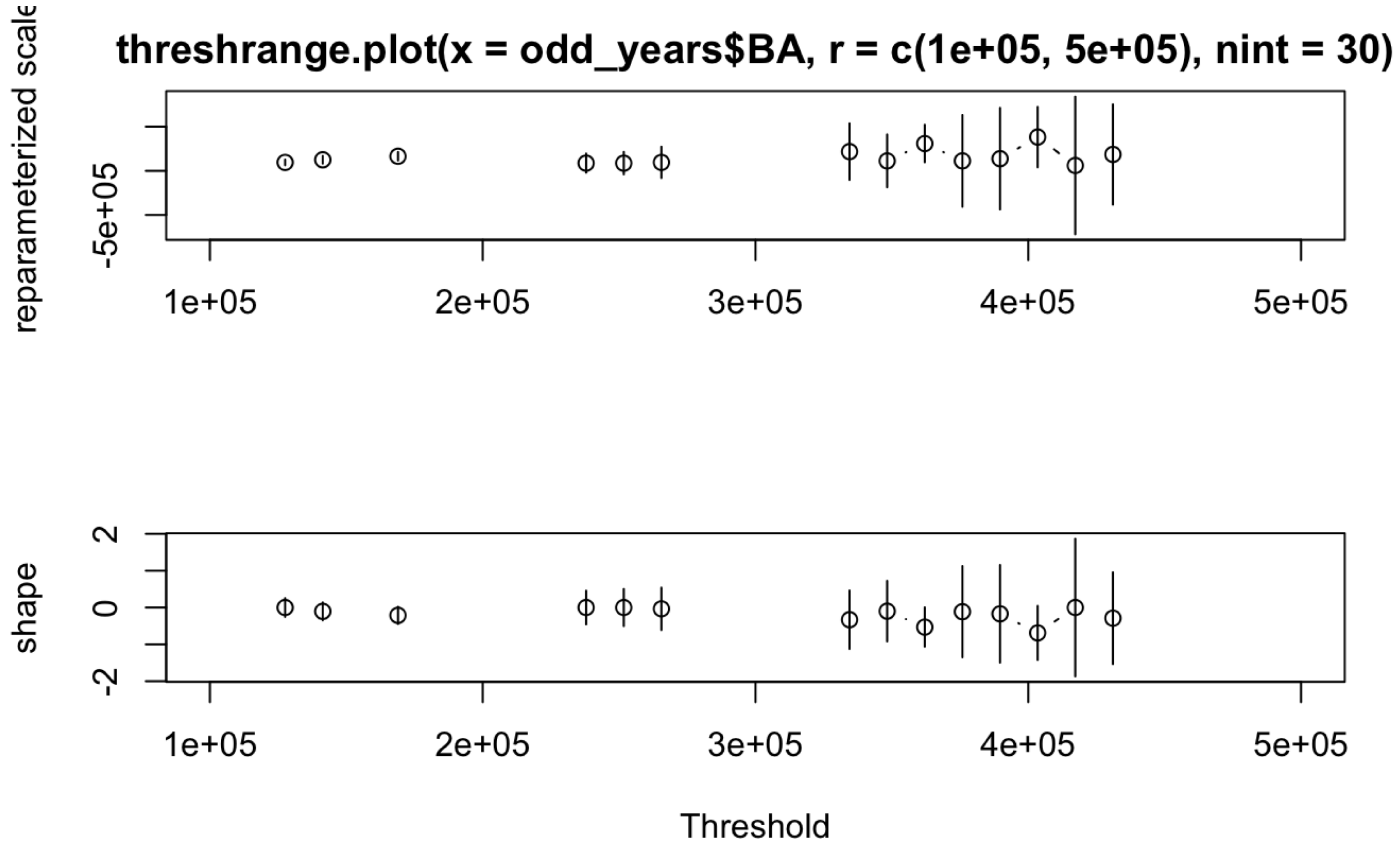


Mean Residual Life Plot of BA (odd years and BA > 0)

