Threshold estimation in marginal modelling of spatially-dependent non-stationary extremes

Philip Jonathan
Shell Technology Centre Thornton, Chester
philip.jonathan@shell.com

Paul Northrop
University College London
paul@stats.ucl.ac.uk

Environmental Extremes
Royal Statistical Society
April 2011

Outline

- Motivation and application.
- Threshold modelling using quantile regression.
- Implications of QR threshold for PP model parameterisation.
- Adjusting for spatial dependence.
- Results for application.
- Initial theoretical & simulation studies.
- Conclusions.

Motivation: Rational design of marine structures

- **Covariate** effects:
  - Location, direction, season, ...
  - Multiple covariates in practice.
- **Cluster** dependence:
  - e.g. storms independent, observed (many times) at many locations.
  - e.g. dependent occurrences in time.
- **Scale** effects:
  - Modelling H_S^2 gives different estimates cf. modelling H_S.
- **Threshold** estimation; **parameter** estimation.
- **Measurement** issues:
  - Field measurement uncertainty greatest for extreme values.
  - Hindcast data are simulations based on pragmatic physics, calibrated to historical observation.

Motivation: Rational design of marine structures

- **Multivariate** extremes:
  - Waves, winds, currents, ...
  - Componentwise maxima $\Leftrightarrow$ max-stability $\Leftrightarrow$ regular variation:
    - Assumes **all** components extreme.
    - $\Rightarrow$ Perfect independence or asymptotic dependence **only**.
  - Extremal dependence:
    - Assumes regular variation of joint survivor function.
    - $\Rightarrow$ Asymptotic dependence, asymptotic independence (with +ve, -ve association).
  - Conditional extremes:
    - Assumes, given one variable being extreme, convergence of distribution of remaining variables.
    - Allows some variables not to be extreme.
  - Inference:
    - ... *a huge gap in the theory and practice of multivariate extremes* ... (Beirlant et al. 2004)

Aim: **Useful** models with rigourous assessment of model performance, **especially** in extreme quantiles.

Motivation: Good threshold estimation critical

- Considerable **empirical** evidence from applications that careful estimation of threshold including covariate effects important for satisfactory modelling.
- Often reasonable to assume some (or all) extreme value parameters are **independent** of (some or all) covariates following good thresholding, greatly simplifying model form.
- Quantile thresholds as functions of covariate(s) produce near **constant rates** of threshold exceedence (appealing from design perspective).

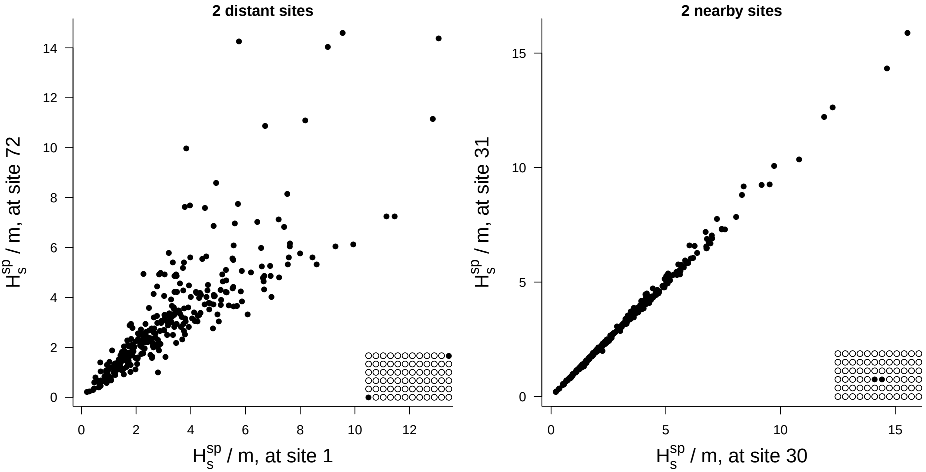
Application: Marginal estimation of extreme H_S^{SP}

- Data from hindcast of Y storm peak significant wave height (in metres) in the Gulf of Mexico.
  - **Wave height**, h: trough to the crest of the wave.
  - **Significant wave height**, H_S: the average of the largest 1/3 wave heights h in given period (usually 3 hours).
  - **Storm peak** H_S^{SP}: largest value of H_S from a storm (cf. declustering).
- 6×12 grid of 72 sites ($\approx$ 14 km apart).
- Sep 1900 to Sep 2005 : 315 storms in total.
- Average of 3 observations (storms) per year, at each site.

