G3 – SDI
2022–2023

Bayesian Analysis
Case study of a dike construction

1 Problem Statement

The flow for the Garonne river is very variable. In the past, may floods occurred. It has been decided to build a dike to protect the surroundings of the bank near a specific point called *Mas d'Agenais* (`http://fr.wikipedia.org/wiki/Le_Mas-d'Agenais`). The question under consideration is then to determine the height of this dike. In order to do so, flow data have been recorded there over several decades and are available. You are asked to answer to this question.

2 The data

The flow has been recorded each day from 1913 to 1977 (hence over 65 years). Experts consider a flood when the flow is higher than 2500 m^3s^{-1}. The figure 1 shows the flows for the year 1974. Several overflows can actually be observed. You can open the file *MA1974.csv* to see the original file from which these data have been extracted. This file can be downloaded from the database HYDRO available on Internet (`http://www.hydro.eaufrance.fr`). As only flows over 2500 m^3s^{-1} are of interest, the data collection that is now considered is the one built from the flows over this threshold. Then, for simplicity's sake, in the following, a data x referring to these last flows will only represent the part over the threshold. For instance, $x = 500$ means a flow equal to $2500 + 500 = 3000$ m^3s^{-1}.

Remark 1. *Data being flows we are going to derive the dike height as an equivalent flow d_h. Indeed experts can give an real height knowing the corresponding flow at a given point.*

3 Probabilistic Modeling

The model described below is called POT: Peaks over Threshold.

