Bayesian Credibility Model with heavy tail random variables: calibration of the prior and application to natural disasters and cyber insurance

Antoine Heranval, Olivier Lopez, Maud Thomas

Bayesian Credibility Model with heavy tail random variables: calibration of the prior and application to natural disasters and cyber insurance.

Antoine Heranval[1],Olivier Lopez[1], Maud Thomas[2]

January 29, 2024

Abstract

The Bayesian credibility approach is a method for evaluating a certain risk of a segment of a portfolio (such as policyholder or category of policyholders) by compensating for the lack of historical data through the use of a prior distribution. This prior distribution can be thought as a preliminary expertise, that gathers information on the target distribution. This paper describes a particular Bayesian credibility model that is well-suited for situations where collective data are available to compute the prior, and when the distribution of the variables are heavy-tailed. The credibility model we consider aims to obtain a heavy tailed distribution (namely a Generalized Pareto distribution) at a collective level and provides a closed formula to compute the credibility premium at an individual level. Two cases of application are presented: one related to natural disasters and the other to cyber insurance. In the former, a large database on flood events is used as the collective information to define the prior, which is then combined with individual observations at a city level. In the latter, a classical database on data leaks is used to fit a model for the volume of data exposed during a cyber incident, while the historical data on a given firm is taken into account to consider individual experience.

Key words: Bayesian credibility theory; extreme value analysis; natural disasters; cyber insurance.

Short title: Parametric insurance and extreme risks.

[1] CREST Laboratory, CNRS, Groupe des Écoles Nationales d'Économie et Statistique, Ecole Polytechnique, Institut Polytechnique de Paris, 5 avenue Henry Le Chatelier 91120 PALAISEAU, France

[2] Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005 Paris, France, E-mails: antoine.heranval@ensae.fr, olivier.lopez@ensae.fr, maud.thomas@sorbonne-universite.fr

1 Introduction

Insurers often face the challenge of dealing with scarce data when assessing extreme risks for pricing or reserving purposes. Due to the rarity of such events (otherwise insurability of the risk would be at stake), experience can only be gained at a collective level, leading to the use of structurally heterogeneous databases. This article presents a methodology for pricing (or, more generally analyzing and predicting) a risk whose consequences are particularly severe, using a combination of (few) individual information and collective data. By severe, we mean that the distributions of the random losses associated with the risk are heavy-tailed. Additionally, the events being targeted may not have occurred for a given policyholder or may have occurred only a few times. This history is taken into account to mitigate the prior evaluation of the claim.

Bayesian credibility theory [see e.g. Heilmann, 1989, Bühlmann and Gisler, 2005] is a classical way to deal with such issues. In this frameworks, a policyholder is represented by an unobserved risk factor, i.e. a hidden random variable whose distribution reflects the heterogeneity of the population. The prior distribution is computed from a preliminary analysis based on an experience made on the whole portfolio. The individual information is collected on a given policyholder and is then used to form a posterior distribution, leading to an estimate of the corresponding risk factor based on both individual and collective experience. Among many other use cases, credibility theory has recently found applications in agricultural insurance [see Zhu et al., 2019], in health insurance [see Chiroque-Solano and Moura, 2022], or in motor insurance [see Pechon et al., 2021], where credibility theory is used as an insurance pricing tool. Diao and Weng [2019] have adapted regression trees to the definition of credibility factors when covariates can be used to classify data to perform different levels of credibility modeling. A special focus on heavy tail distributions has been considered in [Chiroque-Solano and Moura, 2022, Gómez-Déniz et al., 2022, for example] where the authors present a family of distributions that have nice properties regarding to heavy tail modeling.

The approach considered in this paper is based on the fundamental result of Extreme Value Theory (EVT) by Balkema and de Haan [1974], Pickands [1975], which states that the tails of heavy-tailed distributions can be approximated by a Generalized Pareto (GP) distribution. Heavy-tailed distributions are a large class of distributions, known as the Fréchet domain, which includes usual distributions such as Student, log-gamma and Cauchy distributions. This GP fit can be observed in many situations where the losses are highly volatile, such as in the two applications we are considering, natural disasters [see e.g. Rohrbeck et al., 2018] and cyber [see e.g. Maillart and Sornette, 2010, Farkas et al., 2021b]. In this case, Bayesian framework is constrained by the fact that the collective data can be viewed as a mixture of individual distributions with different risk factors, which is expected to be distributed as a GP distribution. Furthermore, a simple model is preferred to enable explicit computation of the posterior

distribution and of the premium.

The rest of the paper is organized as follows: Section 2 provides a general overview of the methodology, starting with fundamental results in EVT. Then, it expresses our Bayesian credibility model and describes how to determine the prior from collective data. Section 3 and Section 4 illustrate two different use cases of the methodology in the context of cyber risk and of flood insurance, respectively.

2 Data and methodology

This paper analyzes the loss associated with a severe insurance claim, meaning that the distribution of the loss variable is heavy tailed. Section 2.1 reviews key results on EVT, and their consequences on the credibility approach being developed. The key idea is to consider the GP distribution as a mixture of exponential distributions, as shown in 2.2. Section 2.3 presents the Bayesian credibility model, while Section 2.4 presents the computations of priors.

2.1 Extreme value theory and Bayesian credibility

EVT is the branch of statistics developed to handle extreme events, such as extreme floods or extreme data breaches. EVT allows for prediction of risks of episodes outside of the observed range.

Consider independent and identically distributed (i.i.d) observations $Y_1, Y_2, \ldots$ with an unknown survival function $\overline{F}$ (that is $\overline{F}(y) = P(Y_1 > y)$). A natural way to define extreme events is to consider the values of Y_i that have exceeded some high threshold u. The excesses above u are then defined as the variables $Y_i - u$ given that $Y_i > u$. Extreme events are by definition located in the tail of the distribution and thus rare events. The asymptotic behavior of extreme events is characterized by the distribution of the excesses which is given by

$$\overline{F}_u(z) = P[Y_1 - u > z \mid Y_1 > u] = \frac{\overline{F}(u+z)}{\overline{F}(u)},\ z > 0\,.$$

Pickands [1975] showed that, if $\overline{F}$ satisfies the following property

$$\lim_{t\to\infty} \frac{\overline{F}(ty)}{\overline{F}(y)} = y^{-1/\gamma},\ \forall y > 0, \tag{2.1}$$

with $\gamma > 0$, then

$$\lim_{u\to\infty} \sup_{z>0} |\overline{F}_u(z) - \overline{H}(z;\sigma,\gamma)| = 0 \tag{2.2}$$

for some $\sigma > 0$ and $\overline{H}(\cdot;\sigma,\gamma)$ necessarily belongs to the GP distribution family whose distribution function is of the form

$$\overline{H}(z;\sigma,\gamma) = \left(1 + \gamma\frac{z}{\sigma}\right)^{-1/\gamma},\ z > 0,$$

where $\sigma > 0$ is a scale parameter and $\gamma > 0$ is a shape parameter, reflecting the heaviness of the tail of the distribution. In particular, if $\gamma \in (0, 1)$, the expectation of Y_1 is finite whereas if $\gamma \geq 1$ the expectation of Y_1 is infinite. More details on these results can be found in e.g. [Coles, 2001, Beirlant et al., 2004]. Note that in full generality, the shape parameter $\gamma \in \mathbb{R}$. However, the applications we have in mind, such as in Sections 3 and 4, concern cyber events with severe data breaches and natural catastrophes which fall into the domain of heavy-tailed distributions, that is distributions for which $\gamma > 0$. We therefore choose here to focus on the case $\gamma > 0$.

The so-called Peaks over Threshold (PoT) method, widely used [see Davison and Smith, 1990, Coles, 2001], consists in choosing a high threshold u and fitting a GP distribution to the excesses above this threshold u. The parameters σ and γ can be estimated by maximizing the GPD likelihood. The choice of the threshold u can be understood as a trade-off between bias and variance: the smaller the threshold, the less valid the asymptotic GP approximation, leading to a large bias; on the other hand, a threshold that is too large will generate few excesses to fit the model, leading to a high variance. In practice, threshold selection is a challenging task. Existing methods for choosing the threshold u rely on graphical diagnostics or computational approaches based on additional conditions (depending on unknown parameters) on the underlying distribution function F [see Scarrott and MacDonald, 2012]. However, it is worth mentioning that some recent works model the upper tail of GP distribution (with $\gamma > 0$) and the rest of the full distribution in one step, which allows to overcome the challenging issue of threshold selection [Tencaliec et al., 2020, Huang et al., 2019].

