The integral of a stochastic process:
tail probability and high quantile

Chen Zhou

Spatial extreme value theory and properties of max-stable processes

9 Nov 2012
Fighting with the water

- The Netherlands
  - A low-lying flat country
  - Threatens: high sea level, storm, extreme rainfall
The Netherlands
- A low-lying flat country
- Threatens: high sea level, storm, extreme rainfall

In the northwest (the province of North Holland)
- There are 32 stations recording rainfall data.
- Rainy season: the fall (Sept, Oct and Nov)

Research questions:
- What is total amount of precipitation in this area on one day that is exceeded once in 100 years?
- What is the chance of having a daily total precipitation exceeding a certain threshold?
Mathematical model

- It is a typical extreme value problem.
- Model the rainfall process
  - Denote the area as $S$ (compact subset of $\mathbb{R}^2$)
  - Denote the daily rainfall process as $X = \{X(s)\}_{s \in S}$
  - $X$ is a stochastic process with continuous sample path
- We study the tail distribution of $\int_S X(s)\,ds$,
- The two research questions are on

$$p = \text{Pr}\left(\int_S X(s)\,ds > x\right)$$

- Tail probability: for known $x$, estimate $p$
- High quantile: for known $p$, estimate $x$
General plans

- Simulation based
  - Model the “extreme” of the process $X(s)$
  - Estimate the parameters from data
  - Simulate a large dataset of daily rainfall
  - The main difficulty: 
    How to connect simulated extreme rainfall process with observed rainfall process?
General plans

- **Simulation based**
  - Model the “extreme” of the process $X(s)$
  - Estimate the parameters from data
  - Simulate a large dataset of daily rainfall
  - The main difficulty: *How to connect simulated extreme rainfall process with observed rainfall process?*

- **Calculation based**
  - Relate the tail distribution of the integral to
    - Tail properties of marginal distributions
    - Spatial dependence structure
  - Estimate the components
  - Estimate the tail distribution of the integral
Max-stable process v.s. Domain of attraction

Simulation based solution
Calculation based solution
Comparison

Max-stable process v.s. Domain of attraction
Definitions

Extreme value theory for stochastic processes

- Consider independent copies \( X_1, X_2, \cdots \) of the process \( X \).
- There exist some positive functions \( a_s(n) \) and real functions \( b_s(n) \) such that

\[
\max_{1 \leq i \leq n} \frac{X_i(s) - b_s(n)}{a_s(n)} \to \eta(s) \quad \text{in } C[0, 1].
\]

Representations (after normalization)

- Max-stable process \( \eta(s) = \bigvee_{i=1}^{+\infty} Z_i V_i(s) \).
- Example in the domain of attraction (GPP): \( X = YV(s) \).
- \( V, V_1, \cdots \) are i.i.d. processes satisfying

\[
E \sup_{s \in S} V(s) < +\infty \quad \text{or} \quad \sup_{s \in S} V(s) = c, \quad \text{a.s.}
\]
Which process to use?

- How to use them for modeling the rainfall process?
Which process to use?

- How to use them for modeling the rainfall process?
  - Neither of them can be directly used.
Which process to use?

- How to use them for modeling the rainfall process?
  - Neither of them can be directly used.
  - They might be used for modeling the extreme part, and then “connect to” the moderate part.
  - Transformations to accommodate marginal extreme value parameters are necessary.
Which process to use?

- How to use them for modeling the rainfall process?
  - Neither of them can be directly used.
  - They might be used for modeling the extreme part, and then “connect to” the moderate part.
  - Transformations to accommodate marginal extreme value parameters are necessary.

- Choice between the two
  - MSP has a fully parametric marginal distribution
  - Processes in the domain of attraction are flexible
  - Technical reason for simulation: Simulating MSP: simulating many $V$ processes.