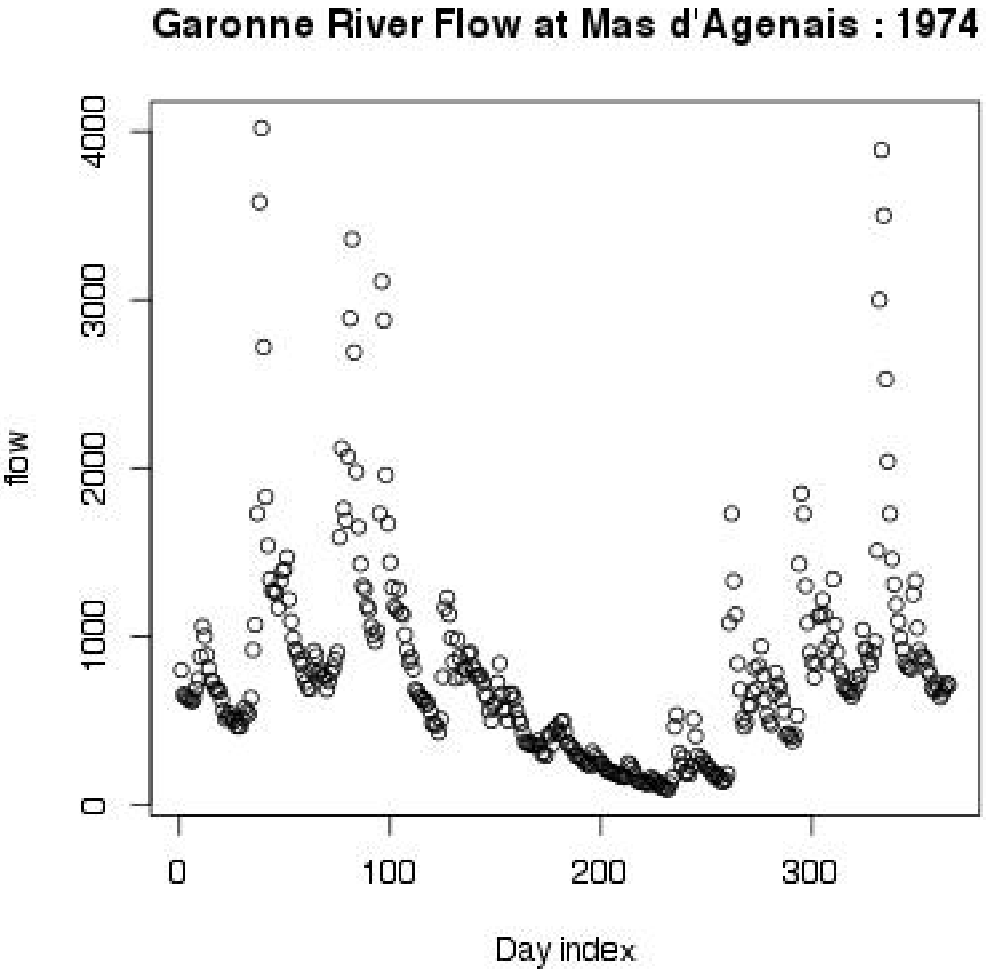

Figure 1: Flow data : Year 1974

3.1 Likelihood

Let us suppose that n independents floods have occurred in the past over r successive years. x_j ($j \in \{1, \dots, n)\}$) denotes the amplitude of the j^{th} flood (over the threshold 2500 m^3s^{-1}). k_i denotes the number of floods for year i ($i \in \{1, \dots, r)\}$). We thus have:

$$\sum_{i=1}^{r} k_i = n \tag{1}$$

The POT model considers that the number of floods occurring each year i is a Poisson random variable denoted K_i in the following. We thus have:

$$P(K_i = k|\mu) = \frac{\mu^k e^{-\mu}}{\Gamma(k+1)} \tag{2}$$

Remark 2. *We suppose here that the flood phenomenon is stationary. The same law is thus considered over the r years and therefore for the coming years. It results that the parameter μ is the the same each year.*

Remark 3. *From the properties of a Poisson random variable, μ is the average number of floods per year.*

Let us now consider the amplitude of a flood. The POT model supposes that each X_j (the random variable corresponding to the amplitude of a flood whose realization is denoted x_j) follows an exponential law:

$$P(X_j < x|\rho) = \int_0^x \rho e^{-\rho z}\, dz = 1 - e^{-\rho x} = G(x) \tag{3}$$

Remark 4. *As said above, the flood phenomenon being supposed stationary, the same law is considered over the r years. It results that the parameter ρ is the same each year.*

Remark 5. *From the properties of an exponential random variable, $\frac{1}{\rho}$ is the mean value of X_j and therefore of an overflow.*

We now make the hypothesis that the years are statistically independent. Let us call $\mathbf{x}$ the vector built from all the x_j. From the results and hypotheses above it comes:

$$p(\mathbf{x}|\mu, \rho) = \prod_{i=1}^{r} \frac{\mu^{k_i} e^{-\mu}}{\Gamma(k_i+1)} \prod_{j=1}^{n} \rho e^{-\rho x_j} \tag{4}$$

Using equation (1) this last equality can still be written:

$$p(\mathbf{x}|\mu, \rho) = \frac{\mu^n e^{-\mu r} \rho^n e^{-\rho \sum_{j=1}^{n} x_j}}{\prod_{i=1}^{r} \Gamma(k_i+1)} \tag{5}$$

Remark 6. *For those interested to go further into this modeling, the POT model is a specific case of a wider class of model related to extreme values modeling.*

3.2 Prior

The state of the world is here sum up by the parameter $\theta = [\mu, \rho]^T$. Both components of θ being positive, we choose a gamma prior for each. We thus have:

$$p(\mu) = \frac{c_\mu^{d_\mu}}{\Gamma(d_\mu)} \mu^{d_\mu - 1} e^{-\mu c_\mu} \tag{6}$$

$$p(\rho) = \frac{a_\rho^{b_\rho}}{\Gamma(b_\rho)} \rho^{b_\rho - 1} e^{-\rho a_\rho} \tag{7}$$

c_μ, d_μ, a_ρ and b_ρ are the hyper parameters of the problem. We suppose also that μ and ρ are independent.

4 Return Period

4.1 General results

Definition 1. *Let us call* $\{Y_1, Y_2, \ldots, Y_n, Y_{n+1}\}$ $n+1$ *random variables independent and identically distributed (i.i.d.). Let us now call* y_l *a given value (or threshold) and* $Z(y_l)$ *the random variable whose value is equal to the time range (given in index) between two times at which the* Y_j *are greater than* y_l*. We have:*

$$P(Z(y_l) = n) = P(Y_2 < y_l, \ldots, Y_n < y_l, Y_{n+1} > y_l | Y_1 > y_l) \tag{8}$$

The ***return period*** *is the expected value of* $Z(y_l)$*. It is therefore the mean time (given in index) between two times at which the* Y_j *are greater than* y_l*.*

