Extreme events on graphs

Rita Maatouk[1,2], Thomas Opitz[2], Mike Pereira[1]

[1] STIM Center, Mines Paris - PSL
[2] BioSP/INRAE

March 2026

Statistiques au sommet de Rochebrune

1. Context

Which European rivers are more likely to flood *simultaneously*?

Understanding the spatial dependence of extreme river flows


L’objectif est que les crues
simultanées ne soient plus
une surprise, mais qu’elles
coulent comme de l’eau
sur le dos d’un canard.

PhD objective: Build a model to generate extreme river flows

New graphical model
Nodes : weather and
river flow variables

PhD objective: Build a model to generate extreme river flows

Weather variables
at different locations

→

New graphical model
Nodes : weather and
river flow variables

PhD objective: Build a model to generate extreme river flows

Weather variables at different locations → New graphical model Nodes : weather and river flow variables → Extreme river flow simulations at different stations

PhD objective: Build a model to generate extreme river flows

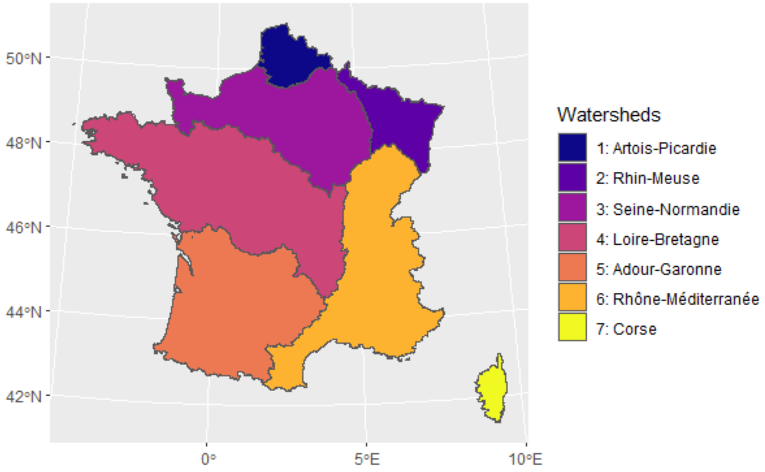
Weather variables at different locations → New graphical model Nodes : weather and river flow variables → Extreme river flow simulations at different stations

- × **Causality**
- ✓ **Explain how weather variables and river flows interact in extreme situations**

2. Approach

Studying extreme precipitation events across watersheds

Daily rainfall data for seven watersheds in France (1970–2024)



Methodological pipeline

Data

Rainfall
daily data
1970–2024
7 watersheds
$(X_1, \ldots, X_7)$

01

Methodological pipeline

Data

Rainfall
7 watersheds
$(X_1, \ldots, X_7)$
1970–2024

01

Marginal transformation

$X_i^* = \frac{1}{1-F_i(X_i)} \sim \text{Pareto}(1)$

02

Why this matters

- Each $X_i \sim F_i$ has its **own marginal distribution**
- $X_i^* = \frac{1}{1-F_i(X_i)} \sim \text{Pareto}(1)$
- $\Rightarrow$ All X_i^* share a **common reference distribution** $\Rightarrow$ variables are **comparable**

Methodological pipeline

Data
Rainfall
7 watersheds
$(X_1, \ldots, X_7)$
1970–2024
01

Marginal transformation
$X_i^* = \frac{1}{1 - F_i(X_i)} \sim \mathrm{Pareto}(1)$
02

Declustering
Extract independent extreme events $\mathbf{Y}$ based on a threshold u
03

Methodological pipeline

Data

Rainfall
7 watersheds
$(X_1, \ldots, X_7)$
1970–2024

01

Marginal transformation

$X_i^* = \frac{1}{1 - F_i(X_i)} \sim \text{Pareto}(1)$

02

Declustering

Extract independent extreme events **Y** based on a threshold u

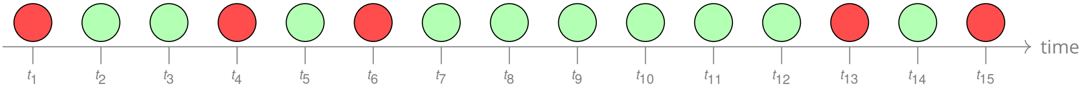
03

Goal of declustering

Retain **independent** extreme events - consecutive extremes may be part of the **same episode**

