

Extreme Value Statistics and Robust Filtering for Hydrological Data

WBS Herbstseminar 2014



Extreme Value Statistics and Robust Filtering for Hydrological Data

WBS Herbstseminar 2014

Bernhard Spangl¹

Peter Ruckdeschel^{2,3}

¹ **BOKU – Univ. of Natural Resources and Life Sciences, Vienna**,
Dept. of Landscape, Spatial and Infrastructure Sciences
Inst. of Applied Statistics and Computing

Peter-Jordan-Straße. 82, 1190 **Vienna**, Austria

² **Fraunhofer ITWM**, Dept. of Financial Mathematics,
Fraunhofer-Platz 1, 67663 **Kaiserslautern**, Germany

³ **TU Kaiserslautern**, Dept. of Mathematics,
Erwin-Schrödinger-Straße, Geb 48, 67663 **Kaiserslautern**, Germany

Bernhard.Spangl@boku.ac.at

Wien, Nov. 06, 2014

The Need for Robustness in Extreme Value Theory: An Illustrative Example

- want to estimate an extreme (here 99.5%) quantile
- ideal data: 1000 obs. from $\exp(\mathcal{N}(\mu = 3, \sigma = 2))$
- true value in this example: 3470*
- contamination: modify first 7 observations to $\sim 10^7$
- naïve estimation by empirical quantile:
2960 (ideal), but 8910000 (contaminated)
- parametric (Max-Likelihood) estimation:
3390 (ideal), but 7580 (contaminated)
- robust estimation (by rmx-procedure):
3440 (ideal), but 3710 (contaminated)

*: all numbers rounded to 3 significant digits

The Need for Robustness in Extreme Value Theory: An Illustrative Example

- want to estimate an extreme (here 99.5%) quantile
- ideal data: 1000 obs. from $\exp(\mathcal{N}(\mu = 3, \sigma = 2))$
- true value in this example: 3470*
- contamination: modify first 7 observations to $\sim 10^7$
- naïve estimation by empirical quantile:
2960 (ideal), but 8910000 (contaminated)
- parametric (Max-Likelihood) estimation:
3390 (ideal), but 7580 (contaminated)
- robust estimation (by rmx-procedure):
3440 (ideal), but 3710 (contaminated)

*: all numbers rounded to 3 significant digits

The Need for Robustness in Extreme Value Theory: An Illustrative Example

- want to estimate an extreme (here 99.5%) quantile
- ideal data: 1000 obs. from $\exp(\mathcal{N}(\mu = 3, \sigma = 2))$
- true value in this example: 3470*
- contamination: modify first 7 observations to $\sim 10^7$
- naïve estimation by empirical quantile:
2960 (ideal), but 8910000 (contaminated)
- parametric (Max-Likelihood) estimation:
3390 (ideal), but 7580 (contaminated)
- robust estimation (by rmx-procedure):
3440 (ideal), but 3710 (contaminated)

*: all numbers rounded to 3 significant digits

The Need for Robustness in Extreme Value Theory: An Illustrative Example

- want to estimate an extreme (here 99.5%) quantile
- ideal data: 1000 obs. from $\exp(\mathcal{N}(\mu = 3, \sigma = 2))$
- true value in this example: 3470*
- contamination: modify first 7 observations to $\sim 10^7$
- naïve estimation by empirical quantile:
2960 (ideal), but 8910000 (contaminated)
- parametric (Max-Likelihood) estimation:
3390 (ideal), but 7580 (contaminated)
- robust estimation (by rmx-procedure):
3440 (ideal), but 3710 (contaminated)

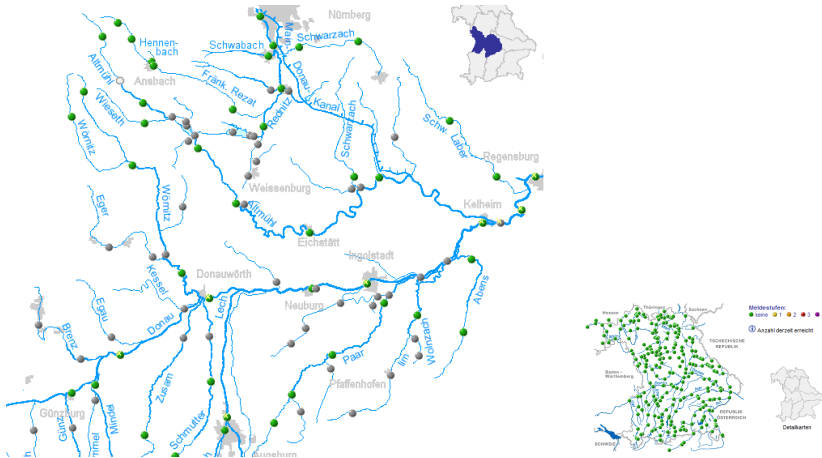
*: all numbers rounded to 3 significant digits