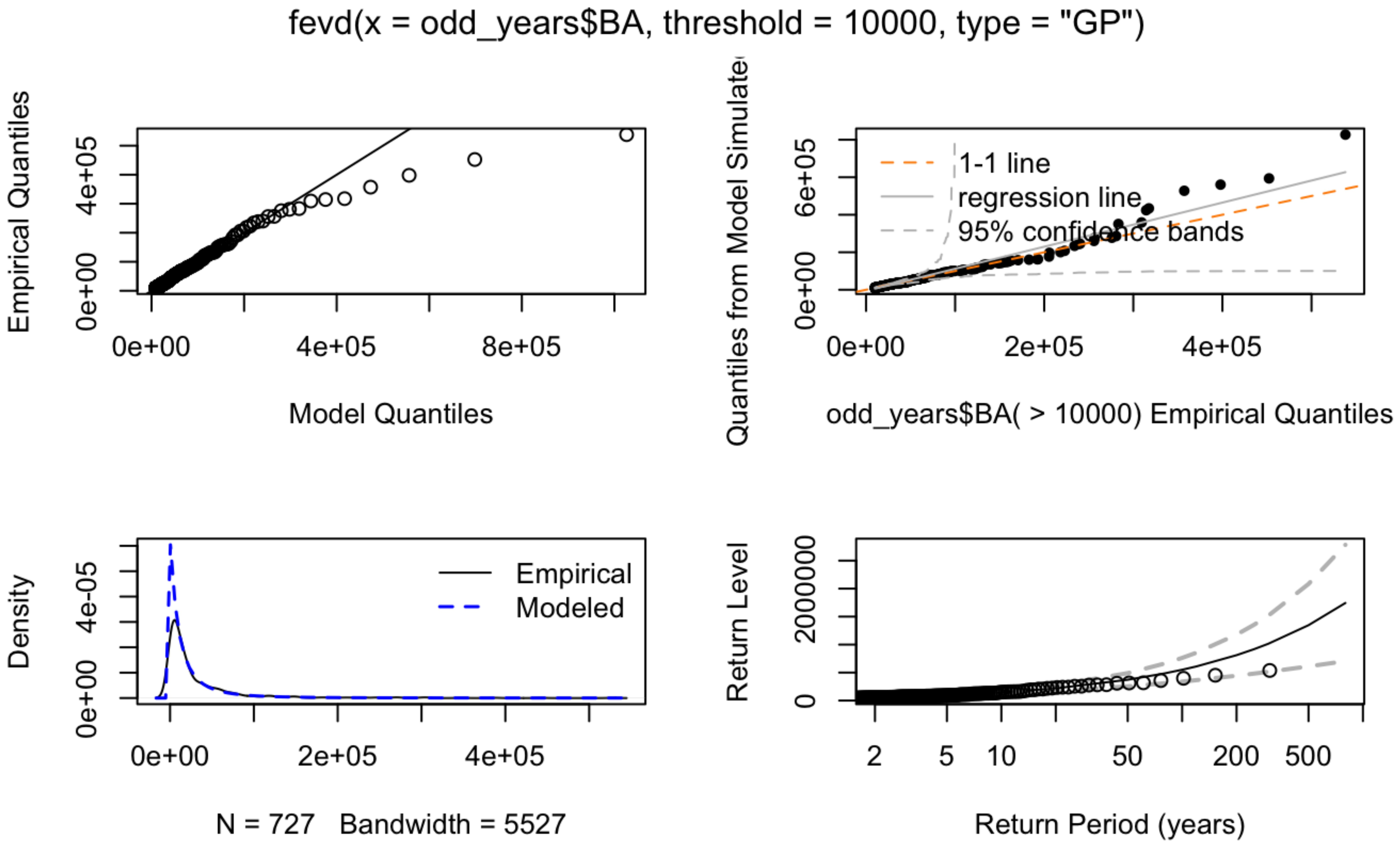
Library “extRemes”
“mrlplot” function



Threshold range plot (odd years and BA > 0)



Fitting generalised Pareto distribution with threshold of 10000:



Resources

Coles, Stuart. An Introduction to Statistical Modeling of Extreme Values. Springer, 2011.

Gilleland E, Katz RW (2016). “extRemes 2.0: An Extreme Value Analysis Package in R.” *Journal of Statistical Software*, **72**(8), 1–39. doi: 10.18637/jss.v072.i08.