Declustering: extracting independent extreme events

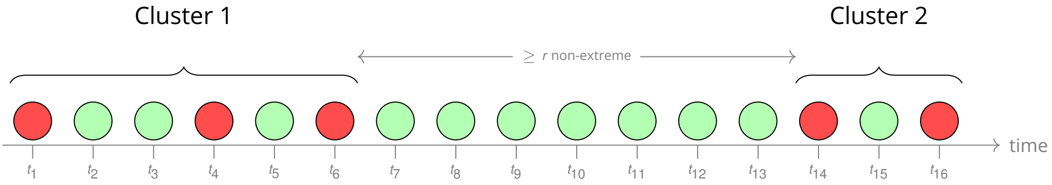
- Each observation = one row $(X_1^*, \ldots, X_7^*)$ in the dataset
- Threshold u typically chosen as a high quantile



● **Extreme:** $\exists\, i : X_i^* > u$; ● **Non-extreme:** $\forall\, i : X_i^* \leq u$

Declustering: extracting independent extreme events

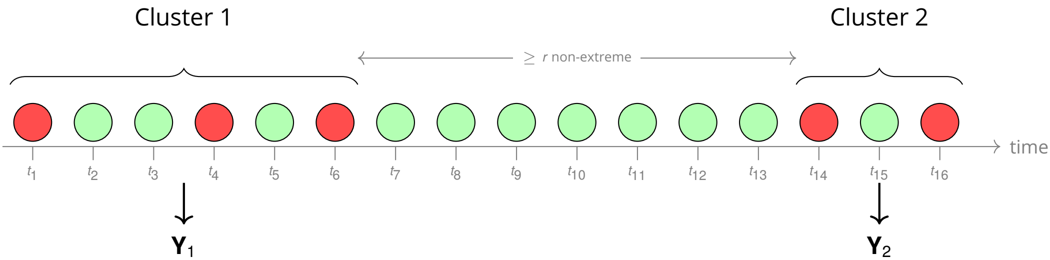
- We fix a constant r (here $r = 5$)
- Two extremes $\rightarrow$ **same cluster** if $< r$ non-extreme obs. between them
- Two clusters $\rightarrow$ **distinct** if $\geq r$ non-extreme obs. between them



● **Extreme:** $\exists\, i : X_i^* > u$; ● **Non-extreme:** $\forall\, i : X_i^* \leq u$

Declustering: extracting independent extreme events

- We fix a constant r (here $r = 5$)
- Two extremes $\rightarrow$ **same cluster** if $< r$ non-extreme obs. between them
- Two clusters $\rightarrow$ **distinct** if $\geq r$ non-extreme obs. between them



- **Extreme:** $\exists\, i : X_i^* > u$; - **Non-extreme:** $\forall\, i : X_i^* \leq u$
- $\mathbf{Y}^k = \left(\frac{\max_{t \in C_k} X_{t1}^*}{u}, \ldots, \frac{\max_{t \in C_k} X_{t7}^*}{u} \right) = (\mathbf{Y}_1^k, \ldots \mathbf{Y}_7^k) \in \mathbb{R}^7$
- $\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^n$ are assumed i.i.d. and follow a **MGPD** (Multivariate Generalized Pareto Distribution)

Methodological pipeline

Data

Rainfall
7 watersheds
$(X_1, \ldots, X_7)$
1970–2024

01

Marginal transformation

$X_i^* = \frac{1}{1-F_i(X_i)} \sim \text{Pareto}(1)$

02

Declustering

Extract independent extreme events $\mathbf{Y}$ based on a threshold u

03

04

Graphical model

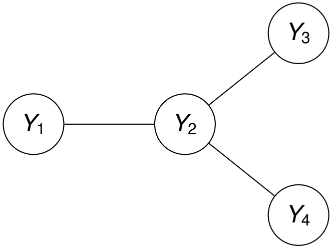
Fit Hüsler–Reiss on extremes $\mathbf{Y}$

