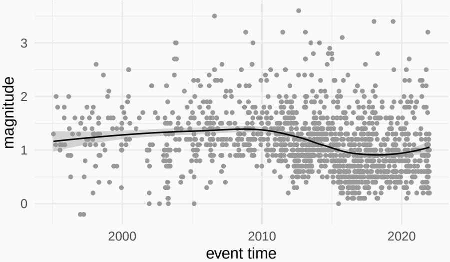
Automated threshold selection and associated inference uncertainty for univariate extremes

Conor Murphy
Jonathan Tawn, Zak Varty, Peter Atkinson, Ross Towe

Lancaster University, Shell

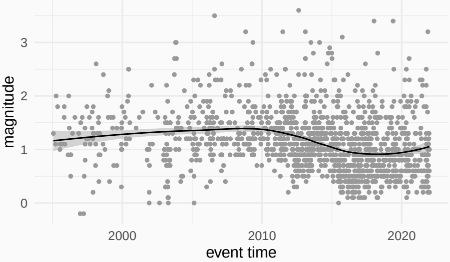
Motivation

- Production of oil/gas can cause earthquakes.
- Low magnitude events at shallow depths.
- Similar characteristics at CO_2 storage sites.



Challenges

- Partial censoring due to development of geophone network.
- Network too sparse/insensitive to detect low magnitude events.



Threshold stability property:

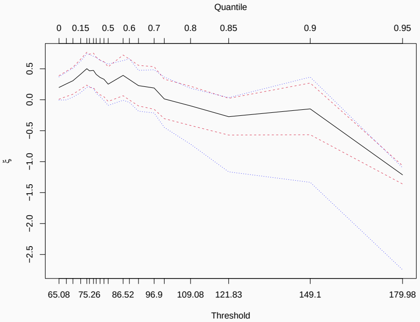
GPD(σ_u, ξ) above $u \Rightarrow$ for $v > u$,
$Y - v | Y > v \sim$ GPD($\sigma_u + \xi(v - u), \xi$)

Goal:

$\Rightarrow$ Forecast hazards under future extraction scenarios...

Why is threshold selection important?

- Parameter estimates
- Quantiles/Return levels
- Uncertainty

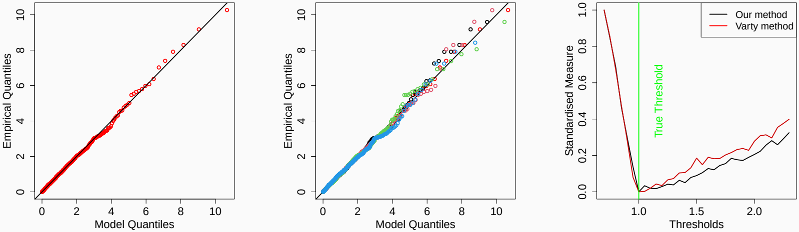


Challenge:

⇒ Bias-variance trade-off

Expected Quantile Discrepancy (EQD)

- Compares the deviation from the line of equality on a QQ-plot across replications for each threshold.
- Result: A set of metric values corresponding to each proposed threshold.



$$d_{(i)} = \frac{1}{m} \sum_{j=1}^{m} |M_{(i)}(p_j) - Q_{(i)}(p_j)| \qquad EQD = \frac{1}{k} \sum_{i=1}^{k} d_{(i)}.$$

