TP1 Statistical learning with extreme values

Anne Sabourin, Antoine Doizé

October, 9, 2025

Objectives. At the end of this lab (parts 0–2), you will know how to : (i) fit a GEV to annual maxima (Port Pirie), (ii) choose a threshold and fit a GPD for exceedances (Rain), (iii) read the main diagnostic plots and estimate return levels.

— Part 0-1-2 will be done in class.

— Part 3 will be your homework.

Brief theoretical reminders

GEV (block maxima). If $M_n = \max(X_1, \ldots, X_n)$ and if there exist $a_n > 0, b_n \in \mathbb{R}$ such that $\Pr\,(M_n - b_n)/a_n \leq z \to G(z)$ non-degenerate, then G is of GEV type :

$$G(z) = \exp\left\{-\left[1 + \gamma \tfrac{z-\mu}{\sigma}\right]^{-1/\gamma}\right\}, \qquad 1 + \gamma \tfrac{z-\mu}{\sigma} > 0.$$

Limit cases : $\gamma = 0$ (Gumbel), $\gamma > 0$ (Fréchet), $\gamma < 0$ (Weibull).

Return levels (GEV). For an annual probability p (e.g. $p = 1/T$),

$$z_p = \mu + \frac{\sigma}{\gamma}\big([-\log(1-p)]^{-\gamma} - 1\big) \quad (\gamma \neq 0), \qquad z_p = \mu + \sigma \log\!\left(\frac{1}{-\log(1-p)}\right) \quad (\gamma = 0).$$

POT/GPD (exceedances). Let x_F be the right endpoint of F, the c.d.f. of X. For $u < x_F$,

$$F_u(y) := \Pr(X - u \leq y \mid X > u) \xrightarrow[u \uparrow x_F]{} 1 - (1 + \gamma y/\beta(u))^{-1/\gamma},$$

for $y \geq 0$ with $1 + \gamma y/\beta(u) > 0$ (i.e., $F_u \Rightarrow \mathrm{GPD}(\gamma, \beta(u))$).

Part 0 — Setup (5 min)

1. Create a Python environment with `numpy`, `scipy`, `matplotlib`, `pandas`.
2. We will use the public datasets *Port Pirie* (annual maxima of sea level) and *Rain* (daily rainfall, SW England), provided as *.csv* files.

NB. SciPy notations are `genextreme(c, loc, scale)` with $c = -\gamma$, and `genpareto(c, loc, scale)` with $c = \gamma$, scale $= \beta$.

Part 1 — Annual maxima (Port Pirie) GEV (35 min)

1A. Data reading and visualization

1. Load the CSV file *portpirie* (columns : `year`, `sea`).
2. Plot $(\text{year}, \text{annual maximum})$ and comment : trend, dispersion, visually extreme values.

1B. GEV fit by MLE

1. Fit a GEV via `scipy.stats.genextreme.fit`. Reminder : returns (c, μ, σ) with $\hat{\gamma} = -c$.
2. Report $(\hat{\mu}, \hat{\sigma}, \hat{\gamma})$ and interpret the sign of $\hat{\gamma}$ (heavy/bounded/light tail).

1C. Diagnostics

1. QQ-plot : empirical quantiles vs theoretical quantiles of the fitted GEV.
2. *Return-level plot* : plot z_T for $T \in 2, 5, 10, 20, 50, 100$ (logarithmic horizontal scale). Comment on the shape and the plausibility of the extrapolations.

Part 2 — Exceedances above a threshold (Rain) GPD (45 min)

2A. Exploration

1. Load the CSV *rain* (column : `rain_mm`). Visualize a snippet of the series and the overall histogram.
2. Discuss the presence of a right tail (intense but rare rainfall values).

2B. Threshold choice

1. Plot the *Mean Residual Life* curve $u \mapsto \mathbb{E}[X - u \mid X > u]$ for a grid of quantiles (0.80–0.999) and the number of exceedances.
2. Identify an *approximately linear* zone with enough exceedances.
3. Check the stability of the parameters $(\hat{\gamma}, \hat{\beta})$ for several candidate thresholds.

2C. GPD fit and diagnostics

1. Fix u : $u = 30$ mm). Fit a GPD on the exceedances $X - u$.
2. Display the GPD QQ-plot and comment on the fit.
3. Estimate the annual return levels : convert T years into $T_{obs} = 365, T$ and use $\hat{\lambda} = N_u/N$. The formula is

$$z_T = u + \frac{\beta}{\gamma}(\lambda T_{obs})^{\gamma} - 1 \ (\gamma \neq 0), \qquad z_T = u + \beta \log(\lambda T_{obs}) \ (\gamma = 0).$$

General advice. Always check adequacy (QQ-plot, return-level, MRL, stability) before strong extrapolation. Document the threshold choice and discuss uncertainty.