MATH5805 – A Modern Introduction to Extreme Value Theory

Week 1 - Introduction

Dr Boris Béranger
School of Mathematics and Statistics
UNSW Sydney

CRICOS Provider Code 00098G

Lecture outline

1. General information
2. Motivation
3. Course Overview

1. General information



Contact

Lecturer

Dr Boris Béranger
Room 2050, Anita B. Lawrence Centre (H13)
Email: b.beranger@unsw.edu.au
Webpage: borisberanger.com

Consultation

Please use email to arrange an appointment.

Classes

Classes are held on Weeks 1-5 and 7-10.

Lectures

- Tuesdays: 9-11am, UNSW Business School room 105 (K-E12-105)
- Wednesdays: 2-3pm, UNSW Business School room 205 (K-E12-205)

Tutorials/Lab

- Wednesdays: 3-4pm, UNSW Business School room 205 (K-E12-205)

Attendance is not compulsory but highly recommended!

Pre-requisites

Theoretical knowledge

You should have a good understanding of statistical theory (not limited to):

- the central limit theorem
- probability theory
- likelihood-based inferential techniques
- multivariate statistical methods.

Computational knowledge

- Good data analysis skills in a suitable programming language (like `R` or Python).
- This course itself will use `R` through `RStudio`.

R and RStudio Learning Support

- **Step 1** Download and install R 
- **Step 2** Download and install RStudio 
- **Step 3** Open RStudio and install the following packages: extRemes, ismev, evd, ExtremalDep, SpatialExtremes.

Assessment structure

There will be **4** assessments throughout the term:

Assessment	Weight	When	Where
Quiz	10%	Week 4	In person
Test	15%	Week 7	In person
Assignment	20%	Week 9	Home
Final Exam	55%	TBC	In person

Information specific to each assessment will be provided in Moodle and during the lectures.

References



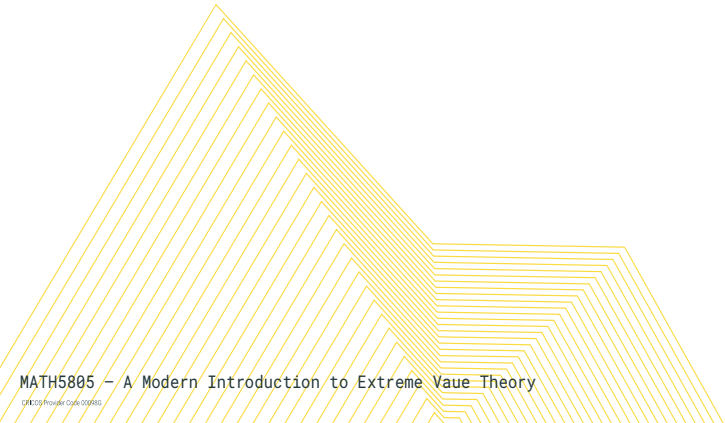
Sources and recommended reading:

1. [BGST04] Beirlant, J., Goegebeur, Y., Segers, J. & Teugels J. (2004) *Statistics of Extremes: Theory and Applications*. John Wiley & Sons.
2. [C01] Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*. Springer.
3. [dHF06] de Haan, L. & Ferreira, A. (2006) *Extreme Value Theory: An Introduction*. Springer.

2. Motivation



2.1 Motivating examples

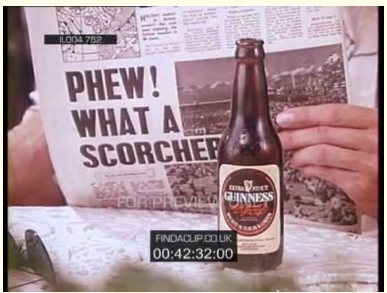


Why are we interested in modelling extremes?

Modelling and predicting extremes is important in many areas.

Few **examples** include:

- 1953 North Sea floods
- 1980 Sinking of the M.V. Derbyshire
- 1999 The Vargas tragedy
- 2003 European heatwave
- 2009 The risk of terrorism
- 2021 Bitcoin prices and exchange rates
- 2021 Fastest 100m sprint times



1953 North Sea floods

On night of 31 Jan–1st Feb 1953 **an extreme storm tide** in the North Sea was caused by a **combination** of:

- high spring tide
- severe European windstorm.

