Extreme Value Theory :

With Applications to Temperature Data

Antoine Pissoort
Supervised by Johan Segers

February 22, 2017

ISBA Université Catholique de Louvain

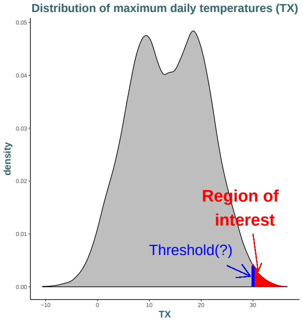
Table of contents

Introduction

Why do we need $\boxed{\text{Extreme Value Theory}}$?

A lot of applications in various domains :

- **Financial** : Risk analysis, insurance, stock fluctuations, ...
- $\boxed{\textbf{Environmental}}$ most "important" application : heatwaves, floods, drought, hurricanes, ...
- ...



⇒ **Extreme Value Theory** allows a relevant and efficient modelling of these **extremes** located at the **tail**(s) of the distribution.

- Low frequency of occurences (**small samples**)
- Can be harder to grasp, to define, ...

⇒ large uncertainty

☐ 2 main methods :

⇒ **Block-maxima**
⇒ **Peaks-Over-Threshold**
⇒ ...

Whereas *TCL* deals with $\overline{X}$, we look here for a non-degenerate distribution in the limit for $X_{(n)} = \max(X_1, \ldots, X_n)$.

Theorem (*Extremal Type* from Fisher-Tippett (1928))

Let $X_i \overset{iid}{\sim} F$ *and let* $a_n > 0$, $b_n \in \mathbb{R}$ *be sequences of constants, then*

$$Pr\Big\{a_n^{-1}(X_{(n)} - b_n) \leq z\Big\} = F^n(a_n z + b_n) \longrightarrow G(z), \qquad n \to \infty.$$

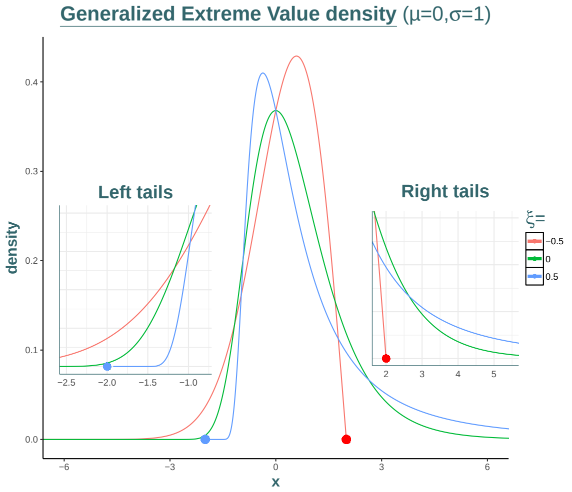
where G *is a non-degenerate distribution function.*

G is known as the *Generalized Extreme Value* (GEV) distribution :

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

3 parameters : • $\mu \in \mathbb{R}$: *location* • $\sigma > 0$: *scale* • $\xi \in \mathbb{R}$: ***shape***

▷ Which block length ? (bias vs variance)



GEV distribution has 3 "child distributions" as special case

ξ determines distribution

- **Weibull** family : $\boxed{\xi < 0}$ right-endpoint, left heavy-tailed
- **Gumbel** family : $\boxed{\xi = 0}$ light-tailed
- **Fréchet** family : $\boxed{\xi > 0}$ left-endpoint, right heavy-tailed

- E.g. in our temperature data : $\xi \approx -0.25$ $\Rightarrow$ right-endpoint (why ?)

Same principle as for block-maxima. But here, we don't look for only one value per block but all values which exceed a fixed(?) threshold u.

▷ We deal now with the excess $\boxed{Y = X - u}$.

Theorem (Pickands - Balkema - de Haan (1974))

$X_i \overset{iid}{\sim} F$ and $x_* = \sup\{x : F(x) < 1\}$ is the right-endpoint of F. We have

$$Pr\{X - u \leq y \mid X > u\} = \frac{F(y+u) - F(u)}{\overline{F}(u)} \longrightarrow H_{\xi,\sigma_u}(y), \qquad u \to x_*,$$

where H is a (non-degenerate) Generalized Pareto Distribution (GPD).

