An Introduction to Extreme Value Analysis

Whitney Huang

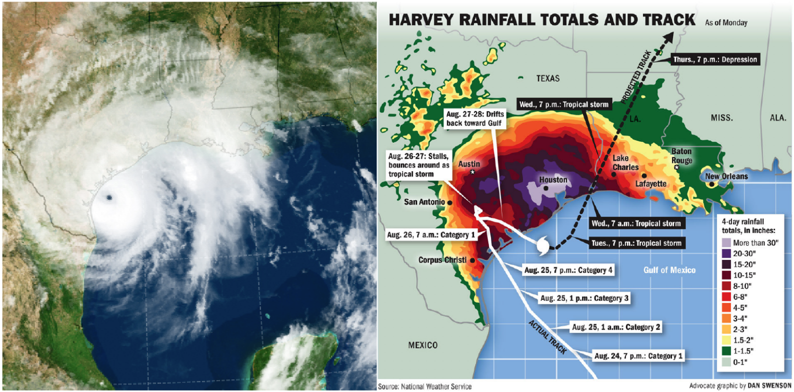
CLEMSON UNIVERSITY

wkhuang@clemson.edu

Clemson EVA Group, January 29, 2020

Outline

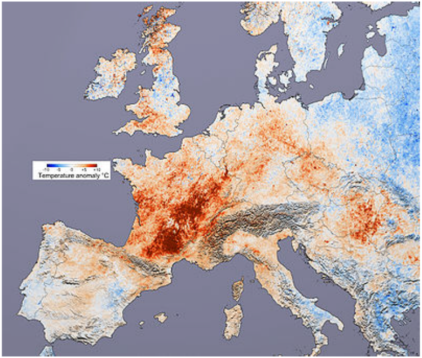
Extreme Rainfall During Hurricane Harvey



Source: NASA (Left); National Weather Service (Right)

- *"A storm forces Houston, the limitless city, to consider its limits"* – The New York Times (8.31.17)

Environmental Extremes: Heatwaves, Storm Surges, etc.



- **Heat wave**: The 2003 European heat wave led to the hottest summer on record in Europe since 1540 that resulted in at least 30,000 deaths
- **Storm Surge**: Hurricane Katrina produced the highest storm surge ever recorded (27.8 feet) on the U.S. coast

Scientific Questions

- How to estimate the magnitude of extreme events (e.g. 100-year rainfall)?
- How extremes vary in space?
- How extremes may change in future climate conditions?

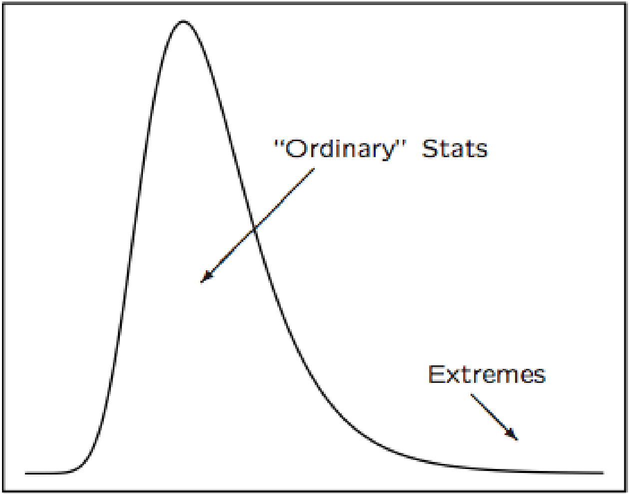
Outline

Motivation

Extreme Value Theorem & Block Maxima Method

Peaks–Over–Threshold (POT) Method

Usual vs Extremes



Probability Framework

Let $X_1, \cdots, X_n \overset{\text{iid}}{\sim} F$ and define $M_n = \max\{X_1, \cdots, X_n\}$
Then the distribution function of M_n is

$$\begin{aligned}\mathbb{P}(M_n \le x) &= \mathbb{P}(X_1 \le x, \cdots, X_n \le x)\\ &= \mathbb{P}(X_1 \le x) \times \cdots \times \mathbb{P}(X_n \le x) = F^n(x)\end{aligned}$$

Remark

$$F^n(x) \xlongequal{n\to\infty} \begin{cases} 0 & \text{if } F(x) < 1 \\ 1 & \text{if } F(x) = 1 \end{cases}$$

$\Rightarrow$ the limiting distribution is degenerate.