Water levels **locally exceeded 5.6m above (mean) sea level**:

- overwhelmed sea defences
- flooded areas of the Netherlands, England, Belgium, Denmark and France
- ∼2,500 people died.



Watersnoodramp, 1953

1953 North Sea floods

Dutch government set up **Delta Commission**.

Set up flood defences such that the **acceptable period over which complete failure would occur** is:

- North and South Holland:
  1 per 10,000 years
- Other areas at risk from sea flooding:
  1 per 4,000 years
- Transition between high and low land:
  1 per 2,000 years



Erith, London, 1953

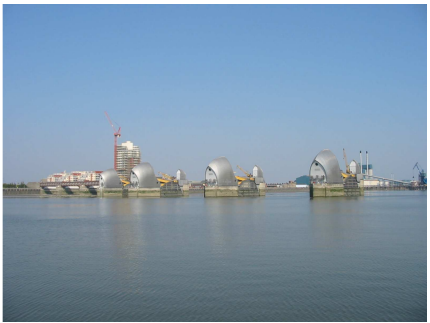
Based on data observed over a much shorter period!!

1953 North Sea floods

Maeslantkering, South Holland



Thames Barrier, London



Defences completed in 1997.

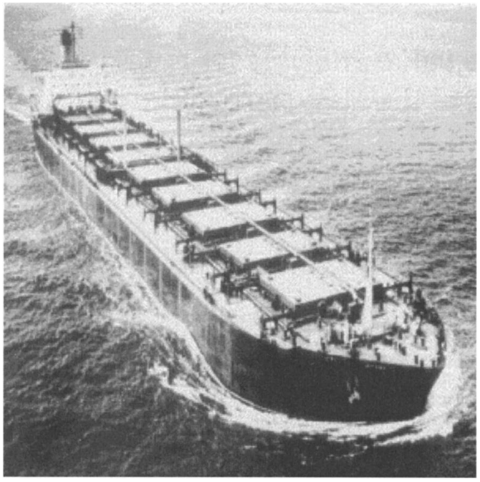
1980 Sinking of the M.V. Derbyshire

In 1980 the bulk carrier **M.V. Derbyshire sank in the Pacific Ocean** during a typhoon.

- All 44 people on board died.
- Remains largest ship the U.K. lost at sea.
- No mayday was received, no lifeboats were launched.

1994: wreckage was recovered

- 2.5 miles under the sea surface.
- Photos suggested: hatch cover on Hold 1 failed during the typhoon.
- This would lower the ship and make flooding in other holds more likely, increasing risk of sinking.



M.V. Derbyshire: Hold 1 is at the front.

1980 Sinking of the M.V.~Derbyshire

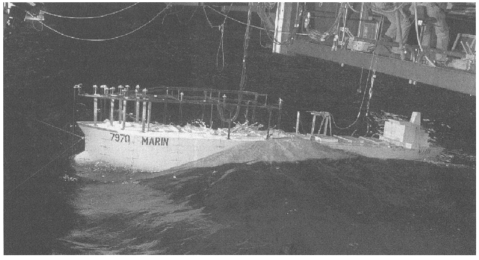
2000: Inquiry (+ court challenges)

- Safety standards (at the time) required hatch to withstand impact of 42 kPa.
- Interest in the risk that the pressure of wave impacts on Hold 1 cover $>$ collapse pressure during the typhoon.

During the court cases:

Assessment of the probability that the maximum impact $>$ 42 kPA for a range of scenarios (relating to the unknown condition of the ship and storm).

Statistical expert witness: How to estimate this probability using best statistical practice?



MARIN test tank with 1:65 scale model of the Derbyshire.