Graphical model for multivariate extreme events: Hüsler–Reiss

- **Hüsler–Reiss:** Gaussian analogue for multivariate extremes
  Multivariate extreme events : $\mathbf{Y}^1, \dots, \mathbf{Y}^n$, with $\mathbf{Y}^k = \left(\mathbf{Y}_1^k, \dots \mathbf{Y}_d^k\right)$
  Variables : $\mathbf{Y}_1, \dots, \mathbf{Y}_d$

Graphical model for multivariate extreme events: Hüsler–Reiss

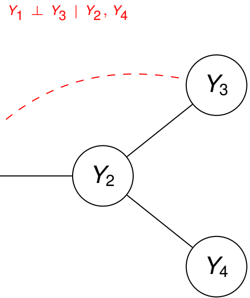
- **Hüsler–Reiss:** Gaussian analogue for multivariate extremes
  Multivariate extreme events : $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$, with $\mathbf{Y}^k = (\mathbf{Y}_1^k, \ldots \mathbf{Y}_d^k)$
  Variables : $\mathbf{Y}_1, \ldots, \mathbf{Y}_d$
- **Graphical model:** encodes dependencies via a graph $G = (V, E)$



Graphical model for multivariate extreme events: Hüsler–Reiss

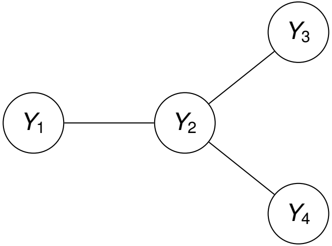
- **Hüsler–Reiss:** Gaussian analogue for multivariate extremes
  Multivariate extreme events : $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$, with $\mathbf{Y}^k = (\mathbf{Y}^k_1, \ldots \mathbf{Y}^k_d)$
  Variables : $\mathbf{Y}_1, \ldots, \mathbf{Y}_d$
- **Graphical model:** encodes dependencies via a graph $G = (V, E)$

$$\text{no edge } (i,j) \iff Y_i \perp Y_j \mid \mathbf{Y}_{-ij}$$



Graphical model for multivariate extreme events: Hüsler–Reiss

- **Hüsler–Reiss:** Gaussian analogue for multivariate extremes
  Multivariate extreme events : $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$, with $\mathbf{Y}^k = (\mathbf{Y}^k_1, \ldots \mathbf{Y}^k_d)$
  Variables : $\mathbf{Y}_1, \ldots, \mathbf{Y}_d$
- **Graphical model:** encodes dependencies via a graph $G = (V, E)$
- Characterised by precision matrix Q: $(i, j) \in E \iff q_{ij} \neq 0$



$$Q = \begin{pmatrix} q_{11} & q_{12} & 0 & 0 \\ q_{12} & q_{22} & q_{23} & q_{24} \\ 0 & q_{23} & q_{33} & 0 \\ 0 & q_{24} & 0 & q_{44} \end{pmatrix}$$

3. Model Validation

First validation criteria

Tail dependence - χ measure

$$\chi_{ij} = \mathbb{P}(X_j^* > u \mid X_i^* > u) \in [0,1]$$

$$= \mathbb{P}(Y_j > 1 \mid Y_i > 1) \in [0,1]$$

χ_{ij} measures the probability that variable *i* is extreme **given that** variable *j* is extreme.

Empirical $\hat{\chi}ij$ vs. χij computed from the fitted model

First validation criteria

Tail dependence - χ measure

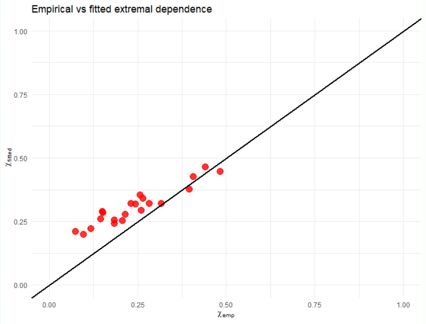
$$\chi_{ij} = \mathbb{P}(X_j^* > u \mid X_i^* > u) \in [0,1]$$

$$= \mathbb{P}(Y_j > 1 \mid Y_i > 1) \in [0,1]$$

χ_{ij} measures the probability that variable i is extreme **given that** variable j is extreme.