We can easily prove it by the link with GEV (...)

$$H(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)_+^{-\xi^{-1}}, & \xi \neq 0; \\ 1 - \exp\left\{-\frac{y}{\sigma_u}\right\}, & \xi = 0. \end{cases}$$

Again, 3 parameters :

- μ : *location*
- $\sigma_u = \sigma + \xi(u - \mu)$: *scale*
- $\boxed{\xi}$: ***shape***

$\boxed{\xi}$ determines distribution

- "**Beta**" type : $\boxed{\xi < 0}$ bound at $-\sigma/\xi$
- **Exponetial** type : $\boxed{\xi = 0}$ light-tailed
- **Pareto** type : $\boxed{\xi > 0}$ heavy-tailed

Research Question(s) ?

- "In-depth" study of univariate Extreme Value Theory (EVT)
  - Characterization of main theoretical models/methods (see earlier)
  - Dealing with stationary and **non-stationary** sequences
- Application of EVT to a new dataset gratefully delivered by the IRM :
  - Assess "*Climate Change*" by trend analysis, extremes variability, ...
  - Make relevant statistical inferences : Return Levels, ...

? Performance simulation study and comparisons of several "advanced" methods, effective in a **non-stationary** context :

- Varying threshold selection methods: **Mixture models**, ...
- **Bayesian** Analysis to better quantify uncertainty. Problem : *prior* ?
- **Neural Networks** : based on R library `GEVcdn`
- **Bootstrap** evaluation to gain precision. E.g.: confidence intervals
- ...

Literature Review

Literature Review : Main Books

Basics for *Extreme Value Theory*

- **Coles (2001)** → short & very comprehensive
- Reiss and Thomas (2007) → more details & statistical derivations
- Embrechts et al. (2011) → finance-insurance oriented
- ...

More (mathematically) strict and extensive

- **Beirlant et al. (2006)** → big coverage + applications : **time-series**
- Falk et al. (2011)
- Haan and Ferreira (2006)
- ...
- **Dey and Yan (2016)** → lots of research areas covered : **non-stationarity**, **mixtures**, **bayesian**,...

Climate-oriented

- Mudelsee (2014) → + ***bootstrap*** applications
- AghaKouchak et al. (2012) "in a **changing climate**" ...
  ⇒ deals with **non-stationarity**
- ...

Variety of interesting articles for each **part**, but mainly :

(Advanced : multivariate)

...

Bayesian

Stephenson and Ribatet (2006) : for *evdbayes* R package

Northrop et al. (2017) : Accounts for uncertainty in threshold selection

⇒ Bayesian model-averaging

Climate - Neural Network

Galiatsatou et al. (2016)

Cannon (2010)

Bootstrap

Wide applications, in many articles

Stationary (clustering)

Ferro and Segers (2003) :

Non-stationary

Cheng et al. (2014) : ...

Mixture Models

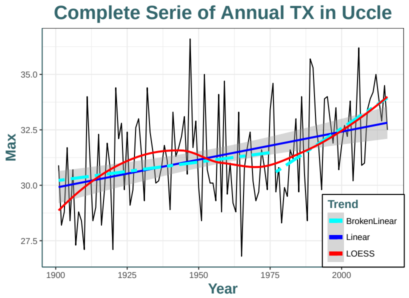
Scarrott and MacDonald (2012) : Review of available methods

Hu (2013) : thesis on *evmix* package

......

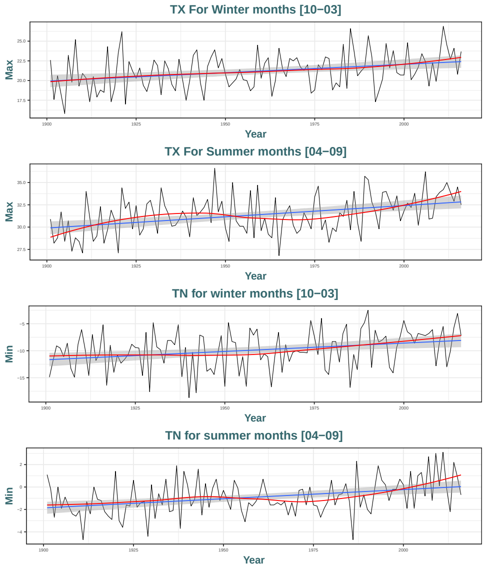
"Methodology" : Next Steps

First Analysis of the Data



- ☐ **TX** and TN in Uccle [1901-2016]
  - Upward trend : significant for all (TX, TN) but heavier for TX in summer
  - Trend heavier in [1976-2016] than [1901-1975] : **climate warming**

- ☐ We considered max(min) with (half-)yearly blocks. Also done with seasonal or monthly blocks



... ...