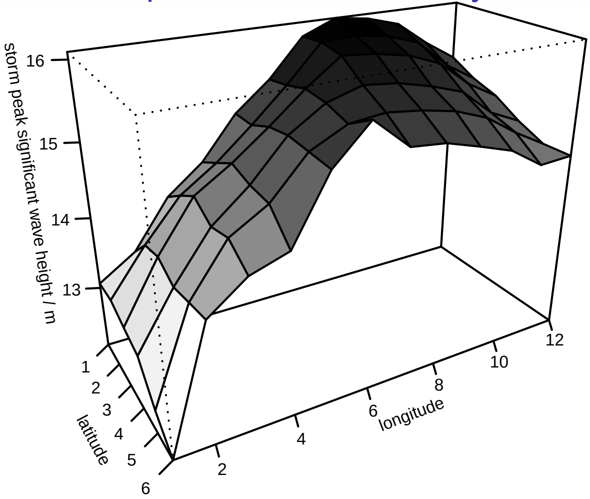
Aim: Quantify the extremal behaviour of Y at each site, making appropriate adjustment for spatial dependence.

Typical hurricane event in Gulf of Mexico

Spatial dependence



Spatial non-stationarity



- From single event ?

Modelling approach

- Spatial non-stationarity:
  - Model threshold as Legendre polynomial in longitude and latitude using **quantile regression**.
  - Model spatial variation of PP parameters as Legendre polynomials in longitude and latitude.
  - Lots of other suitable bases: splines, random fields ...
- Spatial dependence:
  - Estimate parameters assuming **conditional independence** of responses given covariate values.
  - **Adjust standard errors** etc. for spatial dependence.
- Estimate extreme quantiles.

Extreme value regression model

Conditional on covariates $\mathbf{x}_{ij}$ exceedances over a high threshold $u(\mathbf{x}_{ij})$ follow a 2-dimensional **non-homogeneous Poisson process**.

If responses $Y_{ij}, i = 1, \ldots, 72$ (**space**), $j = 1, \ldots, 315$ (**storms**) are **conditionally independent**:

$$L(\theta) = \prod_{j=1}^{315} \prod_{i=1}^{72} \exp\left\{ -\frac{1}{\lambda} \left[1 + \xi(\mathbf{x}_{ij}) \left(\frac{u(\mathbf{x}_{ij}) - \mu(\mathbf{x}_{ij})}{\sigma(\mathbf{x}_{ij})} \right) \right]_+^{-1/\xi(\mathbf{x}_{ij})} \right\}$$
$$\times \prod_{j=1}^{315} \prod_{i: y_{ij} > u(\mathbf{x}_{ij})} \frac{1}{\sigma(\mathbf{x}_{ij})} \left[1 + \xi(\mathbf{x}_{ij}) \left(\frac{y_{ij} - \mu(\mathbf{x}_{ij})}{\sigma(\mathbf{x}_{ij})} \right) \right]_+^{-1/\xi(\mathbf{x}_{ij}) - 1}.$$

λ : mean number of observations per year.

$\mu(\mathbf{x}_{ij}), \sigma(\mathbf{x}_{ij}), \xi(\mathbf{x}_{ij})$: PP parameters at $\mathbf{x}_{ij}$.

θ : vector of all model parameters.