We denote F the cumulative distribution function of each Y_i:

$$P(Y_i < y_l) = F(y_l) \tag{9}$$

Question 1. *Let us call* $T(y_l)$ *the return time. Show that :*

$$T(y_l) = \frac{1}{1 - F(y_l)} \tag{10}$$

4.2 Application to our problem

The previous result cannot applied straight since we usually have several floods per year. We therefore focus on the event *The height of the highest flood of the year* the random variables of which is the Y_i's introduced above. We first need to determine the cumulative distribution function F with respect to the probabilities introduced in the modeling section. For notation simplicity's sake, we *forget* the subscript i in the following since the random variables are *i.i.d.*.

Question 2. *Let us suppose there is s flood a year. Show that:*

$$P(Y < y|s) = G(y)^s \tag{11}$$

Question 3. *Deduce from the previous question that:*

$$F(y) = e^{-\mu(1-G(y))} = e^{-\mu e^{-\rho y}} \tag{12}$$

Hint 1. *s is a Poisson random variable (see equation (2))*

We suppose now that d_h is high enough and therefore $T(d_h)$ as well.

Question 4. *Show that :*

$$T(d_h) \simeq \frac{e^{\rho d_h}}{\mu} \tag{13}$$

5 Classical Decisional Making Model

5.1 Economical Cost

The direct economical cost linked to the dike construction are:

- If we decide to build a dike the height of which is d_h, we suppose that we invest $C(d_h)$ a year (annual deprecation).
- If a flood exceeds d_h, the damage is denoted $D(X, d_h)$ with X the random variable measuring the flow over *2500* m^3s^{-1}.

Question 5. *Show that the annual mean damage is then equal to :*

$$D_a = \mu \int_{d_h}^{+\infty} D(X, d_h)\rho e^{-\rho x} dx \tag{14}$$

Using (14) the annual total mean cost is equal to:

$$W(d_h|\rho, \mu) = C(d_h) + \mu \int_{d_h}^{+\infty} D(X, d_h)\rho e^{-\rho x} dx \tag{15}$$

5.2 A classical approach to d_h derivation

This approach supposes that $C(d_h)$ and $D(x, d_h)$ are linear functions:

$$C(d_h) = C_0 d_h \tag{16}$$

$$D(x, d_h) = D_0(x - d_h) \tag{17}$$

with C_0 and D_0 two known constant cost.

Question 6. *Let us call $d_h^\star$ the optimal height minimizing the annual total mean cost. Find a simple relation between $T(d_h^\star)$, C_0 and D_0.*

Question 7. *Give the value of $d_h^\star$ with respect to C_0, D_0, μ and ρ.*

Question 8. *Suppose that we want to build a protective dike against a once-a-century flood. What is the cost of the annual deprecation rate C_0 with respect to the cost of the damage rate D_0?*

Question 9. *At this stage, μ and ρ are unknown. We denote $S(n) = \sum_{j=1}^{n} x_j$. Using the results above, give a possible value of $d_h^\star$ (within a classical framework) with respect to C_0, D_0, n, r and $S(n)$.*

5.3 Numerical application

The data recorded in the HYDRO database are so that:

- $n = 155$
- $S(n) = 158164$

Question 10. *What the value of $d_h^\star$ corresponding to these data when we want to be protected against a once-a-century flood ?*

6 Bayesian Decisional Making Model

In the result above, uncertainty exists about μ and ρ. We now are going to use a Bayesian approach to take it into account. The priors have been defined in the section 3.2.

Question 11. *Derive the a posteriori probability density function $p(\mu, \rho|\mathbf{x})$. Show that it is the product of two gamma law whose parameters will be derived.*

Question 12. *Using above results, calculate the a posteriori cost $\mathbb{E}_{\mu,\rho|\mathbf{x}}[W(d_h, \rho, \mu)]$.*

Question 13. *Using the above question derive the new value of $d_h^\star$.*

Question 14. *Using Python, draw the probability density function of a gamma law when all the parameters tend towards zero. Which kind of prior do you recognize?*

Question 15. *Write the value of $d_h^\star$ in case of a non informative prior.*

6.1 Numerical application

Question 16. *What the new value of $d_h^\star$ when we want to be protected against a once-a-century flood ?*

Question 17. *Comment on all these results.*