Developing clustering algorithms for conditional extremes models

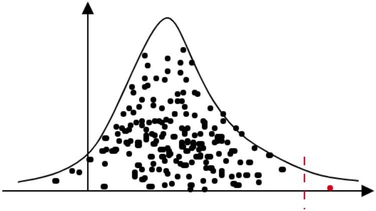
Thesis formulation report (TFR) presentation
9^{th} October, 2024
Paddy O'Toole
Supervised by Christian Rohrbeck and Jordan Richards (University of Edinburgh)

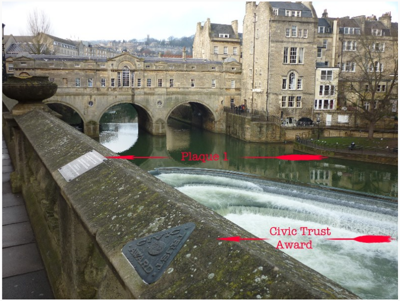
Table of Contents

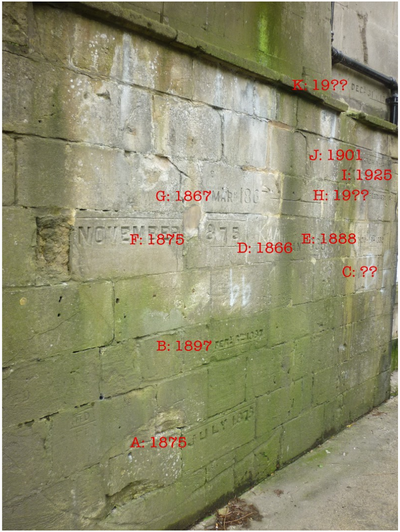
Introduction

Introduction

- Extreme Value Theory models the extreme tails of distributions X
- Only known method that can reliably predict beyond observations in extremal context
- Application fields include finance and insurance, and environmental data, such as extreme precipitation and/or wind speeds.



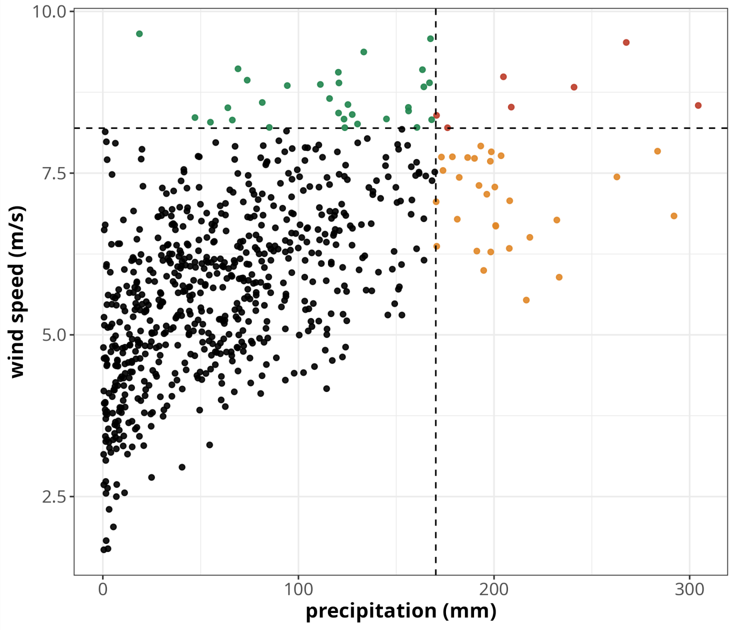

Plaque 1
Civic Trust
Award


K: 19??
J: 1901
I: 1925
G: 1867
H: 19??
F: 1875
D: 1866
E: 1888
C: ??
B: 1897
A: 1875

Multivariate extremes

- Concurrently occurring extremal events can be particularly destructive
- $\mathbb{P}(X \in C) = \sum_{i=1}^{d} \mathbb{P}(X_i \in C_i)$, for some extreme set C_i corresponding to each vector X_i.
- This may refer to different spatial locations, or different variables
- For example, offshore platforms must be built to withstand extreme wind speed and wave height conditions at sea
- Storm defences must withstand extreme rainfall and wind speed conditions, and insurance premiums must take into account these particularly destructive events

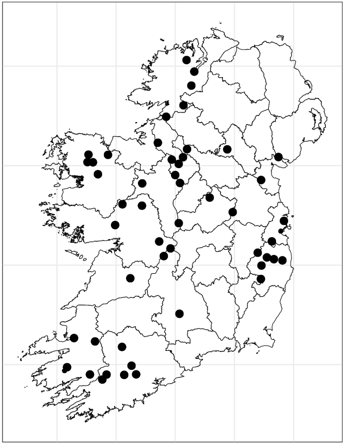
Multivariate extremes



Motivating example

Motivating example - Ireland

- Extreme precipitation and wind speed in Ireland, 1990-2020
- Precipitation data from 52 Met Éireann weather sites across country, wind data from ERA5 reanalysis dataset.
- Take weekly sum of precipitation and mean of daily wind speed maxima for Winter only (Oct-Mar), in line with Vignotto et. al. (2021) [1].