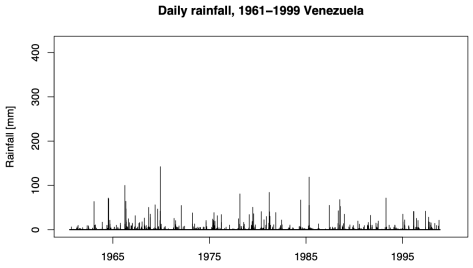
https://en.wikipedia.org/wiki/MV_Derbyshire

1999 The Vargas tragedy

On 14–16 December 1999 **torrential rains caused landslides** through the upper watersheds of the Cerro Grande River in Vargas State on the northern coast of Venezuela.

Major natural catastrophe:

- Caused flash floods and debris flows, mud flows
- Killed tens of thousands of people (10% of Vargas State population)
- Completely destroyed (gone) many coastal tourist towns



Daily rainfall at Maiquetia International Airport

https://en.wikipedia.org/wiki/Vargas_tragedy

1999 The Vargas tragedy

The 3-day storm:

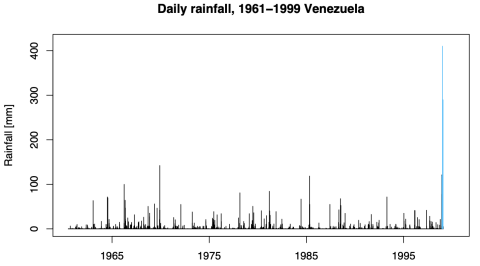
- 1st day = any previously seen maximum
- 2nd day >3 times previous maximum
- 3rd day >2 times previous maxumum

Estimated recurrence time of the 3-day storm range from **250–6 million years!**



How to estimate this? How can we be so uncertain?!

Again: Need to extrapolate events beyond the observable data.

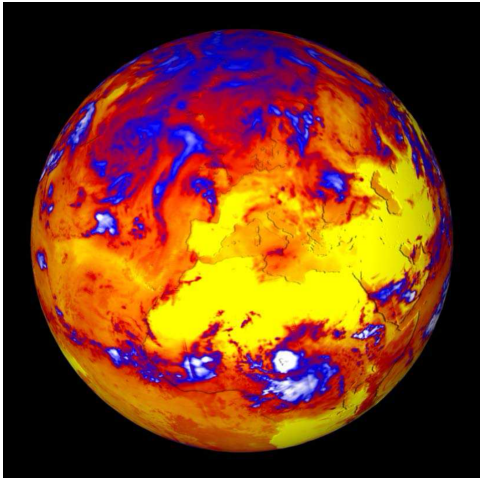




2003 European heatwave

One of the hottest summers on record:

- temperatures $> 40^{\circ}C$ in N. Europe
- almost $50^{\circ}C$ in Portugal
- health crises in many countries (sustained heat)
- combined with drought, created crop shortfall in S. Europe
- linked to forest fires in Portugal, flash floods in the Alps, buckled train tracks in U.K., dried up rivers
- $> 35,000$ *excess deaths* that summer.



2003 European heatwave

Modelling challenges:

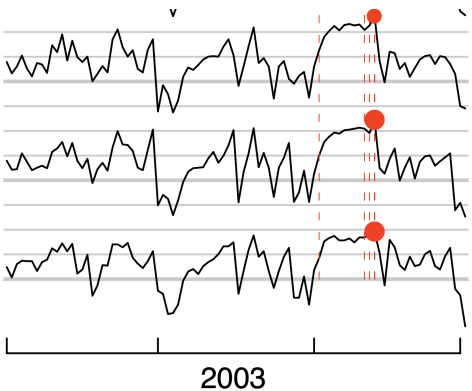
1) Heatwaves

- Not just interested in max temperature (although this is important)
- Heatwaves are sustained high temperatures
- Modelling interest in

$$\max(\min \text{ temperature})$$

over periods of time.

- Questions of temporal dependence, definition of heatwave etc.