Processes in the domain of attraction are better!
**Spatial dependence: the V process**

- **Simulation based solution**
  - Parametric model of $V$
  - Parameter(s) quantifies spatial dependence
  - Estimation of the parameter(s) from data
  - Which type of $V$?
    - $\sup_{s \in S} V(s) = c$, a.s.
      - Restrictive: moving maxima
    - $E \sup_{s \in S} V(s) < +\infty$
      - Flexible: Brown-Resnick

- **Calculation based solution**
  - We tend to not make parametric assumption
  - Non-parametric estimation on the spectral measure
  - Hence, do not need a model on the $V$ process
<table>
<thead>
<tr>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max-stable process v.s. Domain of attraction</td>
</tr>
<tr>
<td>Simulation based solution</td>
</tr>
<tr>
<td>Calculation based solution</td>
</tr>
<tr>
<td>Comparison</td>
</tr>
</tbody>
</table>

**Simulation based solution**
Planned simulation procedure so far...

A peak-over-threshold simulation:

- Randomly pick one day with rainfall observations
- Simulate a extreme rainfall process
  - Simulate the parametric spatial process
  - Transform univariate marginal to GPD by local parameters
- Connect the two as: if at one location the observed rainfall is above the local threshold (location), then change the observed data to the simulated rainfall at this point.
- For the spatial process, we planned to
  - Use a GPP process $YV(s)$ for spatial dependence
  - With a parametric model $V(s)$ s.t. $E \sup_{s \in S} V(s) < +\infty$
A dilemma

- However, there is a theoretical problem
  - The tail of $Y V(s)$ with $E \sup_{s \in S} V(s) < +\infty$ is NOT Pareto.
  - There is no such a problem if we require $\sup_{s \in S} V(s) = c$, a.s..
  - This can not be used in the “connection”: After transformation, the extreme rainfall process do not have GPD univariate marginal.

What to do?
However, there is a theoretical problem

- The tail of $YV(s)$ with $E \sup_{s \in S} V(s) < +\infty$ is NOT Pareto.
- There is no such a problem if we require $\sup_{s \in S} V(s) = c$, a.s.
- This can not be used in the “connection”: After transformation, the extreme rainfall process do not have GPD univariate marginal.

What to do?

- We go back to Max-stable process
  - MSP has a Fréchet marginal
  - We can transform that to standard Pareto.
For the spatial dependence, we use

\[ \eta(s_1, s_2) = \bigvee_{i=1}^{+\infty} Z_i \exp \left\{ W_{i,1}(\beta s_1) + W_{i,2}(\beta s_2) - \beta(|s_1| + |s_2|)/2 \right\} \]

Then we transform as

\[ \xi(s_1, s_2) := \frac{1}{1 - \exp\left\{-\frac{1}{\eta(s_1, s_2)}\right\}} \]

Final transformation to the extreme rainfall process

\[ X(s_1, s_2) := \hat{a}(s_1, s_2)(n/k) \left( \frac{\xi(s_1, s_2)\hat{\gamma}_{n,k} - 1}{\hat{\gamma}_{n,k}} \right) + \hat{b}(s_1, s_2)(n/k). \]

More details on the “connection” to observed rainfall

- Divide the area according to stations (Vertices)
- Extreme v.s. non-extreme Triangles
Estimate the high quantile

- We have observations on $91 \times 30 = 2730$ days
  - No extreme Vertices (no simulations): 2299
  - All Triangles are extreme (full simulations): 44
- Estimate the $1-1/9,100$ quantile of total rainfall
  - Take the 10-th highest from 91K simulations
  - We repeat 60 times. The average is 58.8(mm)
Calculation based solution
General model

- The rainfall process $X$ is in the domain of attraction
  - Not limited to MSP or GPP
- Spatial dependence
  - Exponent measure $\nu$ on $C^+(S)$,
    \[
    \lim_{t \to \infty} t P \left( \left\{ \left( 1 + \gamma(s) \frac{X(s) - b_s(t)}{a_s(t)} \right)^{1/\gamma(s)} \right\}_{s \in S} \in E \right) = \nu(E)
    \]
  - Spectral measure $\rho$ on $\tilde{C}_1^+(S)$
    \[
    \nu(E) = \int \int_{rg \in E} \frac{dr}{r^2} d\rho(g).
    \]
Assumptions