What are we talking about? — Floodings in Donauwoerth



source: <http://www.wwa-don.bayern.de/hochwasser/hochwasserschutzprojekte/donauwoerth>

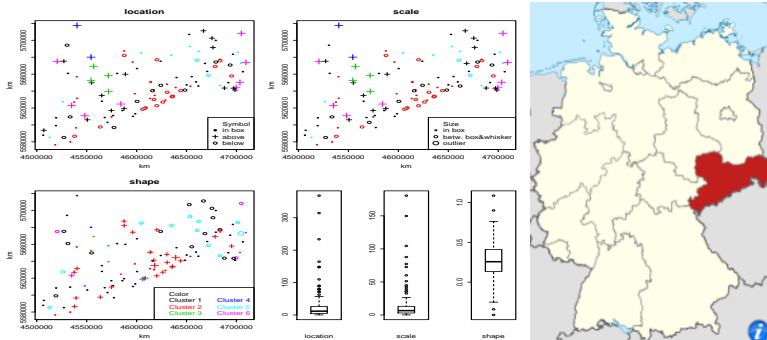
Location of Donauweoerth in Bavaria



source: <http://www.hnd.bayern.de/>; traffic lights for alerts from Dec. 12, 2013

Current Approaches in Hydrology

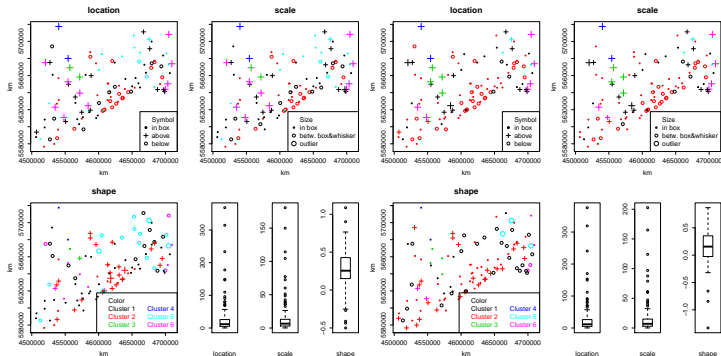
- techniques from **Extreme Value Statistics**
see e.g. Reiß/Thomas[97], Katz et al.[02]
- geostastical aspects \rightsquigarrow **Regionalization** [borrowing strength]
see e.g. Hosking/Wallis[97]



source: Laaha; clustering by GEV parameters (agnes) fit to block maxima in Saxony

Current Approaches in Hydrology (cont.)

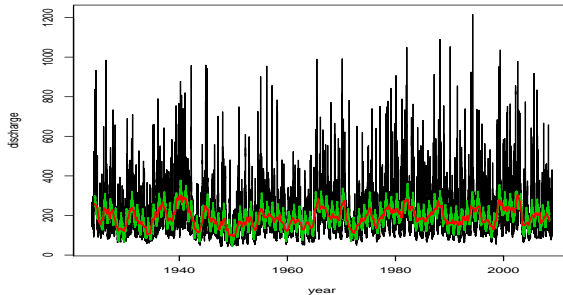
- robust estimation



clustering by ML vs. robust GEV parameters ([agnes](#)) fit to block maxima in Saxony

Issues

- trends and seasonalities:



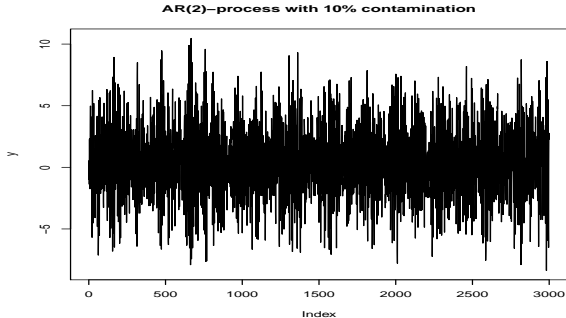
in black: daily discharges; trend & seasonality; non-robust: c.f. Reiß and Thomas[07] /

robust: c.f. Fried et al.[07]