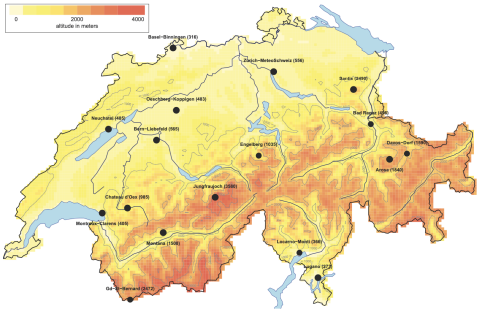


Max daily JJA temperature, Switzerland

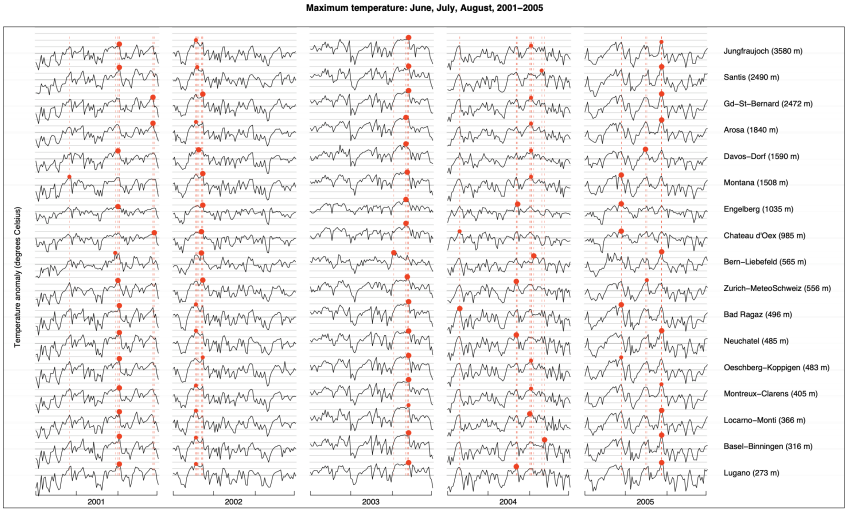
Modelling challenges:

2) Spatial modelling

- Data are inherently spatial (as well as temporal)
- 17 locations over Switzerland
- Could do 17-dimensional multivariate analysis (how?)
- But need to extrapolate predictions to all locations in space.
- Extremes for spatial processes? (Compare: Gaussian processes)



2003 European heatwave



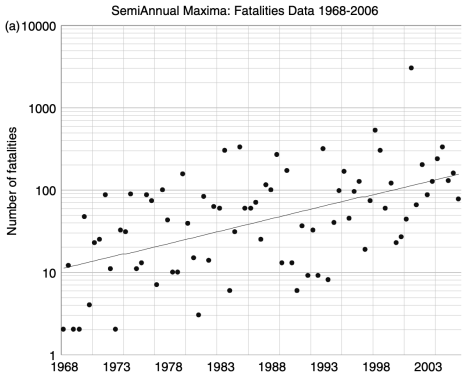
Clear spatial-temporal dependence in the data. How to handle?

2009 The risk of terrorism

Unlike natural disasters, “terrorist threat" is less well understood.

Want to quantify for risk assessment, event prevention, insurance and financial markets.

“Terrorist act” measured by number of fatalities (or casualties). Maximum over 6 months.



Several challenges:

- Clear non-stationarity in the data
  E.g. 9/11 was a clear extreme in 2001, but it is less unlikely now.

Max daily fatalities over 6 months.

2009 The risk of terrorism

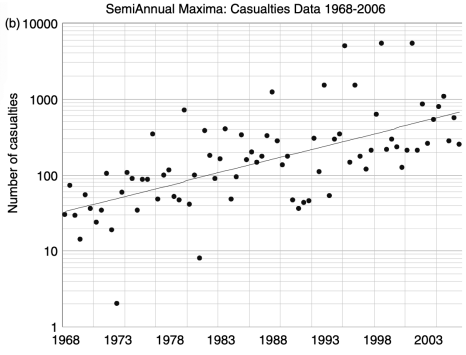
Several challenges:

- Event generation mechanism is non-stationary (E.g. as a response to other events that may (not) be occurring)
- Data are discrete
- Influenced by many factors (country of event, nationality of perpetrators etc.)

If (X_t, Y_t) are (max fatalities, max casualties) for 6-month period t, and want to model jointly, note that X_t and Y_t may correspond to different actual events.



Is this a problem?



Max daily casualties over 6 months.

https://www.start.umd.edu/gtd/

2021 Bitcoin prices

The Basel Committee rules regulate behaviour of banks and financial institutes.