In Bayesian credibility theory, a policyholder is associated with a risk factor θ, distributed according to a prior distribution p. In the simplest framework, the individual losses experienced by this policyholder are assumed to be i.i.d. $(Y_1, ..., Y_n)$ conditionally on $\theta = t$, denoting f_t its density. The prior p is supposed to reflect the distribution of Y_1 without information about individual claims. More precisely, from p, one can retrieve the unconditional distribution of Y_1, whose density is given by the following integral $\int f_t(y)p(t)dt$. A collective database is distributed according to this mixture distribution. Thus, based on the results mentioned above, a credibility framework adapted to the context of extreme risks must ensure that the integral $\int f_t(y)p(t)dt$ corresponds to the density of a GP distribution. A special case is discussed in the next section.

2.2 Generalized Pareto distribution as mixture of exponential random variables

Consider that the risk factor θ follows a Gamma distribution, meaning that the prior density p is given by

$$p_{r,\lambda}(t) = \frac{\lambda^r}{\Gamma(r)} t^{r-1} \mathrm{e}^{-\lambda t} \mathbf{1}_{t \geq 0},$$

where Γ is the Gamma function, with parameters $r > 1$ and $\lambda > 0$. The GP distribution can be viewed as a Gamma-mixture of exponentially distributed random variables. Specifically, if $Y|\theta = t$ is assumed to be exponentially distributed with mean $1/t$, then

$$\mathbb{P}(Y \geq y) = \mathbb{E}\left[\mathbb{P}(Y \geq y|\theta)\right] = \int_0^\infty \mathrm{e}^{-ty} p_{r,\lambda}(t)dt = \left(\frac{\lambda}{\lambda + y}\right)^r,$$

which corresponds to a GP distribution with parameters

$$\sigma = \frac{\lambda}{r} \tag{2.3}$$

$$\gamma = \frac{1}{r}. \tag{2.4}$$

Returning to the credibility framework, suppose that a given policyholder generates the sequence of past claims $(Y_1, \cdots, Y_n)$ that are independent with exponential distribution of parameter t, conditionally to $\theta = t$. Then, assuming that θ is gamma (r, λ) distributed and that all policyholders are independent, the random vector $\mathbf{Z} = (Z_1, \cdots, Z_N)$ of all losses experienced by the insurer consist of identically distributed GP variables Z_i with parameters σ and γ satisfying (2.3) and (2.4). Strictly speaking, the vector $\mathbf{Z}$ is not i.i.d. since it potentially contains more than one claim generated by a single policyholder. However, if the size of the portfolio is large compared to the small depth of historical data, this effect can be neglected, and $\mathbf{Z}$ can be considered as i.i.d. This allows the parameters (σ, γ) to be estimated from the sample $\mathbf{Z}$.

We provide more details on the calibration of this prior distribution in Section 2.4, where we introduce the possibility of adding covariates to the analysis. We first explain how to derive the credibility premium and the posterior distribution from the knowledge of (σ, γ) and the policyholder's experience.

2.3 A simple Bayesian credibility model

Consider a sequence $(Y_1, ..., Y_n)$ representing the past claims of a given policyholder. As in the previous section, we assume that the hidden risk factor θ represents the unknown heterogeneity between policyholders, and is distributed according to a Gamma distribution with density $p_{r,\lambda}$. The random variables $(Y_i)_{1 \leq i \leq n}$ are assumed to be i.i.d. conditionally on θ, and, in the spirit of Section 2.2, we assume that $Y_i|\theta = t \sim \mathcal{E}(t)$, that is, exponentially distributed with mean $1/t$.

Simple calculations show that the posterior distribution of θ is a gamma distribution with parameters $(r + n, \lambda + \sum_{i=1}^n Y_i)$. If $\mathbb{E}[Y] < \infty$, we can compute the credibility (pure) premium from this posterior distribution. Note that, since Y is GP distributed from Section 2.2, the condition $\mathbb{E}[Y] < \infty$ is equivalent to $1/\gamma = r > 1$. In this case, the credibility premium is then

$$\pi_{r,\lambda}(Y_1, \cdots, Y_n) = \mathbb{E}_{r,\lambda}\left[Y_{n+1}|Y_1, \cdots Y_n\right] = \mathbb{E}\left[\frac{1}{\theta}|Y_1, \cdots, Y_n\right] = \frac{\lambda + \sum_{i=1}^n Y_i}{r + n - 1},$$

which can be rewritten as

$$\pi_{r,\lambda}(Y_1,\cdots,Y_n) = c_n(r)\frac{\sum_{i=1}^n Y_i}{n} + (1-c_n(r))\frac{\lambda}{r-1}, \tag{2.5}$$

introducing the credibility factor $c_n(r) = n[r+n-1]^{-1}$. Note that, if $r \leq 1$, the credibility premium is not defined (since the expectation is infinite), but the posterior distribution is still valid and can be used, for example, to derive appropriate quantiles.

For a given $\alpha \in (0,1)$, let $q^{\alpha}_{r,\lambda}(Y_1,\cdots,Y_n)$ denote the $(1-\alpha)$-quantile of the conditional distribution of Y_{n+1} given $(Y_1,\ldots,Y_n)$, that is

$$\mathbb{P}\left(Y_{n+1} \geq q^{\alpha}_{r,\lambda}(Y_1,\cdots,Y_n)|Y_1,\cdots,Y_n\right) = \alpha.$$

The conditional distribution of Y_{n+1} given $(Y_1,\cdots,Y_n)$ is a GP distribution with scale parameter $\lambda + \sum_{i=1}^n Y_i$ and shape parameter $(r+n)^{-1}$. This can be seen from the fact that, for $y > 0$,

$$\mathbb{P}(Y_{n+1} \geq y|Y_1,\cdots,Y_n) = \mathbb{E}[\mathrm{e}^{-\theta y}|Y_1,\cdots,Y_n] = \left(\frac{\lambda+\sum_{i=1}^n Y_i}{\lambda+\sum_{i=1}^n Y_i + y}\right)^{r+n}.$$

Hence,

$$q^{\alpha}_{r,\lambda}(Y_1,\cdots,Y_n) = \left(\sum_{i=1}^n Y_i + \lambda\right)\left(\alpha^{-\frac{1}{r+n}} - 1\right) = (r+n-1)\pi_{r,\lambda}(Y_1,\cdots,Y_n)\left(\alpha^{-\frac{1}{r+n}} - 1\right). \tag{2.6}$$

Returning to the case $r > 1$, the credibility premium is linear, and can be computed from a closed formula. The credibility factor $c_n(r)$ somehow reflects whether one can rely on the policyholder's historical data to correctly evaluate the risk, given the number of observations and the parameters of the prior. If n tends to be large, that is, if one has a long history of successive claims for the policyholder, this factor is close to 1. On the other hand, in the absence of history, the premium is equal to $\lambda/(r-1)$, which is the expectation of θ^{-1} from the prior distribution.

This Exponential/Gamma Bayesian model was chosen because it is compatible with the fact that the distribution of Y is heavy-tailed, which is identified with a GP distribution. Additionally, this models allows for a simple, computable formula for the credibility premium. In full generality, numerical approximation is required to obtain the posterior distribution in a Bayesian credibility model [see e.g. Najafabadi, 2010]. Interpreting the model outputs the outputs of the model can be challenging.

To choose the correct value of r and λ, the idea is to rely on the collective experience. A GP distribution is fitted to the sample $\mathbf{Z}$ of the previous section. The parameters σ and γ can be estimated using various techniques such as maximum likelihood or moments methods [See Beirlant et al., 2004, for more details]. In Section 2.4, both parameters are supposed to be functions of some covariates $\mathbf{X}$, estimated via regression trees adapted to the analysis of extreme events.

Remark 2.1 *Rewritten in terms of σ and γ, the credibility factor c_n becomes (with a slight abuse of notation)*

$$c_n(\gamma) = \frac{n}{\frac{1}{\gamma} + n - 1}.$$

When γ is close to 1, we see that the credibility factor tends to 1. In this case, the situation is so chaotic than the prior does not provide significant information, and relying on the empirical mean is more efficient. The opposite effect occurs when γ tends to zero. In this case, a more important set of observations is required to be confident about the diagnosis obtained from the empirical mean.

2.4 Introducing additional heterogeneity through covariates

We now want to take advantage of the fact that the policyholder also has characteristics $\mathbf{X} \in \mathbb{R}^d$ that can help to affect it to a particular group, with a particular risk. In other words, we want X to have impact on the prior distribution used to determine the credibility premium. We thus assume the existence of functions $\mathbf{x} \to r(\mathbf{x})$ and $\mathbf{x} \to \lambda(\mathbf{x})$ (and thus, functions $\mathbf{x} \to \sigma(\mathbf{x})$ and $\mathbf{x} \to \gamma(\mathbf{x})$) describing heterogeneity between classes of policyholders.