Asymptotic: Classical Limit Laws

Recall the Central Limit Theorem:

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0,1)$$

$\Rightarrow$ rescaling is the key to obtain a non-degenerate distribution

Question: Can we get the limiting distribution of

$$\frac{M_n - b_n}{a_n}$$

for suitable sequence $\{a_n\} > 0$ and $\{b_n\}$?

Asymptotic: Classical Limit Laws

Recall the Central Limit Theorem:

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0,1)$$

$\Rightarrow$ rescaling is the key to obtain a non-degenerate distribution

Question: Can we get the limiting distribution of

$$\frac{M_n - b_n}{a_n}$$

for suitable sequence $\{a_n\} > 0$ and $\{b_n\}$?

Extremal Types Theorem (Fisher–Tippett 1928, Gnedenko 1943)

Define $M_n = \max\{X_1, \cdots, X_n\}$ where $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$. If $\exists\, a_n > 0$ and $b_n \in \mathbb{R}$ such that, as $n \to \infty$, if

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \le x\right) \xrightarrow{d} \mathsf{G}(x)$$

then G must be the same type of the following form:

$$\mathsf{G}(x; \mu, \sigma, \xi) = \exp\left\{ - \left[1 + \xi(\frac{x-\mu}{\sigma})\right]_+^{\frac{-1}{\xi}} \right\}$$

where $x_+ = \max(x, 0)$ and $G(x)$ is the distribution function of the generalized extreme value distribution (GEV(μ, σ, ξ))

- μ and σ are location and scale parameters
- ξ is a shape parameter determining the rate of tail decay, with
  - $\xi > 0$ giving the heavy-tailed case (Fréchet)
  - $\xi = 0$ giving the light-tailed case (Gumbel)
  - $\xi < 0$ giving the bounded-tailed case (reversed Weibull)

Extremal Types Theorem (Fisher–Tippett 1928, Gnedenko 1943)

Define $M_n = \max\{X_1, \cdots, X_n\}$ where $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$. If $\exists\, a_n > 0$ and $b_n \in \mathbb{R}$ such that, as $n \to \infty$, if

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{d} \mathsf{G}(x)$$

then G must be the same type of the following form:

$$\mathsf{G}(x; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{\frac{-1}{\xi}} \right\}$$

where $x_+ = \max(x, 0)$ and $G(x)$ is the distribution function of the generalized extreme value distribution (GEV(μ, σ, ξ))

- μ and σ are location and scale parameters
- ξ is a shape parameter determining the rate of tail decay, with
  - $\xi > 0$ giving the heavy-tailed case (Frechet)
  - $\xi = 0$ giving the light-tailed case (Gumbel)
  - $\xi < 0$ giving the bounded-tailed case (reversed Weibull)

Extremal Types Theorem (Fisher–Tippett 1928, Gnedenko 1943)

Define $M_n = \max\{X_1, \cdots, X_n\}$ where $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$. If $\exists\, a_n > 0$ and $b_n \in \mathbb{R}$ such that, as $n \to \infty$, if

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{d} \mathsf{G}(x)$$

then G must be the same type of the following form:

$$\mathsf{G}(x; \mu, \sigma, \xi) = \exp\left\{ - \left[1 + \xi(\frac{x-\mu}{\sigma})\right]_+^{\frac{-1}{\xi}} \right\}$$

where $x_+ = \max(x, 0)$ and $G(x)$ is the distribution function of the generalized extreme value distribution (GEV(μ, σ, ξ))

- μ and σ are location and scale parameters
- ξ is a shape parameter determining the rate of tail decay, with
    - $\xi > 0$ giving the heavy-tailed case (Fréchet)
    - $\xi = 0$ giving the light-tailed case (Gumbel)
    - $\xi < 0$ giving the bounded-tailed case (reversed Weibull)

Extremal Types Theorem (Fisher–Tippett 1928, Gnedenko 1943)

Define $M_n = \max\{X_1, \cdots, X_n\}$ where $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$. If $\exists\, a_n > 0$ and $b_n \in \mathbb{R}$ such that, as $n \to \infty$, if

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{d} \mathsf{G}(x)$$

then G must be the same type of the following form:

$$\mathsf{G}(x; \mu, \sigma, \xi) = \exp\left\{ - \left[1 + \xi\left(\frac{x - \mu}{\sigma}\right)\right]_{+}^{\frac{-1}{\xi}} \right\}$$

where $x_+ = \max(x, 0)$ and $G(x)$ is the distribution function of the generalized extreme value distribution (GEV(μ, σ, ξ))

