Journal of Statistics and Computer Science
Vol. 3, Issue 1, 2024, pp. 13-40
https://doi.org/10.47509/JSCS.2024.v03i01.02

Modelling the cost of the biggest forthcoming disaster using largest values and next record in insurance

Lucien Diégane GNING [*], Daniel PIERRE-LOTI-VIAUD [†]

Abstract

In this paper, we consider a sequence of independent and identically distributed random variables for which, for $n \geq 2$, only the l largest values are observed (with $2 \leq l \leq n$). For example, this is the case when the excesses over a threshold are only observed. We propose a method to calculate a prediction interval for the value of the next record of this sequence. This method is based on using linear regression models, without doing any parametric assumption on the distribution of the observed variables. The explanatory variable in these models is the rank of the order statistics related to the largest values observed. The obtained result is compared with an exact prediction interval knowing the distribution of the sequence, and with a prediction interval calculated by using the Hill estimator and assuming that the sequence has a Pareto distribution. These comparisons are made on several simulated-data sets, and the linear model method is also applied to insurance data related to catastrophic events.

Keywords : Extremes; Records; Order statistics; Linear model; Pareto distribution.

1 Introduction

A fairly elaborate introduction is necessary to present the context, the purpose, the state of the art, and our approach of the problem addressed. For clarity, it is organized into four sections.

1.1 Probabilistic model.

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables with its distribution denoted by $\mathbb{P}_{X_1}$. For each $n \geq 1$, let $X_{n,n} \leq \cdots \leq X_{n,1}$ denote the order statistics associated with the

[*] Corresponding author Lucien Diégane GNING : luciendesgning@yahoo.fr, LERSTAD, UFR Sciences appliquées et Technologie, BP 234 Université Gaston BERGER de Saint-Louis, Sénégal

[†] daniel.pierre-loti-viaud@upmc.fr, Address Daniel Pierre-Loti-Viaud, LPSM-Sorbonne Université 4, place Jussieu, 75252 Paris, Cedex 05, France

Article History
Received : 10 November 2023; Revised : 17 May 2024, Accepted : 28 May 2024; Published : 30 June 2024

To cite this paper
Lucien Diégane GNING, & Daniel PIERRE-LOTI-VIAUD (2024). Modelling the cost of the biggest forthcoming disaster using largest values and next record in insurance. *Journal of Statistics and Computer Science.* 3(1), 13-40.

sample $X_1, \cdots, X_n$. For an integer $\ell \in [1, n]$, let $Y_{\ell,n} = (X_{n,1} \cdots X_{n,\ell})' \in \mathbb{R}^\ell$ be the vector of the ℓ largest values of this sample. For small ℓ compared to n, $Y_{\ell,n}$ represents the upper extreme values of the sample $X_1, \ldots, X_n$.

Two hypotheses allow to complete our framework. Firstly, on the observed data: for a realization ω, an unknown n and a small ℓ, the observations are given by the vector $Y_{\ell,n}(\omega)$. Secondly, on the probability law $\mathbb{P}_{X_1}$: the right end of its support is infinite and the related upper tail is heavy, a typical situation in insurance.

1.2 Aim

In such a framework, the quantity $M_n(\omega) = X_{n,1}(\omega) = \max_{i \in \{1,\ldots,n\}} X_i(\omega)$ is known, and the aim of this presentation is to predict the first value that exceeds $M_n(\omega)$, to be explicit, the value of $X_{n+j}(\omega)$, for j the smallest positive integer such that $X_{n+j}(\omega) > M_n(\omega)$. The quantity $X_{n+j}(\omega)$ is the value of the next record of the sequence $X_1(\omega), X_2(\omega), \ldots$ when it is observed until the index n. We observe that it would be also interesting to predict the value j for which this record emerges, nevertheless this will not be discussed here. In what follows, the next-record value will be denoted by $R(\omega)$.

An example in insurance of next record value is the next largest economic cost of natural disasters that will arrive in the World after the Sendai earthquake. Another example is the next largest claim cost for insurers and re-insurers that will arrive in the World after Hurricane Katrina.

1.3 What does exist

The study of records is related to extreme-value theory and to $\mathbb{P}_{X_1}$ upper tail properties. For example, as n goes to infinity the asymptotic behavior of M_n depends on the upper-tail behavior of $\mathbb{P}_{X_1}$, it is the Fisher-Tippett Theorem. The same is true for the sequence of upper-record values, it is the Resnick duality Theorem.

An R prediction interval (RPI) of the next-record value is often easily obtained if the probability law $\mathbb{P}_{X_1}$ is entirely specified, for example, as one member of a parametric family of distributions such as the normal, log-normal, and Pareto families, or still several others. In spite of that, if the hypothesis is just that $\mathbb{P}_{X_1}$ belongs to such a parametric family of distributions, an RPI then depends on parameters having to be estimated on the basis of Y, and, in general, this is impossible. Among others, for the normal and log-normal families, an RPI depends on all the model parameters, and estimators of these parameters are necessarily functions of n that is unknown. On the contrary, it works differently for the Pareto family, owing to the fact that the RPI depends only on M_n and on the shape parameter of the Pareto distribution, and the latter can be estimated from $Y_{\ell,n}$, for example by using the Hill estimator. However, experience (for instance, see the applications of this method on simulated-data sets discussed in Section 4.2) shows that the resulting RPI is highly dependent on such an assumption about $\mathbb{P}_{X_1}$. What to do if one has no *a priori* information on $\mathbb{P}_{X_1}$? That is a question to which this presentation seeks to answer.

For more accuracy on this issue, observe in addition that the upper-tail behaviour of $\mathbb{P}_{X_1}$ may not be precisely determined for some applications of extreme-value theory. It is what can regularly be observed in the area of insurance, where the log-normal and Pareto distributions are among the most used and are competing as models for $\mathbb{P}_{X_1}$. As it is emphasized and reiterated several times in Embrechts *et al.* (2008), we then can highlight the difficulty of the task by this aphorism: *it is an uncertain mission than predicting the unpredictable.*

To continue with the consequences of the extreme-value theory, in addition to Hill estimator there is another tool adapted to the observation of $Y_{\ell,n}$. Indeed if the vector $Y_{\ell,n}$ lists the peaks over a threshold, the eponym method (named by its acronym POT) allows to make a distinction between the different upper-tail behaviours of $\mathbb{P}_{X_1}$, using the limit distribution given by the Balkema, De Haan, and Pickands Theorem. This limit distribution is the generalized Pareto distribution (GPD), and, particularly, we will use the maximum likelihood estimator of the shape parameter of the GPD, in combination with its asymptotic confidence interval, to estimate whether this parameter is zero or strictly positive. The first case is related to the fact that $\mathbb{P}_{X_1}$ is, for instance, a normal or a log-normal distribution, and the latter to the fact that $\mathbb{P}_{X_1}$ is, for instance, a Pareto distribution. Nevertheless, this method is not very accurate for an unknown n and a small ℓ, it often not allowing distinguishing between both cases, or even not admitting a numerical solution.

General references on order statistics, extreme values, and records, are Arnold *et al.* (1992), Beirlant *et al.* (1996), Deheuvels (2010), Embrechts *et al.* (2008), Feller (1970), Galambos (1978), Nevzorvov (2001), Reiss (1989), Resnick (1987). Some references that focus on extreme-value theory are Embrechts *et al.* (2008), De Haan *et al.* (2006), Neves and Fraga Alves (2008), Reiss and Thomas (1987), Resnick (2007), and on records, see Gulati and Padgett (2003), or Arnold *et al.* (1998), or Kukush *et al.* (2004). Except for Pareto type distributions, with the use of the Hill estimator or of its extensions (Embrechts *et al.* (2008), Resnick (2007)), we do not know a reference addressing the problem of predicting the next-record value in our framework (many references use the sequence of record values instead of $Y_{\ell,n}$, see Arnold *et al.* (1998), Gulati and Padgett (2003), including a censored case, see Mirmostafaee and Ahmadi (2010)). Observe that our problem is related to that of determining the upper-tail behavior of $\mathbb{P}_{X_1}$ from the observation of $Y_{\ell,n}$.It is a problem for which we have no more knowledge of a reference that addresses this problem under our non-asymptotic conditions.

1.4 Approach by linear models

In our framework, when the largest values $X_{n,k}$ are represented as a function of k, it can be observed that they have the appearance of a convex decreasing function (see, for instance, the graphs presented in Section 2). We then propose to study the possibility of applying a linear regression model to explain the observed variable $X_{n,k}$ by the explanatory variable k, because prediction is an easy process in such a model. More precisely, as for the studied examples, the curvature due to the convexity is generally more important for some values of k and can be large, it is proposed the use of two linear regression models with different curvatures here. We therefore consider the Model 1 given by:

$$X_{n,k} = a_0 + a_1 k + a_2 e^{-k} + \epsilon_k, \quad k \in \{1, \dots, \ell\}, \tag{1}$$

where the curvature is observed for small values of k, and also the Model 2 given by:

$$X_{n,k} = a_0 + a_1 k + a_2 k^2 + \epsilon_k\,, \;\; k \in \{1,\ldots,\ell\}\,, \tag{2}$$

where the curvature is observed for larger values of k. Note that in both models a_2 should be a positive parameter to satisfy the convexity property of the observations. Note also that the next-record value R is easily predicted in both models by taking the explanatory variable equals to 0. This leads to potentially propose two quantities:

$$\hat{R}_1 = \hat{a}_0 + \hat{a}_2\,,$$

in Model 1, and:

$$\hat{R}_2 = \hat{a}_0\,,$$

in Model 2, as predictions of R. In addition, from the statistical properties of linear regression models, it is not much more difficult to provide a prediction interval for R, referred to *prediction interval of linear models*. We thus propose to use such an interval for the construction of an RPI. In light of a comparison of the functions of k involved in (1) and (2), we already do observe that Model 1 will tend to overestimate the prediction of R compared to the one of Model 2.

Let us now mention that the use of these linear models is unconventional. With the order statistics as observed variables, *a priori* the residuals ϵ_k do not satisfy the classical assumptions of independence and identical distribution. In addition, these residuals do not follow a normal distribution. Despite that, the data sets considered in this presentation show that the residuals obtained by applying Models 1 and 2 are generally rather small. And this may lead one to think that a prediction obtained in this way for the next record value is credible. In practice, observation of small residuals results in a determination coefficient, or R^2, close to 1.

A difficulty however arises since the prediction intervals obtained by Model 1 or by Model 2 turn out to be of low confidence levels (see Tables 3 and 4 in Section 4.2, where these confidence levels are calculated for several sets of simulated data). After having determined an important part of the low confidence level origin, it is proposed in this paper to involve both models to construct a prediction interval for R by retaining the upper endpoint obtained from Model 1 and the lower endpoint obtained from Model 2. With such a construction, the RPI shows a better confidence level, and, if necessary, a central tendency feature attached to it will therefore serve as a prediction of R. The RPI obtained by this linear model method (LMM) is the subject studied here.

