Modeling Extreme Events Part II

Hao Zhang

Department of Statistics
Purdue University
USA

February 2015
\[ M_n = \max\{X_1, \ldots, X_n\} \text{ for } i.i.d. \ X_1, \ldots, X_n \text{ with a cdf } F. \]

- For large \( n \), \( M_n \) approximately has a cdf

\[
G(z) = \exp \left\{ - \left[ 1 + \xi \left( \frac{z - \mu}{\sigma} \right) \right]_+ \right\}^{-1/\xi}
\]

for some parameters \( \sigma > 0, \xi \) and \( \mu \). The latter two depend on \( n \).

- \[
P(X_i > u) \approx p_n = \frac{1}{n} \left[ 1 + \xi \left( \frac{u - \mu}{\sigma} \right) \right]_+^{-1/\xi}
\]

- \[
P(X_i > u+y|X_i > u) \approx \frac{1}{n} \left( 1 + \xi \frac{y}{\tilde{\sigma}} \right)_+^{-1/\xi}, \text{ for } y > 0, \tilde{\sigma} = \sigma + \xi(u-\mu).
\]
Marked Point Process Model

Let $X_i, i = 1, \ldots, n$, i.i.d $\sim F$. For a large $u$ such that $F(u) < 1$. Consider the point process

$$N_n = \{i/(1 + n) : X_i > u, i = 1, \ldots, n\}.$$  

Because $X_i$ are i.i.d, $N_n$ is approximately a homogeneous Poisson process with intensity $\lambda_0$, where

$$\lambda_0 = E(\text{Number of event in } (0, 1)) \approx np_n = \left[1 + \xi \left(\frac{u - \mu}{\sigma}\right)\right]^{-1/\xi}$$

where $p_n = \frac{1}{n} \left[1 + \xi \left(\frac{u - \mu}{\sigma}\right)\right]^{-1/\xi} \approx P(X_i > u)$.  

For each incident $X_i > u$, define the positive mark $Y_i = X_i - u$. Conditional on the incident, the mark has a distribution

$$P(Y_i > y | X_i > u) \approx \left(1 + \xi \frac{y}{\tilde{\sigma}}\right)_{+}^{-1/\xi}, \ y > 0, \ \tilde{\sigma} = \sigma + \xi (u - \mu).$$
Observe i.i.d. $X_i \sim F, i = 1, \ldots, mn$. Interested in estimating the parameters in the GEV distribution of block maximum $\max\{X_1, \ldots, X_n\}$. Then

$$N_n = \{i/(1 + n) : X_i > u, i = 1, \ldots, mn\}.$$ 

is approximately a marked Poisson process in $[0, m]$ with a intensity function $\lambda = p_n \approx (1/n) \left[1 + \xi \left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}$ and mark $(X_i - u)$. 
The likelihood function for this marked point process is, due to the independence of points and marks

$$
\exp(-mnp_n)\left(\prod_{x_i > u} p_n \right) \prod_{x_i > u} f(x_i - u)
$$

$$
= \exp \left( -m \left[ 1 + \xi \left( \frac{u - \mu}{\sigma} \right) \right]^{1/\xi} \right) \sum_{x_i > u} \frac{1}{n\sigma} \left( 1 + \xi \frac{x_i - \mu}{\sigma} \right)^{-1/\xi-1}
$$

where $f(y) = \left( \frac{1}{\tilde{\sigma}} \right) \left( 1 + \xi \frac{y}{\tilde{\sigma}} \right)^{-1/\xi-1}$ is the pdf of the mark. We have used the following equation

$$
p_n f(x - u) = \frac{1}{n} \left[ 1 + \xi \left( \frac{u - \mu}{\sigma} \right) \right]^{1/\xi} \frac{1}{\tilde{\sigma}} \left( 1 + \xi \frac{y}{\tilde{\sigma}} \right)^{-1/\xi-1}
$$

$$
= \frac{1}{n\sigma} \left( 1 + \xi \frac{x - \mu}{\sigma} \right)^{-1/\xi-1}
$$
The likelihood (2) and that in the threshold excess model are very close to each other, because