Empirical $\hat{\chi}ij$ vs. χij computed from the fitted model

χ measure - ✓ Good



Second validation criteria

Conditional normality (Hüsler–Reiss property)

Model property:
Under the Hüsler–Reiss model, for any conditioning variable m,

$$\mathbf{Z} = (\log Y_i - \log Y_m)_{i \neq m} \,\Big|\, Y_m > 1$$

$$\sim \mathcal{N}\left(-\tfrac{1}{2}\,\mathrm{diag}(\Sigma^{(m)}),\ \Sigma^{(m)}\right)$$

$\Rightarrow$ Define **standardized residuals**:

$$\mathbf{R} = (\Sigma^{(m)})^{-1/2}\left[Z + \tfrac{1}{2}\,\mathrm{diag}(\Sigma^{(m)})\right]$$

$$\sim \mathcal{N}(0, I)$$

Second validation criteria

Conditional normality (Hüsler–Reiss property)

Model property:
Under the Hüsler–Reiss model, for any conditioning variable *m*,

$$\mathbf{Z} = (\log Y_i - \log Y_m)_{i \neq m} \,\Big|\, Y_m > 1$$

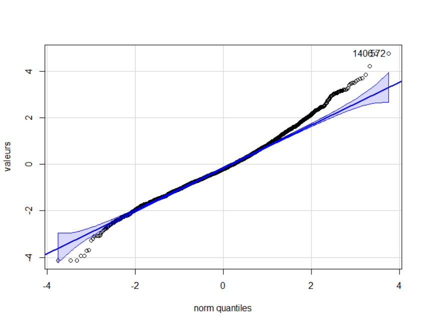
$$\sim \mathcal{N}\left(-\tfrac{1}{2}\,\mathrm{diag}(\Sigma^{(m)}),\ \Sigma^{(m)}\right)$$

$\Rightarrow$ Define **standardized residuals**:

$$\mathbf{R} = (\Sigma^{(m)})^{-1/2}\left[Z + \tfrac{1}{2}\,\mathrm{diag}(\Sigma^{(m)})\right]$$

$$\sim \mathcal{N}(0, I)$$

Q-Q plots - $\times$ Not satisfactory



Idea: if the model is correct, *R* should be Gaussian. Points on diagonal $=$ good fit

Splitting into geographic sub-groups

Why split?

- 7 watersheds spread across France $\rightarrow$ some pairs may be **geographically distant**
- Distant watersheds $\rightarrow$ **weaker dependence** $\rightarrow$ model struggles
- Idea: work on subsets of **spatially close** watersheds

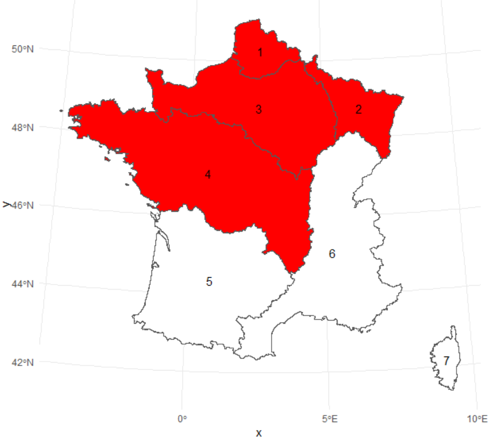
Splitting into geographic sub-groups

Why split?

- 7 watersheds spread across France $\rightarrow$ some pairs may be **geographically distant**
- Distant watersheds $\rightarrow$ **weaker dependence** $\rightarrow$ model struggles
- Idea: work on subsets of **spatially close** watersheds

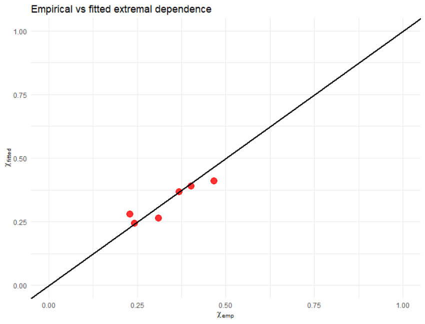
Two groups:
Group 1: watersheds $1, 2, 3, 4$
Group 2: watersheds $5, 6, 7$