Hence, the purpose of this presentation is to evaluate the practical interest of this linear model approach by considering its application to several examples of simulated data according to usual distribution models in extreme-value theory, and to real data. For each simulated-data set, using the computing power, the next-record value was also simulated and the probability that this value is included in an RPI was empirically evaluated by generating a large number of these samples. All the computer processing necessary for the progress of this work was done with the R software (http://www.R-project.org (2008)), and some general references on linear models are Jobson (1991), Searle (1997) and Seber (1977).

The rest of the paper is organized as follows. The sets of simulated and real data, that underlie this presentation, are introduced in Section 2. Some known results to estimate next-record values are then recalled in Section 3. Application of the LMM to the sets of simulated and real data is the subject of Section 4, a comparison of the results with those coming from the Pareto assumption is also provided. We conclude the paper with a brief discussion of the results obtained, some development prospects of this work being also included.

2 Simulated and real data

The sets of simulated data and of real data underpinning this presentation are successively presented.

2.1 Simulated data

Under the model assumptions, we simulate samples taking $\mathbb{P}_{X_1}$ in three parametric families of probabilities with different upper-tail behaviors. These families are representative of those used in extreme-value theory, particularly as a model of loss distributions in insurance. Thus, let us consider the normal family, the log-normal family, and the Pareto family. These families represent, respectively, distributions with not very heavy tails, distributions with heavy tails, and distributions with very heavy tails. The two first families lie in the Gumbel extremal domain of attraction while the last family lies in the Fréchet one. These examples have appeared to be sufficiently illustrative to not have to consider other families of distributions, as gamma, log-gamma, Weibull, or Burr families, for instance.

For the sake of allowing for comparisons within these families, the distributions were chosen with the same mean arbitrarily equals to 1, and with the same variance denoted by σ^2. Therefore, in the presented simulations, the probability distribution $\mathbb{P}_{X_1}$ will be a normal distribution denoted by $\mathscr{N}(1,\sigma^2)$, or a log-normal one denoted by $\log\mathscr{N}(-(1/2)\log(1+\sigma^2),\log(1+\sigma^2))$, or else a Pareto one denoted by $\mathscr{P}(1-1/(1+\sqrt{1+1/\sigma^2}),1+\sqrt{1+1/\sigma^2})$. For σ^2, the values 0.5 and 2 were retained to consider cases of probabilities with large and very large variances (these values have to be compared with respect to the mean value 1).

For vectors $Y_{\ell,n}$ resulting from sample simulations of such distributions, Figures 1, 2, and 3 show several graphs of the coordinates $X_{n,k}$ as a function of k. Each figure is related to a different parametric family, and consists of three sub-figures for the following choices of n and ℓ values: $n = 1000$ and $\ell = 40$ left sub-figure, $n = 100$ and $\ell = 10$ center sub-figure, and $n = 20$ and $\ell = 10$ right sub-figure. The ℓ values were fixed on the basis of the real-data ones. Both choices of σ^2 are represented in each sub-figure. In addition, for the sake of visual comparability, the vertical-axis scales were taken identical for the three sub-figures in each figure, thus, as n becomes smaller, we must observe a decrease in the level of the sample extreme values.

The following comments shed light on some aspects of Figures 1 to 3.

- For the samples presented, a higher level of the extreme values is detected for the log-normal distributions compared to the normal distributions, but not for the Pareto distributions compared to the log-normal distributions despite a heavier tail for the Pareto distributions. This can be explained by looking

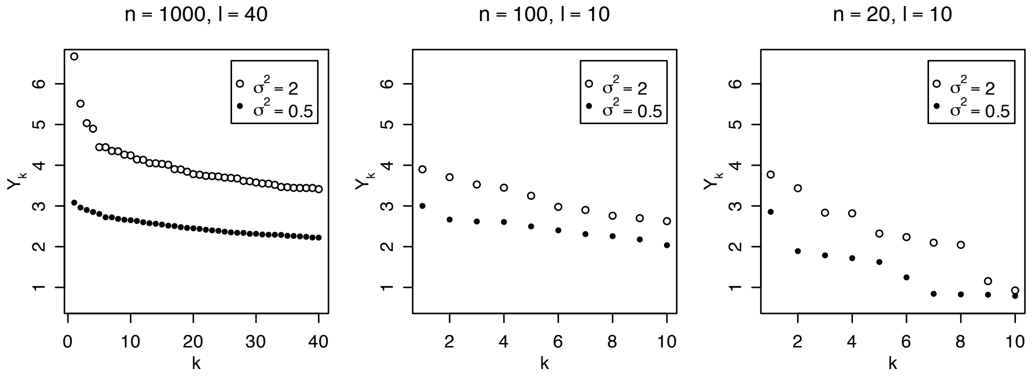

Figure 1. Six simulations of Y for normal distributions

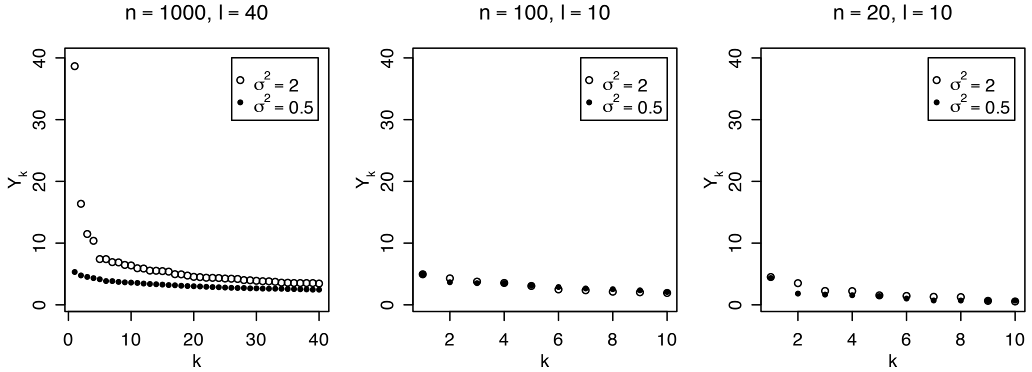

Figure 2. Six simulations of Y for log-normal distributions

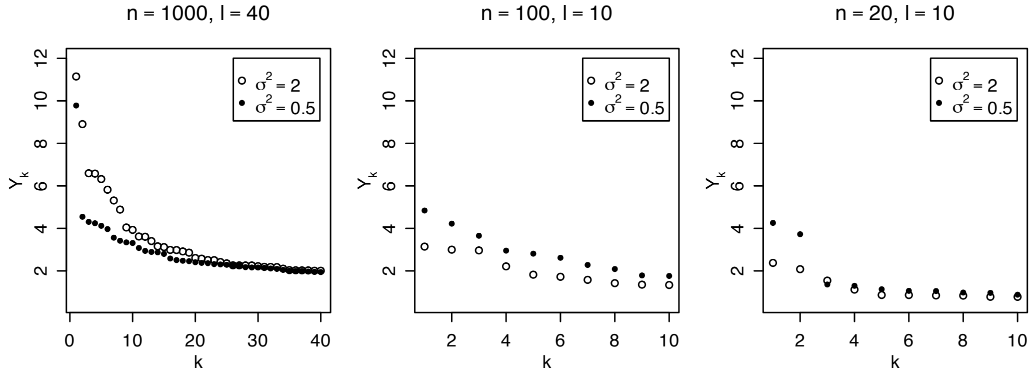

Figure 3. Six simulations of Y for Pareto distributions

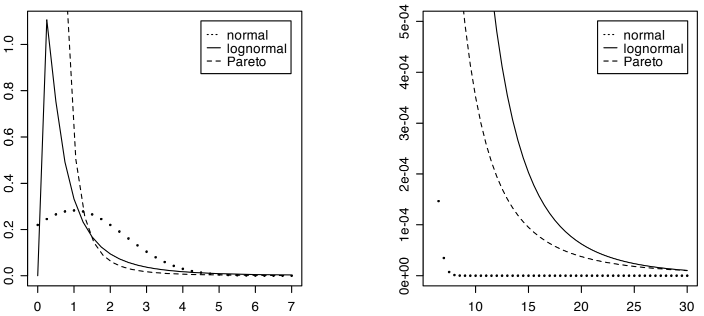

Figure 4. Comparison of the simulated densities for $\sigma^2 = 2$ on two areas

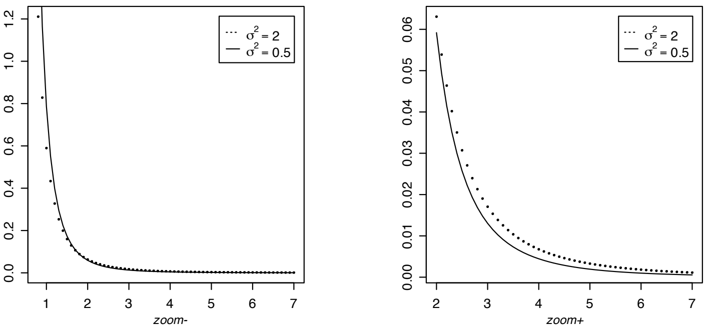

Figure 5. Comparison of the two simulated Pareto densities on two areas

at the densities of these distributions. For example, if $\sigma^2 = 2$, the density of the related log-normal distribution is above or very close to that of the related Pareto distribution in the area of the observed large values as indicated by Figure 4 representing the three densities on two areas. Consequently, the value of n should be greatly increased to observe a change of the situation.

- A higher level of the extreme values is also detected when the variance increases for the normal family, and, if the sample size is large, for the log-normal family and, partly, for the Pareto family. Otherwise, the fluctuations due to a change of samples prevail over those due to an increase in variance. In addition, the densities of the two Pareto distributions $\mathscr{P}(0.551, 2.225)$ (case $\sigma^2 = 2$) and $\mathscr{P}(0.634, 2.732)$ (case $\sigma^2 = 0.5$) are very close in the area of the observed large values as shown by Figure 5.