\[(1 - p_n)^n \approx \exp(-np_n)\].
The marked point process corresponds to a two dimensional Poisson point process. The corresponding point process in $\mathbb{R}^2$ is

$$N_n = \left\{ \left( \frac{1}{1 + n}, X_i \right) : X_i > u, i = 1, \ldots, n \right\}.$$ 

is approximately a Poisson point process on the set $[0, 1] \times [u, \infty)$ with a joint intensity function

$$\lambda f(x - u) = \frac{1}{n\sigma} \left( 1 + \xi \frac{x - \mu}{\sigma} \right)^{-1/\xi - 1}$$

by (4). The intensity measure is therefore

$$\Lambda(A) = (t_2 - t_1) \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-1/\xi}$$

for any $A = [t_1, t_2] \times [x, \infty)$ (Pickands, 1971).
The i.i.d. assumption is not realistic for climate data because of the global warming/increasing mean. For example, let $X_i$ be daily maximum temperature and we are interested in modeling the annual maximum temperature. In light of global warming, we should not treat $X_i$’s as i.i.d. distributed.

However, we are still interested in block maximum $M_n$ in a block of size $n$. We assume that $X_i$’s within each block are identically distributed but may be dependent. Under some regularity conditions, the block maximum still has approximately a GEV. We assume the parameters in the GEV distributions vary from block to block but in some parametric way. For example, the $t$th block might have the location parameter depending on $t$, $\mu_t = \mu(t, \beta) = \beta_0 + \beta_1 t + \beta_2 t^2$. 
Denote by $M_{n,t}$ the maximum of the $t$th block of size $n$, i.e.,

$M_{n,t} = \max\{X_{(t-1)n+j}, j = 1, \ldots, n\}$. Assume $M_{n,t}$ is GEV($\mu_t, \sigma_t, \xi_t$).

Consider the point process

$$\{i/(n+1) : X_i > u\}$$

is now an inhomogeneous marked Poisson process with piece-wise constant intensity, and a mark $x_i - u$. Suppose $\mu_t, \sigma_t$ and $\xi_t$ are functions of $t$ depending on a vector of parameter $\theta$. The likelihood function is

$$L(\theta) = \exp \left\{ - \sum_{t=1}^{T} \left[ 1 + \xi_t \left( \frac{u - \mu_t}{\sigma_t} \right) \right]^{-1/\xi_t} \prod_{t=1}^{T} \prod_{j=1}^{N_t} \frac{1}{\sigma_t} \left[ 1 + \xi_t \left( \frac{x_{tj} - \mu_t}{\sigma_t} \right) \right]^{-1/\xi_t-1} \right\}$$

where $x_{tj}$ is the $j$th exceedance in block $t$, and $N_t$ is the number of exceedance in the $t$th block.
Example 1: Weather Data

750+ national weather stations observed since 1951 with some having missing values.

Variables: Daily maximum temperature, precipitation.

Only look at the July maximum temperature.
An Example

Residual Probability Plot
## 20 Year Return Level of Maximum July Temperature at 2011

<table>
<thead>
<tr>
<th>province</th>
<th>RL</th>
<th>Provs</th>
<th>RL</th>
<th>Provs</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>beijing</td>
<td>41.56</td>
<td>guangdong</td>
<td>38.60</td>
<td>gansu</td>
<td>36.22</td>
</tr>
<tr>
<td>tianjin</td>
<td>40.55</td>
<td>jiangsu</td>
<td>38.56</td>
<td>jilin</td>
<td>35.63</td>
</tr>
<tr>
<td>shanghai</td>
<td>40.47</td>
<td>zhejiang</td>
<td>38.53</td>
<td>guizhou</td>
<td>35.63</td>
</tr>
<tr>
<td>hebei</td>
<td>40.29</td>
<td>shanxi</td>
<td>38.37</td>
<td>sichuan</td>
<td>34.25</td>
</tr>
<tr>
<td>jiangxi</td>
<td>39.53</td>
<td>fujian</td>
<td>38.26</td>
<td>yunnan</td>
<td>32.88</td>
</tr>
<tr>
<td>henan</td>
<td>39.27</td>
<td>guangxi</td>
<td>38.02</td>
<td>qinghai</td>
<td>29.32</td>
</tr>
<tr>
<td>neimenggu</td>
<td>39.11</td>
<td>shandong</td>
<td>37.83</td>
<td>xizang</td>
<td>27.53</td>
</tr>
<tr>
<td>xinjiang</td>
<td>39.06</td>
<td>chongqing</td>
<td>37.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hunan</td>
<td>39.06</td>
<td>ningxia</td>
<td>37.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shanxi</td>
<td>38.69</td>
<td>hainan</td>
<td>37.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hubei</td>
<td>38.67</td>
<td>liaoning</td>
<td>36.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anhui</td>
<td>38.62</td>
<td>heilongjiang</td>
<td>36.51</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How to compare the frequency of extreme events at two time points $t_2 > t_1$?