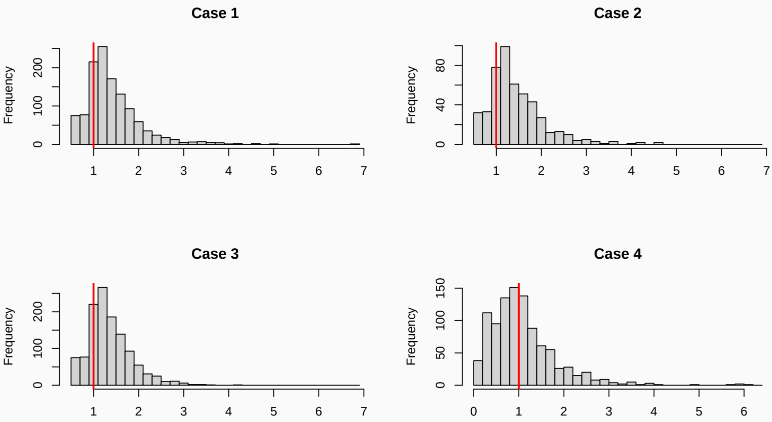
Automated methods

- Wadsworth (2016) utilises the asymptotic joint distribution of MLEs:
  - Consider $\hat{\xi}_i^* = \frac{\hat{\xi}_i - \hat{\xi}_{i+1}}{\nu_i}$ the standardised increments.
  - Main result: $(\hat{\xi}_1^*, \ldots, \hat{\xi}_{k-1}^*)^T \to \boldsymbol{Z}$ where $\boldsymbol{Z} \sim N_{k-1}(\boldsymbol{0}, \boldsymbol{1}_{k-1})$.
  - Changepoint model and likelihood ratio test if $\hat{\xi}_i^* \sim N(\beta, \gamma)$.

- Northrop et al. (2017) use leave-one-out cross-validation in a Bayesian framework:
  - Compare predictive ability above v for all u.
  - Average inferences over posterior distribution of parameters.

Simulation study

Examples of simulated datasets:



Results

	Our method	*Wadsworth**	*Northrop*
Case 1	**5.3**	41.3	52.7
Case 2	**5.5**	43.9	54.5
Case 3	**7.2**	13.7	42.7
Case 4	**10.2**	38.5	48.9

-> Our method achieves RMSEs **between 1.90 and 7.98 times smaller** than the Wadsworth (2016) method, always with **lower variance** and in 3 out of 4 cases, is **the least biased**.

Tables have been scaled by a factor of 100

*Results for Wadsworth are calculated only on the samples where a threshold was estimated. The method failed to estimate a threshold for 2%, 28%, 0.2%, 4% of the simulated datasets in Cases 1-4.

p	*Our method*	*Wadsworth**	*Northrop*	*Our method*	*Wadsworth**	*Northrop*
	Case 1			**Case 2**		
$1/n$	**5.8**	6.1	7.4	**6.2**	**6.2**	7.4
$1/10n$	**13.3**	14.7	20.8	**15.3**	15.8	26.4
$1/100n$	**26.2**	28.9	52.9	**32.2**	33.9	93.6
	Case 3			**Case 4**		
$1/n$	**2.0**	2.0	2.5	**7.0**	7.7	8.5
$1/10n$	**3.3**	3.4	4.8	**16.5**	19.4	26.6
$1/100n$	**4.9**	5.0	8.2	**33.3**	40.1	84.9

-> Our method achieves smallest RMSEs in all cases again!

Tables have been scaled by a factor of 10

*Results for Wadsworth are calculated only on the samples where a threshold was estimated.The method failed to estimate a threshold for 2%, 28%, 0.2%, 4% of the simulated datasets in Cases 1-4.

Uncertainty

⇒ Sample original dataset.

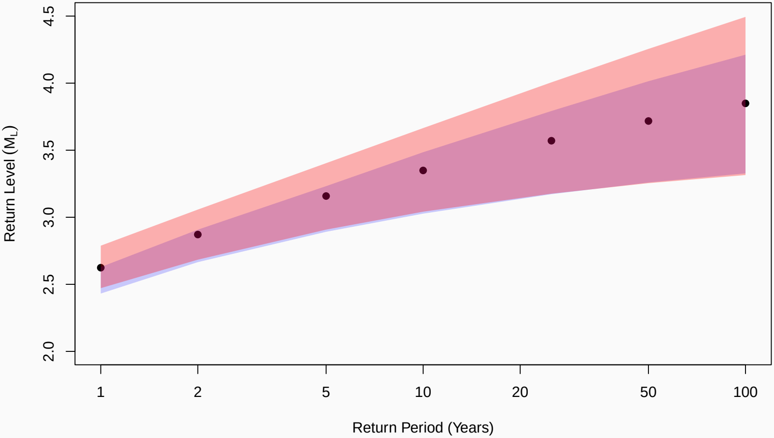
⇒ Estimate threshold for all samples.