- Marginal parameters
  - extreme value index: homogeneity $\gamma(s) = \gamma$
  - scales: comparability

$$\sup_{s \in S} \left| \frac{a_s(t)}{a(t)} - A(s) \right| \to 0, \quad \text{as} \quad t \to \infty.$$ 

- Spectral measures
  - For $\gamma \leq 0$,

$$\rho\{g \in \tilde{C}^+_1(S) : \inf_{s \in S} g(s) = 0\} = 0.$$ 

  - For $\gamma > 0$, $X \geq 0$
Main probability result

- Tail distribution of the integral

$$\lim_{t \to \infty} t P \left( \frac{\int_S X(s) ds - \int_S b_s(t) ds}{a(t)} > y \right) = \theta_\gamma (1 + \gamma y)^{-1/\gamma}$$

- The areal coefficient

$$\theta_\gamma := \int_{\tilde{C}_1^+(S)} \left( \int_S A(s) g^\gamma(s) ds \right)^{1/\gamma} d\rho(g).$$

- Under our assumptions, $\theta_\gamma > 0$

For statistics, we need to estimate each component!
Estimating the tail probability

- Suppose we have \( n \) observations of the full processes.
- We estimate \( p_n = P \left( \int_S X(s) \, ds > x_n \right) \) for some \( x_n \).
- We know ex ante that \( p_n = O(1/n) \).
- By taking \( t = n/k \) in the probability relation

\[
\hat{p}_n = \frac{k}{n} \hat{\theta} \left( 1 + \hat{\gamma} \frac{x_n - \int_S \hat{b}_s(n/k) \, ds}{\hat{a}(n/k)} \right)^{-\gamma}
\]

- Marginal: \( \hat{\gamma} := \int_S \hat{\gamma}(s) \, ds / |S|, \hat{a}(n/k) := \int_S \hat{a}_s(n/k) \, ds \).
- Areal coefficient:

\[
\hat{\theta} = \int_{\tilde{C}_1^+(S)} \left( \int_S \tilde{A}(s) g^{\hat{\gamma}}(s) \, ds \right)^{1/\hat{\gamma}} \, d\hat{\rho}(g)
\]

- Main theorem: under proper conditions

\[
\frac{\hat{p}_n}{p_n} \to^P 1 \quad \text{as} \quad n \to \infty.
\]
Estimate the tail probability

- With the same data, estimate $P\left( \frac{1}{|S|} \int_S X(s) \, ds > 58.8 \right)$
- Only observations on stations: linearization
  - Marginal scale and location functions
  - Functions in spectral measure
- Estimates regarding various $k$ choice

The integral of a stochastic process
Comparison

Max-stable process v.s. Domain of attraction

Simulation based solution

Calculation based solution

The integral of a stochastic process
### Comparing the two methods

<table>
<thead>
<tr>
<th></th>
<th>Simulation</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target</strong></td>
<td>high quantile</td>
<td>tail probability</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>Parametric</td>
<td>Non-parametric</td>
</tr>
<tr>
<td><strong>Marginal</strong></td>
<td>homogeneous extreme value index</td>
<td></td>
</tr>
<tr>
<td><strong>Spatial</strong></td>
<td>Brown-Resnick</td>
<td>spectral measure</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td>Point and Std</td>
<td>Point only</td>
</tr>
<tr>
<td><strong>Risk</strong></td>
<td>model misspecification</td>
<td>estimation variation</td>
</tr>
<tr>
<td><strong>Extension</strong></td>
<td>The targets can be swapped.</td>
<td></td>
</tr>
<tr>
<td><strong>Speed</strong></td>
<td>Slow</td>
<td>Fast</td>
</tr>
</tbody>
</table>
The simulation method:

The calculation method: