Overview of Extreme Value Analysis (EVA)

Brian Reich
North Carolina State University
July 26, 2016

Rossbypalooza
Chicago, IL
Importance of extremes in climate research

- From heat waves to hurricanes, often the environmental processes that are the most critical to understand probabilistically are extreme events.

- There is a large literature on EVA.

- There are some beautiful mathematical results.

- The statistical methodology is unique.

- There are many open statistical problems.
Common objectives in EVA

- Estimate the 1,000 year return level, i.e., the value that occurs on average once every 1,000 years

- Identify environmental covariates that drive extremes

- Test the hypothesis that the likelihood of an extreme event is changing over time

- Determine if two locations are asymptotically dependent

- Project the change in the 99.9th percentile in 2050
Unique statistical challenges

- Most of our intuition and methodology are built around the mean and deviation from the mean.

- In EVA concepts like mean and variance are irrelevant because they don’t speak to the tail of the distribution.

- Similarly, correlation isn’t the best measure of dependence because it is based on deviation around the means.

- We need new ways to describe distributions and dependence between random variables.
Isolating the extremes

- The first step in classic EVA is to separate extreme observations from the bulk of the distribution.

- For example, in a daily time series of precipitation in FL, these correspond to very different weather regimes.

- Bulk: Thunderstorms

- Tails: Hurricanes

- Mean regression focus on thunderstorms and treats hurricanes as outliers.

- If you want to estimate the 100-year storm, you should focus only on hurricanes.

- Two common ways to isolate the extremes: block maxima and points above a threshold.
Block maxima (BM) in Cheeseboro

Block is a year and the block maximum is annual maximum.
Points above a threshold (POT) for Cheeseboro
The *threshold* is 50 and we analyze the points in red.

![Hourly Wind Speed Graph](image-url)
Pros/cons of analyzing BM

Pros:
- Can evoke EVA theory and use a simple model
- It removes dependence with block

Cons:
- Excludes some large values (second highest each year)
- Must pick the block size: too big and you lose data; too small and you can’t use EVA theory
Pros/cons of analyzing POT

Pros:
- Can evoke EVA theory and use a simple model
- Retains all large values in the analysis

Cons:
- Must deal with dependence within block
- Setting the threshold is really difficult
Let $Y_1, \ldots, Y_n$ be the $n$ independent and identically distributed values in a block (say $n = 365$ days in a year)

In certain conditions, for large $n$ the sample (annual) mean

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

is approximately normally distributed

Holds with some forms of dependence and nonstationarity

The underlying data $Y_i$ do not have to be Gaussian

For example, the mean of 10 uniforms is $\approx$ normal

So if you are analyzing data that are constructed as means, then a normal distribution is a good start

You should still check the assumption of normality
A similar result holds for the block maximum

\[ \tilde{Y}_n = \max\{ Y_1, \ldots, Y_n \} \]

Under certain conditions, for large \( n \) \( \tilde{Y}_n \) approximately follows the Generalized Extreme Value (GEV) distribution.

Holds with some forms of dependence and nonstationarity.

So if you are analyzing data that are constructed as say annual maximums, then GEV is a good start.

You should still of course check the GEV fit.
BM: GEV distribution

- The GEV has three parameters:
  - Location: $\mu$
  - Scale: $\sigma > 0$
  - Shape: $\xi$

- The shape defines three special cases:
  - Weibull: $\xi < 0$ and the distribution is bounded above
  - Gumbel: $\xi = 0$ and the distribution is unbounded
  - Frechet: $\xi > 0$ and the distribution is bounded below

Explore the shape of the distribution: http://teaching.stat.ncsu.edu/shiny/bjreich/GEV/
Say $Y_1, \ldots, Y_m \overset{iid}{\sim} \text{GEV} (\mu, \sigma, \xi)$

Estimates of the three GEV parameters and the standard errors can be obtained with usual MLE

The \texttt{fgev} package in \texttt{R} does this

\textbf{CLIMDEX example:} \url{http://www4.stat.ncsu.edu/~reich/Rossby/CLIMDEX_GEV.html}
BM: Model checking

- Just because data are block maxima doesn’t necessarily mean they fit the GEV perfectly

- Why?