- μ and σ are location and scale parameters
- ξ is a shape parameter determining the rate of tail decay, with
  - $\xi > 0$ giving the heavy-tailed case (Fréchet)
  - $\xi = 0$ giving the light-tailed case (Gumbel)
  - $\xi < 0$ giving the bounded-tailed case (reversed Weibull)

Extremal Types Theorem (Fisher–Tippett 1928, Gnedenko 1943)

Define $M_n = \max\{X_1, \cdots, X_n\}$ where $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$. If $\exists\, a_n > 0$ and $b_n \in \mathbb{R}$ such that, as $n \to \infty$, if

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{d} \mathsf{G}(x)$$

then G must be the same type of the following form:

$$\mathsf{G}(x; \mu, \sigma, \xi) = \exp\left\{ - \left[1 + \xi(\frac{x-\mu}{\sigma})\right]_{+}^{\frac{-1}{\xi}} \right\}$$

where $x_+ = \max(x, 0)$ and $G(x)$ is the distribution function of the generalized extreme value distribution (GEV(μ, σ, ξ))

- μ and σ are location and scale parameters
- ξ is a shape parameter determining the rate of tail decay, with
  - $\xi > 0$ giving the heavy-tailed case (Fréchet)
  - $\xi = 0$ giving the light-tailed case (Gumbel)
  - $\xi < 0$ giving the bounded-tailed case (reversed Weibull)

Extremal Types Theorem (Fisher–Tippett 1928, Gnedenko 1943)

Define $M_n = \max\{X_1, \cdots, X_n\}$ where $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$. If $\exists\, a_n > 0$ and $b_n \in \mathbb{R}$ such that, as $n \to \infty$, if

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{d} \mathsf{G}(x)$$

then G must be the same type of the following form:

$$\mathsf{G}(x; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{\frac{-1}{\xi}} \right\}$$

where $x_+ = \max(x, 0)$ and $G(x)$ is the distribution function of the generalized extreme value distribution (GEV(μ, σ, ξ))

- μ and σ are location and scale parameters
- ξ is a shape parameter determining the rate of tail decay, with
  - $\xi > 0$ giving the heavy-tailed case (Fréchet)
  - $\xi = 0$ giving the light-tailed case (Gumbel)
  - $\xi < 0$ giving the bounded-tailed case (reversed Weibull)

Extremal Types Theorem (Fisher–Tippett 1928, Gnedenko 1943)

Define $M_n = \max\{X_1, \cdots, X_n\}$ where $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$. If $\exists\, a_n > 0$ and $b_n \in \mathbb{R}$ such that, as $n \to \infty$, if

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{d} \mathsf{G}(x)$$

then G must be the same type of the following form:

$$\mathsf{G}(x; \mu, \sigma, \xi) = \exp\left\{ - \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{\frac{-1}{\xi}} \right\}$$

where $x_+ = \max(x, 0)$ and $G(x)$ is the distribution function of the generalized extreme value distribution (GEV(μ, σ, ξ))

- μ and σ are location and scale parameters
- ξ is a shape parameter determining the rate of tail decay, with
  - $\xi > 0$ giving the heavy-tailed case (Fréchet)
  - $\xi = 0$ giving the light-tailed case (Gumbel)
  - $\xi < 0$ giving the bounded-tailed case (reversed Weibull)

Max-Stability and GEV

Definition

A distribution G is said to be max-stable if

$$G^k(a_k x + b_k) = G(x), \qquad k \in \mathbb{N}$$

for some constants $a_k > 0$ and b_k

- Taking powers of a distribution function results only in a change of location and scale
- A distribution is max-stable $\iff$ it is a GEV distribution

Quantiles and Return Levels

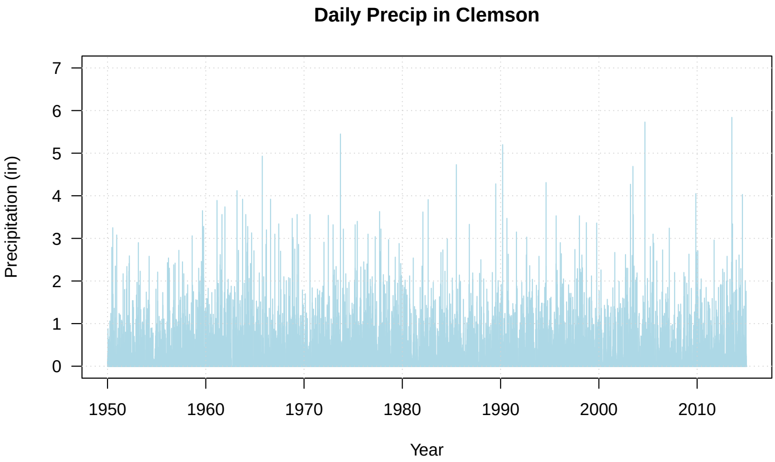
- Quantiles of GEV

$$G(x_p) = \exp\left\{ - \left[1 + \xi \left(\frac{x_p - \mu}{\sigma} \right) \right]_+^{\frac{-1}{\xi}} \right\} = 1 - p$$

$$\Rightarrow x_p = \mu - \frac{\sigma}{\xi} \left[1 - \{ - \log(1-p)^{-\xi} \} \right] \qquad 0 < p < 1$$

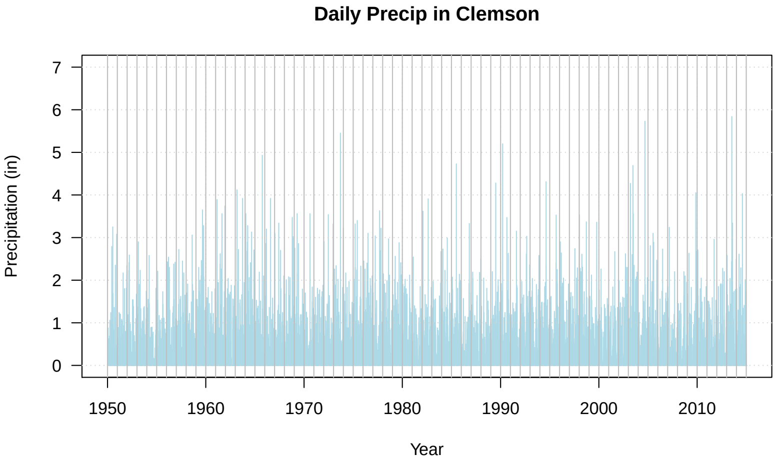
- In the extreme value terminology, x_p is the return level associated with the return period $\frac{1}{p}$

Clemson Daily Precipitation [Data Source: USHCN]



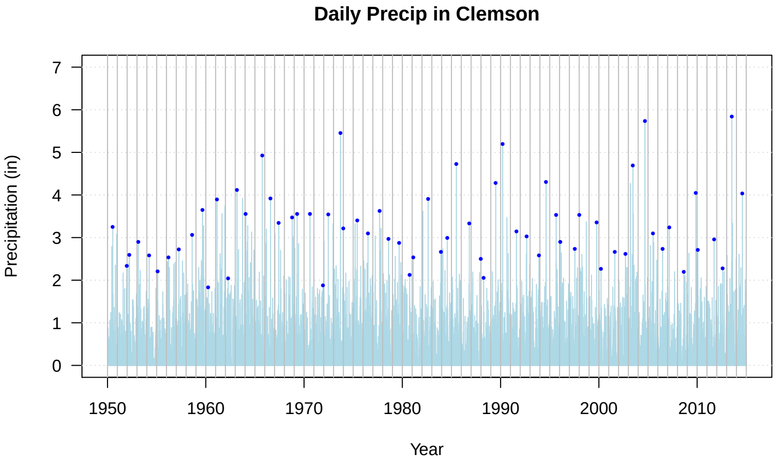
Block Maxima Method (Gumbel 1958)

1. Determine the block size and extract the block maxima



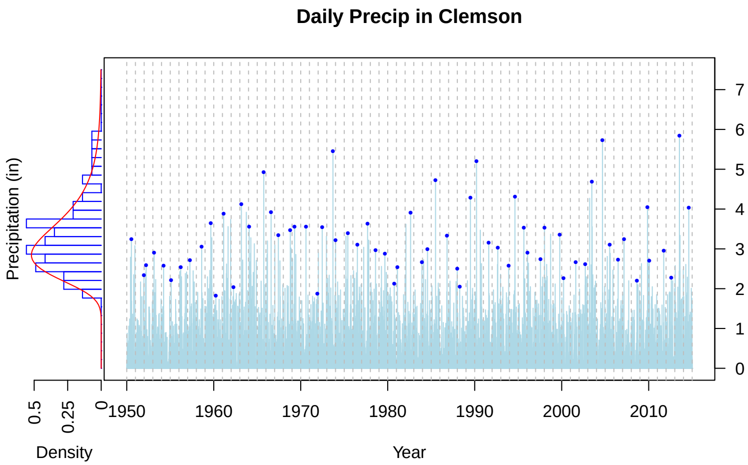
Block Maxima Method (Gumbel 1958)