Neighboring watersheds share similar rainfall patterns $\Rightarrow$ stronger tail dependence



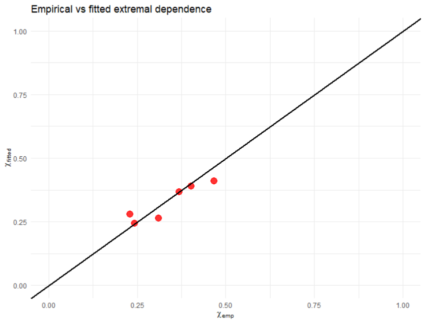
Validation: watersheds 1 – 4

χ measure - ✓

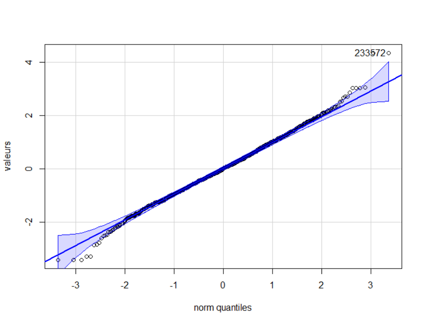


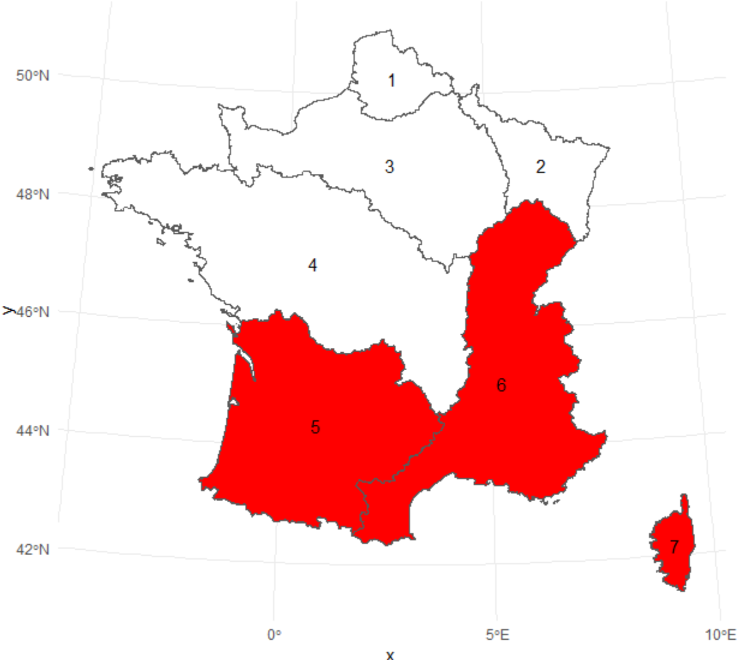
Validation: watersheds 1 – 4

χ measure - ✓



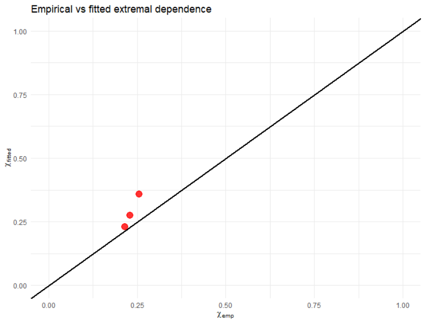
Q-Q plots - ✓




50°N
48°N
46°N
44°N
42°N
y
0°
5°E
10°E
x
1
2
3
4
5
6
7

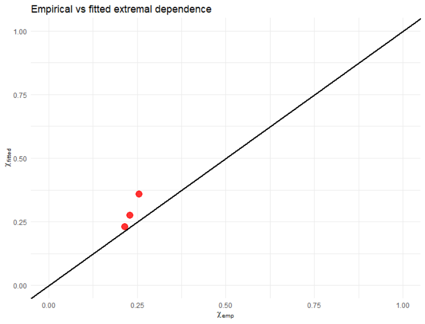
Validation: watersheds 5 – 7

χ measure - ✓

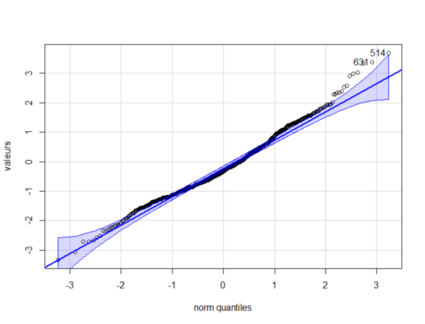


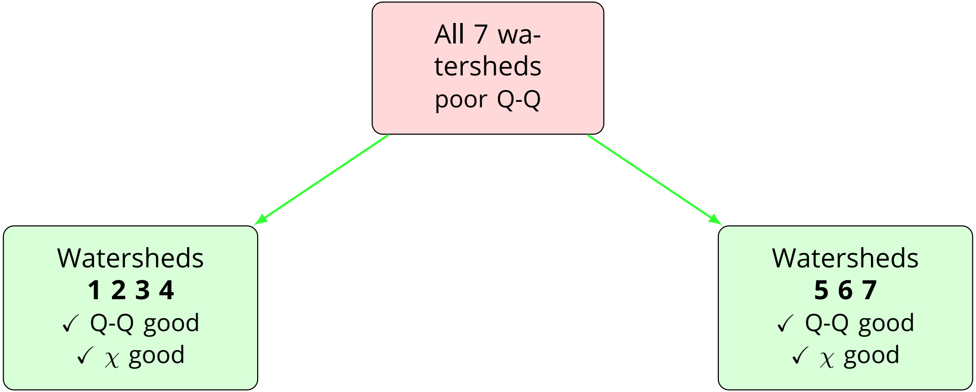
Validation: watersheds 5 – 7

χ measure - ✓



Q-Q plots - ✓





- The model did not perform well on all 7 watersheds simultaneously
- Likely because some pairs have **weak or no tail dependence**
- Splitting into **spatially close** groups $\rightarrow$ stronger dependence within each group $\rightarrow$ good model performance

4. Perspectives

Perspectives

Ongoing - Loire & Danube river flow daily data with more than 20 variables (stations): same conclusion - χ good but Q-Q not satisfactory when we work on all variables together

Perspectives

Ongoing - Loire & Danube river flow daily data with more than 20 variables (stations): same conclusion - χ good but Q-Q not satisfactory when we work on all variables together

Next step - automatic grouping: partition variables into homogeneous groups, fit one model per group

Perspectives

Ongoing - Loire & Danube river flow daily data with more than 20 variables (stations): same conclusion - χ good but Q-Q not satisfactory when we work on all variables together

