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

6 Extreme-Value Theory #1

In the last R programming exercise, we worked with annual maximum storm surge data from Portland, ME. These data are shown in Figure 1 for 68 years from 1950 to 2020 (with 3 missing years). We saw that for such type of data, both the Gamma and Gaussian distributions do not provide an adequate "fit". For data which can be considered extremal compared to the bulk of the sample measurements we have collected, better tools are needed, i.e., specific probability distributions tailored to capturing the behavior of such values. **Extreme-value theory (EVT)** provides a robust statistical framework to model extreme events, which are rare measurements located in the **tails** of the underlying distribution (Figure 2) - EVT also allows us to extrapolate beyond our measured data to unobserved extremal values. In general, there are two ways to define extreme values: 1) when time is divided into predefined blocks, commonly years, the extreme values are the maximum measurements recorded within each block, and 2) when a threshold is selected to separate extreme from typical values, the extreme values consist of all measurements exceeding that threshold.

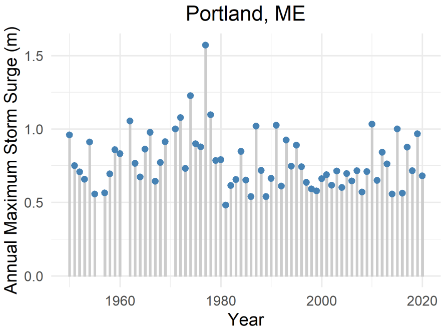

Figure 1: Annual maximum storm surge data at Portland, ME.

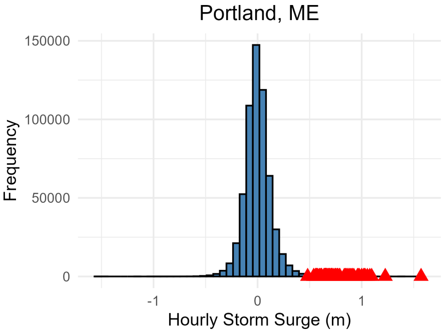

Figure 2: Histogram of hourly storm surge data at Portland, ME, between 1950 and 2020. Red triangles indicate the values of the annual maximum storm surge data.

6.1 Generalized Extreme Value (GEV) Distribution

EVT suggests that for $i.i.d.$ extreme values that originate from block maxima, e.g., annual maxima $(z_1, z_2, \ldots, z_n)$, the **Generalized Extreme Value (GEV)** distribution is a suitable candidate distribution. The cumulative distribution function (cdf) of the GEV distribution is given by the following formula:

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

where $-\infty < \mu < \infty$ is the location parameter, $\sigma > 0$ denotes the scale parameter, and ξ is the shape parameter. The shape parameter is typically $-1 < \xi < 1$. In reality, the GEV distribution is not a single distribution, but rather a family of distributions. Specifically, when $\xi > 0$ we call it Fréchet type, while when $\xi < 0$ we call it Weibull type. For the limiting case where $\xi \to 0$, the GEV distribution reduces to the Gumbel type:

$$F(z) = \exp(-\exp(-(\frac{z-\mu}{\sigma}))). \tag{2}$$

Figure 3 shows the pdf and the cdf of a GEV distribution with $\mu = 0.69$, $\sigma = 0.14$, and $\xi = 0.08$. Recall from **Chapter 3** that the cdf (F) of a continuous random variable Z is defined as:

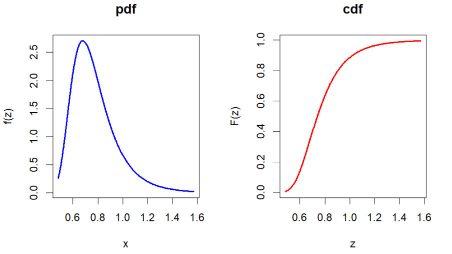

Figure 3: Probability density function (pdf) and cumulative distribution function (cdf) of a GEV distribution with $\mu = 0.69$, $\sigma = 0.14$, and $\xi = 0.08$.

$$F(z) = P(Z \leq z), \tag{3}$$

and is connected to the idea of the probability density function (pdf, also denoted as f) as follows:

$$F(z) = \int_{-\infty}^{z} f(v)dv \implies f(z) = \frac{dF}{dz}. \tag{4}$$

Given that probability is always between 0 and 1, recall also the following property:

$$\lim_{z\to\infty} F(z) = 1 \implies P(Z \geq z) = 1 - P(Z \leq z) = 1 - F(z). \tag{5}$$

When working with annual maxima $(z_1, z_2, \ldots, z_n)$, the quantity $p = P(Z \geq z)$ gives the probability that, in any given year, the maximum value will be greater than z, i.e., the so-called **exceedance probability** of z. The **return**

period (T), or else, **recurrence interval**, is the *average* waiting time (years) until z is exceeded, and is expressed as:

$$T = \frac{1}{p} = \frac{1}{P(Z \geq z)} = \frac{1}{1 - P(Z \leq z)} = \frac{1}{1 - F(z)}. \tag{6}$$