1. Determine the block size and extract the block maxima



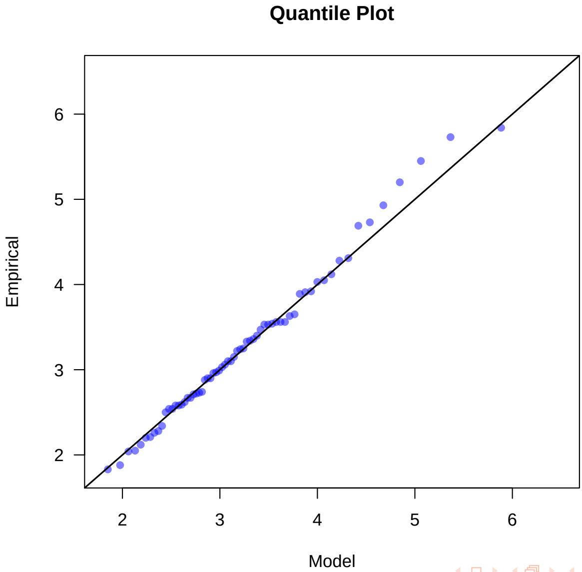
Block Maxima Method (Gumbel 1958)

2. Fit the GEV to the maximal and assess the fit



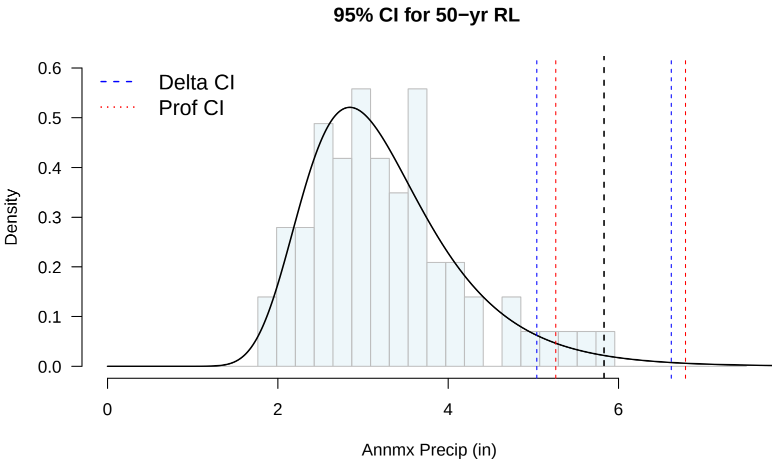
Block Maxima Method (Gumbel 1958)

2. Fit the GEV to the maximal and assess the fit



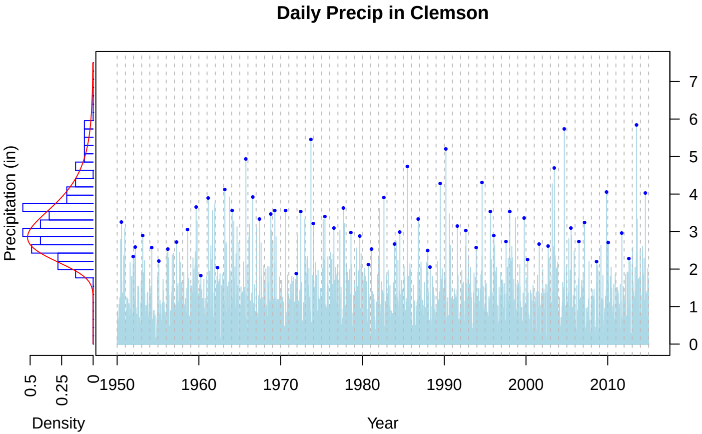
Block Maxima Method (Gumbel 1958)