Covariate-dependent thresholds

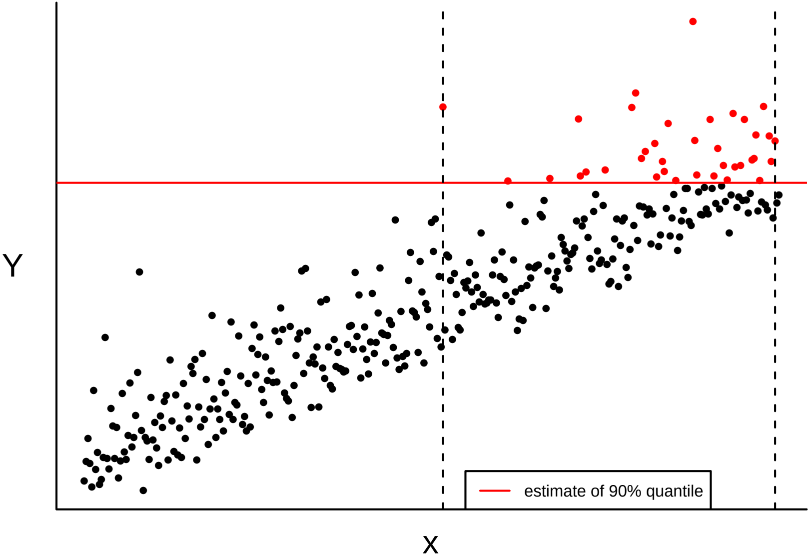
Arguments for:

- Asymptotic justification for EV regression model : the threshold $u(\mathbf{x}_{ij})$ needs to be high for each $\mathbf{x}_{ij}$.
- Design : spread exceedances across a wide range of covariate values.

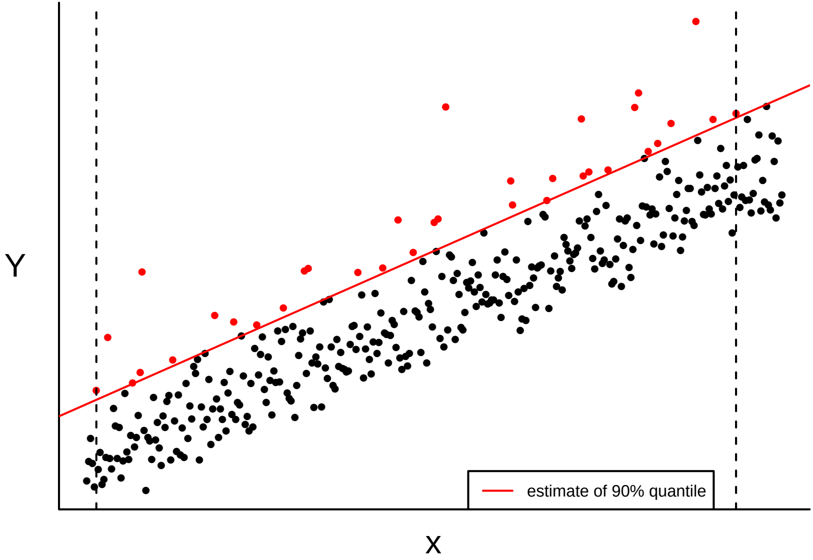
Set $u(\mathbf{x}_{ij})$ so that $P(Y > u(\mathbf{x}_{ij}))$, is approx. constant for all $\mathbf{x}_{ij}$.

- Set $u(\mathbf{x}_{ij})$ by trial-and-error or by discretising $\mathbf{x}_{ij}$, e.g. different threshold for different locations, months etc.
- **Quantile regression (QR)** : model quantiles of a response Y as a function of covariates.

Constant threshold



Quantile regression



Simple quantile regression in outline

- Data $\{x_i, y_i\}_{i=1}^n$
- τ^{th} conditional quantile function $Q_y(\tau|x) = x\phi(\tau)$ estimated by solving:

$$\min_{\phi} \sum_{i=1}^{n} \rho_\tau(y_i - x_i\phi)$$

where $\rho_\tau(r) = \tau r - r\, I(r < 0)$, or (with $r_i = r_i(\phi) = y_i - x_i\phi$):

$$\min_{\phi} \{\tau \sum_{r_i \geq 0}^{n} |r_i| + (1-\tau) \sum_{r_i < 0}^{n} |r_i|\}$$

- As a linear program:

$$\min_{\phi,u,v} \{\tau 1_n^T u + (1-\tau) 1_n^T v \,|\, x\phi + u - v = y\}$$

where $\{u_i\}$ and $\{v_i\}$ are **slack** variables corresponding to (absolute values of) positive and negative residuals.

