Overdispersion in peak over threshold (POT) flow data and its effect on flood frequency practice

Ilaria Prosdocimi - ilapro@ceh.ac.uk / @ilapros
with Gianni Vesuviano
Centre for Ecology & Hydrology, Wallingford
Motivation

32003 - Harpers Brook at Old Mill Bridge
Motivation

T=50
T=500
T=1000

GEV
47.1
77.4
88.0

GLO
47.7
90.2
108.2
Motivation

GEV has a theoretical justification and statisticians love it.

GLO has practical advantages and practitioners love it.
Motivation – more theory!

There are some relationships between the distribution of peaks over the threshold (POT) and block maxima (AMAX).

$X$ indicates POT values; $N$ the number of exceedances/year. $Q$ indicates the AMAX data values.

When $Y = (X - u) \sim GPD(\tilde{\sigma}, \xi)$ and $N \sim \text{Pois}(\lambda)$ \quad $Q \sim \text{GEV}(\mu, \sigma, \xi)$

If $N \sim \text{Pois}(\lambda)$: $E(N) = \lambda$; $\text{Var}(N) = \lambda$

When $Y = (X - u) \sim GPD(\tilde{\sigma}, \xi)$ and $N \sim \text{NB}(\lambda, 1)$ \quad $Q \sim \text{GLO}(\mu, \sigma, \xi)$

If $N \sim \text{NB}(\lambda, 1)$: $E(N) = \lambda$; $\text{Var}(N) = 2\lambda$

Does the data agree on this? Can this be useful?

See Eastoe and Tawn (2010), WRR
Data and variables

Take the .PT files from the WINFAP-FEH (v.3.3.4).
Create dataset of annual counts.
Stations with average counts in $[2,7]$, otherwise discarded.
AMAX series “suitable for QMED” and non-urban catchments.

Record length between 10 and 127 years. Median length = 39

Define dispersion coefficient $D$: $D = \frac{\text{Var}(N)}{E(N)}$.

For Poisson $D = 1$.
Overdispersion when $D > 1$; larger variability, i.e. clustering.
For Negative Binomial $D=2$
Dispersion coefficient

Test if the dispersion coefficient $D$ gives indication of Poisson or Negative Binomial distribution.

Build test as in Eastoe and Tawn (2010).
Dispersion coefficient

- $D$ too large: 35 (6.8%)
- $D = 2$: 272 (52.9%)
- $D = 2 \& D = 1$: 145 (28.2%)
- $D = 1$: 62 (12.1%)

Record length

[Graph showing the distribution of dispersion coefficients]
Relationship to AMAX distribution

Compare with accepted AMAX distribution as chosen by the Kjeldsen and Prosdocimi (2014) test, based on pooled regions.

<table>
<thead>
<tr>
<th></th>
<th>$D$ large</th>
<th>$D=2$</th>
<th>$D=2 &amp; D=1$</th>
<th>$D=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEV</td>
<td>5</td>
<td>87</td>
<td>41</td>
<td>22</td>
</tr>
<tr>
<td>GLO</td>
<td>20</td>
<td>127</td>
<td>61</td>
<td>27</td>
</tr>
<tr>
<td>GNO</td>
<td>5</td>
<td>29</td>
<td>26</td>
<td>10</td>
</tr>
<tr>
<td>PE3</td>
<td>0</td>
<td>13</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>none</td>
<td>5</td>
<td>16</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>272</td>
<td>145</td>
<td>62</td>
</tr>
</tbody>
</table>

Kjeldsen and Prosdocimi (2014), poster at BHS
But why?

Is the theory completely wrong? It relies on “some regularity assumptions” and asymptotic. Real data are short records of a non-stationary process.

For each series we take the annual median of daily catchment average rainfall. Use the annual median rainfall as an explanatory variable in a Poisson regression. Estimate $D$ after the regression.

CEH GEAR presented tomorrow at 10:30 in the Water resources session
Conclusions

Information on the dispersion of the exceedance counts can be useful when testing for AMAX distribution.

There is a clear source of non-stationarity which breaks the assumptions underlying the statistical theory.

Eventually, in $n \to \infty$ years time, theory and practice might agree.

Thank you!

ilapro@ceh.ac.uk