This involves estimating the Value At Risk (Var).

(Very) simplistically this is an extreme quantile of a distribution.

For $X_1, X_2 \ldots$, where X_t is the closing price of a quantity on day t, it is common to model the log daily returns:

$$Y_t = \log(X_t / X_{t-1}).$$

To quantify risk, interest is in extremely large or small values of log daily returns.

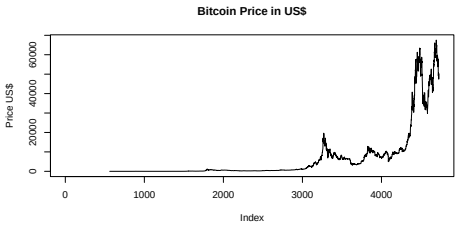
2021 Bitcoin prices

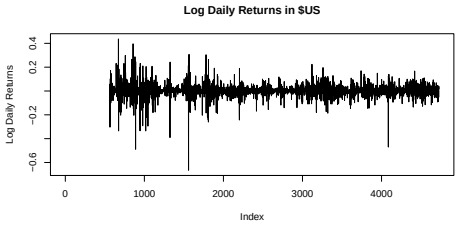
Plots show 4,727 observations of

- Closing price (X_t) in US$
- Log daily returns (Y_t)

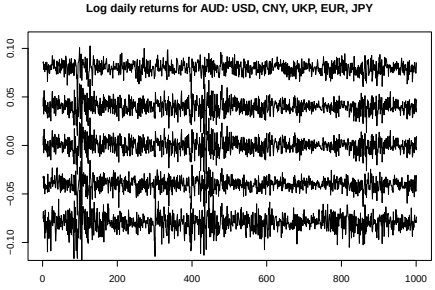
from 3rd Jan 2009 – 12th Dec 2021.

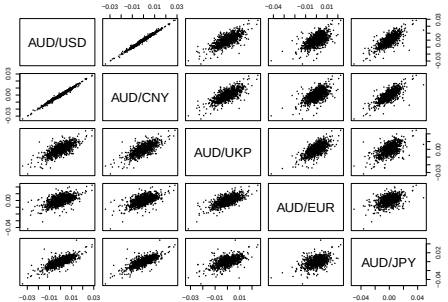
- Extreme high/low changes in Y_t don't correspond to obvious big changes in X_t.
- Big changes in Y_t correspond to relatively large changes in X_t
- How to model extreme high/low changes in Y_t?





2021 Exchange rates





Daily exchange rates: How to handle multivariate extremes and risk measures?

2021 Fastest 100m sprint times

Fastest 100m times show clear improvement since the 1970s.

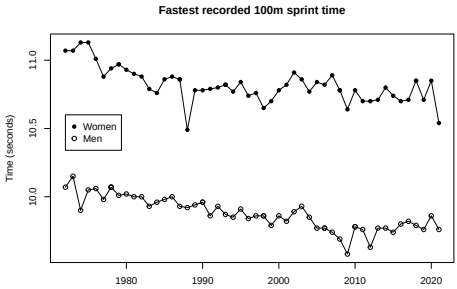
Possibly constant performance since ∼1990 (women) and ∼2005 (men)?

Various factors affect these times (wind speed and direction, hand timing etc).

Have we reached the limit of performance?

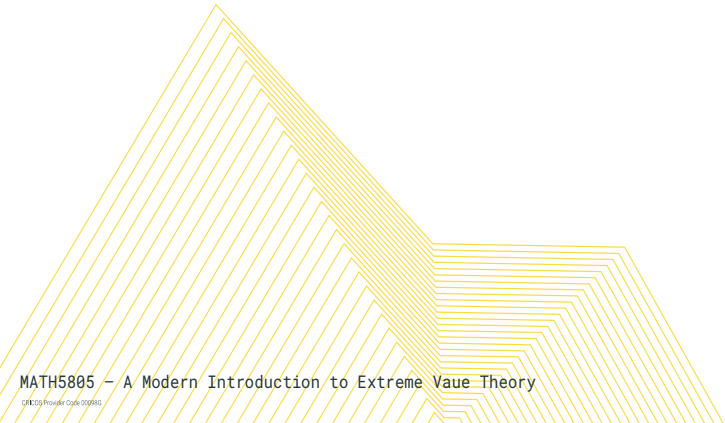
Can we calculate the probability that:

- The world record will be broken next year?
- The world record will hold for the next 50 years?



https://en.wikipedia.org/wiki/100_metres

2.1 Motivation for Extreme Value Theory



Motivation for Extreme Value Theory

Summary:

- Interested in rare (e.g. 1 in 500 years) but very high impact events (impact $ cost, insurance $ cost, lives, time). E.g. The Vargas tragedy.
- Need to estimate the probability of occurrance of future extreme events in order to take preventative measures. E.g. sea wall height, design features of ships etc.
- Events of interest are typically beyond range of observed data, so no empirical estimates available. E..g. Failure of sea defences in 1 per 4,000 years with < 100 years of data.

How might we model extreme values using standard methods?

Motivation for Extreme Value Theory

Option 1: Fit a model to the data and then extrapolate

Want to estimate how long (how many observations) before we observe $X > 5$.

$X \sim F$

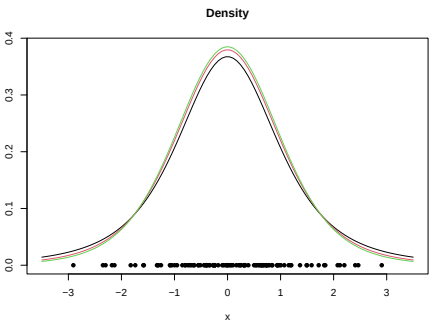
Data could come from any model:

- several could fit data well.
- $\Rightarrow$ can't justify any choice of F.

But model fit is driven by non-extreme data: $\bar{X}$.

Choice of F completely determines the tail behaviour.

$\Rightarrow$ implications for tail prediction are large.



Motivation for Extreme Value Theory

Write the upper tail probability as:

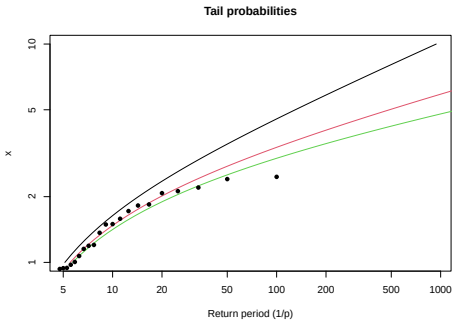
$$p(x) = 1 - F(x) = \mathbb{P}(X > x)$$

$1/p(x)$ is the number of observations we need before we expect to see a value of $X > x$.

$1/p(x)$ is known as the **return period** of x.

- return period for $X > 5 = 130$ obs
- return period for $X > 5 = 487$ obs
- return period for $X > 5 = 1{,}278$ obs

How can we trust any of these estimates when we are unable to justify any of them?



Conclusion: Estimating F to quantify extremes seems inherently problematic.

Option 2: Model the extremes directly

E.g. Week 1 Tutorial:

If $X_1, \ldots, X_n \sim F_X$ then the distribution function of $Y = \max(X_1, \ldots, X_n)$ is

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(\max(X_1, \ldots, X_n) \leq y) \\ &= \mathbb{P}(X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y) \\ &= \mathbb{P}(X_1 \leq y)\mathbb{P}(X_2 \leq y) \ldots Pr(X_n \leq y) \\ &= \mathbb{P}(X \leq y)^n = F_X(y)^n. \end{aligned}$$

This gives a way for us to model the extremes directly but requires independence and knowledge of F.

Motivation for Extreme Value Theory



How can we know about $F_X(y)^n$ as $n \to \infty$ if we don't know F?

EVT is **specifically designed** to model extremes of processes:

- beyond the range of observed data
- without knowing distribution of F
- in a mathematically justified way.
- is far more reliable than the above approaches.
- univariate, multivariate, spatial, ...