- For σ^2 fixed, the convergences of the empirical mean and of the empirical variance to the values 1 and σ^2 are slower for the log-normal distribution than for the normal, and for the Pareto distribution than for the log-normal. In fact, in the Pareto case, the empirical variance may remain substantially smaller than the theoretical variance, as shown in Figure 6 where the evolutions of the empirical moments are given for a sample of size 100,000 when $\sigma^2 = 2$. The lack of very large values in a Pareto sample also appears to be related to the smallness of the empirical variance. In fact, in a sample, the nearer are the empirical and theoretical variances, the larger are the extreme values, however this too scarcely arrives, it seems. This issue will not be discussed further here, but we will keep in mind that there may be a problem on the

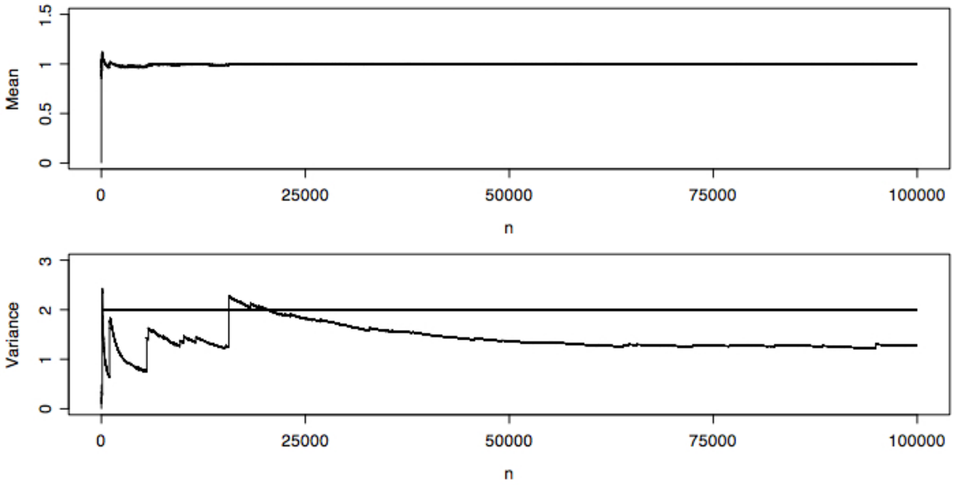

Figure 6. Evolutions of the empirical moments of a 100,000 Pareto sample for $\sigma^2 = 2$

Pareto simulations to eventually limit the scope of the results obtained.

2.2 Real data

In Figure 7, we present the coordinates $X_{n,k}$ of $Y_{\ell,n}$ as a function of k for the two following examples. First, the largest economic costs for the natural disasters in the World. These costs were observed since 1976 (the Sendai disaster included), and are expressed in 2010 US dollars (source De Haan *et al.* (2006)). There is a total of 10 observations; the three largest, in descending order, being the Sendai earthquake (2011), the Kobe earthquake (1995), and Hurricane Katrina (2005). Then, the largest claim-costs for insurers and reinsurers in the World. These costs were observed since 1970 (the Sendai disaster not included), and are also expressed in 2010 US dollars (source http://www.swissre.com/sigma (2011)). There is a total of 40 observations; the three largest, in descending order, being Hurricane Katrina(2005), Hurricane Andrews (1992), and the terrorist attacks of September 11, 2001. Economic cost and cost for insurance are obviously different, since the latter usually only covers a (small) fraction of the damage caused by a natural disaster. In addition, both the costs do not take into account the “cost” of lost human lives, and so, some of the recent disasters do not appear in these rankings.

Now, it can effectively be seen on the graphs in Figures 1-3 and 7 the convex decreasing shape of the representations of the largest values $X_{n,k}$ as a function of k, with a noise more important if n is small. Moreover, taking into account the different scales for the vertical axes, these graphs seem to show great similarities as ℓ remains fixed.

3 Some known facts

Results dealing with records, extreme-value theory, or asymptotic of order statistics are recalled now. All results quoted are in relation to the objective of estimating R starting from the observation of $Y_{\ell,n}$. Three subsections

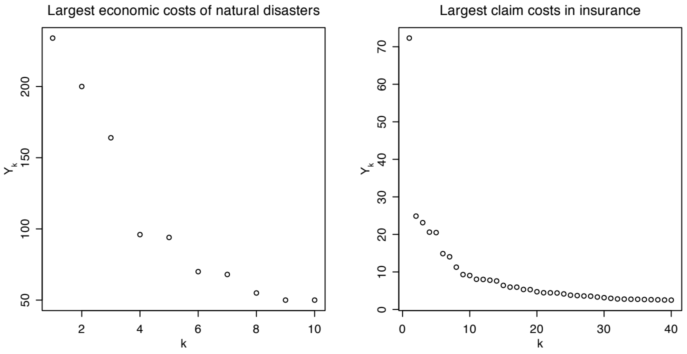

Figure 7. Observed data in billions of dollars

present successively elementary facts, asymptotic results and complementary observations on these issues.

3.1 Elementary facts

Let F be the distribution function of $\mathbb{P}_{X_1}$, and fix in this section the quantities n and ℓ. The right end of the support of $\mathbb{P}_{X_1}$ being infinite, we can make the following remarks.

Firstly, for $y \in \mathbb{R}^\ell$ fixed, if the maximum of the coordinates of y is denoted by s, it is easily seen that, for $x \geq s$:

$$\mathbb{P}(R > x / \, Y_{\ell,n} = y) = \mathbb{P}(R > x / \, M_n = s) = \sum_{\ell=1}^{\infty} F^{\ell-1}(s)(1 - F(x)) = \frac{1 - F(x)}{1 - F(s)} \,. \tag{3}$$

Thus, the best prediction of R is the conditional expectation:

$$\mathbb{E}(R / Y_{\ell,n} = y) = \mathbb{E}(R / M_n = s) = \frac{1}{1 - F(s)} \int_{]s,\infty]} (1 - F(x)) \, \mathrm{d}(x) = \mathbb{E}(X_1 - s / X_1 > s) \,, \tag{4}$$

that is the expected cost in excess of the threshold s. Moreover, an expression for an RPI can generally be obtained if F is known. Note that all these quantities depend on the observations of M_n, however, their expressions are not always explicit and may require the use of numerical solutions.

For $a > 0$ and $\alpha \in]0, \infty[$, the Pareto distribution $\mathscr{P}(a, \alpha)$ has the distribution function $1 - (x/a)^{-\alpha}$ for $x > a$, and 0 otherwise. For this distribution and $0 < \delta < 1$, an RPI with confidence level $1 - \delta$ is easily obtained from (3) as:

$$I_{Pareto}(R) = \left[(1 - \delta/2)^{-1/\alpha} M_n \,, \, (\delta/2)^{-1/\alpha} M_n\right]. \tag{5}$$

Similarly, in the cases of a normal or a log-normal distribution an RPI is expressed in terms of the distribution and quantile functions of the standard normal. Note that, to obtain an exact RPI with a fixed confidence level in the case of simulated data, we will use such an expression calculated from the conditional distributions (3). Note also that, in general, there exists no estimator of F that is a function of $Y_{\ell,n}$, which prevents to use such an estimator instead of F in (3) and (4). So that, in such a case it cannot be obtained by this mean a prediction or a prediction interval for R. However, in the Pareto case, using (5), it is sufficient to know an estimator of α that is a function of $Y_{\ell,n}$ to obtain a prediction interval for R. If necessary, we will use in this presentation the Hill estimator $\hat{\alpha}_{Hill}$ given by:

$$\hat{\alpha}_{Hill} = \Big(\frac{1}{\ell}\sum_{k=1}^{\ell} \log X_{n,k} - \log X_{n,\ell}\Big)^{-1}. \tag{6}$$

And, if we want to be more precise, the asymptotic normality of the Hill estimator (see, for example, Theorem 6.4.6 in Embrechts *et al.* (2008)) provides a confidence interval at 95% asymptotic confidence level of the form:

$$I_{Hill}(\alpha) = \big[\hat{\alpha}_{Hill}(1-1.96/\sqrt{\ell}),\, \hat{\alpha}_{Hill}(1+1.96/\sqrt{\ell})\big]. \tag{7}$$

Nevertheless, this way to proceed is equivalent to assume that $\mathbb{P}_{X_1}$ follows a Pareto distribution, and, as we will see in what follows, this assumption may lead to an RPI with an upper endpoint much higher than required if it is not verified.

3.2 Asymptotic results

We focus here on two asymptotic results in extreme value theory. Recall that, under the assumption of the infinite right end for the support of $\mathbb{P}_{X_1}$, the Fisher-Tippett theorem shows that the limiting distribution of M_n is in the Fréchet extremal domain of attraction or in the Gumbel one. In fact, if $\mathbb{P}_{X_1}$ is in the Fréchet domain, as are Pareto distributions, then the limiting distribution of M_n is a Fréchet distribution, while if $\mathbb{P}_{X_1}$ is in the Gumbel domain, as are normal and log-normal distributions, then this limiting distribution is the Gumbel distribution. This dichotomy is reflected in the results we state now, the first, that can be used to test between these two asymptotic situations, and the second, that we use to show that the LMM applies quite similarly in both circumstances. The first deals with the convergence to a GPD, and the second with an asymptotic of order statistics.

The GPD with shape parameter $\xi \in [0,\infty[$ and scaling parameter $\beta \in]0,\infty[$ has its distribution function defined by:

$$1 - G_{\xi,\beta}(x) = (1+\xi x/\beta)^{-1/\xi},\ \text{if } x \in [0,\infty[,\ \text{and } 0, \text{otherwise},$$

with the convention $(1+\xi x/\beta)^{-1/\xi} = e^{-x/\beta}$ if $\xi = 0$ (Embrechts *et al.* (2008)). Then (see, for example, Theorem 3.4.13 in Embrechts *et al.* (2008)), for a certain function $\beta(s)$, the functions $(1-F(s+\cdot))/(1-F(s))$ and $1-G_{\xi,\beta(s)}(\cdot)$ come closer (in uniform norm) as the threshold s goes to infinity. Here, the parameter ξ is null if $\mathbb{P}_{X_1}$ is in the Gumbel domain, while the parameter ξ is strictly positive if $\mathbb{P}_{X_1}$ is in the Fréchet one. Besides, we have in the latter case the relation $\xi = 1/\alpha$ where α is the shape parameter of the limiting distribution.

For observed data, the method POT may then be applied to distinguish between both asymptotic cases the one related to these data. For that purpose, it is recalled that, for a sample of a GPD distribution, the maximum likelihood estimator $\hat{\xi}$ of ξ is such that $\sqrt{\ell}(\hat{\xi}-\xi)$ converges in distribution to the normal $\mathcal{N}(1,(1+\xi)^2)$, as ℓ goes to infinity (Embrechts *et al.* (2008), section 6.5.1). The Slutsky Theorems then allow to obtain a confidence interval for ξ. At a 95% asymptotic confidence level, this interval is given by:

$$I_{POT}(\xi) = \big[\hat{\xi} - 1.96(1+\hat{\xi})/\sqrt{\ell},\, \hat{\xi} + 1.96(1+\hat{\xi})/\sqrt{\ell}\big]. \tag{8}$$

Therefore, assuming that Y comes from observations over a large threshold s, and assuming, from what was just pointed out, that the random variables $X_{n,k} - s$, for $k \in \{1,\ldots,\ell\}$, form approximately a sample of a GPD distribution, the interval in (8) may be evaluated. And, if this interval does not approach the value 0, we are led to believe that $\mathbb{P}_{X_1}$ is in the Fréchet domain, while, otherwise, the assumption that $\mathbb{P}_{X_1}$ is in the Gumbel domain cannot be rejected. Note, however, that there is not always a maximum likelihood estimator of ξ which is non-negative (the strictly negative case is related with the assumption that $\mathbb{P}_{X_1}$ has a finite support). In addition, if $\mathbb{P}_{X_1}$ is a Pareto distribution, the accuracy obtained on ξ by (8) is significantly smaller than that obtained on $\alpha = 1/\xi$ by (7) ($1+\hat{\xi}$ should be replaced by $\hat{\xi}$ in (8) to obtain the same accuracy when ℓ is large).

As n tends to infinity with ℓ remaining fixed, and using centering and normalizing constants, the multivariate limiting distribution of Y is known (Theorem 4.2.8 and Example 4.2.9 in Embrechts *et al.* (2008), see also Resnick (2007)) and depends on whether $\mathbb{P}_{X_1}$ is in the Fréchet domain or in the Gumbel one. The expression of the mean $(m_1 \cdots m_\ell)'$ of this multivariate limiting distribution is given by:

$$m_k = \frac{1}{(k-1)!}\int_0^\infty y^{k-1-1/\alpha} e^{-y}\,\mathrm{d}y = \frac{\Gamma(k-1/\alpha)}{\Gamma(k)},\ k \in \{1,\ldots,\ell\},$$

if $\mathbb{P}_{X_1}$ is in the Fréchet domain with a shape parameter $\alpha > 0$, and:

$$m_k = \frac{1}{(k-1)!}\int_0^\infty (-\log y)\, y^{k-1} e^{-y}\,\mathrm{d}y,\ k \in \{1,\ldots,\ell\},$$

if $\mathbb{P}_{X_1}$ is in the Gumbel domain. This provides vectors of "asymptotic data" to which can be applied the linear models defined by (1) and (2).

The results of the application of Model 1 on these asymptotic data sets are detailed in Figure 8 for $\ell = 10$, and in Figure 9 for $\ell = 40$. The two sub-figures in each of these figures expound the two cases related to both values of variance, which correspond to two different values of the shape parameter α in the Fréchet case. The Gumbel case is represented in both sub-figures, its representation being facilitated by using the relation $m_k = m_{k-1} - 1/(k-1)$, for $k \in \{2,3,\ldots\}$. Also, observe that the m_k can take negative values in the Gumbel case because the centering constants are non-null.

For Model 2, taking into account the asymptotic-data shapes, its fits will be slightly less efficient than those obtained by using Model 1. That model only serves to capture an estimate of the lower endpoint of the RPI. This estimate is better than the one that would be obtained simply by considering M_n and its application on the asymptotic data sets. This last one is not given here.

Model 1 fits almost as well in both the Fréchet and the Gumbel cases, and the fit is better for a small value of ℓ. It is an element in favour of the use of the LMM to predict R.

3.3 Complementary observations

In the context studied, n is unknown or in a non-asymptotic background, and ℓ is rather small. Thus, depending on the data, and on misunderstandings about the data setting, another approach than the one recalled in Sections 3.1 and 3.2 must be sought in some cases. This presentation will show that two basic models of linear

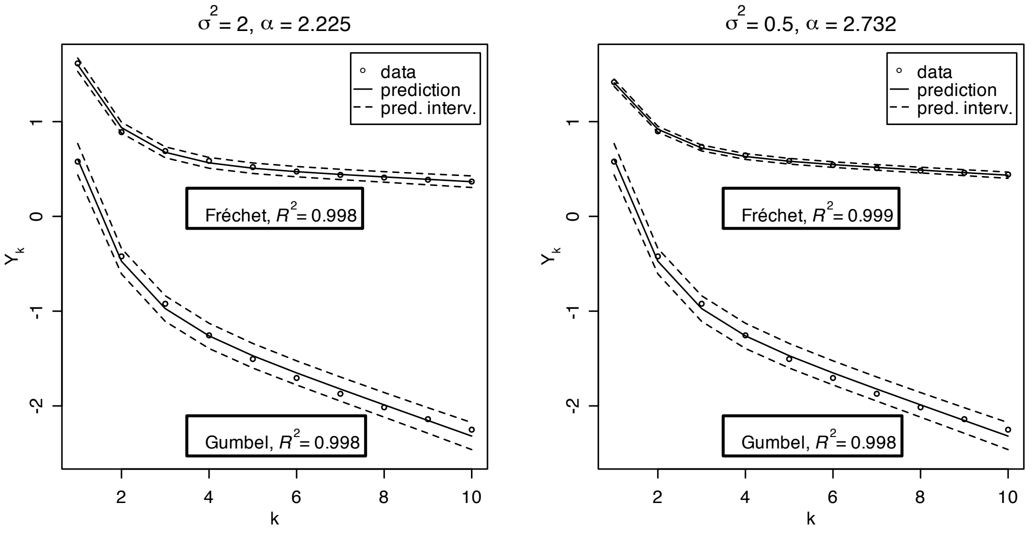

Figure 8. Applications of Model 1 to the asymptotic data for $\ell = 10$

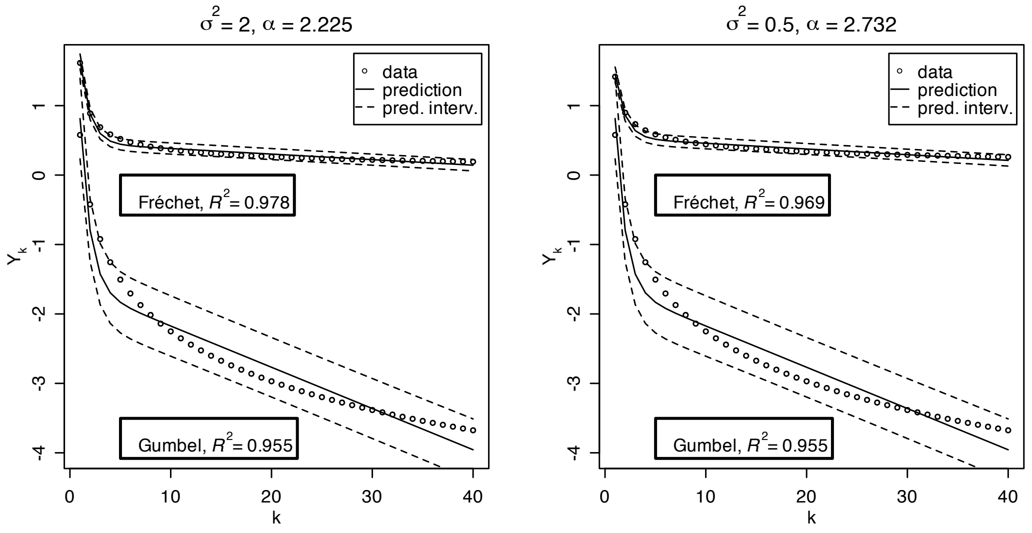

Figure 9. Applications of Model 1 to the asymptotic data for $\ell = 40$

regression type may serve as prototypes for the dependence of R in Y. And this approach is non-parametric, in the sense that it does not search to estimate $\mathbb{P}_{X_1}$ in a parametric family of distributions, or to impose a type of decrease for the tail of $\mathbb{P}_{X_1}$. Furthermore, it now becomes clear that the use of the LMM is related to a low belief in the fact that $\mathbb{P}_{X_1}$ follows a Pareto distribution, or, more generally, belongs to the Fréchet domain (anyway, its tail is then of Pareto type). In this presentation, we propose to assess by the Monte-Carlo method the mean of the confidence levels for the prediction intervals constructed by the LMM. In fact, this is possible in the case of simulated data, where a large number of samples $X_1, \ldots, X_{n+j}$ can be generated for j very large compared to n. Then, on each sample, the first n values determine a vector Y on which the LMM can be applied, while the next j values allow determining the next-record value. In this way, an estimate of the probability that this next-record value lies in its prediction interval can thus be obtained. However, we will also calculate the confidence levels of the RPI obtained for the simulated-data sets. Moreover, on each of our examples of simulated data, we will compare the prediction interval obtained by the LMM with the exact one calculated using the conditional distribution (3), and with that calculated by (5) and (6) therefore assuming that $\mathbb{P}_{X_1}$ follows a Pareto distribution.

4 Application of the linear model method

Three subsections present the RPI obtained for the simulated data sets by using the LMM, different evaluations of the accuracy of these prediction intervals, and the RPI obtained for the real-data sets by using the LMM.

4.1 Application on the simulated data

For all sets of simulated data presented in Section 2.1, the results obtained by the LMM are detailed in Figures 10 to 18. The two sub-figures in each of these figures expound the two cases related with both values of variance, the other parameters remaining fixed. All the graphs in these figures incorporate the description of the values $X_{n,k}$ as a dotted line, the description of the predictions by Models 1 and 2 as two continuous lines, and the description of the related prediction intervals as two broken lines. And, even if we are only interested in the prediction interval for $k = 0$, the latest curves are entirely represented for clarity. Recall also that the upper endpoints of these prediction intervals are the ones of the 95% prediction intervals obtained from Model 1, while their lower endpoints are the ones of the 95% prediction intervals obtained from Model 2. In addition to each graph, a legend specifies the values of the two predictions $\hat{R}_1$ and $\hat{R}_2$ (in the legend the first value is for Model 1 and the second for Model 2), the RPI, denoted by PI in the legend, and the values of the determination coefficients, or R^2, for both models (in the legend the first value is for Model 1 and the second for Model 2). It has to be noted that, for the expressions of $\hat{R}_2$ and of the lower endpoint of the RPI, they are taken equal to M_n every time the ones calculated by Model 2 are found smaller than M_n.

The following comments shed light on some aspects of Figures 10 to 18.