To calibrate the prior, we then assume that we have at our disposal $(Z_1, \mathbf{X}_1, \cdots, Z_N, \mathbf{X}_N)$, i.i.d. replications of $(Z_1, \mathbf{X}_1)$. Calibrating the prior is then a matter of estimating the regression functions $(\sigma(\mathbf{x}), \gamma(\mathbf{x}))$, assuming that $Z_1|\mathbf{X}_1 = \mathbf{x}_1$ is distributed according to a GP distribution with parameters $(\sigma(\mathbf{x}_1), \gamma(\mathbf{x}_1))$. This falls into the context of GP regression, or tail index regression [see e.g. Davison and Smith, 1990, Goegebeur et al., 2015, Smith, 1989].

In the following applications, we will rely on a specific GP regression method, GP regression trees [see Farkas et al., 2021b,a]. An attractive feature of this method is that it allows us to construct a finite number of risk classes where the values $(\sigma(\mathbf{x}), \gamma(\mathbf{x}))$ are constant.

The output of the procedure is as follows, denoting $\hat{\sigma}(\mathbf{x}), \hat{\gamma}(\mathbf{x})$ the obtained estimators,

$$(\hat{\sigma}(\mathbf{x}), \hat{\gamma}(\mathbf{x})) = \sum_{j=1}^{K} (\gamma_j, \sigma_j) \times r_j(\mathbf{x}),$$

where the multiplication $\times$ applies to both components of a vector (γ, σ), and where $(r_j)_{1 \leq j \leq K}$ are the rules used to assign an individual to one of the K risk classes determined by the regression tree fitting procedure. More specifically, these functions are such that $r_j(\mathbf{x}) \in \{0, 1\}$ for all j, with $r_j(\mathbf{x}) r_{j'}(\mathbf{x}) = 0$ for all (j, j') with $j \neq j'$, and $\sum_{j=1}^{K} r_j(\mathbf{x}) = 1$. This means that a policyholder with characteristics $\mathbf{x}$ is assigned to exactly one risk class (and no more), based only on the value of its characteristics $\mathbf{x}$. The number of classes K adapted to the dataset is selected within the estimation procedure itself (called the "pruning" step of the regression tree, [see Farkas et al., 2021a, for more details], and does not need to be specified.

This results in a more intelligible pricing procedure, compared to a situation where two individuals with different characteristics $\mathbf{x}$ and $\mathbf{x}'$ would be assigned to two different values of the regression function. Another nice feature is the fact that the procedure can be applied to both discrete and non-discrete covariates, which is not the case with for example smoothing based methods such as [Beirlant and Goegebeur, 2004], while avoiding overly restrictive parametric assumptions such as in [Beirlant et al., 1999]. However, it is of course possible to use competing methods to construct the prior, such as [Chavez-Demoulin et al., 2016].

3 Cyber claim analysis

Our first illustration of this methodology relates to cyber insurance. Cyber risk is one of the top threats identified by many reports on emerging risks, such as the AXA 2022 Future Risk Report[1] or the Swiss Re Institute SONAR Report 2023[2]. Its systemic nature and the increasing dependence of all the sectors of the economy on digital tools give rise huge potential costs, which have been documented by many authors and reports[3]. Predicting the total cost of a cyber event is usually a difficult task, as the consequences of such an event may vary greatly from one victim to another (and potentially from one policy to another, as terms and conditions may vary). However, the specific case of data breaches is easier to track than other consequences of cyber such as business interruption, image deterioration, loss of business...

In this domain, Maillart and Sornette [2010] identified the heavy-tailed characteristic of data breach size from a public database collected by the Privacy Rights Clearinghouse (PRC) team[4]. These conclusions have been corroborated by several authors, such as Carfora and Orlando [2019] or Edwards et al. [2016]. The application of this study of data breaches to cyber insurance has been carried out by, for example, Eling and Loperfido [2017], Farkas et al. [2021b], Li and Mamon [2023].

In this section, we first introduce the PRC database in Section 3.1. In our application, this database is used to study the risk at a collective level, and therefore to perform prior calibration. As in [Farkas et al., 2021b], we use a GP regression tree to take into account the heterogeneity of the events in the database. This procedure is explained in Section 3.2. This allows us to determine the prior distribution, which is then used in the credibility theory approach developed in Section 3.3, where we show how to individually price a data breach cyber contract based on the past claims of a given policyholder.

[1] https://www-axa-com.cdn.axa-contento-118412.eu/www-axa-com/a1398464-1f28-4e5a-b503-a47f5acf30c0_AXA_Future_Risks_Report_2023_Francais.pdf

[2] https://www.swissre.com/institute/research/sonar/sonar2023.html

[3] https://permanent.access.gpo.gov/gpo89296/The-Cost-of-Malicious-Cyber-Activity-to-the-U.S.-Economy.pdf

[4] https://privacyrights.org/data-breaches

3.1 Privacy Rights Clearinghouse database

The Privacy Rights Clearinghouse (PRC) is a non-profit organization establised in 1992 to protect the privacy of American citizens. Since 2005, the PRC has maintained a database of businesses involved in data breaches affecting American citizens. The information in this database is based on publicly available reports on breaches and cannot be considered a complete and accurate representation of all data breaches in the United States. It reports breaches that affect US citizens that are made public by government entities and by other sources.

This database is valuable for insurance purposes because it provides not only a list of incidents, but also precise indications of their severity through the "Number of records" metric, which measures the volume of exposed data. However, it is important to note that this metric does not measure the true economic loss. To approximate this amount, Jacobs [2014] established a relationship between the number of records (Y) and the financial loss (L):

$$\log(L) = \alpha + \beta \log(Y). \tag{3.1}$$

Section 3 presents an analysis based on the extraction of the PRC database from March 2023. The database contains 11,222 cyber events that primarily affected American businesses. Only events for which the number of records is available were considered.

For this analysis, we focus only on the impact of two variables on the severity of the breach: the sector of activity and the type of breach. A description of the modalities of these different variables is given in Tables 1 and 2. We also take into account the source that led to the report of the breach in the database, since this gives an indication of the severity: incidents reported by the medias, for example, are more likely to be more severe than others. There are of four type of sources in the PRC database: media, non-profit organization, and two U.S. legal sources, at the state or federal level. A more detailed presentation of this database and this problem of the sources can be found in [Farkas et al., 2021b].

Label	Description
BSF	Businesses in Financial Services, Banking, Insurance Services
BSR	Businesses in Retail/Merchant including Grocery Stores, Online Retailers, Restaurants
BSO	Businesses in Manufacturing, Technology, Communications
EDU	Educational Institutions (Schools, Colleges, Universities)
GOV	Government & Military (State & Local Governments, Federal Agencies)
MED	Healthcare and medical providers (Hospitals, Medical Insurance Services)
NGO	Non-profits (Charities and Religious Organizations)
UNKN	Unknown

Table 1: Sectors of activity of the victims, as reported in the PRC database.

Label	Description
CARD	Fraud involving debit and credit cards not via hacking (skimming devices at point-of-service terminals, etc.)
HACK	Hacked by an outside party or infected by a malware
INSD	Insider (employee, contractor or customer)
PHYS	Physical (paper documents that are lost, discarded or stolen)
PORT	Portable device (lost, discarded or stolen laptop, PDA, smartphone, memory stick, CDs, hard drive, data tape, etc.)
STAT	Stationary computer loss (lost, inappropriately accessed, discarded or stolen computer or server not designed for mobility)
DISC	Unintended disclosure not involving hacking, intentional breach or physical loss (sensitive information posted publicly, mishandled or sent to the wrong party via publishing online, sending in an email, sending in a mailing or sending via fax)
UNKN	Unknown (not enough information about breach to know how exactly the information was exposed)

Table 2: Type of breach as reported in the PRC database.

3.2 Generalized Pareto Regression Tree

The response variable Y that we want to study is the number of records, as an indicator of the size of a data breach. A brief descriptive analysis of the quantiles and empirical mean of this variable in the PRC database is provided in Table 3.