Next step - automatic grouping: partition variables into homogeneous groups, fit one model per group

Conditional simulations: generate realistic extreme scenarios given one or more variables

Perspectives

Ongoing - Loire & Danube river flow daily data with more than 20 variables (stations): same conclusion - χ good but Q-Q not satisfactory when we work on all variables together

Next step - automatic grouping: partition variables into homogeneous groups, fit one model per group

Conditional simulations: generate realistic extreme scenarios given one or more variables

Richer models: include climate variables and river flows as nodes - extend to larger European networks

Thank You

Rita Maatouk
Extreme events on graphs
March 2026

References

1974 Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. The Annals of probability, 2(5) :792–804.

1975 Pickands III, J. (1975). Statistical inference using extreme order statistics. the Annals of Statistics, pages 119–131.

2006 Rootzén, H. and Tajvidi, N. (2006). Multivariate generalized pareto distributions. Bernoulli, 12(5) :917–930.

2006 Haan, L. and Ferreira, A. (2006). Extreme value theory : an introduction, volume 3. Springer.

2015 Asadi, P., Davison, A. and Engelke, S. (2015). Extremes on river networks

2020 Gardes, L. (2020). Théorie des valeurs extrêmes. Université de Strasbourg.

2022 de Fondeville, R. and Davison, A. C. (2022). Functional peaks-over-threshold analysis. Journal of the Royal Statistical Society Series B : Statistical Methodology, 84(4) :1392–1422.