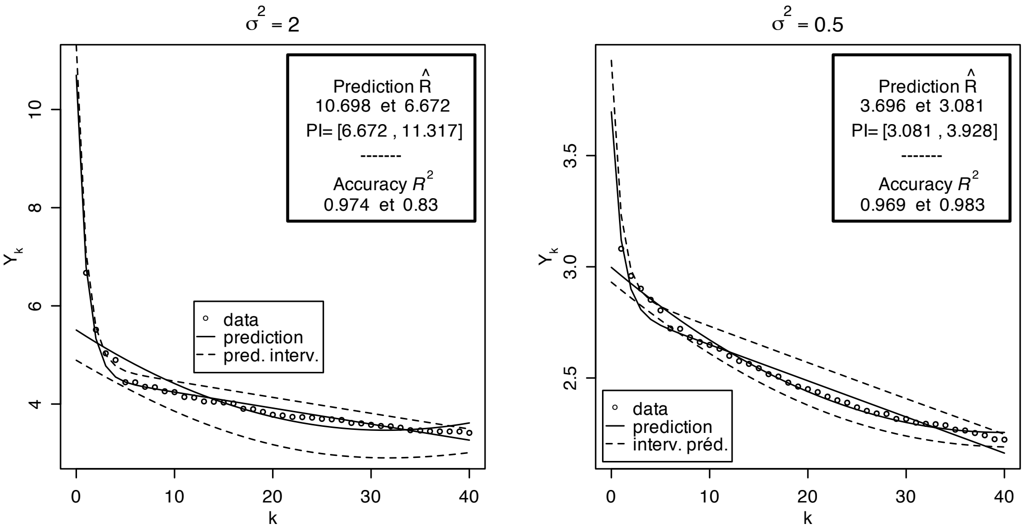

Figure 10. *R* prediction intervals for the normal data when $n = 1000$ and $\ell = 40$

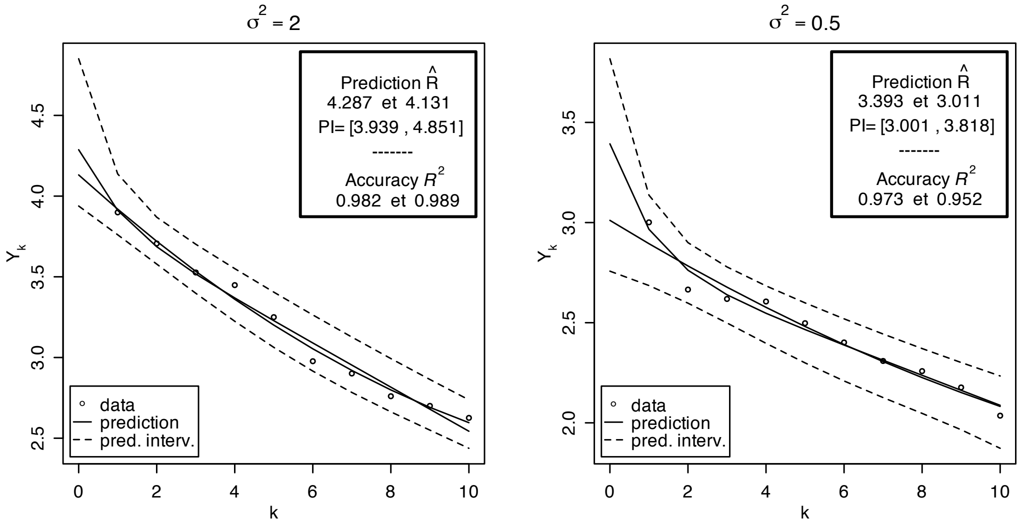

Figure 11. *R* prediction intervals for the normal data when $n = 100$ and $\ell = 10$

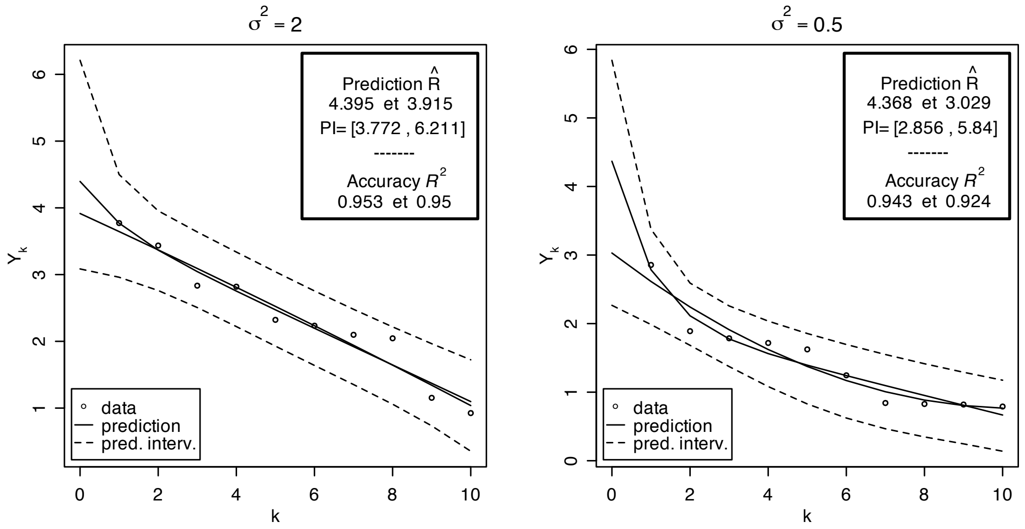

Figure 12. *R* prediction intervals for the normal data when $n = 20$ and $\ell = 10$

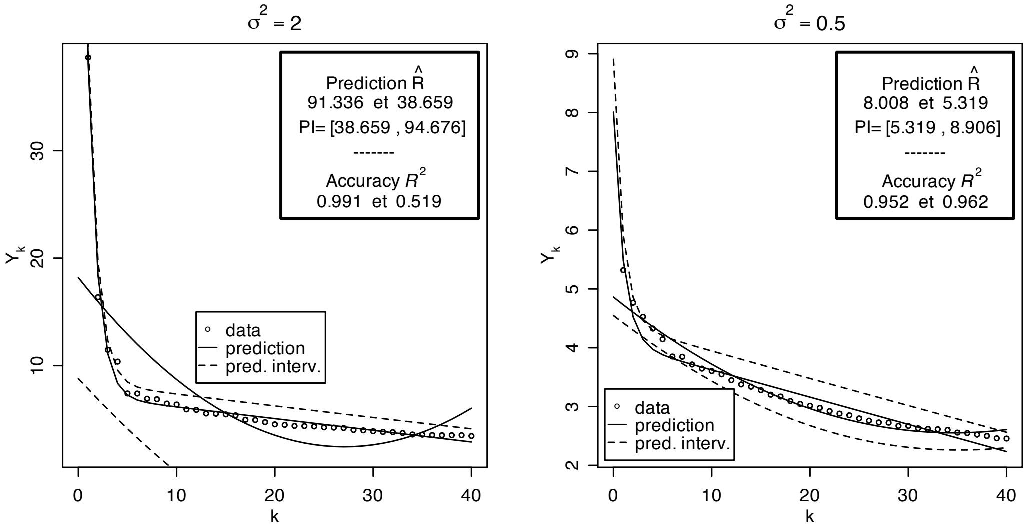

Figure 13. *R* prediction intervals for the log-normal data when $n = 1000$ and $\ell = 40$

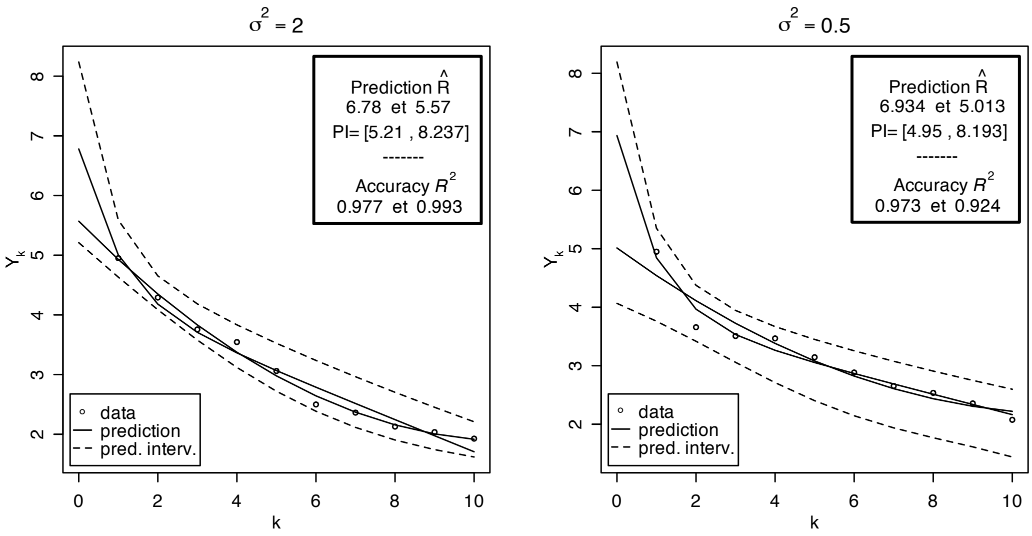

Figure 14. R prediction intervals for the log-normal data when $n = 100$ and $\ell = 10$

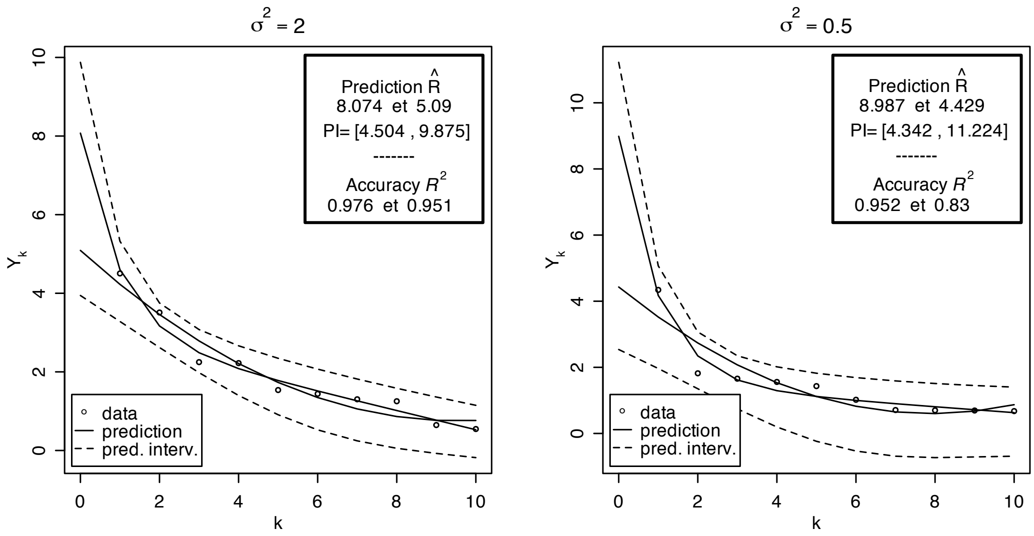

Figure 15. R prediction intervals for the log-normal data when $n = 20$ and $\ell = 10$

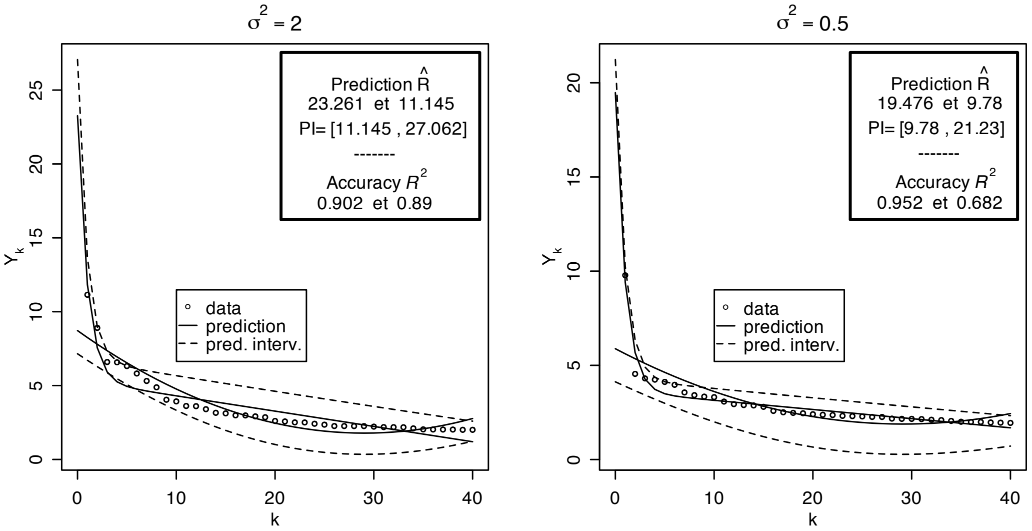

Figure 16. R prediction intervals for the Pareto data when $n = 1000$ and $\ell = 40$