Variable	Min	1st Q	Median	Mean	3rd Q	Max
Records	0	9	880	17,6670	5,000	250,000,000

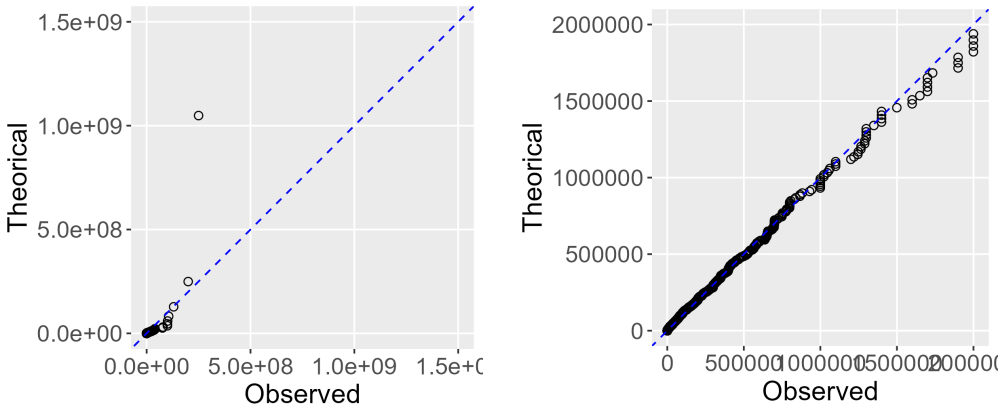
Table 3: Empirical statistics for the response variable "Number of records" in the PRC database, as extracted in March 2023. The empirical variance is $1.67e+13$.

An obvious observation, when looking at Table 3, is the wide range of values for this variable, and the significant gap between the median and the mean, with the mean being driven by some very large claims. This is not surprising from previous studies of Maillart and Sornette [2010], since its distribution is expected to be Pareto tailed.

With respect to (2.2), we only consider observations that are above a threshold u large enough so that a GP approximation seems reasonable. This threshold u was set to 500 , leading to 6,600 events above this threshold. Figure 1 shows the Quantile-Quantile plots that confirms a reasonable fit above the threshold $u = 500$. Figure 1 a) shows all the data while Figure 1 b) corresponds to the same graph zoomed on the data below 2×10^6 (representing 99,09% of the data).

Since the situations covered by the database are heterogeneous, we consider different values of the parameters σ and γ of the GP distribution depending on the characteristics of the event. For this purpose, we determine classes of events using the GP regression tree approach of [Farkas et al., 2021b,a]. This leads to the definition of 8 risk classes shown in Figure 2. Quantile-quantile plots can be found in Section 6.1 (see Figure 6) and show a reasonable fit. We see that all the shape parameters γ are larger than 1, but let us recall that we fitted the tree to the Number of Records and not to the associated financial loss. If we rely on (3.1), the tail index of Y must be multiplied by β to obtain the tail index of L. We can observe that the majority of cases correspond to situations where the tail index of Y is estimated to be less than 2 (in 82% cases), while a minority (2%) have a tail index larger than 3.

Although the tail index of L can be derived from the previous regression tree, the scale parameter σ corresponding to L does not derive directly from (3.1). One way to avoid this issue would be to directly use the formula that transforms a number of records into a cost, and to fit the regression tree to this transformed variable. We decided not to do this to reflect the fact that there is no consensus on (3.1). Thus, the risk classes are built without the need for a formula linking L to Y. This link is used when, after fitting the tree, we estimate the parameter



a) b)

Figure 1: Quantile-quantile plots of the fit with a threshold $u = 500$. a) shows all the data and b) corresponds to the same graph zoomed on the data below 2×10^6 (representing 99,09% of the data)

σ in each leaf by fitting a GPD with parameters σ and γ to $(\alpha + \beta \log Y_i)$ for i belonging to the given leaf.

Remark 3.1 *Note that the variable "Source" (referring to the source of information from which we learned about the incident) plays a special role. We included it to fit the regression tree in order to take into account the bias caused by the particular structure of these sources. In practice, an insurer will rely on a single source coming from its own information system, and a decision has to be made to determine which of the four sources in the PRC database is the closest to the insurer's source.*

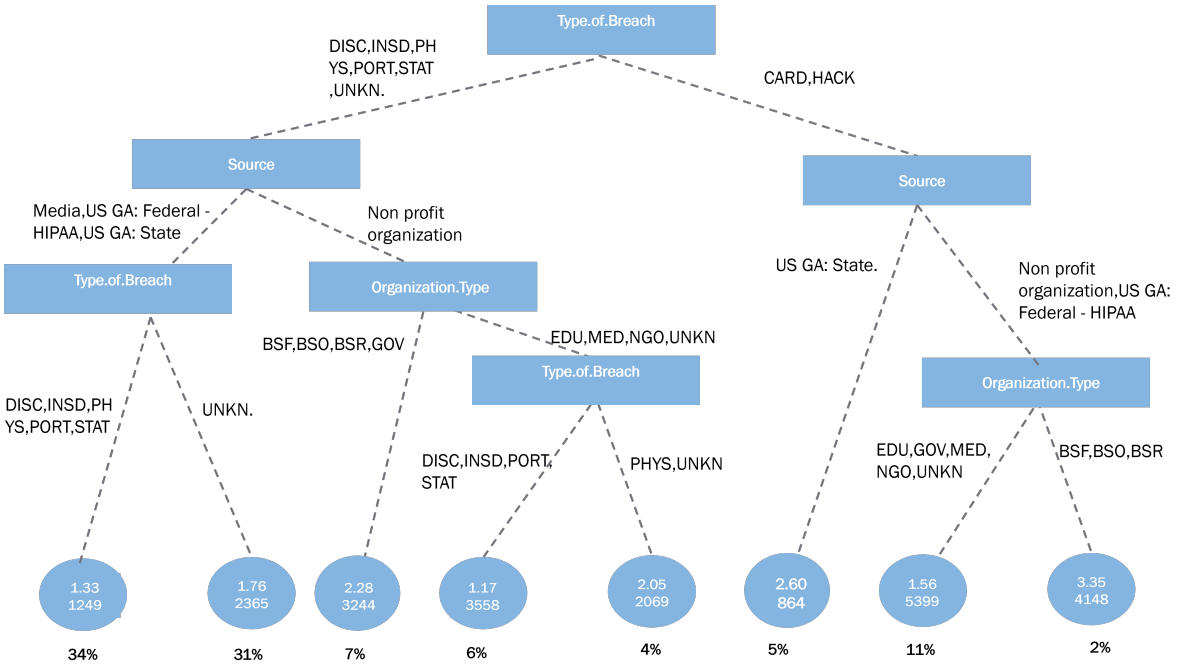


Figure 2: Generalized Pareto regression tree fitted to the PRC database (extracted in March 2023) using the technique described in [Farkas et al., 2021b]. In each leaf, the first (resp. second) line corresponds to the estimated value of γ (resp. σ). The percentage of observations associated with each leaf is given below. The variable "Source" refers to the source of information that reported the breach to the PRC association (non-profit organization, media, US GA: Federal - HIPPA for the federal level government source, US GA: State for the corresponding state level source).

3.3 Application of Credibility Theory

We consider three illustrations from the PRC database. We consider three victims that have repeated occurrences in the database (all reported by an official government source, namely "US GA: Federal"). We consider two types of claims: hacking (HACK, for two different victims with a different depth of historical data) and unintended disclosure of information (DISC). In each case, we want to determine what should be the price of a guarantee for this specific type of risk for this specific policyholder, based on the prior and the historical data. Note that this price is defined up to some constant related to the frequency: we here only have data on severity, and information on frequency would be required to compute a premium.

We use the formula (3.1) to transform the information on the records into a price, with $\alpha = 9.59$ and $\beta = 0.57$. These values correspond to the approach of [Farkas et al., 2021b], which updates the values of the parameters taken by Jacobs [2014] in order to take into account the mega breaches that occurred between the publication of these two papers.

The summary of the historical data of these two potential policyholders is given in Table 4.

Policyholder / Sector	Hazard	Number of past claims	Average past loss	Tail index	Scale parameter	Prior premium	Credibility Factor	Credibility Premium
A / MED	HACK	2	103 357	0.8892	136 525	1 232 179	0.94	169 561
B / MED	HACK	4	134 197	0.8892	136 525	1 232 179	0.96	167 367
C / MED	DISC	3	82 528	0.7182	99 245	352 183	0.88	113 717

Table 4: Examples of three victims from the medical sector, and for two incident types. The "Prior premium" is the premium (up to a multiplication by the frequency) that would be paid if no previous claims were known.

Policyholder / Sector	95% quantile	Ratio 95% quantile-premium	99% quantile	Ratio 99% quantile-premium
A / MED	579 445	3.42	1 212 595	7.15
B / MED	548 279	3.28	1 005 291	6.01
C / MED	377 241	3.31	714 916	6.29

Table 5: Quantiles based on the credibility model from Equation (2.6) for the three policyholders of Table 4, with $\alpha = 0.05$ and $\alpha = 0.01$. The third and fifth columns show the ratio between this quantile and the credibility premium.