- outliers:
- dynamics:

Issues

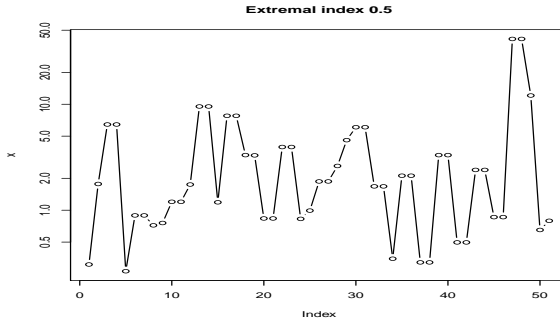
- trends and seasonalities:
- outliers:



- dynamics:

Issues

- trends and seasonalities:
- outliers:
- dynamics:



toy data acc. to [Coles\[01\]](#), Ex. 5.1.

Research Questions and Challenges

- find models and **robust** procedures which
 - capture *extreme* behaviour
 - provide a simple & parsimonious, yet flexible *dynamics*
 - possibly account for *regional* effects
- address the question:

inter-arrival time distribution of extremes

no new question: see, e.g., [Khaliq et al.\[06\]](#)

State Space Models and Filtering Problem

Linear, Time–Discrete, Time-Invariant Euclidean Setup

ideal model:

$$\begin{aligned}x_t &= Fx_{t-1} + v_t, & v_t &\overset{\text{indep.}}{\sim} (0, Q), \\y_t &= Zx_t + \varepsilon_t, & \varepsilon_t &\overset{\text{indep.}}{\sim} (0, V), \\x_0 &\sim (a_0, Q),\end{aligned}$$

$\{v_t\}, \{\varepsilon_t\}, x_0$ indep. as processes

(hyper-parameters Z known, F, Q, V to be estimated)

Generalizations also covered

- non-linear SSM's (by *Extended K.F.*, *Unscented K.F.*)
- time-varying F_t, Q_t, V_t depending on time-inv. param. θ

Algorithms for Filtering and Parameter Fitting

Filter/Smoothing Problem (for known hyp.-param.'s)

$$E |x_i - f_t(y_{1:j})|^2 = \min_{f_t} !, \quad \text{with } y_{1:j} = (y_1, \dots, y_j), \quad y_{1:0} := \emptyset$$

class. solution: Kalman–Filter and –Smoother— Kalman[60]

optimal among linear [Gaussian setting: among all] filters & smoothers:

Initialization: $x_{0|0} = a_0$

Prediction: $x_{i|i-1} = Fx_{i-1|i-1}, \quad [\Delta x_i = x_i - x_{i|i-1}]$

Correction: $x_{i|i} = x_{i|i-1} + M_i^0 \Delta y_i, \quad [\Delta y_i = y_i - Z_i x_{i|i-1}]$

Smoothing: $x_{i|T} = x_{i|i} + J_i(x_{i+1|T} - x_{i|i}), \quad [J_i = \Sigma_{i|i} F^T \Sigma_{i+1|i}^{-1}]$

+ corresp. recursions for predict-/filter-/smoothing error cov.'s $\Sigma_{i|i-1}, \Sigma_{i|i}$ and Kalman gain M_i^0

route: Init. \rightarrow "forward-loop" = {Prediction, Correction} \rightarrow

\rightarrow "backward-loop" = Smoothing

Algorithms for Filtering and Parameter Fitting

EM-Algorithm for SSM (unknown hyp.-param.'s)

application of EM-Algo to **SSMs** (with X_t as missings) by [Shumway/Stoffer](#)[82];
improvements by several authors since, see [Durbin/Koopman](#)[01].

Initialization: get initial estimators for F, Q, V ,
e.g. by moment-type-estimator

E-Step: reconstruct unobserved states by Kalman filter and smoother

M-Step: parameter estimation, e.g. by (conditional) ML estimator

route: **Init.** → “EM-loop” = {**E-Step**, **M-Step**}

Elements of Extreme Value Statistics

two settings — consider

(a) *block maxima* or (b) *exceedances over some threshold*

(a) Fisher-Tippett-Gnedenko Theorem:

possible limit distributions of $\max(X_i)$ have

cdf $H_\theta(x) = \exp(-(1 + \xi(x - \mu)/\beta)^{-1/\xi})$ (**GEVD** [= Gen. Extreme Value Distrib.])