Univariate extremes

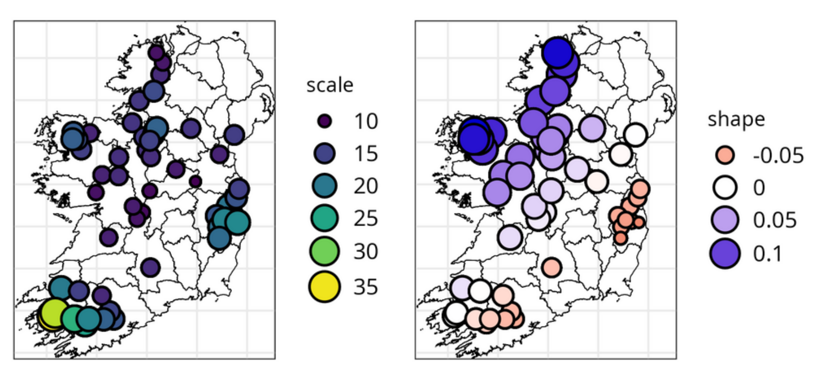
Generalised Pareto distribution

The Generalised Pareto distribution (GPD) has survival function:

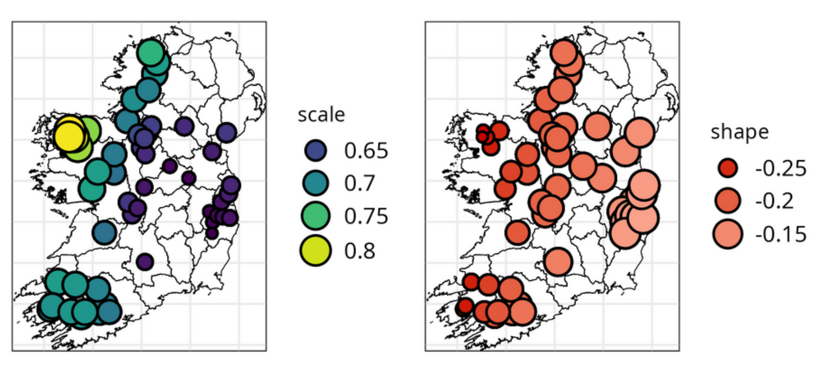
$$\mathbb{P}(X > x + u \mid X > u) = \left(1 + \xi \frac{x}{\sigma}\right)_+^{-1/\xi},$$

- $(x)_+ = \max(0, x)$
- Scale σ, shape ξ, threshold u
- For $\xi < 0$, (increasingly small) finite upper end point
- Can model parameters as function of spatial location, i.e. $\sigma(s), \xi(s)$, with $s =$ (longitude, latitude)
- Threshold often taken as high quantile of data, can also model as $u(s)$

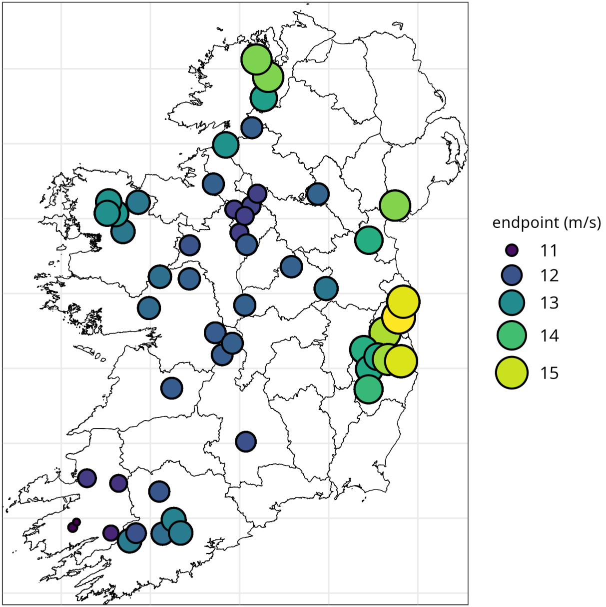
Motivating example - Precipitation



Motivating example - Wind speed



Motivating example - Wind speed



Conditional extremes

Introduction

- Gets name from modelling *conditional* on single variable, e.g. precipitation conditional on observing extreme wind speed
- Has simple marginal and dependence components
- Can model all varieties of extremal dependence, from asymptotic independence (Probability of X_i being extreme is not effected by X_j being extreme) to perfect dependence (X_i is extreme only where X_j is extreme)
- Some definitions:
    - X_{-i}: Vector X without i^{th} component
    - $\alpha_{|i}$: Vector of parameters $\alpha_{j|i}$ conditional on X_i, $j \neq i$