2024 Engelke, S., Hentschel, M., Lalancette, M., and Röttger, F. (2024). Graphical models for multivariate extremes. arXiv preprint arXiv :2402.02187.

5. Annexe

Representative observation per cluster

- A cluster C_k contains several observations:

$$\mathbf{X}^*_{t_1},\ \mathbf{X}^*_{t_2},\ \dots,\ \mathbf{X}^*_{t_m}$$

with

$$\mathbf{X}^*_t = (X^*_{t1}, \dots, X^*_{td})$$

Representative observation per cluster

- A cluster C_k contains several observations:

$$\mathbf{X}^*_{t_1},\ \mathbf{X}^*_{t_2},\ \ldots,\ \mathbf{X}^*_{t_m}$$

  with

$$\mathbf{X}^*_t = (X^*_{t1}, \ldots, X^*_{td})$$

- We summarize the cluster by taking the maximum reached by each variable

$$\begin{aligned} \mathbf{X}^*_{t_1} &= (X^*_{t_1 1}, \ldots, X^*_{t_1 d}) \\ \mathbf{X}^*_{t_2} &= (X^*_{t_2 1}, \ldots, X^*_{t_2 d}) \\ &\vdots \\ \mathbf{X}^*_{t_m} &= (X^*_{t_m 1}, \ldots, X^*_{t_m d}) \end{aligned} \longrightarrow \mathbf{Y}_k = \frac{1}{u}\left(\max_{t \in C_k} X^*_{t1}, \ldots, \max_{t \in C_k} X^*_{td}\right)$$

As $u \to \infty$, the vectors $\mathbf{Y}_k$ follow a Multivariate Generalized Pareto Distribution (MGPD).

Graphical model for extreme events : Hüsler–Reiss HR(Γ)

Summary	Gaussian model	**Hüsler–Reiss model**
Density	$\propto \exp(-\frac{1}{2}\|\mathbf{x}-\mu\|_Q^2)$	$\propto \exp(-\frac{1}{2}\|\log(\mathbf{y})-\mu_Q\|_Q^2)$
Parameter matrix	Σ (covariance matrix)	Γ (extremal variogram)

where $\mu_Q = P\left(-\frac{\Gamma}{2}\right)\mathbf{1}$ with $P = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$

$\Gamma_{ij} = \mathrm{Var}\big(\log Y_i - \log Y_j \mid Y_j > 1\big)$

Graphical model for extreme events : Hüsler–Reiss HR(Γ)

Summary	Gaussian model	**Hüsler–Reiss model**
Density	$\propto \exp(-\frac{1}{2}\|\mathbf{x}-\mu\|_Q^2)$	$\propto \exp(-\frac{1}{2}\|\log(\mathbf{y})-\mu_Q\|_Q^2)$
Parameter matrix	Σ (covariance matrix)	Γ (extremal variogram)
Precision matrix Q	Σ^{-1}	$\left(P(-\frac{1}{2}\Gamma)P\right)^{+}$

where $\mu_Q = P\left(-\frac{\Gamma}{2}\right)\mathbf{1}$ with $P = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$

$\Gamma_{ij} = \mathrm{Var}\big(\log Y_i - \log Y_j \mid Y_j > 1\big)$

Graphical model for extreme events : Hüsler–Reiss HR(Γ)

Summary	Gaussian model	**Hüsler–Reiss model**
Density	$\propto \exp(-\frac{1}{2}\|\mathbf{x}-\mu\|_Q^2)$	$\propto \exp(-\frac{1}{2}\|\log(\mathbf{y})-\mu_Q\|_Q^2)$
Parameter matrix	Σ (covariance matrix)	Γ (extremal variogram)
Precision matrix Q	Σ^{-1}	$\left(P(-\frac{1}{2}\Gamma)P\right)^{+}$
Graphical structure	$Q_{ij} \neq 0$ if edge $(i,j) \in E$	$Q_{ij} \neq 0$ if edge $(i,j) \in E$

where $\mu_Q = P\left(-\frac{\Gamma}{2}\right)\mathbf{1}$ with $P = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$