Model parameterisation

Let $p(\mathbf{x}_{ij}) = P(Y_{ij} > u(\mathbf{x}_{ij}))$. Then, if $\xi(\mathbf{x}_{ij}) = \xi$ is constant,

$$p(\mathbf{x}_{ij}) \approx \frac{1}{\lambda}\left[1 + \xi\left(\frac{u(\mathbf{x}_{ij}) - \mu(\mathbf{x}_{ij})}{\sigma(\mathbf{x}_{ij})}\right)\right]^{-1/\xi}.$$

If $p(\mathbf{x}_{ij}) = p$ is constant then:

$$u(\mathbf{x}_{ij}) = \mu(\mathbf{x}_{ij}) + c\,\sigma(\mathbf{x}_{ij}), \text{ for some constant } c.$$

The form of $u(\mathbf{x}_{ij})$ is determined by the extreme value model:

- if $\mu(\mathbf{x}_{ij})$ and/or $\sigma(\mathbf{x}_{ij})$ are linear in $\mathbf{x}_{ij}$: linear QR.
- if $\log(\mu(\mathbf{x}_{ij}))$ and/or $\log(\sigma(\mathbf{x}_{ij}))$ is linear in $\mathbf{x}_{ij}$: non-linear QR.

Adjustment for spatial dependence

- **Independence** log-likelihood:

$$\ell_{IND}(\theta) \;=\; \underset{\text{(storms)}}{\sum_{j=1}^{k}} \; \underset{\text{(space)}}{\sum_{i=1}^{72}} \log f_{ij}(y_{ij};\theta) = \sum_{j=1}^{k} \ell_j(\theta)$$

- If **correct** model specification:

$$\widehat{\theta} \to N(\theta_0, I^{-1})$$

- If **model mis-specified**, in regular problems, as $k \to \infty$:

$$\widehat{\theta} \to N(\theta_0, I^{-1}\, V\, I^{-1})$$

- I = Expected information: $-\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}\,\ell_{IND}(\theta_0)\right)$.
- $V = \mathrm{var}\left(\frac{\partial}{\partial\theta}\,\ell_{IND}(\theta)\right)$.

Adjustment of $\ell_{IND}(\theta)$

- Idea: Adjust $\ell_{IND}(\theta)$ to have correct curvature near $\widehat{\theta}$ using sandwich estimate.

$$\begin{aligned}\ell_{ADJ}(\theta) &= \ell_{IND}(\widehat{\theta}) \\ &+ \frac{(\theta-\widehat{\theta})'\left(-\widehat{I}^{-1}\,\widehat{V}\,\widehat{I}^{-1}\right)^{-1}(\theta-\widehat{\theta})}{(\theta-\widehat{\theta})'(-\widehat{I})(\theta-\widehat{\theta})}\left(\ell_{IND}(\theta)-\ell_{IND}(\widehat{\theta})\right)\end{aligned}$$

- Estimate I by observed information at $\widehat{\theta}$.
- Estimate V by $\sum_{j=1}^{k} U_j^2\left(\widehat{\theta}\right)$, $U_j(\theta) = \frac{\partial \ell_j(\theta)}{\partial\theta}$.
- **Vertical** adjustment preserves asymptotic distribution of likelihood ratio statistic.
- See Davison (2003), Chandler and Bate (2007).