(b) Pickands-Balkema-de Haan Theorem:

possible limit distr. of threshold exceedances have

cdf $F_\theta(x) = 1 - (1 + \xi(x - \mu)/\beta)^{-1/\xi}$ (**GPD** [= Gen. Pareto Distrib.])

- FTG-Thm \iff PBdH-Thm
- linked by same **Parameter** $\theta = (\xi, \beta, \mu)^T$:
 - shape ξ (≥ 0) (tail behavior)
 - scale β
 - location/threshold μ ($\leq x$)

Interplay of SSM and EVT

- basic extreme value theorems cover i.i.d. situation
- in a dynamic, time-dependent setting:
use concept of **extremal index**
see [Embrechts et al.\[97\]](#)
 - non-parametric approach: is $1/\text{limiting mean cluster size}$
 - compare [Drees\[03,08\]](#), [Janßen/Drees\[13\]](#), [Janßen\[10\]](#), ERCIM 13

here: dynamics captured by SSM
extremes modeled in the i.i.d. innovations

thus: flexible, parametric DGP to study inter-arrival times of
exceedances (*—not only in the limit*)

Outliers

- Outliers and extremes – a contradiction? Dell’Aquila/Embrechts[06]
- What makes an obs. an outlier? (—and not a regular extreme)
 - occur **rarely** (usually, 5%–10%)
 - **uncontrollable**, from **unknown** distr. (may vary obs.-wise), **unpredictable**
 - have **no predictive power**
 - usually: no error-free separation from ideal obs.
- In dynamic setting

exogenous outliers affecting only singular observations

$$\text{AO} \quad :: \quad \varepsilon_t^{\text{re}} \sim (1 - r_{\text{AO}})\mathcal{L}(\varepsilon_t^{\text{id}}) + r_{\text{AO}}\mathcal{L}(\varepsilon_t^{\text{di}})$$

$$\text{SO} \quad :: \quad y_t^{\text{re}} \sim (1 - r_{\text{SO}})\mathcal{L}(y_t^{\text{id}}) + r_{\text{SO}}\mathcal{L}(y_t^{\text{di}})$$

endogenous outliers / structural changes

$$\text{IO} \quad :: \quad \xi_t^{\text{re}} \sim (1 - r_{\text{IO}})\mathcal{L}(\xi_t^{\text{id}}) + r_{\text{IO}}\mathcal{L}(\xi_t^{\text{di}})$$

but also

trends, level shifts

Outliers

- Outliers and extremes – a contradiction? Dell’Aquila/Embrechts[06]
- What makes an obs. an outlier? (—and not a regular extreme)
 - occur **rarely** (usually, 5%–10%)
 - **un**controllable, from **un**known distr. (may vary obs.-wise), **un**predictable
 - have **no predictive power**
 - usually: no error-free separation from ideal obs.
- In dynamic setting

exogenous outliers affecting only singular observations

$$\text{AO} \quad :: \quad \varepsilon_t^{\text{re}} \sim (1 - r_{\text{AO}})\mathcal{L}(\varepsilon_t^{\text{id}}) + r_{\text{AO}}\mathcal{L}(\varepsilon_t^{\text{di}})$$

$$\text{SO} \quad :: \quad y_t^{\text{re}} \sim (1 - r_{\text{SO}})\mathcal{L}(y_t^{\text{id}}) + r_{\text{SO}}\mathcal{L}(y_t^{\text{di}})$$

endogenous outliers / structural changes

$$\text{IO} \quad :: \quad \xi_t^{\text{re}} \sim (1 - r_{\text{IO}})\mathcal{L}(\xi_t^{\text{id}}) + r_{\text{IO}}\mathcal{L}(\xi_t^{\text{di}})$$

but also

trends, level shifts

Outliers

- Outliers and extremes – a contradiction? [Dell'Aquila/Embrechts\[06\]](#)
- What makes an obs. an outlier? (—and not a regular extreme)
 - occur **rarely** (usually, 5%–10%)
 - **uncontrollable**, from **unknown** distr. (may vary obs.-wise), **unpredictable**
 - have **no predictive power**
 - usually: no error-free separation from ideal obs.
- In dynamic setting