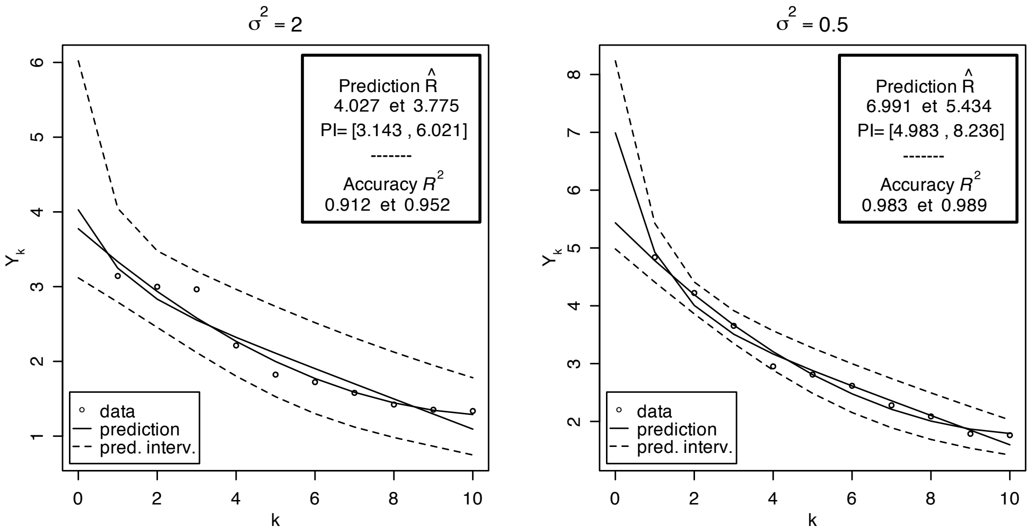

Figure 17. R prediction intervals for the Pareto data when $n = 100$ and $\ell = 10$

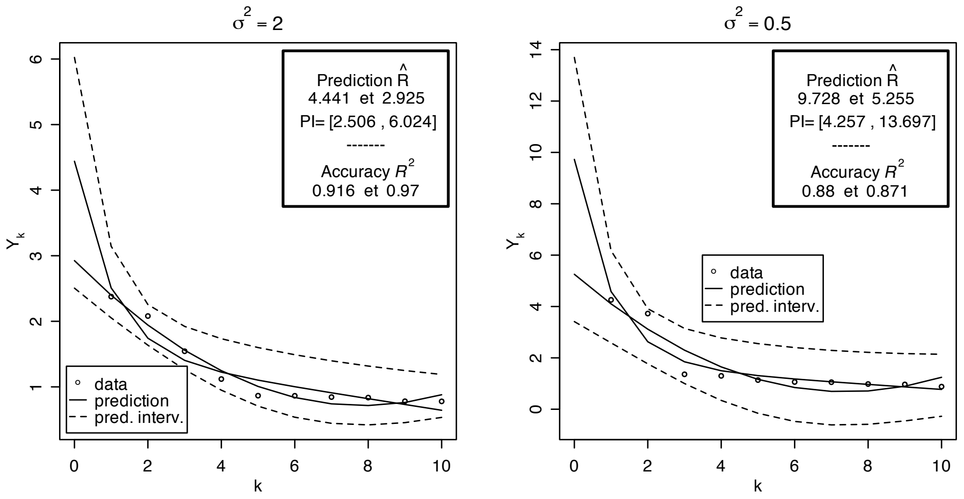

Figure 18. R prediction intervals for the Pareto data when $n = 20$ and $\ell = 10$

- Excluding in the left sub-figure in Figure 13, the determination coefficients are quite good and often greater than 0.9.
- The lower endpoints of the RPI calculated by Model 2 do not seem very efficient, appearing only four times in place of M_n on the data sets presented.
- The appearance of the different curves show that Model 1 has often a better fit to the data than Model 2, which results in a better determination coefficient for the first model. In such cases, Model 2 requires to be less close to the data in order to correctly predict the small values of R. It is also clear that the lower endpoints of the RPI calculated by Model 1 are inappropriate to this objective. Indeed, these lower endpoints can easily be imagined from the upper endpoints by the symmetries parallel to the vertical axis around the continuous lines of prediction of Model 1.
- When $\ell = 10$ several data representations suggest that the simple linear model ($a_2 = 0$ in Models 1 and 2) fits rather well on the data. However, in many of these cases, this model will underestimate the large values of R since the enlargements, which appear to the left of all the prediction interval graphs, are necessary to handle all the R variability if we assume that the RPI fairly measure the R variability (see the next subsection). In particular, with the aim to predict the next-record value, it seems not appropriate to use the common practice of eliminating a parameter that is eventually not significant in a statistical model.

4.2 Evaluation of the LMM for the simulated-data sets

The results given in the following eight tables are intended to analyze the interest of the statistical process being studied. The first four tables use the Monte-Carlo method to estimate the means of the determination coefficients and of the confidence levels of different prediction intervals obtained by applying the LMM when $\mathbb{P}_{X_1}$,

n, and ℓ are fixed as in the simulated data sets. Then, the exact confidence levels of the RPI obtained for the simulated data sets are calculated in the fifth table. Finally, these RPI are compared with those obtained by an exact calculation and by using the Pareto assumption. This last assumption is tested in the last three tables. For each $\mathbb{P}_{X_1}$, n and ℓ fixed as in the simulated data sets, and for $j = 100,000$, the simulation of a thousand samples of size $n + j$ according to $\mathbb{P}_{X_1}$ is generated. For each of these samples, the determination coefficients and the prediction intervals for Models 1 and 2 were calculated on the first n values, and the value of the next record was observed on the last j values (with a few exceptions that were counted). Thus, the empirical assessments of the following quantities were obtained:

- the probability of observing R in its prediction interval constructed by the LMM, this probability being an approximation of the mean of the confidence levels of such prediction intervals, it appears in Table 1;
- the mean of the determination coefficients obtained when applying Models 1 and 2, given in Table 2;
- the probabilities that R is above the prediction intervals calculated by both Models 1 and 2, given in Table 3;
- the probabilities that R is below the prediction intervals calculated by both Models 1 and 2, given in Table 4.

Table 1. Means of the confidence levels of the RPI obtained by the LMM, estimated on 1000 samples

Distributions	Normal		log-normal		Pareto	
σ^2	2	0.5	2	0.5	2	0.5
$n = 1000$, $\ell = 40$	0.968	0.972	0.925	0.948	0.826	0.86
$n = 100$, $\ell = 10$	0.899	0.907	0.837	0.872	0.782	0.821
$n = 20$, $\ell = 10$	0.925	0.928	0.845	0.879	0.76	0.796

The results in Table 1 seem rather satisfactory. To go further in the discussion, since one assesses the probabilities of certain events, and taking into account the number of samples simulated, we may use the usual asymptotic confidence interval of the parameter in a Bernoulli sample to obtain an order of magnitude for the accuracy of the results presented. In particular, at a 95% confidence level, the accuracy is about ± 0.0248 if the probability equals 0.8 and about ± 0.0186 if the probability equals 0.9. Thus, some comparisons between the results in Table 1 that may appear to be surprising are not significant in fact. For example, it is the case if we compare for the normal distribution the results when $n = 100$ and when $n = 20$, or if we compare the results when $\sigma^2 = 2$ and when $\sigma^2 = 0.5$, the other parameters remaining fixed. In a significant way, this time, the confidence levels are better when $n = 1000$, and they decrease when the tail of the distribution becomes increasingly heavy.

Table 2. Means of the determinant coefficients in applying models 1 and 2, estimated on 1000 samples

Distributions	Normal		Log-normal		Pareto	
σ^2	2	0.5	2	0.5	2	0.5
$n = 1000$, $\ell = 40$, Model 1	0.947	0.945	0.925	0.93	0.925	0.927
$n = 1000$, $\ell = 40$, Model 2	0.937	0.935	0.831	0.877	0.75	0.773
$n = 100$, $\ell = 10$, Model 1	0.959	0.961	0.959	0.96	0.959	0.958
$n = 100$, $\ell = 10$, Model 2	0.941	0.942	0.908	0.924	0.888	0.894
$n = 20$, $\ell = 10$, Model 1	0.962	0.963	0.96	0.962	0.957	0.958
$n = 20$, $\ell = 10$, Model 2	0.948	0.946	0.901	0.922	0.885	0.898

Table 3. Probabilities to be above the prediction intervals calculated by models 1 and 2, estimated on 1000 samples

Distributions	Normal		Log-normal		Pareto	
σ^2	2	0.5	2	0.5	2	0.5
$n = 1000$, $\ell = 40$, Model 1	0.022	0.017	0.065	0.044	0.165	0.131
$n = 1000$, $\ell = 40$, Model 2	0.916	0.921	0.977	0.961	0.989	0.983
$n = 100$, $\ell = 10$, Model 1	0.066	0.06	0.146	0.099	0.209	0.163
$n = 100$, $\ell = 10$, Model 2	0.243	0.254	0.428	0.379	0.517	0.446
$n = 20$, $\ell = 10$, Model 1	0.038	0.048	0.14	0.098	0.226	0.191
$n = 20$, $\ell = 10$, Model 2	0.189	0.179	0.454	0.352	0.521	0.497

The results in Table 2 are satisfactory. Observe only that there is no detectable changes in function of the distribution families for Model 1, unlike for Model 2 that fits less well when the distribution tail becomes increasingly heavy.

The results in Tables 3 and 4 justify the approach retained to select the upper and lower endpoints of the RPI in the LMM.

The confidence levels of the prediction intervals obtained by the LMM are calculated numerically using (3) for the eighteen simulated data sets since for these data $\mathbb{P}_{X_1}$ is known. These confidence levels are detailed in Table 5. Observe that in this table the value 1 means a level greater than $1 - 10^{-4}$.

