Data Analytics for Coastal Systems

Georgios Boumis, Ph.D.
Assistant Professor
Department of Civil and Environmental Engineering
The University of Maine

Abstract. This course introduces statistical methods for analyzing environmental data, where students develop R programming skills through hands-on work on processing real-world data from Maine's coastal environments. Emphasis is given in datasets such as sea level observations, temperature and sea-surface pressure measurements, as well as seasonal wind speed records. The course advances systematically from basic data visualization and descriptive statistics to predictive modeling, probability distribution fitting via maximum likelihood estimation, and parametric uncertainty quantification. It culminates with the application of extreme-value theory and trend detection for evaluating rare coastal events, particularly within the context of non-stationarity due climate change-driven sea-level rise.

Contents

7 Extreme-Value Theory #2

In the previous chapter, we saw that the Generalized Extreme Value (GEV) distribution can be used to statistically model annual maximum measurements $(z_1, z_2, \ldots, z_n)$. We saw that extracting block maxima, in general, is one of the two ways to define extreme values in our sample $(x_1, x_2, \ldots, x_n)$. The other way to extract extreme values that we briefly mentioned is the **peaks-over-a-threshold** approach. In this chapter, we will focus on probability distributions that allow us to model "$Y = X - u$" as a random variable, where u is a pre-defined threshold such that $Y > 0 \implies X > u$.

7.1 Generalized Pareto (GP) Distribution

Extreme-value theory suggests that for $i.i.d.$ peaks-over-a-threshold ($y_j = x_j - u$), where $y_j > 0$, the **Generalized Pareto (GP)** distribution is a suitable candidate distribution. An example of peaks-over-a-threshold for our storm surge data at Portland, ME, is shown in Figure 1. The cumulative distribution function (cdf) of the GP distribution is given by the following formula:

$$F(y) = 1 - (1 + \xi \frac{y}{\sigma})_+^{-1/\xi}, \tag{1}$$

where $\sigma > 0$ is the scale parameter, while ξ denotes the shape parameter. Usually, $-1 < \xi < 1$. The GP distribution reduces to a simple **Exponential** distribution with parameter $1/\sigma$ for the limiting case where $\xi \to 0$:

$$F(y) = 1 - \exp(-y/\sigma). \tag{2}$$

When dealing with peaks-over-a-threshold, the **return period (T)**, i.e., the *average* waiting time (years) until y is exceeded, is given by:

$$T = \frac{1}{\lambda p}, \tag{3}$$

where λ is the **recurrence rate**, that is, the average number of threshold exceedances per year, while $p = P(Y > y) = P(X - u > x - u) = P(X > x)$.

Note that when we work with annual maxima and the GEV distribution instead, by definition, $\lambda = 1$, and therefore p is the annual exceedance probability. However, when we work with peaks-over-a-threshold and the GP distribution, the annual exceedance probability is now λp, which accounts for the fact that, on average, we have multiple exceedances per year. This small clarification is an important difference between the two models, but it should not change our general interpretation of the return period: the event with, e.g., a return period of $T = 100$ years is expected to occur, on average, every 100 years, or else, it has a 1% chance of occurring in any given year.

The return level y_T associated with return period T can be computed by inverting either Equation 1 or 2, noting that $p = 1 - F(y)$, as follows:

$$y_T = \begin{cases} \frac{\sigma}{\xi}((\lambda T)^{\xi} - 1) & \text{for } \xi \neq 0 \\ \sigma \log(\lambda T) & \text{for } \xi \to 0. \end{cases} \tag{4}$$

In practice, after we estimate the parameters (σ and ξ) of the GP distribution with Maximum Likelihood Estimation (MLE), we can compute any return level of interest x_T with return period T, by simply using Equation 4 and adding back the threshold u (since $y_T = x_T - u$).

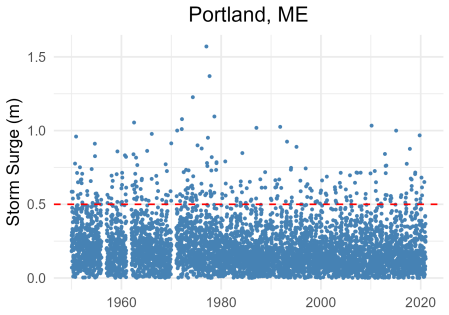

Figure 1: Peaks-over-a-threshold ($u = 0.50$ m) at Portland, ME. Note that independency has been ensured by first extracting the maximum hourly storm surge measurements within +/- 3-day windows.