exogenous outliers affecting only singular observations

$$\text{AO} \quad :: \quad \varepsilon_t^{\text{re}} \sim (1 - r_{\text{AO}})\mathcal{L}(\varepsilon_t^{\text{id}}) + r_{\text{AO}}\mathcal{L}(\varepsilon_t^{\text{di}})$$

$$\text{SO} \quad :: \quad y_t^{\text{re}} \sim (1 - r_{\text{SO}})\mathcal{L}(y_t^{\text{id}}) + r_{\text{SO}}\mathcal{L}(y_t^{\text{di}})$$

endogenous outliers / structural changes

$$\text{IO} \quad :: \quad \xi_t^{\text{re}} \sim (1 - r_{\text{IO}})\mathcal{L}(\xi_t^{\text{id}}) + r_{\text{IO}}\mathcal{L}(\xi_t^{\text{di}})$$

but also

trends, level shifts

Robustness

component-wise robustification to tackle outlier issue

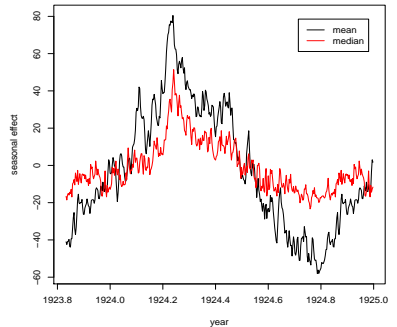
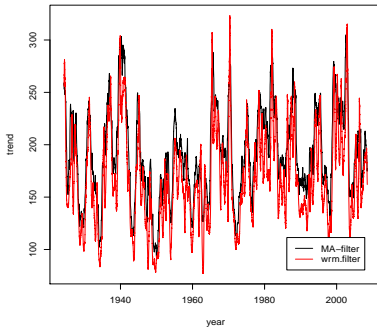
- **Init-EM:** use robust autocovariances, see Higham[02], S.[10]
- **E-Step-EM:** use *rLS-Filter* (Ruckdeschel[01,10]),
i.e., in *Corr.*, replace $M_i^0 \Delta y_i$ by $H_{b_i}(M_i^0 \Delta y_i)$, $H_b(x) = x \min\{1, b/|x|\}$
use *rLS-Smoother* (Ruckdeschel, S., Pupashenko[14]),
i.e., in *Smooth.*, replace $J_i(x_{i+1|T} - x_{i|i})$ by $H_{\tilde{b}_i}(J_i(x_{i+1|T} - x_{i|i}))$
- **M-Step-EM:** use robust multiv. regression and scale est.'s
(see Croux/Joossens[08], Agullo et al.[08], ERCIM 10)
... work in progress—soon some more on this ...
- **EVT:** use optimally-robust *RMXES* to fit GPD, GEVD (Ruckdeschel/Horbenko [12,13]); extends and improves, a.o. Hosking et al.[85]

Real Data Set

- daily average discharge data of *Danube river* in [m³/s]
- location: gauge at *Donauwörth* (see initial pictures)
- start: 1923-11-01, end: 2008-12-31 (> **30,000** days)
- currently collected and provided by
*Hochwassernachrichtendienst (HND),
Bayerisches Landesamt für Umwelt (LfU)*
[translated: $\hat{=}$ *Flooding news service by the Bavarian Environmental Office*]
- provided to us by *G. Laaha* within project “*Robust Risk Estimation*”

Model Specification

- the original data y_t^h is first **detrended** and **deseasonalized** by moving averages to series y_t according to **Reiß and Thomas[07]**, i.e., with *yearly trend* m_t and *yearly season* s_t , $y_t^h = m_t + s_t + y_t$
- robust trend/season extraction: c.f. **Fried et al.[07]**



Model Specification

- the original data y_t^{d} is first **detrended** and **deseasonalized** by moving averages to series y_t according to **Reiß and Thomas[07]**, i.e., with *yearly trend* m_t and *yearly season* s_t , $y_t^{\text{d}} = m_t + s_t + y_t$
- robust trend/season extraction: c.f. **Fried et al.[07]**