$\Gamma_{ij} = \mathrm{Var}\big(\log Y_i - \log Y_j \mid Y_j > 1\big)$

Parameter estimation strategies

Fixed graph

Graph is imposed
Estimate model parameters

Parameter estimation strategies

Fixed graph	**Data-driven**
Graph is imposed Estimate model parameters	Estimate the graph Estimate model parameters

Parameter estimation strategies

Fixed graph

Graph is imposed
Estimate model parameters

Data-driven

Estimate the graph
Estimate model parameters

Parameter estimation strategies

Estimation of HR(Γ) model parameters

↓

Maximize surrogate log-likelihood

$$l(Q) = \log\det(Q) + \frac{1}{2}\mathrm{Tr}(\hat{\Gamma} Q)$$

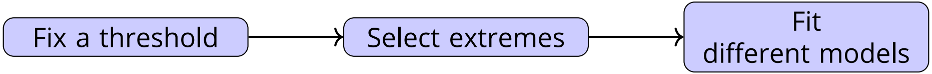
↙ ↘

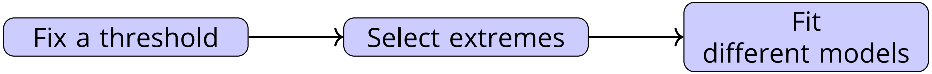
Approach with a fixed graph

1. Zero pattern of Q is imposed
2. Maximization over non-zero entries

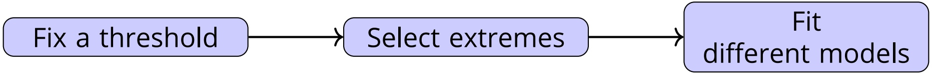
Data-driven approach

1. Zero pattern of Q is estimated
2. Maximization over non-zero entries


Fix a threshold
Select extremes
Fit
different models



→ Iterate over different model specifications and threshold levels



→ Iterate over different model specifications and threshold levels

→ Select and validate an appropriate model

Validation Criteria

1. The number of selected extreme events is reasonable (bias-variance tradeoff)

Validation Criteria

1. The number of selected extreme events is reasonable (bias-variance tradeoff)
2. Check the conditional multivariate normality suggested by the model:

$$(\log X_i^* - \log X_m^*)_{i \neq m} \big| \, X_m^* > u \sim \mathcal{N}\Big(-\tfrac{1}{2}\mathrm{diag}(\Sigma^{(m)}), \Sigma^{(m)} \Big)$$

   Transformed remaining variables | one variable is extreme $\approx$ Gaussian distribution

3. Another way to assess the reliability of the model is to compare the dependence observed in the data with that predicted by the model using the χ measure

Validation Criteria

1. The number of selected extreme events is reasonable (bias-variance tradeoff)
2. Check the conditional multivariate normality suggested by the model:

$$(\log X_i^* - \log X_m^*)_{i \neq m} \big| \; X_m^* > u \sim \mathcal{N}\Big(-\tfrac{1}{2}\mathrm{diag}(\Sigma^{(m)}), \Sigma^{(m)} \Big)$$

   Transformed remaining variables $\big|$ one variable is extreme $\approx$ Gaussian distribution

3. Another way to assess the reliability of the model is to compare the dependence observed in the data with that predicted by the model using the χ measure
4. The χ measure indicates whether two variables tend to experience extreme events simultaneously

Validation Criteria

1. The number of selected extreme events is reasonable (bias-variance tradeoff)
2. Check the conditional multivariate normality suggested by the model:

$$(\log X_i^* - \log X_m^*)_{i \neq m} \big| \; X_m^* > u \sim \mathcal{N}\Big(-\tfrac{1}{2}\mathrm{diag}(\Sigma^{(m)}), \Sigma^{(m)} \Big)$$

   Transformed remaining variables $\big|$ one variable is extreme $\approx$ Gaussian distribution

3. Another way to assess the reliability of the model is to compare the dependence observed in the data with that predicted by the model using the χ measure
4. The χ measure indicates whether two variables tend to experience extreme events simultaneously
5. For each pair of variables, we compare:
   - the empirical χ estimated from the data
   - the fitted χ predicted by the model