3. Perform inference for return levels, probabilities, etc



Outline

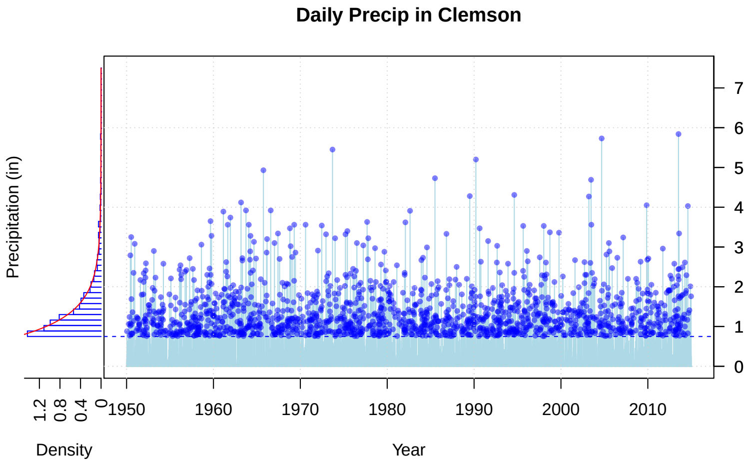
Recall the Block Maxima Method



Question: Can we use data more efficiently?

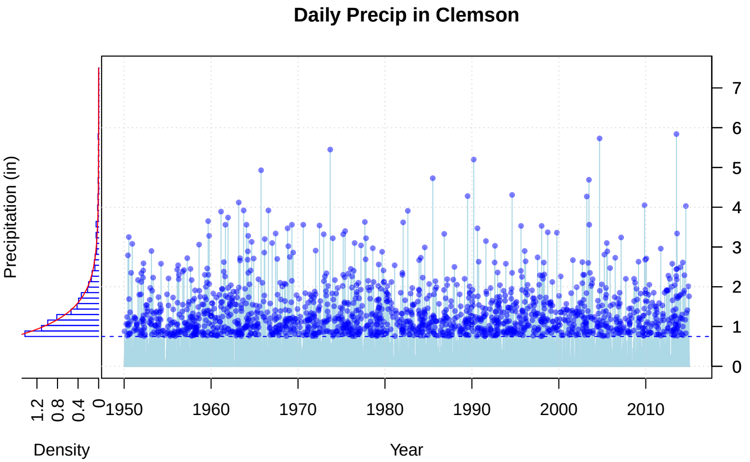
Peaks–over–threshold (POT) method [Davison & Smith 1990]

1. Select a "sufficiently large" threshold u, extract the exceedances



Peaks–over–threshold (POT) method [Davison & Smith 1990]

2. Fit an appropriate model to exceedances



GPD for Exceedances

If $M_n = \max_{i=1,\cdots,n} X_i$ (for a large n) can be apprximated by a GEV(μ, σ, ξ), then for sufficently large u,

$$
\begin{aligned}
\mathbb{P}(X_i > x + u | X_i > u) &= \frac{n\mathbb{P}(X_i > x + u)}{n\mathbb{P}(X_i > u)} \\
&\to \left(\frac{1 + \xi \frac{x+u-b_n}{a_n}}{1 + \xi \frac{u-b_n}{a_n}} \right)^{\frac{-1}{\xi}} \\
&= \left(1 + \frac{\xi x}{a_n + \xi(u - b_n)} \right)^{\frac{-1}{\xi}}
\end{aligned}
$$

$\Rightarrow$ Survival function of generalized Pareto distribution

Pickands–Balkema–de Haan Theorem (1974, 1975)

If $M_n = \max_{1 \le i \le n}\{X_i\} \approx \text{GEV}(\mu, \sigma, \xi)$, then, for a "large" u (i.e., $u \to x_F = \sup\{x : F(x) < 1\}$),

$$\mathbb{P}(X > u) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{\frac{-1}{\xi}}$$

$F_u = \mathbb{P}(X - u < y | X > u)$ is well approximated by the generalized Pareto distribution (GPD). That is:

$$F_u(y) \xrightarrow{d} H_{\tilde{\sigma},\xi}(y) \qquad u \to x_F$$

where

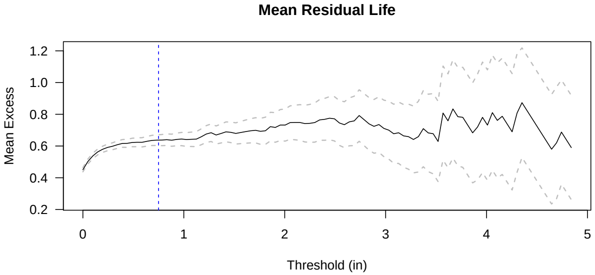
$$H_{\tilde{\sigma},\xi}(y) = \begin{cases} 1 - (1 + \xi y/\tilde{\sigma})^{-1/\xi} & \xi \neq 0; \\ 1 - \exp(-y/\tilde{\sigma}) & \xi = 0. \end{cases}$$

and $\tilde{\sigma} = \sigma + \xi(u - \mu)$

How to Choose the Threshold?

Bias–variance tradeoff:

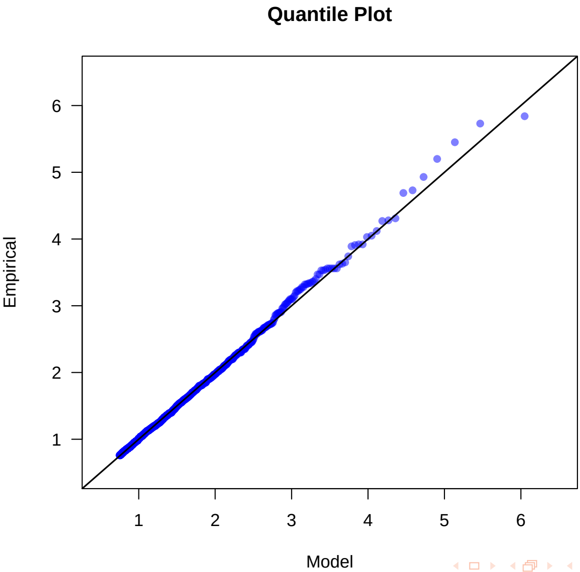
- Threshold too low $\Rightarrow$ bias because of the model asymptotics being invalid
- Threshold too high $\Rightarrow$ variance is large due to few data points



Task: To choose a u_0 s.t. the Mean Residual Life curve behaves linearly $\forall u > u_0$

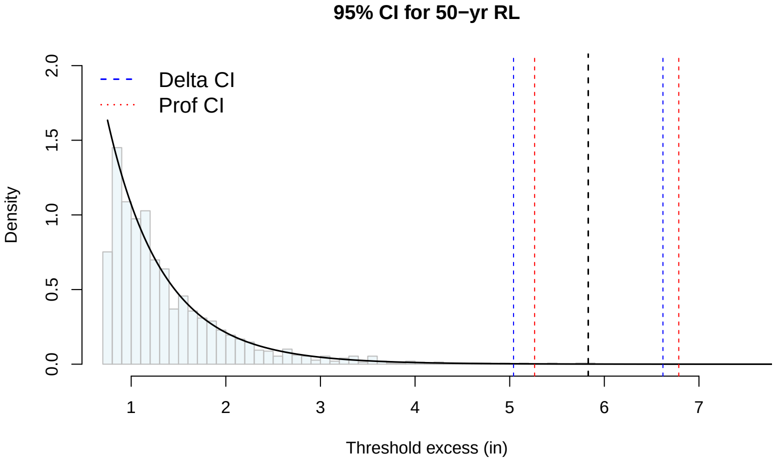
Peaks–over–threshold (POT) method [Davison & Smith 1990]

2. Fit an appropriate model to exceedances and assess the fit



Peaks–over–threshold (POT) method [Davison & Smith 1990]

3. Perform inference for return levels, probabilities, etc



Summary & Discussion

- Extreme value theory provides a framework to model extreme values
  - GEV for fitting block maxima
  - GPD for fitting threshold exceedances
  - Return level for communicating risk
- Practical Issues: seasonality, temporal dependence, non-stationarity, ...

Summary & Discussion

- Extreme value theory provides a framework to model extreme values
    - GEV for fitting block maxima
    - GPD for fitting threshold exceedances
    - Return level for communicating risk
- Practical Issues: seasonality, temporal dependence, non-stationarity, ...

For Further Reading

- S. Coles
  An Introduction to Statistical Modeling of Extreme Values.
  Springer, 2001.
- J. Beirlant, Y Goegebeur, J. Segers, and J Teugels
  Statistics of Extremes: Theory and Applications.
  Wiley, 2004.
- L. de Haan, and A. Ferreira
  Extreme Value Theory: An Introduction.
  Springer, 2006.
- S. I. Resnick
  Heavy-Tail Phenomena: Probabilistic and Statistical Modeling.
  Springer, 2007.