⇒ Fit GPD to each sample of excesses of chosen thresholds.

⇒ Generate GPD samples using fitted parameters.

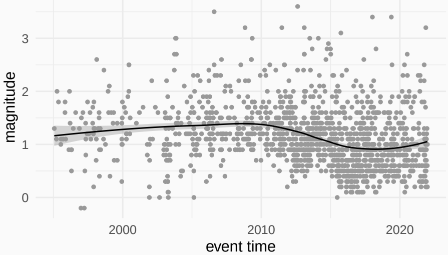
⇒ Refit GPD and obtain summary $s(\hat{\sigma}, \hat{\xi}, \hat{\lambda}_{\hat{u}})$.

Applied to Earthquake data



Previous work

- Varty et al. (2021) incorporated time-varying data quality into threshold.
- Behaviour comes from changing geophone network.
- Network does not change uniformly across space!

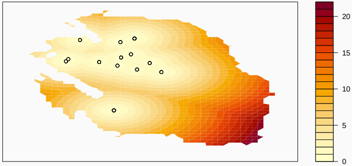


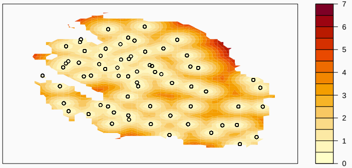
Spatio-temporal threshold

- Spatial variability also present.
- Relationship between distance from geophones and probability of detection.

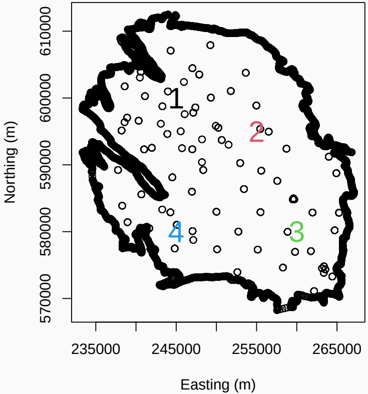
$$u(x,t) = \theta V_{\text{geo}}(x,t)$$

$$Y - u(x,t) | Y > u(x,t) \sim \text{GPD}(\sigma + \xi u(x,t), \xi)$$

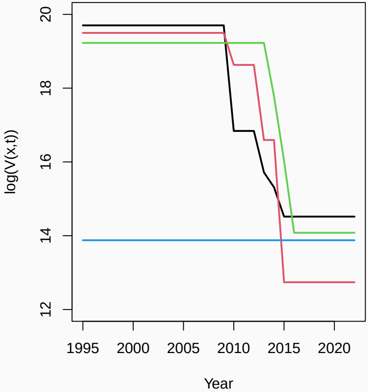




$V_{geo}(x,t)$

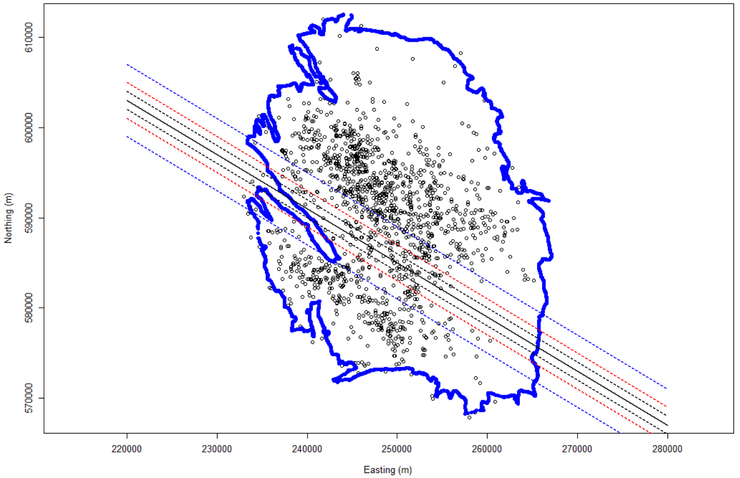


Northing (m)
570000
580000
590000
600000
610000
235000
245000
255000
265000
Easting (m)
1
2
3
4



log(V(x,t))
12
14
16
18
20
1995
2000
2005
2010
2015
2020
Year

Should the underlying parameters $Y \sim \text{GPD}(\sigma, \xi)$ vary spatially?