- QQ-plots are a good diagnostic

- KS goodness-of-fit tests can be constructed
Say the data are annual maxima

The $n$-year return level is the value exceeded once every $1/n$ years

This is the $1 - 1/n$ quantile of the GEV distribution

In $\mathbb{R}$, the $n$-year return level is

$$ RLn = qgev(1-1/n, \mu.\text{est}, \sigma.\text{est}, \xi.\text{est}) $$

Standard errors account for uncertainty in the GEV parameters, and can be found using the delta method
Until now we have assumed the distribution is stationary, i.e., constant over time.

However, the GEV parameters (usually $\mu$, but sometimes $\sigma$ and $\xi$) can vary with time.

**Linear GEV location:** $Y_t \sim \text{GEV}(\beta_0 + t\beta_1, \sigma, \xi)$

**Add a covariate:** $Y_t \sim \text{GEV}(\beta_0 + t\beta_1 + X_t\beta_2, \sigma, \xi)$

**Linear GEV location and scale:**

$Y_t \sim \text{GEV}([\beta_0 + t\beta_1, \exp(\alpha_0 + t\alpha_1), \xi)]$

These models can be fit in `evd/fgev` using MLE.
A POT analysis begins by selecting a threshold $T$ that separates the bulk from the extremes.

The dataset for analysis then becomes only the values that exceed $T$.

For most distributions, for large enough $T$ the tail of the distribution matches the Generalized Pareto Distribution (GPD).

Picking the threshold too low leads to bias because the GPD doesn’t fit well.

Picking the threshold too high leads to high variance because the number of observations is small.
POT: GDP distribution

- The GPD has three parameters:
  - Location/lower bound/threshold: $T$
  - Scale: $\sigma > 0$
  - Shape: $\xi$

- The shape defines three special cases:
  - $\xi < 0$ and the distribution is bounded above
  - $\xi > 0$ and the distribution is unbounded

Say we set the threshold at $T$ and $Y_1, \ldots, Y_{m_T}$ are the $m_T$ observations above $T$.

The model is $Y_1, \ldots, Y_{m_T} \overset{iid}{\sim} \text{GPD}(T, \sigma_T, \xi_T)$.

Estimates of the two GPD parameters $\sigma_T$ and $\xi_T$ and the standard errors can be obtained with usual MLE.

The `fpot` package in R does this.

CLIMDEX example: [http://www4.stat.ncsu.edu/~reich/Rossby/CLIMDEX_GPD.html](http://www4.stat.ncsu.edu/~reich/Rossby/CLIMDEX_GPD.html)
POT: Model checking

- The biggest challenge is picking the threshold $T$

- For the GPD, for any $u > T$ the mean residual life is:
  \[ E(Y - u | Y > u) = \frac{\sigma}{1 - \xi} + \frac{\xi}{1 - \xi} u \]

- The mean residual life (MRL) plot plots the sample mean estimate of $E(Y - u | Y > u)$ versus $u$

- The smallest $u$ which the MRL plots is linear above $u$ is a reasonable threshold choice

- In R/evd this is the \texttt{mrlplot} function
POT: Return levels

- Now the data are daily data
- The $n$-year return level is the value exceeded once every $1/n$ years, which is $1/(365n)$ days
- Let $\rho_T$ be the probability below the threshold
- On a given day the probability of being below $u > T$ is $\rho_T + (1 - \rho_T)F_{GPD}(u)$
- In $\mathbb{R}$, the $n$-year return level is
  $$q = 1 - \frac{1/(365n) - \rho_T}{1 - \rho_T}$$
  $$R_{Ln} = qgpd(q, \text{thresh}, \text{sigma.est}, \text{xi.est})$$
- Standard errors account for uncertainty in the GPD parameters, and can be found using the delta method
As with the GEV, the GPD parameters can vary over space and time following covariates.

Allowing the threshold to vary with covariates is probably a good idea, but really tricky.

Another departure for the simple model assumptions is serial dependence in the daily data.

This can be handled by declustering.

For example, if 5 consecutive days exceed the threshold then only the largest is retained.

Declustering is implemented in fpot.
Other extensions

- Multivariate extremes
- Time series of extremes and heat waves
- Spatial extremes
- Detection and attribution
- Methods to handle large $n$
- Many more!
Resources

▶ Book on applied EVA: Coles (2001)

▶ Book on theory: de Haan and Ferreira (2006)

▶ Book on recent methods: Dey and Yan (2016)

▶ More computing in R: evd; extRemes; SpatialExtremes

▶ My info: http://www4.stat.ncsu.edu/~reich/