Summary of modelling of wave height data

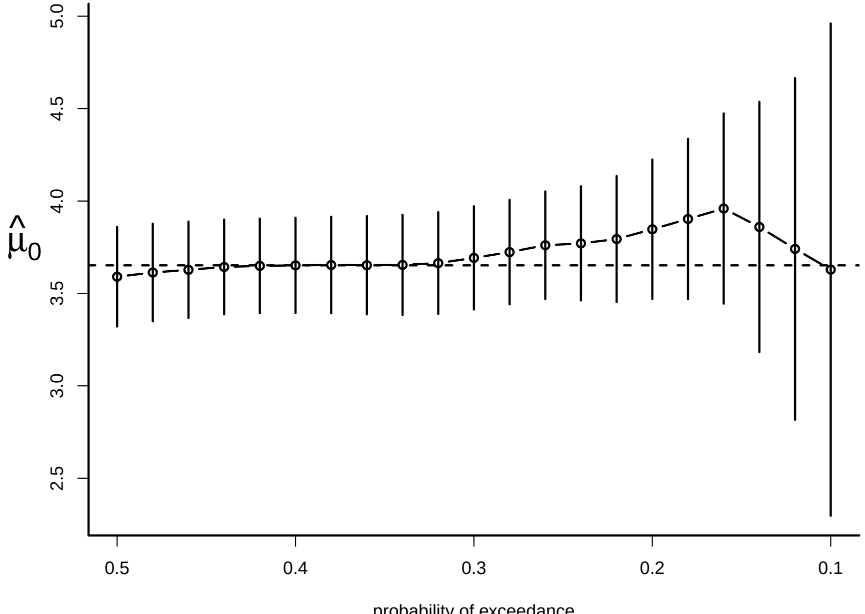
- Threshold selection:
  - Choice of p: look for stability in parameter estimates.
  - Based on μ (and u) quadratic in longtiude and latitude, σ and ξ constant ...
- Spatial model:

$$\mu = \sum_{i=0}^{q_x} \sum_{j=0}^{q_y} \mu_{i+jq_y} \phi_{xi}(l_x) \phi_{yj}(l_y)$$

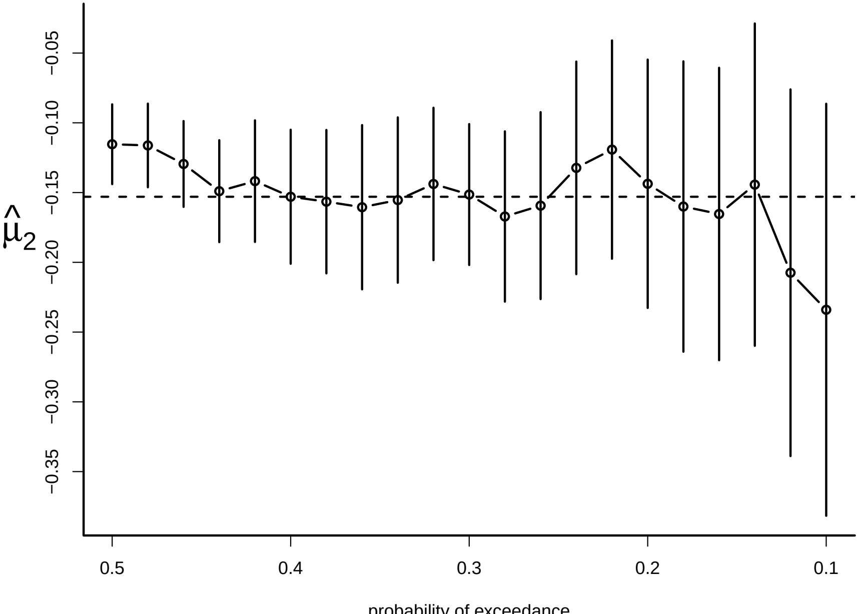
where:

- $\phi_{\cdot 0}(\cdot) = 1$.
- $\phi_{x1}(l_x) = \frac{1}{5.5}(l_x - 6.5)$, $\phi_{y1}(l_y) = \frac{1}{2.5}(l_y - 3.5)$.
- $\phi_{\cdot 2}(\cdot) = \frac{1}{2}(3\phi_1^2(\cdot) - 1)$, for $l_x, l_y \in [-1, 1]$.