Likelihood ratio tests

Thresholds	Models	GPD(σ, ξ)	GPD(σ_R, ξ)
$u = 1.07$	GPD(σ_R, ξ)	0.000	NA
	GPD(σ_R, ξ_R)	0.000	**0.036** $\uparrow$
$u = 1.318$	GPD(σ_R, ξ)	0.158	NA
	GPD(σ_R, ξ_R)	0.357	0.797
$(u_U, u_L) = (1.2, 0.876)$	GPD(σ_R, ξ)	0.001	NA
	GPD(σ_R, ξ_R)	0.001	**0.064** $\uparrow$

$\Rightarrow$ **Evidence to suggest GPD scale parameter varies over region.**

Next steps:

- Compare above models using appropriate thresholds for all cases... How?

Further work

Combined threshold & model selection

1. Adjust threshold selection method for all desired models.
2. Transform to common margins and record d_{min}.
3. Compare d_{min} and select model which minimises overall EQD.

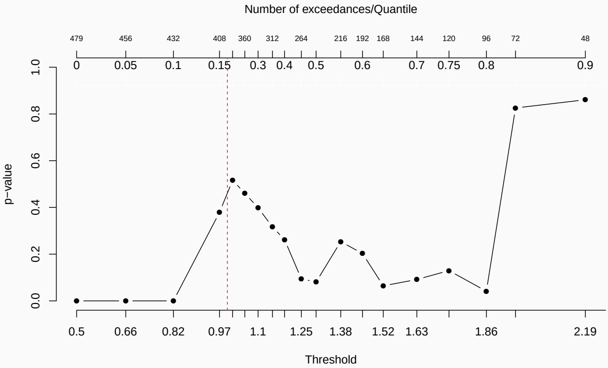
Spatio-temporal modelling

- Assess variability in GPD parameters with spatio-temporal threshold.
- Utilise relevant covariates for GPD parameters.
- Explore relationship with $V_{\text{geo}}(x, t)$.

Thanks for listening!

Developments from parameter stability plots

- Northrop and Coleman (2014) developed a multiple-threshold GPD model -> likelihood ratio and score tests to assess stability.



-> **Clearly automated methods are needed!**

Simulation Study

Simulated from two distributions:

$$F_1(x) = \begin{cases} \frac{x-0.5}{3}, & 0.5 \leq x \leq 1 \\ \frac{1}{6} + \frac{5}{6}\left[H(x-1; 0.5, 0.1)\right], & x > 1. \end{cases}$$

$$F_2(x) = \begin{cases} \int_0^x h(x; 0.5, 0.1)\mathbb{P}(B < x)\mathrm{d}x, & 0 \leq x \leq 1 \\ q + (1-q)\left[H(x-1; 0.5, 0.1)\right], & x > 1. \end{cases}$$

where $q = \int_0^1 h(x; 0.5, 0.1)\mathbb{P}(B < x)\mathrm{d}x$.

True quantiles from the simulated distributions can be calculated as follows:

$$x_p = 1 + \frac{\sigma_1}{\xi}\left[\left(\frac{6p}{5}\right)^{-\xi} - 1\right], \qquad y_p = 1 + \frac{\sigma_1}{\xi}\left[\left(\frac{p}{1-q}\right)^{-\xi} - 1\right].$$

Simulation study

Breakdown of RMSE:

- Bias and variance of threshold choice for GPD data.

	Our method			*Varty method*		
n	RMSE	Bias	Variance	RMSE	Bias	Variance
1000	**9.4**	4.7	0.7	10.7	5.0	0.9
10000	**13.2**	3.5	1.6	13.3	3.8	1.6
40000	**5.8**	2.7	0.2	8.1	3.3	0.5

- Bias and variance of quantile estimation for Gaussian data.

	Our method			*Varty method*		
n	RMSE	Bias	Variance	RMSE	Bias	Variance
1000	**72.8**	62.6	13.9	79.3	70.3	13.5
10000	**38.0**	25.2	8.1	42.0	30.5	8.3
40000	**23.6**	16.6	2.8	24.8	18.1	2.9

Table values have been scaled by a factor of 100

Results

	Our method			*Wadsworth**			*Northrop*		
	RMSE	Bias	Variance	RMSE	Bias	Variance	RMSE	Bias	Variance
Case 1	**5.3**	**3.4**	**0.2**	41.3	15.1	14.8	52.7	25.7	21.1
Case 2	**5.5**	**3.0**	**0.2**	43.9	18.8	15.8	54.5	26.9	22.5
Case 3	**7.2**	4.6	**0.3**	13.7	**3.9**	1.7	42.7	22.9	12.9
Case 4	**10.2**	**6.8**	**0.6**	38.5	7.2	14.3	48.9	15.0	21.7

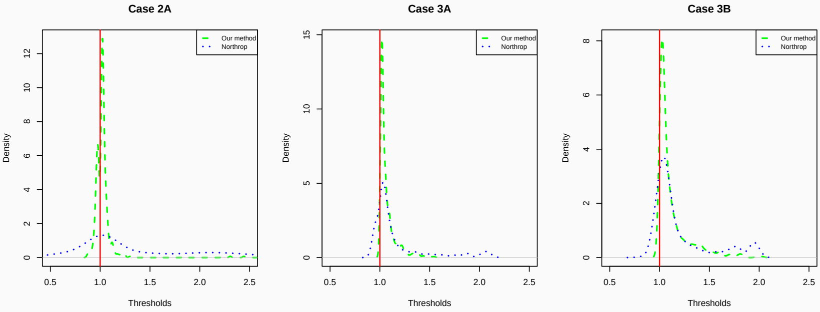
-> Our method achieves RMSEs **between 1.9 and 8 times smaller** than the Wadsworth (2016) method, always with **lower variance** and in 3 out of 4 cases, is **the least biased**.

Tables have been scaled by a factor of 100

*Results for Wadsworth are calculated only on the samples where a threshold was estimated. The method failed to estimate a threshold for 2%, 28%, 0.2%, 4% of the simulated datasets in Cases 1-4.

Results

Comparison in cases where Wadsworth (2016) broke down:

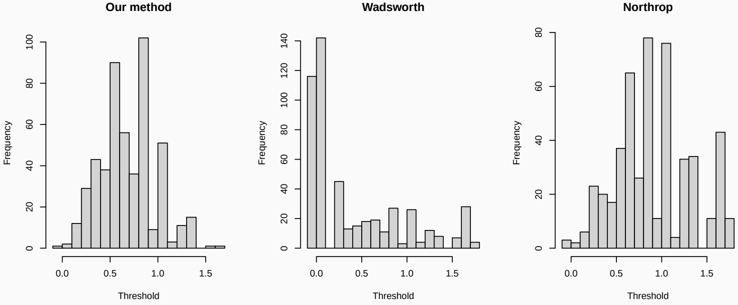


- Small sample of 120
- Same number of thresholds
- Case 3A: $\xi = -0.2$
- Case 3B: $\xi = -0.3$

-> **Our method achieves accurate results in all cases!**

Gaussian data

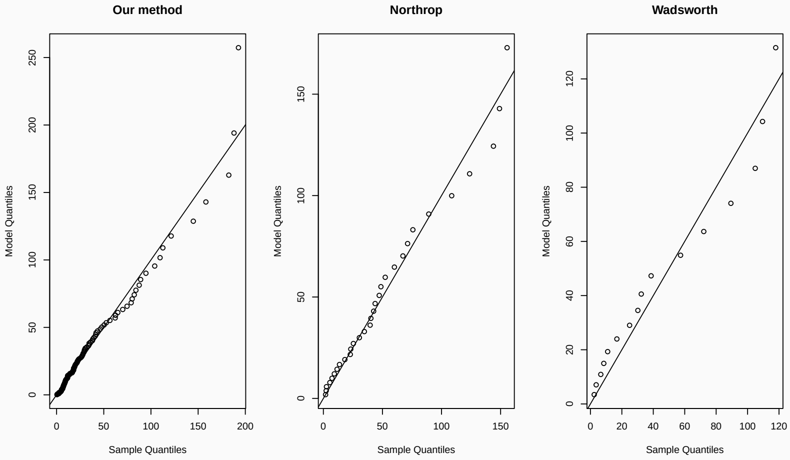
Gaussian Case			
p	*Our method*	*Wadsworth**	*Northrop*
$1/n$	**2.1**	2.5	**2.3**
$1/10n$	**4.3**	5.4	**4.6**
$1/100n$	**7.0**	9.0	7.7



Tables have been scaled by a factor of 10

*Results for Wadsworth are calculated only on the samples where a threshold was estimated. In this case, the method failed to obtain an estimate for 0.4% of the samples.

River Nidd



$u = 69.74$ $u = 109.08$ $u = 149.09$

Algorithm 3: Parameter uncertainty for unknown threshold

Given sample X of size n, obtain threshold choice, fit GPD and obtain point estimates of quantiles of interest.

Non-parametric Bootstrap: Input (X, n, m_2)

For $j = 1, \ldots, m_2$,

1. Obtain sample $\tilde{X}_j^{(b)}$ of size n by sampling n times with replacement from X.
2. Run threshold selection on $\tilde{X}_j^{(b)}$ and obtain threshold choice $\hat{u}_j$ and number of excesses $n_{\hat{u}_j}$, fit a GPD and obtain required parameters.

   Parametric Bootstrap Input $(\hat{\sigma}_{\hat{u}_j}, \hat{\xi}_j, \hat{\lambda}_{\hat{u}_j}, n_{\hat{u}_j}, m_1)$

   For $i = 1, \ldots, m_1$,

   (a) Simulate GPD sample of size $n_{\hat{u}_j}$ of excesses with parameters $(\hat{\sigma}_{\hat{u}_j}, \hat{\xi}_j)$.

   (b) Obtain parameter estimates $(\hat{\sigma}_{(j,i)}, \hat{\xi}_{(j,i)})$ for i^{th} sample and estimate any quantiles of interest.

Uncertainty

Goal: Incorporate parameter and threshold uncertainty into our inference.

CIs for the 10-year level:

River Nidd dataset:

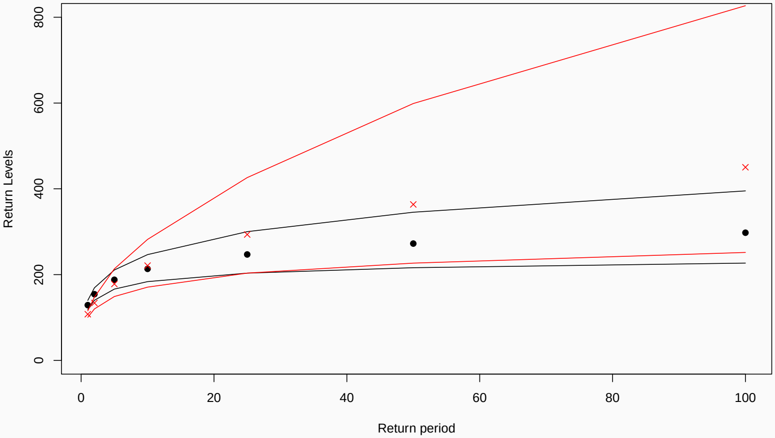
A1: $(349.9, 2214.0)$ *A3*: $(279.3, 2225.4)$

Ratio(95%) = 1.044.

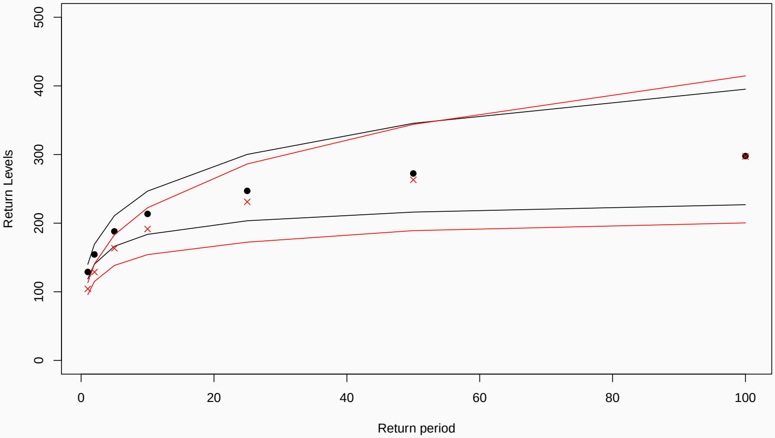
Simulated Gaussian data:

Ratio(50%) = 1.584, Ratio(90%) = 1.599.