As an example, the event with a recurrence interval of 100 years, also known as the 100-year **return level**, has a probability of exceedance in any given year equal to:

$$p = \frac{1}{T} = \frac{1}{100} = 0.01 \quad \text{or} \quad 1\%. \tag{7}$$

Similarly, the 5-year return level, or the event with a return period of 5 years, has an annual probability of exceedance equal to:

$$p = \frac{1}{T} = \frac{1}{5} = 0.2 \quad \text{or} \quad 20\%. \tag{8}$$

The return level associated with an annual probability of exceedance p can be computed by inverting either Equation 1 or 2 (depending on the type of the GEV distribution):

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}(1 - (-\log(1-p))^{-\xi}) & \text{for } \xi \neq 0 \\ \mu - \sigma \log(-\log(1-p)) & \text{for } \xi \to 0. \end{cases} \tag{9}$$

In other words, after we estimate the parameters (μ, σ, and ξ) of the GEV distribution with Maximum Likelihood Estimation (MLE), we can compute any return level of interest with annual probability of exceedance p.

6.2 Short Note: The Fallacy of the Return Period

What is the probability of at least one exceedance of the 100-year return level in 100 years?

For the 100-year return level, every year there is a 1% chance of exceedance, or else, a 99% chance of non-exceedance. Therefore:

$$P(\text{at least one exceedance in 100 years}) = 1 - P(\text{no exceedance in 100 years}), \tag{10}$$

$$P(\text{at least one exceedance in 100 years}) = 1 - (0.99)^{100} \approx 1 - 0.366 \approx 0.634. \tag{11}$$

Many people would intuitively think that a 100-year event should happen exactly once every 100 years, giving a probability of exceedance equal to 1 (100% chance). However, as mentioned before, the return period is the *average* waiting time between exceedances of z, and because events are random, we could have zero, one, or multiple exceedances of the 100-year event within 100 years. In fact, as we computed above, there is $\sim 63\%$ chance that we observe a value greater than or equal to the 100-year event in 100 years!

6.3 Example: Probability of Seawall Overtopping

Let us now consider that we are designing a seawall for a coastal area so that the wall can provide protection from an extreme sea level equal to the 100-year return level. Let us also assume that the seawall has a **design lifetime** of 50 years, i.e., the construction materials of the wall will significantly deteriorate over time, thus it will require reconstruction after 50 years. What is the probability that we observe a sea level greater than the 100-year return level leading to seawall overtopping during its design lifetime? Following the same logic as in Section 6.2, this probability can be computed as follows:

$$P(\textbf{overtopping}) = 1 - (1 - \frac{1}{T})^D = 1 - (1 - \frac{1}{100})^{50} \approx 0.39, \qquad (12)$$

where T is the return period (years) of the return level for which we are designing our structure, while D is the design lifetime (years) of the structure.

6.4 Probability Plots

Imagine that we have fit a GEV distribution with MLE to the annual maximum storm surge data at Portland, ME (Figure 1). For each annual maximum (z) we can then obtain the respective estimate of the probability of non-exceedance $\hat{F}(z)$ with the use of Equation 1. A way to assess the goodness-of-fit of the GEV distribution is to compare $\hat{F}(z)$ with an **empirical** estimate of $F(z)$. The latter can be obtained as follows: let $z_1, z_2, \ldots, z_n$ denote the **ordered** annual maximum storm surge data so that $z_1 \leq z_2 \leq \cdots \leq z_n$. Then, for any z_i there are exactly i measurements which are less than or equal to z_i and therefore an empirical estimate of the non-exceedance probability is $\tilde{F} = i/n$. To avoid the situation where $\tilde{F}(z_n) = 1$, we make an adjustment so that $\tilde{F}(z) = i/(n+1)$. A probability plot is a scatterplot between $\hat{F}(z)$ and $\tilde{F}(z)$. If the points lie on the unit diagonal, or else, the "1-1" line, then this is a good indication that the GEV fits our data well. Figure 4 shows such a probability plot obtained after fitting a GEV distribution to annual maximum storm surge data at Portland, ME.

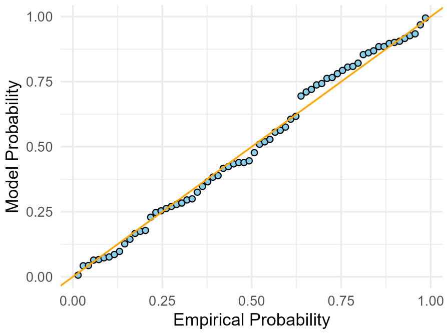

Figure 4: Probability plot obtained after fitting a GEV distribution to annual maximum storm surge data at Portland, ME. The orange line indicates the unit diagonal, or else, the "1-1" line.