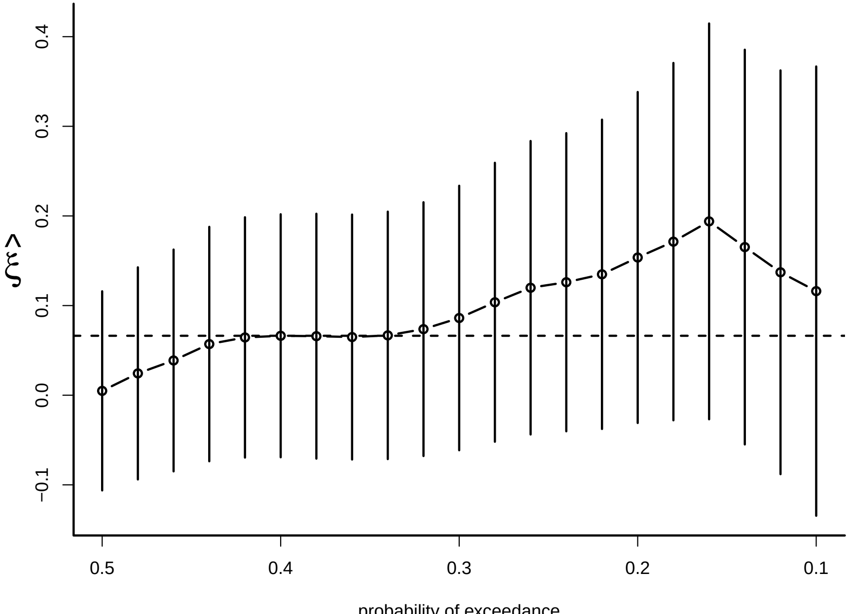
Threshold selection : μ intercept



Threshold selection : μ coefficient of latitude



Threshold selection : ξ

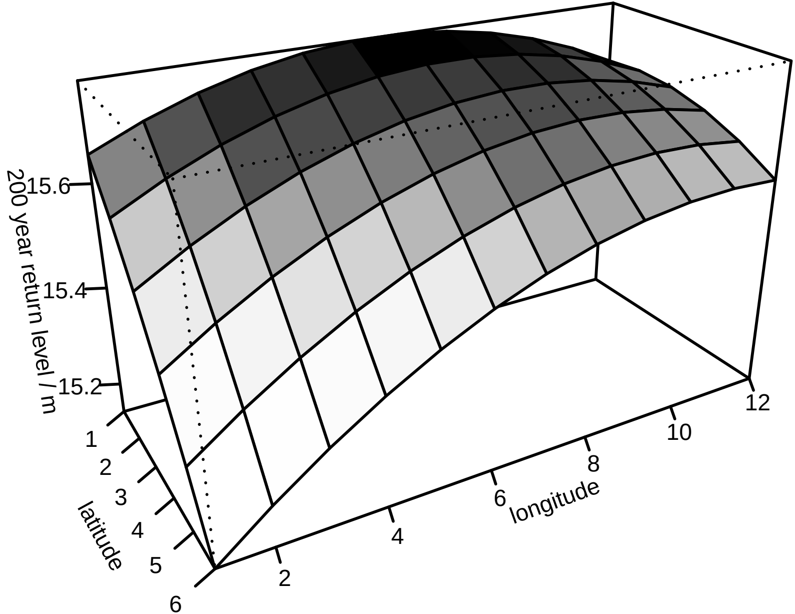


Summary of modelling of wave height data

- Choice of p: look for stability in parameter estimates.
  Use $p = 0.4$.
- $\widehat{\xi} = 0.07$, with 95% confidence interval $(-0.05, 0.22)$.
- Estimated 200 year return level at (long=7, lat=1) is 15.8m with 95% confidence interval $(12.9, 22.3)$m.

- Close agreement between parameter estimates for threshold u and point process mean μ.