One way is to look at the ratio of two return levels, e.g.,

$$\frac{\text{20 year return level at time } t_1}{\text{Corresponding RL at } t_2} = (RL_{t_1}) P(Y_{t_2} > RL_{t_1}).$$

For example, if the 20 year return level at time $t_1$ is $RL_{t_1} = 38$, which becomes a 5 year return level at $t_2$, we say it has increased 4 times. For a particular region (e.g., province), average the ratios over all stations in the region.
<table>
<thead>
<tr>
<th>Province</th>
<th>Return level ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shanghai</td>
<td>19.31</td>
</tr>
<tr>
<td>Sichuan</td>
<td>7.87</td>
</tr>
<tr>
<td>Beijing</td>
<td>7.17</td>
</tr>
<tr>
<td>Gansu</td>
<td>5.40</td>
</tr>
<tr>
<td>Hainan</td>
<td>5.31</td>
</tr>
<tr>
<td>Xizang</td>
<td>5.24</td>
</tr>
<tr>
<td>Shanxi</td>
<td>5.05</td>
</tr>
<tr>
<td>Tianjin</td>
<td>4.29</td>
</tr>
<tr>
<td>Shaanxi</td>
<td>3.92</td>
</tr>
<tr>
<td>Guizhou</td>
<td>3.46</td>
</tr>
</tbody>
</table>
Invasion cover and biomass are observed at 18108 sites. One research problem is to study the maximum invasion cover for a given biomass.

Exactly how do we define a block maximum?

What is a block?
Divide the biomass into intervals each of which has about 100 observations. We will model the maximum of invasion cover within each interval, whose distributional parameters are functions of the center of the interval.
Example 2: Invasion Cover

Model the scale as the quadratic function of biomass, $\sigma$ and $\xi$ as constants.

![Graph showing the relationship between biomass and invasion cover](image-url)
Modeling Frequency and Magnitude Separately

Realization of the 2-dimensional Poisson Process

Earthquake magnitudes and the threshold

Hao Zhang (Purdue University)  zhanghao@purdue.edu
An Issue with Current Model

Current models imply that

\[ p_n = P(X_i > u) \approx \frac{1}{n} \left[ 1 + \xi \left( \frac{u - \mu}{\sigma} \right) \right]^{-1/\xi} \]

\[ z_p = \begin{cases} 
\mu - \frac{\sigma}{\xi} [1 - \{ - \log(1 - p) \}^{-\xi}], & \text{for } \xi \neq 0 \\
\mu - \sigma \log(- \log(1 - p)), & \text{for } \xi = 0 
\end{cases} \]

- The probability \( p_n \) has to do with frequency of the extreme event—The larger \( p_n \) is, the more frequent the extreme event would be. The quantile \( z_p \) has to do with magnitude.
- The models result in these two quantities moving in the same direction.
- There are situations where events may become more extreme in magnitude but not in frequency. A new model?
Charkraborty and Zhang (2015):

Treat \( \{ t/(1 + T) : M_t > u, \, t = 1, \ldots, T \} \) as an inhomogeneous Poisson point process and model its intensity as \( \lambda(t, \theta) \) where \( \theta \) is the parameter.