Main Methods for GEV include

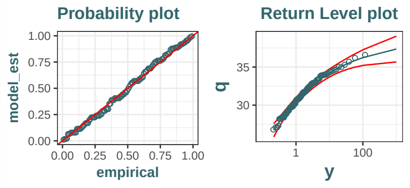
- **Maximum Likelihood** (ML) as usual is a good method but irregularities arise when $\xi < -0.5$
- **Penalized ML** : prior for ξ to penalize values close to irregular region
- **Profile-likelihood** : preferred in EVT due to the usual asymmetries in the likelihood surface of shape parameter
- **Moments** and **Probability-Weighted-Moments**,...

Main Methods for POT include

- **Hill** estimator ($\xi > 0$)
- **Pickands**, **ML**,...
- **Threshold selection** is an issue :
  - for example look at **mean residual life** plot and check for linearity.
  - or vary threshold following seasons.
- **Point Process** approach very useful : unifies the 2 models and provides a natural formulation of non-stationarity in POT

Model diagnostics : Validation

E.g. check fit of the model by **PP** or QQ plot, **return level plot**, density, ...



⚠ Good fit but actually, data are not stationary... $\Rightarrow$ needs to handle this

$D(u_n)$ condition : limited long-range dependence

In words, it says that if the X_i's are not independent (most often) then provided the long-range dependence is limited, the extremal laws still occur in the limit.

Theorem : Leadbetter (1983)

Let $\{X_i^*\}$ be stationary series and $\{X_i\}$ be iid series of n R.V.'s. Then, we have that

$$\Pr\{a_n^{-1}(X_{(n)} - b_n) \leq x\} \longrightarrow G(x), \qquad n \to \infty$$

and

$$\Pr\{a_n^{-1}(X_{(n)}^* - b_n) \leq x\} \longrightarrow G^*(x), \qquad n \to \infty$$

where

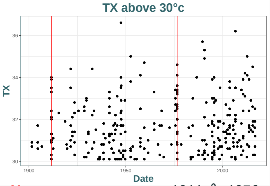
$$G^*(x) = G^\theta(x).$$

$\boxed{\theta}$ is the **extremal index** which quantifies the extent of extremal dependence

- ▷ Parameters of G^* and G are related

⚠ POT : **Clusters of extremes** with mean size θ^{-1}

For threshold of 30°c : we obtain $\boxed{\theta \approx 0.5}$ ($\not\perp$)



Heat waves: **summers 1911 & 1976**

$\Longrightarrow$ Needs for **declustering** (e.g. Ferro and Segers (2003))

Relaxing Independence Assumption (2) : $\boxed{\text{Non-stationary}}$ Extremes

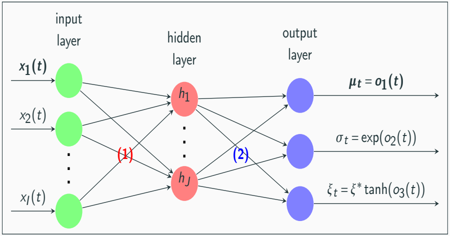
- X_i's almost never $\perp\!\!\!\perp$, and rarely even stationary, especially for temperatures data in a context of **climate change**
- **Non-stationarity** occurs in 2 ways : 1) **Trend** and 2) **seasonality**.
  1) can be handled e.g. by allowing μ to vary while 2) can be handled by various ways (e.g. seasonal varying threshold)
- Inferences such as return levels must account for this likely trend

Comparisons of nested models by the statistic of **deviance**

Trend model in μ	ℓ	df	p-value
constant	-251.8	3	
linear	**-241.8**	**4**	$8 \cdot 10^{-6}$
quadratic	-241.5	5	0.42

- $GEV(\mu(t),\sigma,\xi)$ with $\mu(t) = \beta_0 + \beta_1 \cdot t$ is the preferred model so far.
- Allowing time-varying scale parameter does not seem useful.
- We still have to reinforce this result $\Downarrow$

Neural Networks $\Rightarrow$ Improvements (?)



TikzFig.: ***Neural Network applied to GEV**. helped by Cannon (2010)*

(1) $h_j(t) = m\left(\sum_i^I x_i(t)\cdot w_{ji}^{(1)} + b_j^{(1)}\right)$

(2) $o_k(t) = \sum_j^J h_j(t)\cdot w_{kj}^{(2)} + b_k^{(2)}$, $(k = 1, 2, 3)$

- need to choose relevant # hidden layers
- We rely on (refined?) `GEVcdn` R package

model	AIC_c	BIC	hidden	df
stationary	-19.6	-11.5	0	3
$\boldsymbol{\mu_t}$	**-37.4**	**-26.7**	**0**	**4**
$\mu_t,\ \sigma_t$	-35.4	-22.2	0	5
$\mu_t,\ \sigma_t,\ \xi_t$	-34.2	-18.4	0	6
μ_t	-35.4	-19.6	1	6
$\mu_t,\ \sigma_t$	-36.2	-17.9	1	7
$\mu_t,\ \sigma_t,\ \xi_t$	-34	-13.3	1	8
μ_t	-37.4	-14.3	2	9
$\mu_t,\ \sigma_t$	-32.5	-4.7	2	11
$\mu_t,\ \sigma_t,\ \xi_t$	-38.4	3.9	2	13

(...)

▷ Same model is chosen (so far?) : **validation**

▷ Linear trend in μ seems acceptable but we did not consider all models (yet?)

$\Rightarrow$ **Other covariates than time ?**

$\Rightarrow$ NN is a **powerful** method dealing efficiently with non-stationarity by considering lots of "models", relying on Generalized Maximum Likelihood of Martins and Stedinger (2000)

Bagging **process of averaging ensemble of models fitted on bootstrapped data**

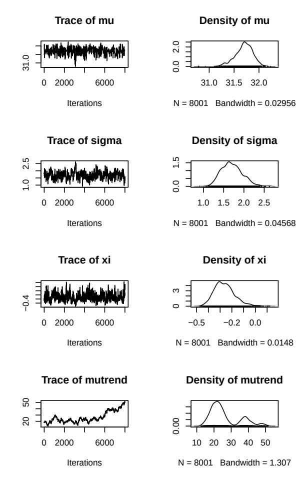
$\Rightarrow$ Decrease variance of the estimates

$$\boxed{\pi(\theta|\boldsymbol{X}) = \frac{\pi(\theta)\cdot L(\theta;\boldsymbol{X})}{\int_\Theta \pi(\theta)\cdot L(\theta;\boldsymbol{X})d\theta} \propto \pi(\theta)\cdot L(\theta;\boldsymbol{X}), \qquad \theta=(\mu,\sigma,\xi).}$$

- Can overcome regularity conditions of usual likelihood inferences.
- Allows better quantification of uncertainty from the posterior (predictive) distribution

▷ Non-informative priors (large variance) leads to ≈ same estimates as others methods (such as ML) in stationary models → kind of **validation**

▷ Can accommodate trend, seasonality or even variable threshold. We need to improve modelling to include that. E.g.: problem in trend here : large variance/autocorrelation,...



? Could we reliably defend a sustainable prior, enhancing the analysis ?

Mixture Models

Mixture Models rely on 2 separate

- Bulk model : below threshold, either parametric or non(semi)parametric
- Excess model : above threshold, is of the GPD family

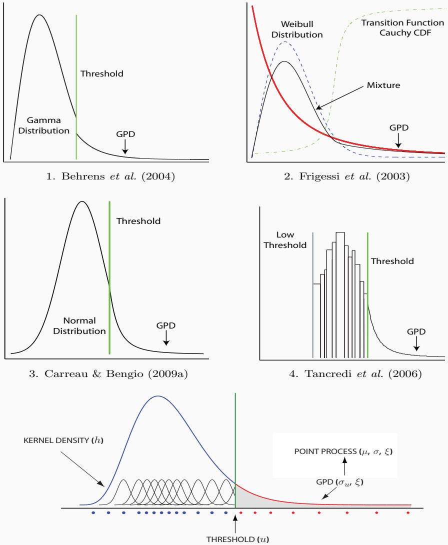
⇒ Put together to obtain full distribution of the data : improve fitting, asses threshold uncertainty, ...

$$\boxed{f(x) = (1-\phi_u)\cdot \boldsymbol{b_t(x)} + \phi_u \cdot \boldsymbol{g(x)},}$$

$\phi_u = \Pr(X > u)$ is the *"tail fraction"*.

- Main R package is `evmix`
- From now, we did not obtained relevant results

? In theory, model seems very interesting... But in practice, is it really worth it?



1. Behrens *et al.* (2004)
2. Frigessi *et al.* (2003)
3. Carreau & Bengio (2009a)
4. Tancredi *et al.* (2006)
5. MacDonald *et al.* (2011a)

Figure **of models taken from Scarrott and MacDonald (2012)**

"Conclusions"

Temporary Conclusions

What I have done :

- Literature review and description of most concepts from univariate EVT
- R implementation of the data (still to enhance?) :
  - preprocessing of data, methods from the various packages in EVT + comparisons, (re)building of functions
  - Stationary + non-stationary analysis, variable threshold
  - ▷ Bayesian analysis, Neural Network
- ▷ Understanding concept of Mixture Models in EVT
- ▷ Bootstrap to improve accuracy or to compare models

Still to do :

- ▷ Aggregate used concepts into a smooth goal-oriented final document

? Build other simulated data in order to make reliable comparisons of the available models.

References

AghaKouchak, A., Easterling, D., Hsu, K., Schubert, S., and Sorooshian, S. (2012). *Extremes in a Changing Climate: Detection, Analysis and Uncertainty*. Springer Science & Business Media.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2006). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons. Google-Books-ID: jqmRwfG6aloC.

Cannon, A. J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes*, 24(6):673–685.

Cheng, L., AghaKouchak, A., Gilleland, E., and Katz, R. W. (2014). Non-stationary extreme value analysis in a changing climate. *Climatic Change*, 127(2):353–369.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer London, London.

Dey, D. K. and Yan, J. (2016). *Extreme Value Modeling and Risk Analysis: Methods and Applications*. CRC Press. Google-Books-ID: PYhUCwAAQBAJ.

Embrechts, P., KlÃijppelberg, C., and Mikosch, T. (2011). *Modelling Extremal Events: for Insurance and Finance*. Springer Berlin Heidelberg. Google-Books-ID: dfZecgAACAAJ.

Falk, M., HÃijsler, J., and Reiss, R.-D. (2011). *Laws of Small Numbers: Extremes and Rare Events*. Springer Basel, Basel.

Ferro, C. A. T. and Segers, J. (2003). Inference for Clusters of Extreme Values. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(2):545–556.

Galiatsatou, P., Anagnostopoulou, C., and Prinos, P. (2016). Modeling nonstationary extreme wave heights in present and future climates of Greek Seas. *Water Science and Engineering*, 9(1):21–32.

Haan, L. d. and Ferreira, A. (2006). *Extreme value theory: an introduction*. Springer series in operations research. Springer, New York ; London. OCLC: ocm70173287.

Hu, Y. (2013). Extreme Value Mixture Modelling with Simulation Study and Applications in Finance and Insurance.

Martins, E. S. and Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36(3):737–744.

Mudelsee, M. (2014). *Climate Time Series Analysis*, volume 51 of *Atmospheric and Oceanographic Sciences Library*. Springer International Publishing, Cham.

Northrop, P., Attalides, N., and Jonathan, P. (2017). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(1):93–120. arXiv: 1504.06653.

Reiss, R.-D. and Thomas, M. (2007). *Statistical analysis of extreme values: with applications to insurance, finance, hydrology and other fields ; [includes CD-ROM]*. BirkhÃďuser, Basel, 3. ed edition. OCLC: 180885018.

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold es-timation and uncertainty quantification. *REVSTATâĂŞStatistical Journal*, 10(1):33–60.

Stephenson, A. and Ribatet, M. (2006). A UserâĂŹs Guide to the evdbayes Package (Version 1.1). *month*.