The results in Table 5 illustrate a rather substantial variability of the confidence levels of the RPI obtained by

Table 4. Probabilities to be below the prediction intervals calculated by Models 1 and 2, estimated on 1000 samples

Distributions	Normal		Log-normal		Pareto	
σ^2	2	0.5	2	0.5	2	0.5
$n = 1000$, $\ell = 40$, Model 1	0.87	0.847	0.809	0.827	0.73	0.729
$n = 1000$, $\ell = 40$, Model 2	0.01	0.011	0.01	0.008	0.009	0.009
$n = 100$, $\ell = 10$, Model 1	0.408	0.387	0.485	0.452	0.474	0.482
$n = 100$, $\ell = 10$, Model 2	0.035	0.033	0.017	0.029	0.009	0.016
$n = 20$, $\ell = 10$, Model 1	0.354	0.363	0.494	0.472	0.456	0.433
$n = 20$, $\ell = 10$, Model 2	0.037	0.024	0.015	0.023	0.014	0.013

Table 5. Exact confidence levels of the *R*PI obtained by the LMM for the simulated data sets

Distributions	Normal		Log-normal		Pareto	
σ^2	2	0.5	2	0.5	2	0.5
$n = 1000$, $\ell = 40$	1	0.989	0.981	0.946	0.861	0.88
$n = 100$, $\ell = 10$	0.774	0.993	0.61	0.937	0.765	0.689
$n = 20$, $\ell = 10$	0.995	1	0.865	0.996	0.762	0.959

the LMM. There are several confidence levels too high, with intervals therefore too large, and there are several confidence levels too low, with intervals therefore not enough large. A comparison of these results with those of Table 1 also allows measuring the influence of the sample on the confidence level variability. Nonetheless, these results often reflect fairly good confidence levels.

Tables 6 and 7 allow comparing the *R*PI obtained by the LLM with the exact prediction intervals based on (3), and with the prediction intervals obtained using (5) and (6), thus assuming that $\mathbb{P}_{X_1}$ is a Pareto distribution. Following, in order to test the Pareto assumption, Table 8 gives the asymptotic confidence intervals calculated by using (8) for the parameter ξ in the POT method. All these calculations were done for the eighteen simulated data sets, and all the confidence levels were taken equals to 95%.Thus, Table 6 gives the exact 95% prediction intervals, while, for the upper endpoints of the prediction intervals, Table 7 gives in percentage the differences between the exact calculation and those obtained by the LMM and by using (5) and (6). For the lower endpoints, it is observed here that the same differences do not exceed a few percentage units, so that they are not specified in this presentation. Finally, Table 8 gives the asymptotic confidence interval for ξ, but only if the maximum likelihood estimator $\hat{\xi}$ of ξ is found non-negative, the negative case being related to the assumption that $\mathbb{P}_{X_1}$ has finite support.

Table 6. Exact prediction intervals for the simulated data sets, rounded to the tenth

Distributions	Normal		Log-normal		Pareto	
σ^2	2	0.5	2	0.5	2	0.5
$n = 1000$, $\ell = 40$	$[6.7, 7.8]$	$[3.1, 3.8]$	$[38.9, 89.3]$	$[5.4, 10]$	$[11.3, 58.5]$	$[9.9, 37.7]$
$n = 100$, $\ell = 10$	$[3.9, 5.7]$	$[3, 3.7]$	$[5, 18.1]$	$[5, 9.5]$	$[3.2, 16.5]$	$[4.9, 18.7]$
$n = 20$, $\ell = 10$	$[3.8, 5.6]$	$[2.9, 3.6]$	$[4.6, 17]$	$[4.4, 8.6]$	$[2.4, 12.5]$	$[4.3, 16.4]$

Table 7. Upper endpoints of the RPI constructed by the LMM and using the Pareto assumption (PA), given by their differences with the exact upper endpoints in percentage

Distributions	Normal		Log-normal		Pareto	
σ^2	2	0.5	2	0.5	2	0.5
$n = 1000$, $\ell = 40$, LMM	45%	3.6%	6%	-11%	-54%	-44%
$n = 1000$, $\ell = 40$, PA	46%	22%	110%	30%	-5.4%	-15%
$n = 100$, $\ell = 10$, LMM	-14%	2.4%	-55%	-14%	-64%	-56%
$n = 100$, $\ell = 10$, PA	35%	58%	24%	111%	-23%	33%
$n = 20$, $\ell = 10$, LMM	12%	61%	-42%	31%	-52%	-17%
$n = 20$, $\ell = 10$, PA	1512%	411%	1254%	336%	-30%	45%

The results in Table 7 reflect deviations from the exact calculations that are often important and may even be very large, especially when using the Pareto assumption, for small n ($n = 20$) but also for n greater in the case of log-normal distributions. The very large deviations are positive, they thus show that the Pareto assumption can lead to provide much larger new record values (up to 16 times) than necessary. As might be expected, for Pareto samples, with the exception of $n = 20$, the method using (5) and (6) works better than the LMM, although it still happens to predict with quite large differences. From this point of view, in all considered cases, the LMM obtains differences of up to 64%, that is to say, differences that remain more reasonable than the ones obtained under the Pareto assumption.

Table 8. Asymptotic confidence intervals of the parameter ξ for the simulated data sets

Distributions	Normal		Log-normal		Pareto	
σ^2	2	0.5	2	0.5	2	0.5
$n = 1000$, $\ell = 40$	$[-0.24, 0.43]$	$[-0.28, 0.37]$	$[0.07, 1.03]$	$[-0.31, 0.31]$	$[0.13, 1.15]$	$[-0.09, 0.73]$
$n = 100$, $\ell = 10$	$\hat{\xi} < 0$	$\hat{\xi} < 0$	$\hat{\xi} < 0$	$\hat{\xi} < 0$	$\hat{\xi} < 0$	$\hat{\xi} < 0$
$n = 20$, $\ell = 10$	$\hat{\xi} < 0$	$\hat{\xi} < 0$	$\hat{\xi} < 0$	$\hat{\xi} < 0$	$[-0.13, 2.72]$	$[-0.30, 1.99]$

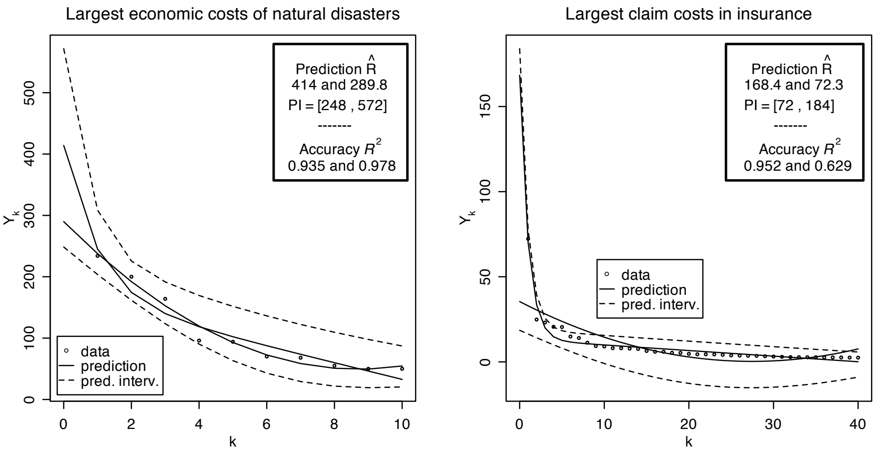

Figure 19. Prediction of the next-record values for the real-data sets

Recall that the asymptotic confidence interval (8) should eventually allow deciding between the Gumbel and the Fréchet asymptotic cases,that is to say, between the normal or log-normal assumptions and the Pareto one for the simulated data sets, and that, depending on whether this interval contains 0 or not. Thus, this interval is given only when the estimator $\hat{\xi} \geq 0$. The case $\hat{\xi} < 0$ being assimilated to the Gumbel case here. Often, the results in Table 8 do not allow to correctly determine the distribution type of $\mathbb{P}_{X_1}$. One error on three found in each row of the table. For example, they never recognize the Pareto assumption for small n ($n = 100$ and $n = 20$).

4.3 Application on the real-data sets

In Figure 19 are summarized the treatments by the LMM of the two examples of real data reported in Section 2.2, with exactly the same conventions as those used to present the figures in Section 4.1.

Thus, considering that the real-data sets come from samples of heavy-tail distributions, and considering that the confidence levels values of Table 1 are significant for these data sets, it can therefore be estimated with a probability of at least 76% that the next-record value will not exceed, for the economic costs of natural disas-

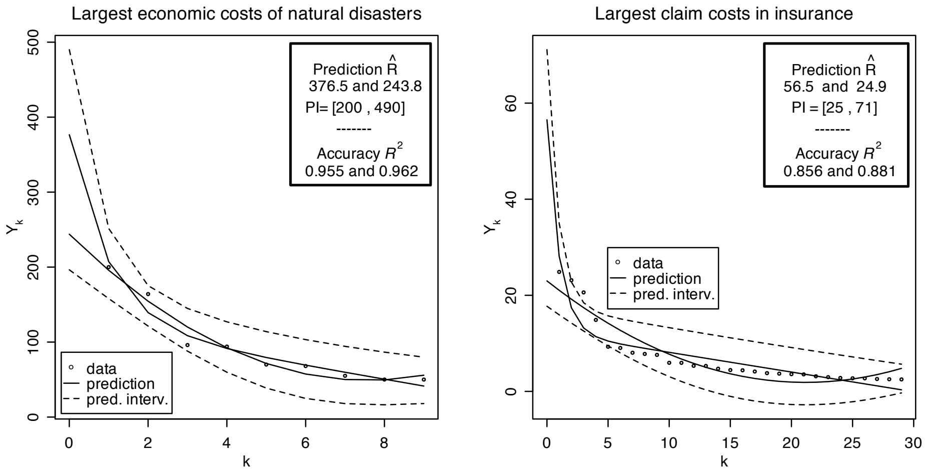

Figure 20. Prediction of the next-record values for the real-data sets when they are observed just before the largest cost

ters, 572 billion of US dollars (remember that the earthquake in Sendai would have created an economic cost of 234 billion). Similarly, we can estimate with a probability of about 82% that the next-record value will not exceed, for the claim costs for insurers and reinsurers, 184 billion (remember that Hurricane Katrina generated a cost of 72 billion). At the time of this writing, we do not know if the claim cost for insurers and reinsurers caused by the earthquake in Sendai will be the next record.The following observation shows that the insured risks should be analyzed precisely before being able to decide. Indeed, the Kobe earthquake is number two on the list of economic costs of natural disasters, and only number twenty-eight for the claim costs for insurers and reinsurers.