Univariate

Univariate component of model: semi-parametric piecewise function of empirical distribution below threshold and GPD (from previous section) above threshold:

$$\hat{F}_{X_i}(x) = \begin{cases} 1 - \{1 - \tilde{F}_{X_i}(u_{X_i})\}\{1 + \xi_i(x - u_{X_i})/\sigma_i\}_+^{-1/\xi_i} & \text{if } x > u_{X_i} \\ \tilde{F}_{X_i}(x) & \text{if } x \leq u_{X_i}, \end{cases}$$

where $\tilde{F}_{X_i}$ is the empirical distribution function of X_i.

Marginal transformation

Marginals transformed to Laplace margins using probability integral transform:

$$Y_i = \begin{cases} \log\{2F_{X_i}(X_i)\} & \text{for } X_i < F_{X_i}^{-1}(0.5), \\ -\log\{2(1 - F_{X_i}(X_i))\} & \text{for } X_i \geq F_{X_i}^{-1}(0.5), \end{cases}$$

- transformation denoted Y_i to differentiate from original marginals X
- Both tails exponential with mean 1

Multivariate

Dependence component of model:

$$Y_{-i} = \alpha_{|i} y_i + y_i^{\beta_{|i}} Z_{|i}, \text{ for } Y_i = y_i > u_{Y_i}.$$

- $\alpha_{j|i} \in [-1, 1]$ is slope parameter for Y_j conditional on Y_i, $\beta_{j|i} \in (-\infty, 1]$ is spread parameter
- Residuals $Z_{|i}$ said to have distribution $G_{|i}$
- $\beta_{|i}$ controls level of stochasticity of relationship between Y_{-i} and large Y_i.
- Special cases:
  - $\alpha_{|i} = 0, \beta_{|i} = 0 \implies Y_{-i}, Y_i$ independent
  - $\alpha_{|i} = -1/1, \beta_{|i} = 0 \implies$ perfect positive/negative dependence
  - $-1 < \alpha_{|i} < 1 \implies$ asymptotic independence

Estimation

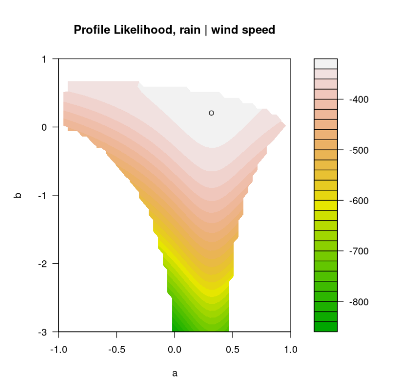
Common estimation procedure:

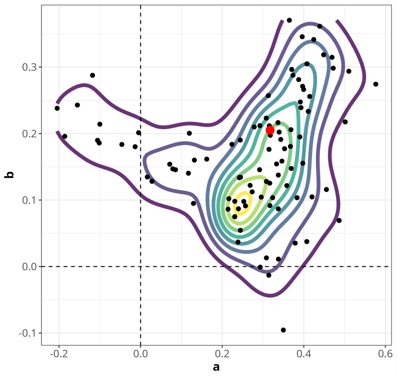
1. First assume $Z_{|i} \sim N(\mu_{|i}, \sigma_{|i})$, generate residuals from this distribution
2. Use likelihood methods to estimate $\hat{\alpha}_{|i}, \hat{\beta}_{|i}$, and nuisance parameters $\hat{\mu}_{|i}, \hat{\sigma}_{|i}$
3. Estimate $\hat{G}_{|i}$ as the empirical distribution of

$$Z_{|i} = \frac{Y_{-i} - \hat{\alpha}_{|i} Y_i}{Y_i^{\hat{\beta}_{|i}}}.$$

This procedure is very simple, and can easily be used to generate MC samples to calculate desired probabilities