In the three considered examples, the victims have a claim history much smaller than the prior premium. The credibility premium is computed from (2.5), and shows a significant reduction of the amount (compared to the prior premium), thanks to the high value of the credibility factor. We see that even a small number of historical data contributes to a high value of this credibility factor. If we compare policyholder A and C, we see that the credibility factor for A is larger than for C, even though the number of past claims is smaller for A. This is due to the fact that the tail index is larger for A, so the information provided by a single observation on the individual risk factor is more important (this must be related with Remark 2.1). On the other hand, we can see from Table 5 that, because of this larger tail for case A, the ratio between the credibility premium and the high quantiles is larger for A.

On the other hand, let us emphasize once again that the so-called premium in Table 4 should not be considered as the final one, since it does not include the information about the frequency. For example, Policyholder B has a larger number of past claims than A, but a smaller premium. This is onlu due to the fact that this large number of claims gives us more information about the severity of incidents that hit A (and this severity is smaller than the average claims of this category). But assuming that policyholders A and B have been observed for the same period, and since the number of past claims for B is twice the number for A, we should have

$$\frac{\pi_A}{\pi_B} = \frac{\pi_A^*}{2\pi_B^*} = 0.506,$$

where π_A (resp. π_B) denotes the final premium paid by A (resp. by B) and π_A^* (resp. π_B^*) the credibility premimum for A (resp. for B) from Table 4. However, since there is no information on the exposure in the PRC database, we cannot readily determine the relative ranking of these two policyholders.

4 Cost prediction of floods in France

The second illustration that we provide is related to the context of flood insurance. Based on physical characteristics of a flood event impacting a city, the goal is to evaluate the amount of the loss. To calibrate the prior distribution, we rely on the SILECC database, provided by France Assureurs, which gathers flood events on the French territory. The large number of events provides valuable information to evaluate the risk, but the final outcome, at a local level, is more difficult to predict due to the significant differences between affected territories. Therefore, our approach combines this collective understanding of the phenomenon with individual historical data, which is at the core of our approach.

Two main applications are being considered. The first one is motivated by the specific context of the French insurance system against natural disasters. The "CatNat Regime" (for "Catastrophe Naturelle") is a public-private partnership that is briefly described in Section 4.1 to provide some context. At the very core of this system is the need of a rapid evaluation of the amount of a catastrophic flood event, in order to trigger a compensation mechanism. The proposed methodology is generally applicable for calibrating scenarios to analyze the cost of a specific type of event in a particular area.

Section 4.2 provides a brief description of the SILECC database used to calibrate our prior. Similar to Section 3, we fit a GP regression tree to analyze the tail with the results presented in Section 4.3. Section 4.4 discusses the application of the credibility approach to predict recent flood events.

4.1 Short description of the French CatNat regime

In France, natural disasters are covered through a public-private partnership, called the CatNat Regime. This specific French framework strongly dictates the management of natural disaster claims. This natural disaster compensation scheme was created by the Law of July 13, 1982, and is based on a solidarity principle. For every contract, the same additional premium insurance rate, fixed by the government, is used to compensate for the losses due to natural disasters. The compensation regime has a broad scope, covering floods, mudslides, earthquakes, and landslides. However, it does not include storms, hail, or snow. Without going into the functional details of this compensation regime, to receive compensation, a government decree must be published

in the "Journal Officiel", which contains all laws and legislative events of the French Republic, acknowledging that a given city is in a state of natural disaster.

This decree is issued by an inter-ministerial commission in response to an official request from the city's mayor to recognize the event as a natural disaster. The commission assessed the exceptional situation of the event at the city level. The prediction of the cost of the flood in each affected city is crucial for decision-making. Therefore, quantitative analysis is necessary to improve this prediction and challenge the analysis provided by Caisse Centrale de Réassurance (CCR). The methods used by CCR are described in [Moncoulon et al., 2014, Moncoulon, 2014, Moncoulon and Quantin, 2013] or in more detail in [Mao, 2019]. In addition to the challenging nature of this evaluation task (see for example [see e.g. Hall and Solomatine, 2008, Eleutério, 2012]), it is also crucial to prioritize the interpretability of the method, advocating for simple tractable formulas in this estimation.

4.2 Flood events database

To analyze floods, we had access to the SILECC database through a partnership with the Mission Risques Naturels (MRN), a technical body of France Assureurs. This database covers approximately 70% of the French non-life insurance market by aggregating the claims of 12 major French insurance companies. The database records the natural disaster claims in France from 1987 to 2019, which each claim has been standardized and geolocalized.

While the database covers several natural hazards, we focused solely on floods events. We used data from to 1999 to 2019 and linked it to the event database, resulting to 3,100 flooding events. This period provides strong representativeness of floods events in France and covers major episodes such as the floods of 2016 in the Seine and Loire, as well as the floods of 2003 in the South of the country.

To study natural hazards, particularly floods, the first step is to categorize them by event, based on the definition made by Bourguignon [2014]. Claims data are received at the communal level and for a given date, but creating events can aid the analysis. Specifically, in our case, grouping by event provides a learning base for estimating the cost when an event occurs. Additionally, this grouping can yield valuable indicators on the affected territories.

In our case, an event is defined by a starting date, an end date and a set of impacted cities. The impacted territories are identified a posteriori according to the decree of natural catastrophe, introduced above. Requests are grouped together to form events according a coherent spatio-temporal perimeter.

The event database comprises almost 140,000 flood decrees grouped in more than 4,300 distinct events between 1982 and 2021. Like many insurance datasets, this database is highly unbalanced, with major events concentrating a large part of the cities. Specifically, the 10 largest

events account for 35% of the database.

The cost of a given flood event is determined by aggregating the cost at the smallest scale, namely the sum of the costs for each affected city. The total cost of a flooding event, is a highly volatile variable, ranging from 0 to 394,376,000 euros with an empirical variance equal to 1.77e + 14. Figure 3 shows the average costs of the top 10% most expensive flooding events in each meteorological region. This statement highlights the heterogeneity of the severity of the most severe events. Furthermore, the top ten most onerous events account for 43% of the total cost of this database and the top hundred account for 80%.

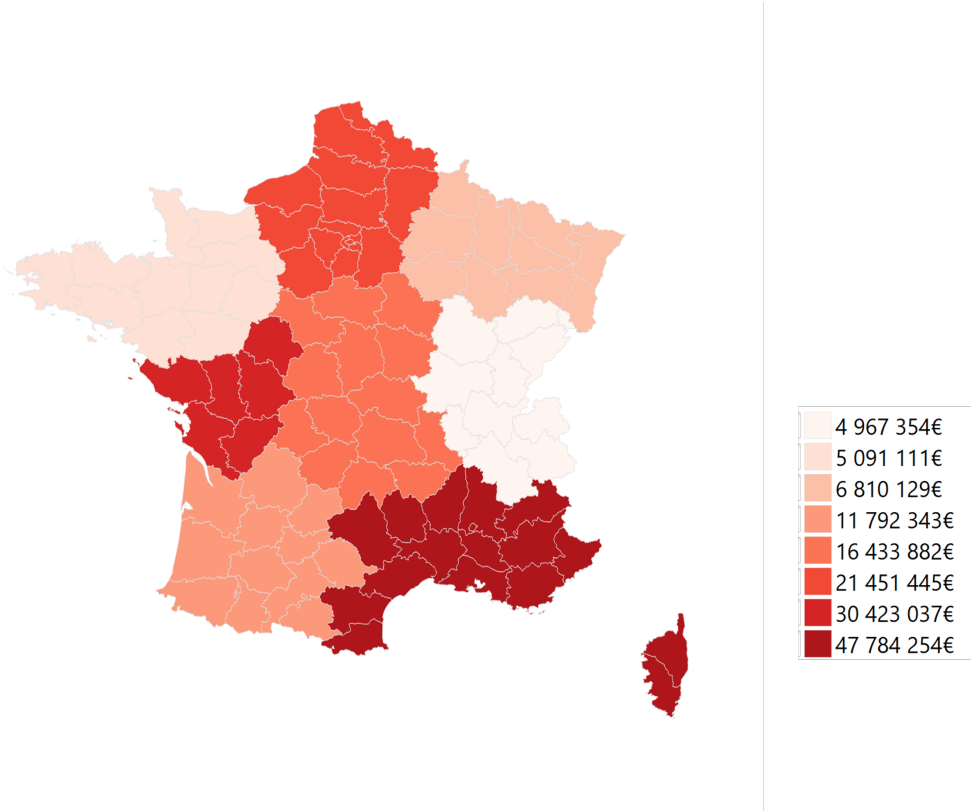


Figure 3: Cartography of the cost of flooding events in France from 1999 to 2019. For each meteorological region, the average of the costs of the 10% more onerous events is shown. The lighter red color suggesting a small cost while a darker color suggests a large cost.