For these real-data sets, the asymptotic confidence intervals for the parameter ξ are equal respectively to $[-0.42, 1.48]$ (economic costs) and to $[0.24, 1.35]$ (claim costs for insurers and reinsurers), and, in the first case, the assumption that $\mathbb{P}_{X_1}$ is in the Gumbel domain cannot be rejected. In addition, assuming that $\mathbb{P}_{X_1}$ follows a Pareto distribution, and using (5) and (6), the prediction intervals of the next record values become then equal in billions respectively to $[237.7, 2253.9]$ (economic costs) and to $[73.88, 1684.9]$ (claim costs for insurers and reinsurers). These results come from estimates of parameter α calculated by (7), and respectively equal to 1.629 and 1.172. The confidence intervals (8) are respectively $[0.619, 2.638]$ and $[0.809, 1.535]$. Observe that the upper endpoints of the last R prediction intervals are considerably more important than those calculated by the LMM. However, these intervals are related to a 95% asymptotic confidence level, which is a priori not contradictory with the results obtained by the LMM where the confidence levels are not exactly known.

To test if the LMM would be able to predict the largest cost observed, we remove from both real-data sets the largest cost and all costs that were occurred after the largest cost. It thus remains 9 observations for the economic costs and 29 observations for the claim costs for insurers and reinsurers. The resulting application of the LMM is then given in Figure 20.

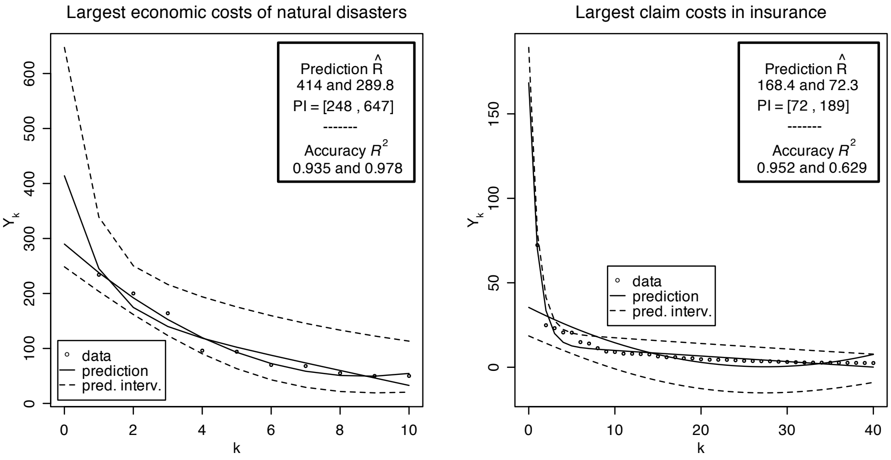

Figure 21. Prediction of the next-record values for the real-data sets with 99% confidence levels to fix the upper endpoints of the prediction intervals

For these "old" real-data sets, the asymptotic confidence intervals for the parameter ξ are now equal respectively to $[0.38, 5.58]$ (economic costs) and to $[-0.01, 1.12]$ (claim costs for insurers and reinsurers). In the second case, the assumption that $\mathbb{P}_{X_1}$ is in the Gumbel domain cannot be rejected. In addition, assuming that $\mathbb{P}_{X_1}$ follows a Pareto distribution, and using (5) and (6), the prediction intervals of the next-record values become then equal in billions respectively to $[202.7, 1316.2]$ (economic costs) and to $[25.36, 421.63]$ (claim costs for insurers and reinsurers). These results come from estimates of parameter α calculated by (7), and respectively equal to 1.958 and 1.303. The confidence intervals (8) are respectively $[0.679, 3.237]$ and $[0.829, 1.778]$. Once again, the upper endpoints of the last R prediction intervals are considerably more important than those calculated by the LMM, and yet the LMM predicts the observed values of the next record from the old real-data sets.

To continue this discussion a little, we observe that the prediction intervals of the linear models can be constructed at a higher confidence level, for example, equal to 99%, in order to increase the confidence levels in the LMM. In this case, it then would be necessary to again estimate the confidence levels of the RPI, as in Table 1. The sensibility of the LMM to this parameter is simply illustrated in Figure 21. The treatments of the real-data sets are also summarized in this Figure when solely the upper endpoints of the prediction intervals are determined at a 99% confidence level. The observed increases of the upper endpoints seem reasonable when comparing Figure 19 and Figure 21.

In conclusion, these studies on real data do not allow to exclude any of the methods studied. However, their consequences are really different in terms of the upper endpoint of the RPI. Indeed, from the points of view of economy or insurance, the amounts involved are considerably different. At this moment, it seems to us that going further in the discussion falls under the belief of a particular tail behavior assumption for the distribution $\mathbb{P}_{X_1}$. Nevertheless, the results observed on the simulated data show that a part of the Pareto case is at least taken into account by the LMM.

5 Conclusion

In a first analysis, the linear model method seems to perform rather well. For example, on the sets of data examined the determination coefficients of the linear regression models (1) and (2) are in almost all cases close to (and often exceed) the value 0.9, which encourages to believe in the method. However, the emergence of a very large extreme value, as generated by heavy-tail probabilities, should bring the next-record value out of the prediction interval obtained. To some extent, the second argument explains the high variability of the problem studied, variability that was illustrated for example when $\mathbb{P}_{X_1}$ is a Pareto distribution and the Hill estimator is used. The main consequence of this presentation lies in the results set out in Table 1 or in Table 5. In fact, they allow deciding between these two arguments by showing that, at least for the range of n and ℓ tested, it is quite possible to predict the next-record value from the largest values of a sample without any parametric assumption on the sample distribution, but with an important margin of error as usual in this domain. From this point of view, the method using linear regression models appears to obtain results more robust than the one using the Pareto assumption when $\mathbb{P}_{X_1}$ is not of Pareto type, but underestimating the upper endpoint of the RPI when $\mathbb{P}_{X_1}$ is of Pareto type.
At this moment, several improvements can be envisaged.

- It seems clear from the results in Table 1 that better R prediction intervals can be constructed by restricting the number of probability families considered in the study. For example, by limiting ourself to the log-normal and Pareto families, we may increase the confidence levels that fix the prediction intervals in the linear models, to obtain RPI pertaining to an intermediate position between both probability families. In doing so, the thickness of the error margin will decrease. Recall also that the log-normal and Pareto families often constitute the first set to be studied in insurance.
- Moreover, even if that does not appear in this work, several other linear regression models were tried before selecting those based on Models (1) and (2), for example, by working on the logarithm of the sample values, or by exploring other functions of the explanatory variable k. Usually, a compromise that is used to select the best model is to fix an acceptable confidence level for the prediction intervals while seeking the smallest width of these intervals. Observing the results in Tables (1) and (5), the latter criterion appears to have excessive weight. Other linear regression models can still be searched, but more systematic research should be undertaken. Theoretical tools as those given in Section 3.2 may help to do this task.
- Furthermore, in the framework studied, it still remains many situations to be tested in order to better specify the range of validity of the LMM, and, clearly, only a small number of situations were considered in this work.

Finally, from a mathematical point of view, this presentation appears very light, but it was necessary to explore the context associated with a first striking observation on one of both examples of real data. And we have been convinced by the preparation of this presentation that there is an interest to continue the study of the linear model approach to predict the next-record value from the largest values of a sample.

References

Amany, E. A., Barakat, H. M. and Magdy, E. El-Adll (2019), Prediction Intervals of the Record-Values Process, *Revstat-Statistical Journal* 17(3): 401-427. DOI: `10.57805/revstat.v17i3.274`

Arnold, B.C., Balakrishnan, N. and Nagaraja, H.N.(1992). A First Course in Order Statistics, *Wiley*, New York. DOI:`10.1137/1.9780898719062`

Arnold, B.C., Balakrishnan, N. and Nagaraja, H.N.(1998). Records, *Wiley*, New York. DOI:`10.1002/9781118150412`

Beirlant, J., Teugels, J.L. and Vynckier, P. (1996). Practical Analysis of Extreme Values, *Leuven University Press*, Leuven. DOI:`10.1016/S0167-6687(97)00022-X`

Courrier International 1065 (2011). *Courrier International SA*, p. 48, Paris.

De Haan, L. and Ferreira, A.(2006). Extreme Value Theory: An Introduction, *Springer*, New York. DOI:`10.1007/0-387-34471-3`

Deheuvels, P. (2010). Extreme Value Theory, Encyclopedia of Quantitative Finance, *Wiley*.

Embrechts, P., Klüppelberg C., and Mikosch, T.(2008). Modelling Extremal Events: For Insurance and Finance, *Springer*, New York. DOI:`10.1007/978-3-642-33483-2`

Feller, W. (1970). An Introduction to Probability Theory and its Applications, vol.II, *Wiley*, New York. DOI:`10.1137/1014119`

Galambos, J. (1978). The Asymptotic Theory of Extreme Order Statistics, *Wiley*, New York. DOI:`10.1080/00401706.1990.10484616`

Gulati, S. and Padgett, W.J. (2003). Parametric and Nonparametric Inference from Record-Breaking Data, *Springer*, New York. DOI: `10.1007/978-0-387-21549-5`

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Ann.Statist.*, 3(1163-1174).
DOI: `http://dx.doi.org/10.1214/aos/1176343247`

Kukush A., Chernikov, Y. and Pfeifer, D. (2004). Maximum Likelihood Estimators in a Statistical Model of Natural Catastrophe Claims with Trend. *Extremes* 7, 309–336.
DOI:`10.1007/s10687-004-3480-0`

J.D. Jobson (1991). Applied Multivariate Analysis, *Spinger*, New York. DOI:`0.1007/978-1-4612-0955-3`

Mirmostafaee, S.M.T.K. and Ahmadi, J. (2010). Prediction of future order statistics coming from Pareto model, *preprint*.

Neves, C., Fraga Alves, M.I.(2008). Testing Extreme Value Conditions: An Overview and Recent Approaches, *Revstat*, 6, 83–100. DOI:`10.57805/revstat.v6i1.59`

Nevzorov, V.B. (2001). Records: Mathematical Theory, *American Mathematical Society*, Providence, RI.

R Development Core Team (2008). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

Reiss, R.D.(1989). Approximate Distributions of Order Statistics, *Springer*, New York. DOI:`10.1007/978-1-4613-9620-8`

Reiss, R.D. and Thomas, M. (1987). Statistical Analysis of Extreme Values, Birkhäuser, Basel. DOI:`10.1007/978-0-387-75953-1`

Resnick, S. (1987). Extremes Values, Regular Variation, and Point Processes, *Springer*, New York. DOI:`10.1007/978-0-387-75953-1`

Resnick, S. (2007). Heavy-Tail Phenomena, Probabilistic and Statistical Modeling, *Springer*, New York. DOI: `10.1007/978-0-387-45024-7`

Searle, S.R. (1997). Linear Models, *Wiley*, New York. DOI: `10.1002/bimj.19740160113`

Seber, G.A.F. (1977). Linear Regression Analysis, *Wiley*, New York. DOI:`10.1002/9780470192610`

Sheather, S.J.(2009). A modern approach to regression with R, *Springer*, New York. DOI: `10.1007/978-0-387-09608-7`

Sigma (2011), Natural catastrophes and man-made disasters in 2010: a year of devastating and costly events, N. 1, Swiss RE, Zurich, URL http://www.swissre.com/sigma.