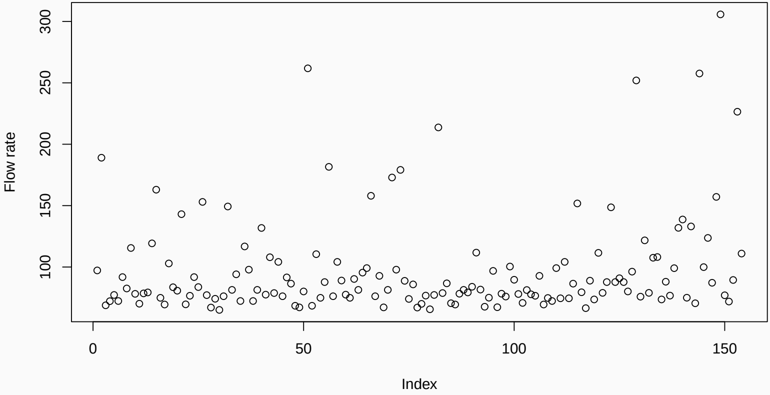
Uncertainty



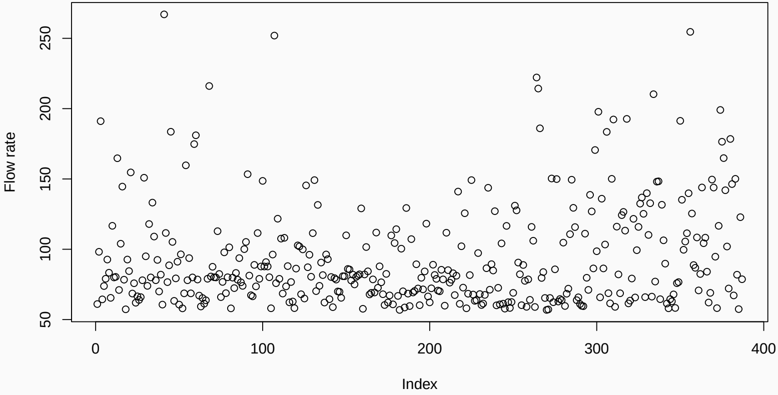
Uncertainty



River Nidd dataset



River Nidd dataset



Further work

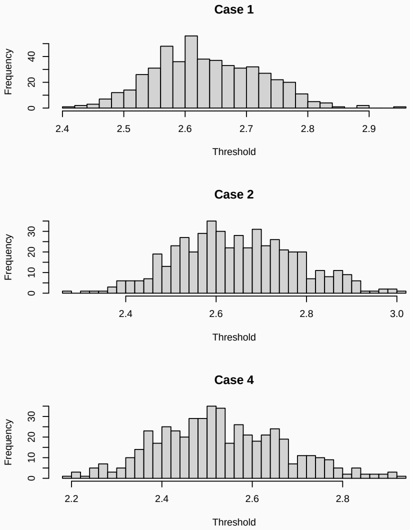
Compare against other existing methods in fixed threshold selection.

Danielsson et al. (2019):

- Quantile-driven approach.
- Maximum distance between empirical and model quantiles.

Applied to River Nidd dataset:

$\Rightarrow$ $u = 189.02$.



References

Danielsson, J., Ergun, L., de Haan, L., and de Vries, C. G. (2019). Tail Index Estimation: Quantile-Driven Threshold Selection. Staff Working Papers 19-28, Bank of Canada.

Northrop, P. J., 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.

Northrop, P. J. and Coleman, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes*, 17(2):289–303.

Varty, Z., Tawn, J. A., Atkinson, P. M., and Bierman, S. (2021). Inference for extreme earthquake magnitudes accounting for a time-varying measurement process. *arXiv preprint arXiv:2102.00884*.

Wadsworth, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics*, 58(1):116–126.