4.3 Prior calibration

As described in Section 3, the first step is to determine the threshold above which the GP approximation seems relevant. The threshold is here set as $u = 100,000$ euros. 820 events are above this threshold. A synthetic description of the database and its characteristics is given in Table 6.

Variable	Min	1st Q	Median	Mean	3rd Q	Max
Cost (in euros)	100,093	199,287	477,943	6,066,835	1,941,047	380,487,161
Number of affected hydrological regions	1	1	2	4	4	35
Number of individual houses in flood risk area	0	5,874	20,692	92,477	71,094	4,097,075
Number of professional business premises in flood risk area	0	2,230	8,163	44,830	26,321	2,050,165

a)

Variable	Category	Number of observations
Meteorological regions	Center	60
	North West	85
	North	135
	North-East	87
	East	96
	South	209
	West	30
	South West	121
Seasons	Spring	272
	Summer	279
	Autumn	187
	Winter	85

b)

Table 6: List of quantitative and categorical variables in the SILECC database (restricted to flood events of amount larger than 100,000 euros) and corresponding descriptive statistics. For the quantitative variables, Table a) shows the minimum, the first quartile, the median, the mean, the third quartile and the maximum, and for the categorical variables, Table b) the number of observations per category.

We can again notice the volatility of the cost variable. Three numerical variables were used: the first, the number of affected hydrological regions specifies the size of the affected area. The other two, the number of individual houses and the number of professional business premises in flood risk area, account for the exposition of the impacted area. It is calculated based on a flood risk cartography done by the MRN, which integrates the risk of floods caused by runoff. We used two categorical variables that explain the situation of the events: the meteorological regions and the seasons. This is linked to the type of floods.

The method produced the tree displayed in Figure 4. The 95% confidence intervals are provided in Tables 10 and 11. The tree has 6 leaves, with splits according to 3 criteria, the number of individual houses in flood risk area, the number of professional business premises in flood risk area, and the number of affected hydro-ecoregions. This seems consistent because the first two covariates represent the exposure to flooding but also the population density of the affected area, the third covariate captures the perimeter of the event.

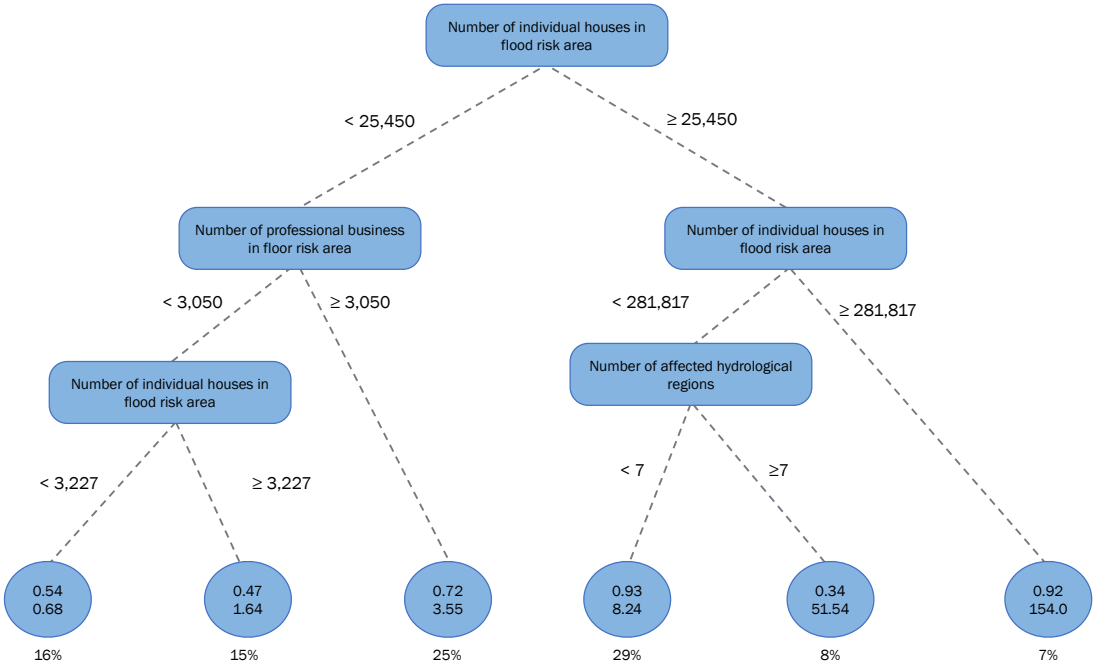


Figure 4: GP regression tree obtained for flooding events. For each leaf, the value of the shape parameter γ (first line) and the scale parameter σ at 10^{-5} (second line) are given. Percentage of observations affected to each leaf is mentioned.

The most extreme case corresponds to the rightmost leaf, with a shape parameter of 0.92, it contains 7% of the events. It corresponds to a large number of affected individual houses and to a large area. Table 7 provides a comparison of the empirical and theoretical medians and means in each leaf. Recall that in the case of a GP distribution with parameter (σ, γ) the theoretical median is equal to $\sigma(2^{\gamma}-1)/\gamma$ and the theoretical mean by $\sigma(1-\gamma)$ for $\gamma < 1$ and to ∞ for $\gamma \geq 1$. For each leaf, the median is well below the mean suggesting that we are indeed dealing with extreme events. Then we observe a very good fit with very close values in all the leaves for the theoretical and empirical medians. For the mean, the theoretical and empirical values are also close, except for leaves 4 and 6 which correspond to the largest shape parameters. The parameters thus seem to fit the distribution in each leaf very well and the classification seems relevant.

Furthermore, one can observe that the correlation between the losses between the classes is empirically small, as shown in Table 12.

Leaf	Shape parameter	Empirical Median	Theoretical Median	Empirical Mean	Theoretical Mean
1	0.54	161 694	157 697	239 923	249 456
2	0.47	226 196	234 764	399 274	410 387
3	0.72	455 663	419 978	1 439 087	1 390 099
4	0.93	950 181	902 387	4 144 876	11 877 446
5	0.34	4 215 647	4 140 879	7 982 445	8 009 145
6	0.92	15 555 487	15 090 137	52 203 995	281 103 859

Table 7: Empirical median and mean, and theoretical median and mean for each leaf (in euros).

4.4 Prediction based on credibility

In the present case, we need to adapt the credibility formula (2.5) to distinguish between two scales in dealing with a flood event. The GP regression model is fitted on the total loss corresponding to a given event, that is at a large scale. On the other hand, the cities affected by such an event do not necessarily present the same depth of historical data: two distinct cities are not necessarily simultaneously stroke by an event, then the number of times they are present in the database is therefore different.

Consequently, we proceed first at the smaller scale, predicting the loss at a city level, and then aggregating back by summing the predictions. To this aim, for city i, we replace the scale parameter σ used in the GP with $p_i\sigma$ where p_i is the proportion of the exposed premium in city i for the city affected by the current event. This leads to

$$\pi_{r,\lambda}(Y_1, \cdots, Y_n) = c_n(\gamma)\frac{\sum_{i=1}^n Y_i}{n} + (1 - c_n(\gamma))\frac{p_i\sigma}{1-\gamma}, \tag{4.1}$$

where $c_n(\gamma)$ is given in Remark 2.1.

The method is illustrated on a database of 48 events considered as major by CCR. Table 8 provides some descriptive statistics on the individual costs of these events. For each event, we extract the cities that are impacted and estimate the loss in each of these cities using (4.1). Note that about half of the cities have no previous experience of such event (for them, $c_n(\gamma) = 0$), and only 20% have more than one historical claim.

These projections are then aggregated to get an estimation of the total loss of this event. This estimation of the loss is compared to the corresponding loss given by CCR, which is considered as a reference. We compare this approach with a projection that would be purely based on the prior distribution (that is, if we imposed $c_n(\gamma) = 0$ for all of the impacted cities), to measure the importance of including historical data in the picture.

The relative gaps between the costs as reported by CCR and those estimated according to the two techniques are reported in Figure 5. We see that the credibility-based technique (that

	Min	1st Q	Median	Mean	3rd Q	Max
Cost	10,710,000	16,110,000	35,680,000	116,900,000	98,270,000	1,056,000,000

Table 8: Summary of the cost variable of the major events of the CCR database.

is, the one based on historical data) performs significantly better than the one based on the sole prior, especially for the most expensive event (CCR cost of 1 billion euros, estimated at 52 million by the prior, and 463 million by the credibility estimator), although both estimators are far below the true value in this extreme case.