- for simplicity, assume an **AR(4)-model** for the y_t ,
—model order chosen by BIAR/IWLS (**Martin/Thomson[82]**) & ACM-type filter (**Martin[79]**)
i.e., in SSM context with the usual embedding

$$X_t = \begin{pmatrix} x_t \\ x_{t-1} \\ x_{t-2} \\ x_{t-3} \end{pmatrix}, F = \begin{pmatrix} \varphi_1 & \varphi_2 & \varphi_3 & \varphi_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, Q = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, V = 0, Z^T = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

... entails that in **M-Step-EM**, we use standard robust MM regression **lmrob**

- to the tails of the obtained (filtered/smoothed) innovations $v_{t|T}$ fit a GPD model

Model Specification

- the original data y_t^{d} is first **detrended** and **deseasonalized** by moving averages to series y_t according to **Reiß and Thomas[07]**, i.e., with *yearly trend* m_t and *yearly season* s_t , $y_t^{\text{d}} = m_t + s_t + y_t$
- robust trend/season extraction: c.f. **Fried et al.[07]**

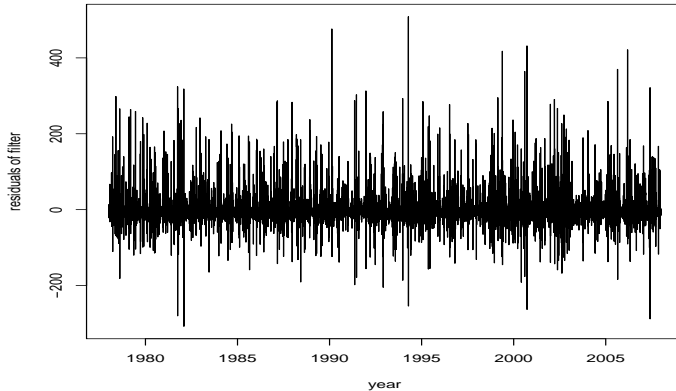
- for simplicity, assume an **AR(4)-model** for the y_t ,
—model order chosen by BIAR/IWLS (**Martin/Thomson[82]**) & ACM-type filter (**Martin[79]**)
i.e., in SSM context with the usual embedding

$$X_t = \begin{pmatrix} x_t \\ x_{t-1} \\ x_{t-2} \\ x_{t-3} \end{pmatrix}, F = \begin{pmatrix} \varphi_1 & \varphi_2 & \varphi_3 & \varphi_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, Q = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, V = 0, Z^T = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

... entails that in **M-Step-EM**, we use standard robust MM regression **lmrob**

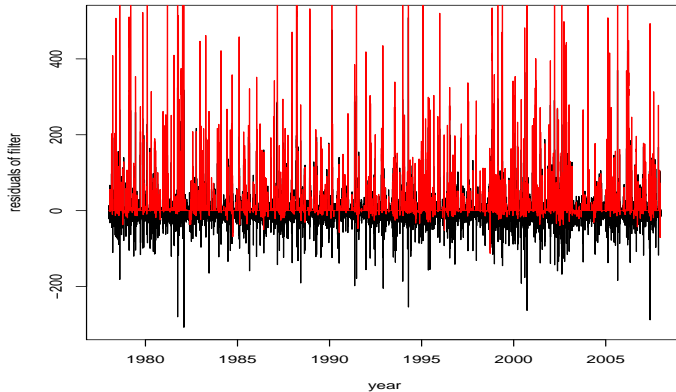
- to the tails of the obtained (filtered/smoothed) innovations $v_{t|T}$ fit a GPD model

Effects of Filtering



detrended and deseasonalized y_t and observation residuals from rob. filter $\hat{\varepsilon}_t = y_t - Zx_{t|t}^{rob}$

Effects of Filtering



detrended and deseasonalized y_t and observation residuals from rob. filter $\hat{\varepsilon}_t = y_t - Zx_{t|t}^{rob}$

Fit Real Data to GPD-Model

- raw data
 - threshold chosen with `gpd.fitrange`: 500

	scale	(SE)	shape	(SE)
MLE	125.68	(22.25)	-0.1766	(0.1245)
RMXE	123.49	(23.00)	-0.1289	(0.1810)

- with filtering

- AR(4)-param's:

	φ_1	φ_2	φ_3	φ_4	μ	σ^2
MLE	1.2804	-0.5692	0.1825	-0.0044	0.1949	1721
rob	1.3078	-0.4492	0.1640	-0.0660	-15.6647	137