Motivating example – Uncertainty

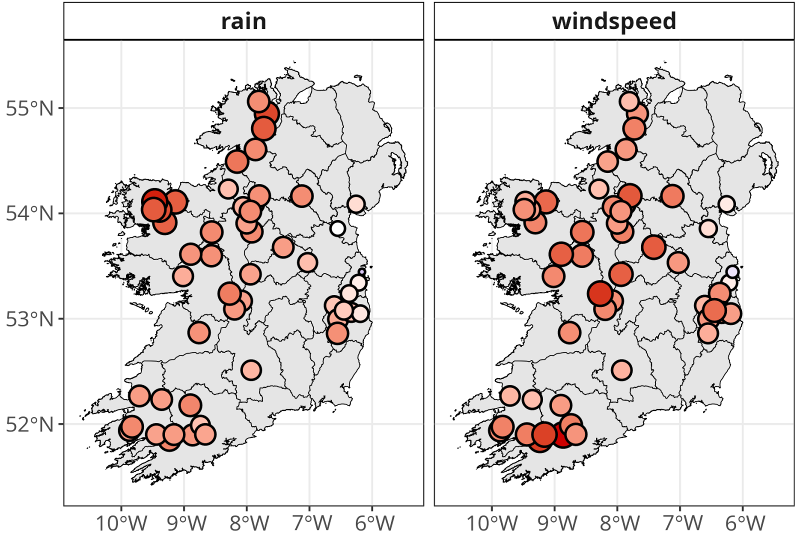




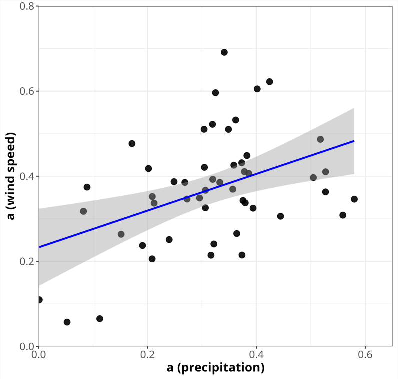
Fixing β

- Problem: Difficult to interpret results while varying both α, β due to uncertainty (and negative, unlikely estimates for β)
- "Hack": Fix $\beta = 0.1$ (allowing some stochasticity), estimate only α
- Idea: By estimating only one parameter, should contain all information about extremal dependence, and be more easily interpretable

Results



Interpretation



- High values of α for wind speed seem to be positively correlated with high values for rainfall, indicating locations where concurrent extremal events are more common
- Extreme rainfall is more likely to occur along with high wind speeds the further West you travel

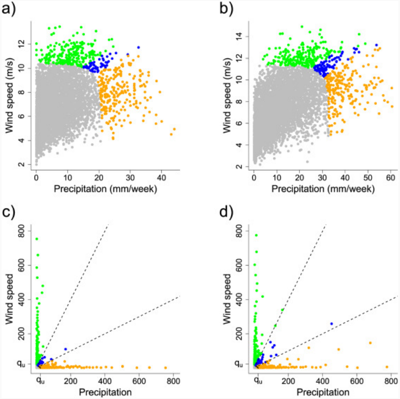
Clustering

Clustering

- There are extensions to CE model which seek to overcome shortcomings of “vanilla” model, such as spatial and spatio-temporal versions.
- Alternative/additional approach: clustering
- Clustering done for two main reasons: to enhance the explainability and interpretation of data, and to improve parameter estimation

Explanatory clustering

- Mainly derives distance metric to perform clustering like k-means, k-mediods, and have user-defined cluster number
- Bernard et. al. (2013) [2] uses the F-madogram (extremal variogram) as distance measure
- Vignotto et. al. (2021) [1] computes KL divergence for risk functions of transformed bivariate rainfall and wind speed observations for different raster locations in UK and Ireland



Hierarchical clustering

- Seeks to find (latent, data-driven) groupings over which parameter inference is optimised
- In Frequestist setting, EM algorithm often used to sequentially maximise likelihood over group membership and within-group parameters (Carreau et. al. (2017) [3], Dupuis et. al. (2023) [4])
- In Bayesian setting, one method includes using Reversible Jump MCMC algorithm which estimates number of clusters and cluster membership as latent variables (Rohrbeck, Tawn (2021) [5])
- Need for priors can be seen as both a strength and a weakness, but the use of distributions improves uncertainty estimation

Conclusions and future work

Conclusions and future work

Conclusions

- Conditional extremes model explained and shown with motivating example
- "Hacks" required to get reasonably interpretable results with "vanilla" model
- Lit review of clustering within extremes , some promising methods found

Future Work

- Perform k-mediods similar to [1] on conditional extremes regression line, where KL divergence appears to be appropriate distance metric, compare results
- Derive Bayesian clustering algorithm similar to [5]
  - Sensible initial prior distribution for Z is the multivariate Normal distribution, will assess suitability in practice
- If successful, expand schemes to extended conditional extremes models

References

Edoardo Vignotto, Sebastian Engelke, and Jakob Zscheischler.
Clustering bivariate dependencies of compound precipitation and wind extremes over great britain and ireland.
Weather and Climate Extremes, 32:100318, 2021.

Elsa Bernard, Philippe Naveau, Mathieu Vrac, and Olivier Mestre.
Clustering of maxima: Spatial dependencies among heavy rainfall in france.
Journal of climate, 26(20):7929–7937, 2013.

Julie Carreau, P Naveau, and Luc Neppel.
Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation.
Water Resources Research, 53(5):4407–4426, 2017.

Debbie J Dupuis, Sebastian Engelke, and Luca Trapin.
Modeling panels of extremes.
The Annals of Applied Statistics, 17(1):498–517, 2023.