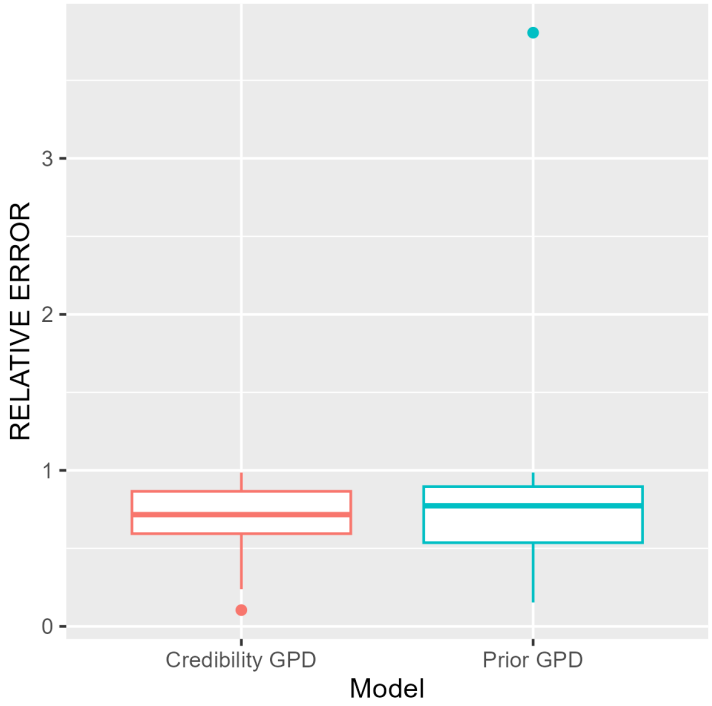


Figure 5: Distribution of the relatives errors $|\hat{C}_i - C_i|C_i^{-1}$ for each event of the CCR database.

We also report in Table 9 the mean absolute error (MAE), that is

$$MAE = \frac{1}{48}\sum_{i=1}^{48}|\hat{C}_i - C_i|,$$

where $\hat{C}_i$ is the estimated value of the total cost of the $i-$th event, and C_i its cost as reported by CCR. We consider two additional indicators, first normalizing the errors compared to the total value of the losses,

$$RE = \frac{\sum_{i=1}^{48}|\hat{C}_i - C_i|}{\sum_{j=1}^{48} C_j},$$

and the average relative error, that is

$$ARE = \frac{1}{48}\sum_{i=1}^{48}\frac{|\hat{C}_i - C_i|}{C_i}.$$

Model	MAE	RE	ARE
Prior	108 426 428	0.84	0.76
Credibility	94 476 224	0.73	0.70

Table 9: Comparison of the errors of the prior model (that is with $c_n(\gamma) = 0$ for all cities), and the one based on the combination with historical data using the credibility formula of (4.1).

We see that the credibility estimator performs better than the one based on the sole prior. In each case, the errors are large, but this magnitude is still considered as reasonable for such type of catastrophe, where the question of estimating the impact of such an event immediately after its occurrence (and from a relatively small amount of data) is considered particularly delicate. This inherent difficulty is increased by the French context, where the process of recognizing a natural catastrophe induces additional uncertainties: the affected cities may not included in the compensation process, increasing the volatility of the final prediction. It should also be noted that, the projection may be significantly improved by enriching the available data with meteorological variables and/or satellite images, which can provide important additional information.

5 Conclusion

In this paper, we have described a Bayesian approach that is particularly adapted to insurance losses that have a Pareto tail. An advantage of a prior distribution is that it can provide a premium even when there are no previous claims, while the particular form of model that we used allows the parameters of this prior to be calibrated from collective data. We have chosen to use a tractable model that is consistent with the fact that, according to results from EVT, the collective distribution of the losses should be approximately GP distributed. This does not mean that the present model is free from misspecification issues such as those considered in [Hong and Martin, 2022, 2020]. Goodness-of-fit procedures could be the next step to validate these techniques in some practical cases, with the difficulty that the lack of data, for the risk being considered lay lead to test procedures with weak power. Another extension is to combine this approach with an approach on frequency, as in [Cheung et al., 2021], while in the present work we considered only the case of severity .

6 Appendix

6.1 Additional empirical results for the application to cyber insurance

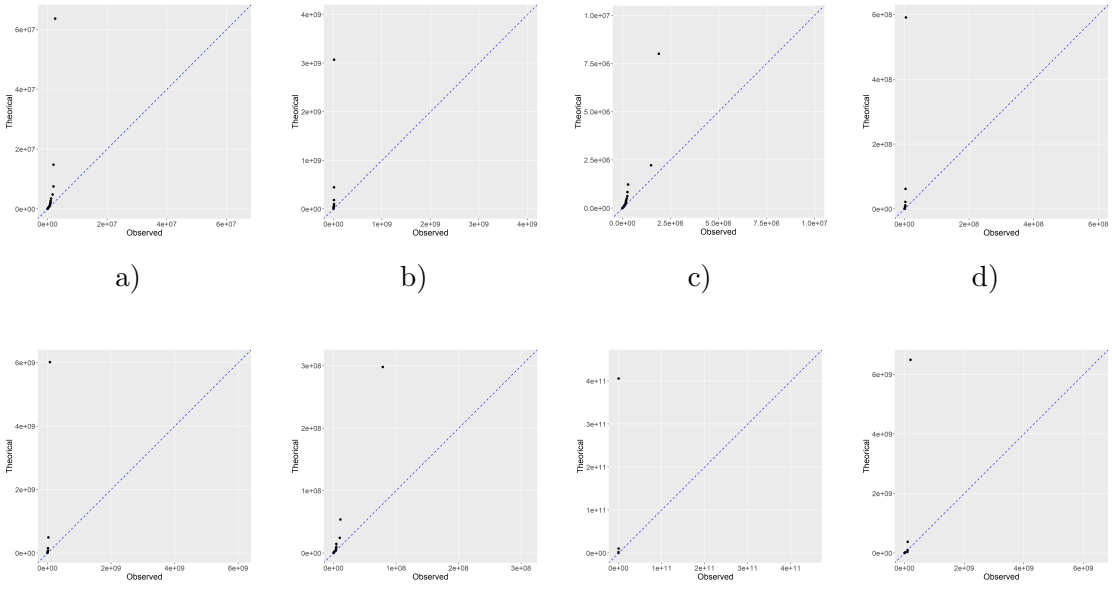


Figure 6: Quantile-quantile plots for each leaf of the regression tree of Figure 2 in the cyber application.

6.2 Additional empirical results for the application to flood insurance

Leaf	Shape parameter estimate	Lower CI	upper CI
1	0.54	0.27	0.82
2	0.47	0.21	0.73
3	0.72	0.50	0.95
4	0.93	0.67	1.19
5	0.34	0.03	0.67
6	0.92	0.38	1.46

Table 10: 95% confidence intervals for the shape parameter γ

Leaf	Scale parameter estimate	Lower CI	upper CI
1	0.68	0.47	0.90
2	1.64	1.15	2.14
3	3.55	2.66	4.44
4	8.24	6.06	10.42
5	51.54	31.36	71.53
6	154.0	69.51	238.33