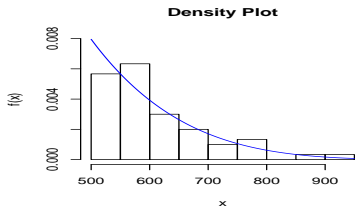
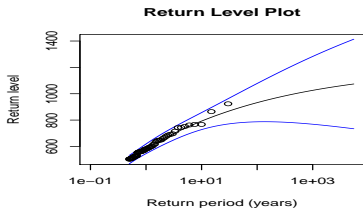
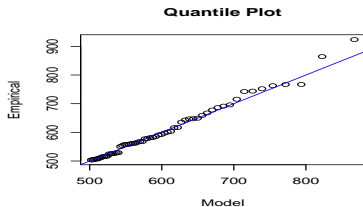
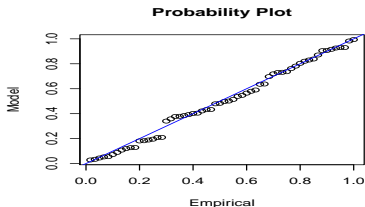
- threshold chosen with `gpd.fitrange`: 250 (non-robust), 400 (robust)

	scale	(SE)	shape	(SE)
MLE	69.49	(17.62)	-0.0551	(0.1892)
RMXE	151.42	(19.94)	-0.0864	(0.1304)


for RMXE used functionality of -pkgs `R0ptEst`, `RobExtremes`



Tail and Fitted GPD on raw data



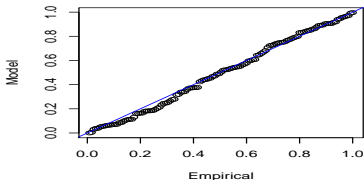
used

function `gpd.diag` of -pkg `ismev`

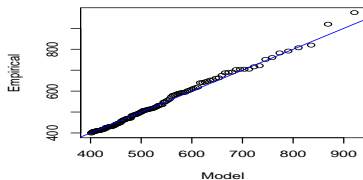


Tail and Fitted GPD on filtered data

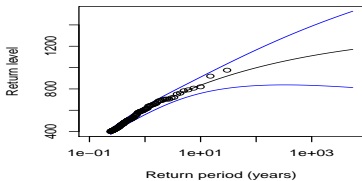
Probability Plot



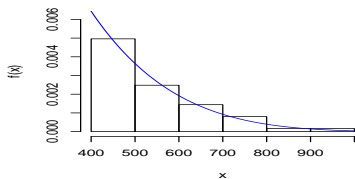
Quantile Plot



Return Level Plot



Density Plot



used

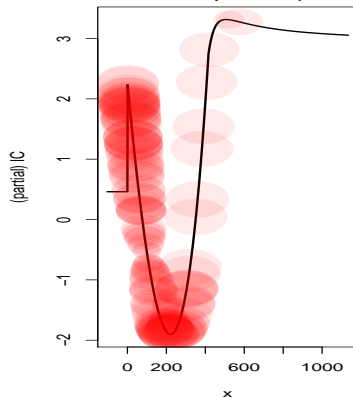
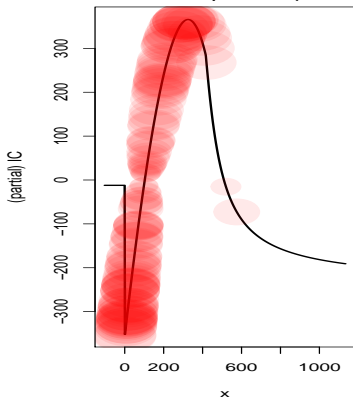
function `gpd.diag` of -pkg `ismev`




IC-Plot

Component 'scale'
of IC of contamination type
with main parameter ('scale' = 151.418, 'shape' = -0.)
and fixed known parameter ('loc' = 0)