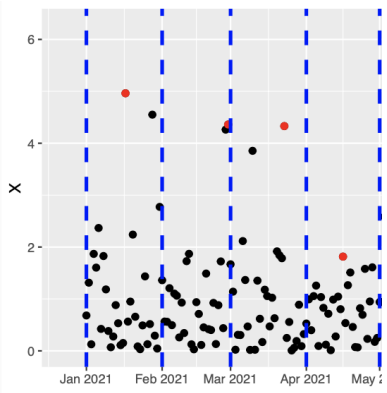
3. Course Overview



Part I: Univariate Extremes

Different ways to define and model extremes.

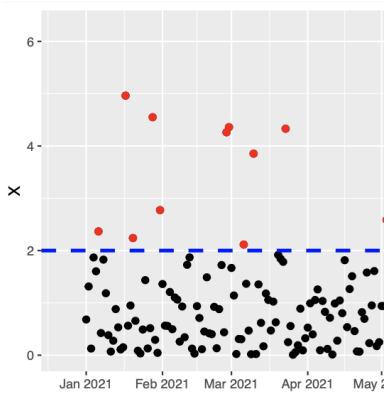
- **Block maxima:** $Y_n = \max\{X_1, \ldots, X_n\}$



Part I: Univariate Extremes

Different ways to define and model extremes.

- Block maxima: $Y_n = \max\{X_1, \ldots, X_n\}$
- **Threshold exceedance:** $X - u | X > u$
- Point process



Course Overview

Part I: Univariate Extremes

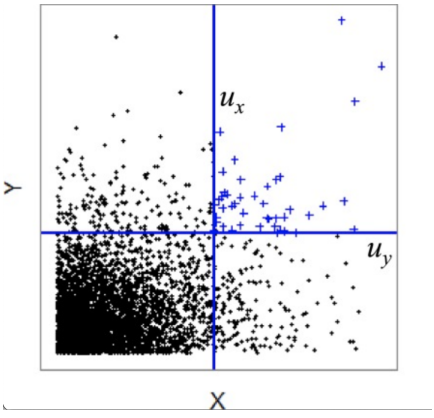
Will cover:

- Mathematical theory and justification
- Statistical implementation
- Applications using R

Univariate extremes is well understood and fairly straightforward.

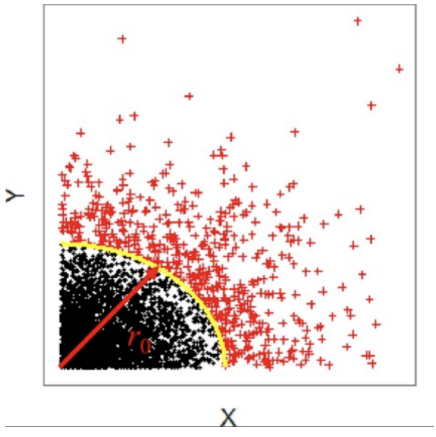
Part II: Multivariate Extremes

- Block maxima: (Y_{1n}, Y_{2n})
  $Y_{1n} = \max\{X_{11}, \ldots, X_{1n}\}$
  $Y_{2n} = \max\{X_{21}, \ldots, X_{2n}\}$
- **Threshold exceedance:**
  $(X - u_x, Y - u_y) | X > u_x, Y > u_y$



Part II: Multivariate Extremes

- Block maxima: (Y_{1n}, Y_{2n})
  $Y_{1n} = \max\{X_{11}, \ldots, X_{1n}\}$
  $Y_{2n} = \max\{X_{21}, \ldots, X_{2n}\}$
- Threshold exceedance:
  $(X - u_x, Y - u_y) | X > u_x, Y > u_y$
- **Point process: pseudo-polar co-ordinates**



Course Overview

Part II: Multivariate Extremes

Will cover:

- Extensions of univariate definitions.
- New extremal dependence concepts.
- Theory, statistics, application.

More involved than univariate.

Course overview

Part III: Spatial Extremes (IF TIME PERMITS)

- Max-stable processes
- r-Pareto processes

Overall: This course is a combination of **theory** (we all love maths!) and **practice** (if its not useful then ...?!)

Final comments

First time this course will have been given:

- Possible (likely?) there will be errors/typos
- Potentially more elegant solutions than those given
- Please be forgiving!
- Help us find the mistakes and correct them as we go along.
- Support each other with your learning.