Table 11: 95% confidence intervals for the scale parameter σ

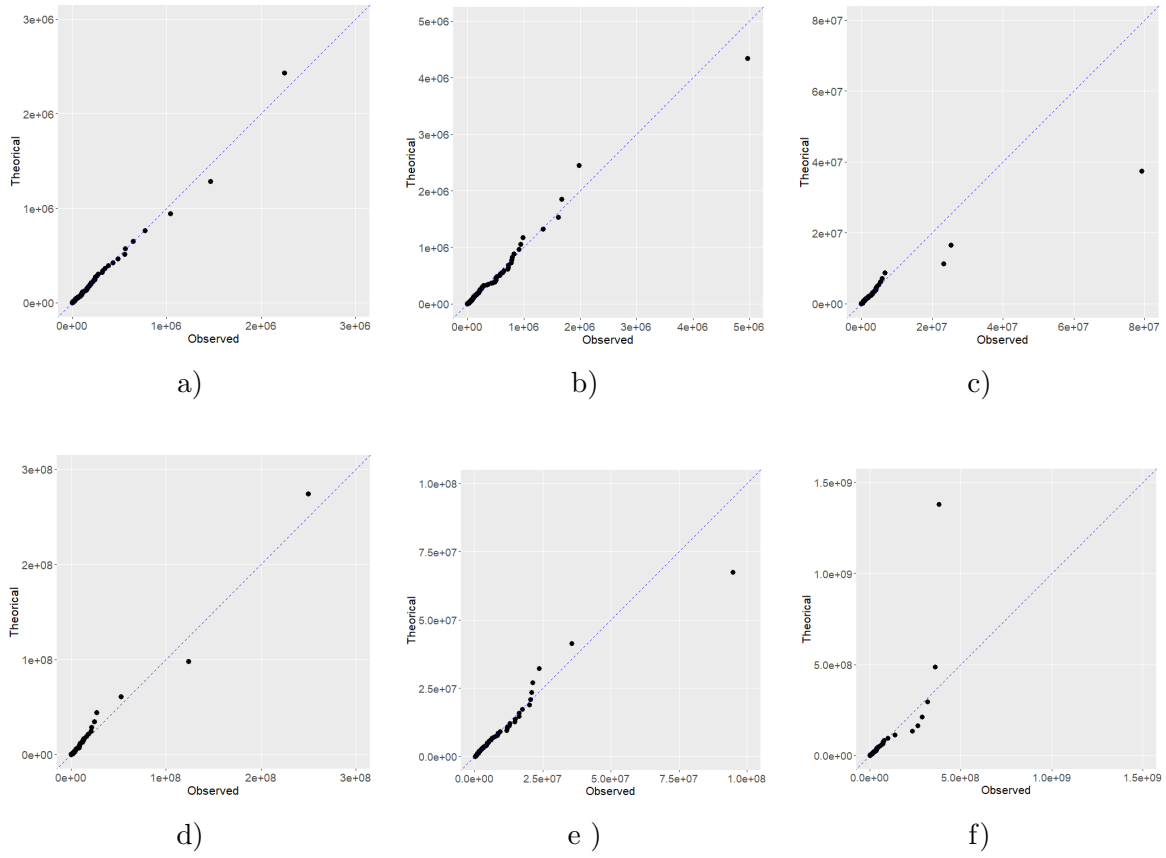


Figure 7: Quantile-quantile plots for each leaf of the regression tree of Figure 4 in the flood event application.

Leaf	1	2	3	4	5	6
1	X	0.33	-0.07	-0.02	0.02	0.12
2	0.33	X	0.04	0.24	0.01	0.44
3	-0.07	0.04	X	0.03	0.26	0.03
4	-0.02	0.24	0.03	X	0.03	0.14
5	0.02	0.01	0.26	0.03	X	0.03
6	0.12	0.44	0.03	0.14	0.03	X

Table 12: Pearson correlation coefficients for the empirical cost of city in each leaf. We compare the average costs in the same city but in different leaf

References

A. A. Balkema and L. de Haan. Residual life time at great age. *The Annals of Probability*, pages 792–804, 1974.

J. Beirlant and Y. Goegebeur. Local polynomial maximum likelihood estimation for pareto-type distributions. *Journal of Multivariate Analysis*, 89(1):97–118, 2004.

J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys. Tail index estimation and an exponential regression model. *Extremes*, 2:177–200, 1999.

J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons, 2004.

D. Bourguignon. *Événements et territoires-le coût des inondations en France: analyses spatio-temporelles des dommages assurés*. PhD thesis, Université Paul Valéry-Montpellier III, 2014.

H. Bühlmann and A. Gisler. *A course in credibility theory and its applications*, volume 317. Springer, 2005.

M. F. Carfora and A. Orlando. Quantile based risk measures in cyber security. In *2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, pages 1–4. IEEE, 2019.

V. Chavez-Demoulin, P. Embrechts, and M. Hofert. An extreme value approach for modeling operational risk losses depending on covariates. *Journal of Risk and Insurance*, 83(3):735–776, 2016.

E. C. Cheung, W. Ni, R. Oh, and J.-K. Woo. Bayesian credibility under a bivariate prior on the frequency and the severity of claims. *Insurance: Mathematics and Economics*, 100:274–295, 2021.

P. M. Chiroque-Solano and F. A. d. S. Moura. A heavy-tailed and overdispersed collective risk model. *North American Actuarial Journal*, 26(3):323–335, 2022.

S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer London, 2001.

A. C. Davison and R. L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):393–425, 1990. doi: https://doi.org/10.1111/j.2517-6161.1990.tb01796.x.

L. Diao and C. Weng. Regression tree credibility model. *North American Actuarial Journal*, 23 (2):169–196, 2019.

B. Edwards, S. Hofmeyr, and S. Forrest. Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, 2(1):3–14, 2016.

J. Eleutério. *Flood risk analysis: impact of uncertainty in hazard modelling and vulnerability assessments on damage estimations*. PhD thesis, Strasbourg, 2012.

M. Eling and N. Loperfido. Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance: mathematics and economics*, 75:126–136, 2017.

S. Farkas, A. Heranval, O. Lopez, and M. Thomas. Generalized pareto regression trees for extreme events analysis. *arXiv preprint arXiv:2112.10409*, 2021a.

S. Farkas, O. Lopez, and M. Thomas. Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, 98:92–105, 2021b.

Y. Goegebeur, A. Guillou, and G. Stupfler. Uniform asymptotic properties of a nonparametric regression estimator of conditional tails. In *Annales de l'IHP Probabilités et statistiques*, volume 51, pages 1190–1213, 2015.

E. Gómez-Déniz, V. Leiva, E. Calderín-Ojeda, and C. Chesneau. A novel claim size distribution based on a birnbaum–saunders and gamma mixture capturing extreme values in insurance: estimation, regression, and applications. *Computational and Applied Mathematics*, 41(4):171, 2022.

J. Hall and D. Solomatine. A framework for uncertainty analysis in flood risk management decisions. *International Journal of River Basin Management*, 6(2):85–98, 2008.

W.-R. Heilmann. Decision theoretic foundations of credibility theory. *Insurance: Mathematics and Economics*, 8(1):77–95, 1989.

L. Hong and R. Martin. Model misspecification, bayesian versus credibility estimation, and gibbs posteriors. *Scandinavian Actuarial Journal*, 2020(7):634–649, 2020.

L. Hong and R. Martin. Imprecise credibility theory. *Annals of Actuarial Science*, 16(1):136–150, 2022.

W. K. Huang, D. W. Nychka, and H. Zhang. Estimating precipitation extremes using the log-histospline. *Environmetrics*, 30(4):e2543, 2019.

J. Jacobs. Analyzing ponemon cost of data breach. *Data Driven Security*, 11:5, 2014.

Y. Li and R. Mamon. Modelling health-data breaches with application to cyber insurance. *Computers & Security*, 124:102963, 2023.

T. Maillart and D. Sornette. Heavy-tailed distribution of cyber-risks. *The European Physical Journal B*, 75(3):357–364, 2010.

G. Mao. *Estimation des coûts économiques des inondations par des approches de type physique sur exposition*. PhD thesis, Université de Lyon, 2019.

D. Moncoulon. *Proposition d'une méthode d'estimation de l'exposition financière aux inondations pour le marché de l'assurance en France: modélisation hydrologique et économique probabiliste spatialisée*. PhD thesis, Toulouse 3, 2014.

D. Moncoulon and A. Quantin. Modélisation des événements extrêmes d'inondation en france métropolitaine. *La Houille Blanche*, (1):22–26, 2013.

D. Moncoulon, D. Labat, J. Ardon, E. Leblois, T. Onfroy, C. Poulard, S. Aji, A. Rémy, and A. Quantin. Analysis of the french insurance market exposure to floods: a stochastic model combining river overflow and surface runoff. *Natural Hazards and Earth System Sciences*, 14 (9):2469–2485, 2014.

A. T. P. Najafabadi. A new approach to the credibility formula. *Insurance: Mathematics and Economics*, 46(2):334–338, 2010.

F. Pechon, M. Denuit, and J. Trufin. Home and motor insurance joined at a household level using multivariate credibility. *Annals of Actuarial Science*, 15(1):82–114, 2021.

J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1): 119–131, 1975.

C. Rohrbeck, E. F. Eastoe, A. Frigessi, and J. A. Tawn. Extreme value modelling of water-related insurance claims. *The Annals of Applied Statistics*, 12(1):246–282, 2018.

C. Scarrott and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical journal*, 10(1):33–60, 2012.

R. L. Smith. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, pages 367–377, 1989.

P. Tencaliec, A.-C. Favre, P. Naveau, C. Prieur, and G. Nicolet. Flexible semiparametric generalized pareto modeling of the entire range of rainfall amount. *Environmetrics*, 31(2):e2582, 2020.

W. Zhu, K. S. Tan, and L. Porth. Agricultural insurance ratemaking: Development of a new premium principle. *North American Actuarial Journal*, 23(4):512–534, 2019.