Marginal 200 year return levels



Toy study 1

Data-generating process: for covariate values $x_1, \ldots, x_n$:

$$Y_i \mid X = x_i \overset{\text{indep}}{\sim} GEV(\mu_0 + \mu_1 x_i, \sigma, \xi).$$

Set threshold:

$$u(x) = u_0 + u_1 x.$$

For each u_1, set u_0 such that the expected proportion of exceedances is kept constant at p.

- Calculate Fisher expected information for $(\mu_0, \mu_1, \sigma, \xi)$.
- Invert to find asymptotic V-C of MLEs $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{\sigma}, \widehat{\xi}$ and hence $\text{var}(\widehat{\mu}_1)$.
- Find the value of u_1 that minimises $\text{var}(\widehat{\mu}_1)$.

Findings of *Toy* study 1

Let $\tilde{u}_1$ be the value of u_1 that minimises $\text{var}(\widehat{\mu}_1)$.

- If covariate values $x_1, \ldots, x_n$ are symmetrically distributed then: $\tilde{u}_1 = \mu_1$ (quantile regression).
- If $x_1, \ldots, x_n$ are positive (negative) skew then $\tilde{u}_1 < \mu_1$ ($\tilde{u}_1 > \mu_1$).

. . . but the loss in efficiency from using $\tilde{u}_1 = \mu_1$ appears to be small.

Simulation study 2

- 30 years of daily data on a spatial grid.
- Spatial dependence : mimics that of wave height data.
- **Temporal** dependence : moving maxima : extremal index 1/2 (**no** declustering)
- Spatial variation: location μ linear in longitude and latitude.

- ξ: $-0.2, 0.1, 0.4, 0.7$.
- Thresholds: 90th, 95th, 99th percentiles.
- SE adjustment: data from distinct years are independent.
- Simulations with no covariate effects and/or no spatial dependence for comparison.

Findings of simulation study 2

- Estimates of regression effects from QR and PP models are very close : both estimate extreme quantiles from the same data.
- Uncertainties in covariate effects of threshold are negligible compared to the uncertainty in the **choice** of threshold level.
- To a large extent fitting the PP model accounts for uncertainty in the covariate effects at the level of the threshold.
- Slight underestimation of standard errors : uncertainty in threshold ignored.

Conclusions

Quantile regression:

- An intuitive and effective strategy to set thresholds for non-stationary EV models.
- Works well in initial applications.
- Supported by initial theoretical and simulation studies.

Ideas:

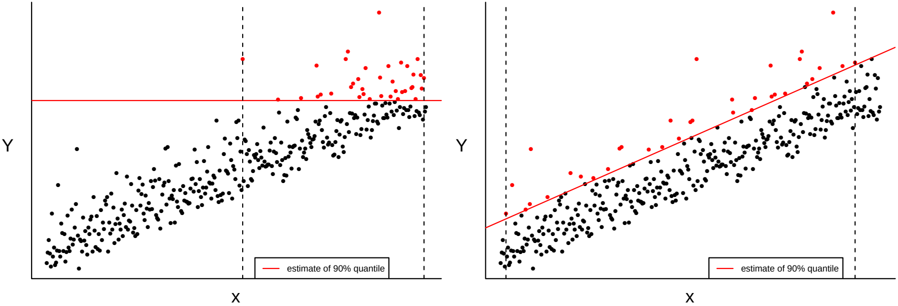
- Kyselý, J., et al. (2010) use quantile regression to set a time-dependent threshold for peaks-over-threshold GP modelling of data simulated from a climate model.
- Simultaneous threshold and PP model would avoid iteration (mixed-integer optimisation; see Beirlant et al. 2004).

References

Chandler, R. E. and Bate, S. B. (2007) Inference for clustered data using the independence loglikelihood. *Biometrika* **94 (1)**, 167–183.

Kyselý, J., Picek, J. and Beranová, R. (2010) Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold *Global and Planetary Change*, **72**, 55-68.

Northop, P. J. and Jonathan, P. Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights. Accepted for *Environmetrics*.



Thank you for your attention.