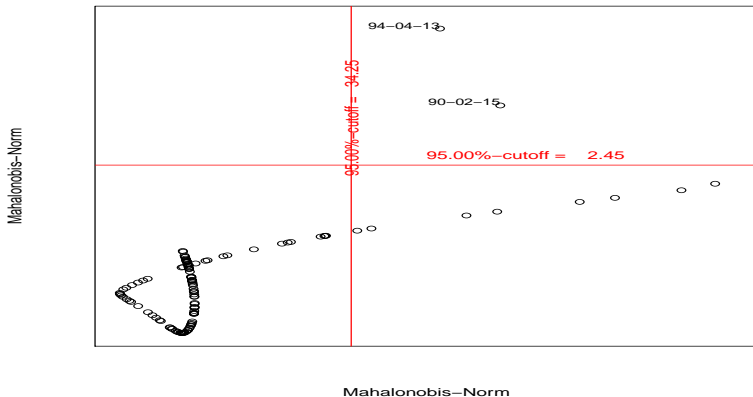
Component 'shape'
of IC of contamination type
with main parameter ('scale' = 151.418, 'shape' = -0.)
and fixed known parameter ('loc' = 0)



used **plot**-method for S4-class IC of -pkg **RobAStBase**



Outlyingness of Extremes



used function `outlyingPlotIC` of -pkg `RobAStBase`



Outlyingness of Extremes II — True Effect?

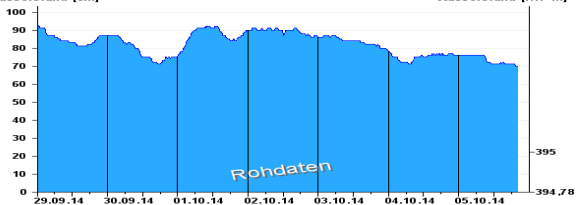
Wasserstands-Grafik Donauwörth / Donau

<http://www.hnd.bayern.de/pegel/wasserstand/peg...>

Stammdaten | **Wasserstand** | Abfluss | Abflusstafel | Hochwassermarken | Mittel- / Höchstwerte
Gebietsdaten / Laufzeiten | Lagekarte | Bild | Jahrbuchseite
Darstellung in Tabellen-Form | Druckversion

Pegel im Donaugebiet: Donauwörth / Donau

Wasserstand [cm]



Vorhersage: keine | 12-Std.-Vorhersage | 2-Tage-Trend

Linien: keine | Meldestufen | Hochwassermarken | historische Ereignisse

- Unsicherheitsbereich der Vorhersage(Erläuterung)
- Vorhersage vom 04.10.14 06:00 Uhr (Publikation: 13:02 Uhr)
- Letzter Messwert vom 05.10.14 19:45 Uhr: 70 cm

Zeitraum auswählen:

Datum von: bis:

- 14.04.1994 Wasserstand: 577 cm
- 16.02.1990 Wasserstand: 553 cm
- 24.05.1999 Wasserstand: 552 cm
- 27.03.1988 Wasserstand: 544 cm
- 01.02.1982 Wasserstand: 543 cm



Conclusion

- presented a flexible, param. dynamic model class for hydrological extremes
- assessment of the inter-arrival distribution of extremes
- provided a step-by-step robustification
- evidence that robustification also enhances analysis of extremes

References

- Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modelling extremal events for insurance and finance*. Springer.
- Fried, R., Einbeck, J., and Gather, U. (2007). Weighted Repeated Median Smoothing and Filtering, *JASA* **102**, 1300–1308.
- Hosking, R.J.M., Wallis, T.J. (1997). *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press.
- Khaliq, M.N., Ouarda, T.B.M.J., Ondo, J.-C., Gachon, P., Bobee, B. (2006). Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review. *Journal of Hydrology*, **329**, 534–552.
- Laaha, G., Blöschl, G. (2006). Seasonality indices for regionalizing low flows. *Hydrological Processes* **20**(18), 3851–3878.
- Reiss, R.-D., Thomas, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser Verlag, Basel.
- Ruckdeschel, P., Horbenko, N. (2013). Optimally-Robust Estimators in Generalized Pareto Models. *Statistics* **47**(4), 762–791.
- Ruckdeschel, P., Horbenko, N. (2012). Yet another breakdown point notion: EFSBP — illustrated at scale-shape models. *Metrika*, **75** (8), 1025–1047.
- Ruckdeschel, P., Spangl, B. and Pupashenko, D. (2014). Robust Kalman tracking and smoothing with propagating and non-propagating outliers. *Statistical Papers* **55**(1), 93–123.
- Spangl, B. (2010). Computing the Nearest Correlation Matrix Which is Additionally Toeplitz. Working Paper.

THANK YOU FOR YOUR ATTENTION!