7.2 Short Note: Properties of the GP Distribution

Before we see how we can choose an appropriate threshold, we need to know two very important properties of the GP distribution:

1) Let us assume that the GP distribution is an appropriate model for peaks-over-a-threshold $Y = X - u$, i.e., $Y \sim GP(\sigma_u, \xi)$. Then, for any other threshold v for which $v > u$ the respective peaks $(Y^* = X - v)$ also follow a GP distribution with scale $\sigma_v = \sigma_u + \xi(v-u)$ and *same* shape ξ, i.e., $Y^* \sim GP(\sigma_v, \xi)$.

2) Let us assume that the GP distribution is an appropriate model for peaks-over-a-threshold $Y = X - u$, i.e., $Y \sim GP(\sigma_u, \xi)$. Then, the expectation of Y^* (as defined above) is $E[Y^*] = \frac{\sigma_u + \xi(v-u)}{1-\xi}$. In other words, the expectation is linear with respect to the threshold since $E[Y^*] = \frac{\sigma_u - \xi u}{1-\xi} + \frac{\xi}{1-\xi} v$. An empirical estimate of $E[Y^*]$ is given by $\bar{y}^* = \frac{\sum_{j=1}^{n_v}(x_j - v)}{n_v}$ (known as mean excess).

7.3 Threshold (u) Selection

Based on the properties above, there are two ways to select an appropriate threshold for the GP distribution:

1) According to the first property, we could iteratively fit a GP to our data using a different threshold each time, and finally select the lowest threshold after which the shape parameter (ξ) remains approximately constant.

2) According to the second property, we could create a plot, where on the x-axis we show the threshold (v) while on the y-axis we show the respective mean excess $\bar{y}^* = \frac{\sum_{j=1}^{n_v}}{n_v}(x_j - v)$. Then, we could "study" the plot and select the lowest threshold after which the plot appears to be linear. Such a plot is called the **Mean Residual Life (MRL)** plot. The MRL plot for our storm surge data at Portland, ME, is shown in Figure 2. *Note that studying the MRL plot and selecting a threshold based on it involves great subjectivity!*

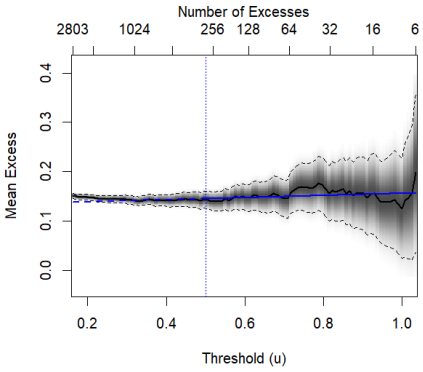

Figure 2: Mean Residual Life plot for our storm surge data at Portland, ME, with a threshold of $u = 0.50$ m marked with a dashed blue line.

7.4 Quantile-to-Quantile (QQ) Plots

Similar to the probability plots we saw in **Chapter 6**, quantile-to-quantile (QQ) plots are just another way to assess the goodness-of-fit of our distribution. Imagine that we have fit a GEV distribution with MLE to the annual maximum (z) storm surge data at Portland, ME, as we did in the previous class. Instead of comparing $\hat{F}(z)$ with $\tilde{F}(z)$, i.e., instead of comparing the model probability with the empirical probability, we now compare the model quantile $\hat{z}_i = \hat{F}^{-1}(\frac{i}{n+1})$ with the actual quantile z_i. A QQ plot is therefore now a scatterplot between $\hat{z}_i$ and z_i. If the points lie on the unit diagonal, or else, the "1-1" line, then this is a good indication that the GEV model fits our data well. Figure 3 shows such a QQ plot obtained after fitting a GEV distribution to annual maximum storm surge data at Portland, ME. *It is important to remember that the probability plot and QQ plot display identical information, just with different axis scales. However, since scale affects how we visually interpret patterns, a model fit that appears adequate on one plot may seem inadequate on the other.*

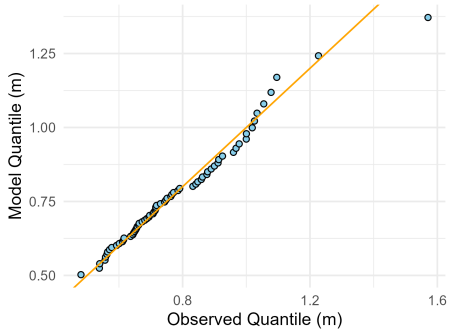

Figure 3: Quantile-to-quantile plot obtained after fitting a GEV distribution to annual maximum storm surge data at Portland, ME. The orange line indicates the unit diagonal, or else, the "1-1" line.